

Developing business analytics to support kick-scooters sharing systems operators

João António Abreu Costa de Sousa

Master's Dissertation

Supervisor: Prof. Vera Miguéis



Mestrado Integrado em Engenharia e Gestão Industrial

2019-07-01

Abstract

Technological change has been shaping our society and, in particular, the mobility industry. Innovative technological concepts integrating new vehicle types and based on new business models have been recently arising to meet new mobility challenges. One main trend is vehicle sharing, which allows users to book a scooter, kick-scooter, car or bicycle through an app. Kick-scooters are especially popular, with sharing systems operators expanding rapidly across the globe and especially in Europe.

Kick-scooters sharing systems operators are struggling with low profitability, mainly due to the intense competitive landscape and low operational efficiency. They are very focused on rapid expansion and, simultaneously, trying to make sure that they make the right strategic decisions and increase the business bottom line. In this context, given the high amount of data they own, artificial intelligence and analytics are set to play an important role. This thesis presents the development of features based on quantitative methods and machine learning to meet operators' needs.

Firstly, a market intelligence study is carried out, as well as a framework to measure the impact of new product features for kick-scooters operators. In order to study the end customers, this thesis develops a clustering algorithm to segment kick-scooters sharing systems users based on variables from historical rides, at any point in time. The conversion from open app events to bookings is analysed, in order to develop an algorithm that converts these events into demand units, taking into account business knowledge. This conversion is crucial for the last developed feature – demand forecast – as it is a censored demand problem. These three algorithms are applied to the city of Málaga, using historical data from one main European operator.

Following this methodology, this work prioritizes feature development based on business impact. The market intelligence study is crucial to acquire the required market expertise to develop the intended algorithms. Customer clustering identifies four main groups based on rides patterns corresponding to different usage purposes. A marketing strategy targeted at each customer segment is developed in order to meet specific needs and wants from each group. The developed algorithm that converts open app events into demand units is capable of breaking down these events by grouping them according to several factors. This conversion into demand units is a key input for demand forecast, given that it solves the censored demand problem, by capturing unmet demand and having a more accurate demand point location. Lastly, the forecast demand feature based on explanatory models, namely Random Forest, identifies which factors determine the demand in every city. It allows forecasting the daily demand of each grid cell with higher accuracy. This feature is crucial for having a more effective fleet rebalancing in order to increase profitability.

This thesis is expected to enable kick-scooters operators to take data-driven decisions, namely on business expansion and marketing, as well as to increase their profitability. Furthermore, this work is the basis for future development of additional analytics features.

Resumo

A evolução tecnológica tem influenciado a sociedade e, em particular, o setor da mobilidade. Conceitos tecnológicos inovadores têm surgido recentemente de forma a responder aos novos desafios da mobilidade, com novos veículos e assentes em novos modelos de negócio. Uma das principais tendências é a partilha de veículos, no qual os utilizadores podem reservar uma *scooter*, trotineta, carro ou bicicleta através de uma *app*. Trotinetas eléctricas são uma das principais tendências, dado que os operadores destes serviços crescem rapidamente por todo o mundo e em especial na Europa.

Os operadores dos serviços de trotinetas eléctricas partilhadas apresentam atualmente uma rentabilidade muito baixa, devido ao intenso ambiente concorrencial e baixa eficiência operacional. Enquanto expandem rapidamente, os operadores querem garantir que tomam as melhores decisões estratégicas e que, simultaneamente, tornam o negócio mais rentável. Neste sentido, dada a quantidade elevada de dados que armazenam, inteligência artificial e *analytics* surgem como potenciais soluções para estes problemas. Este trabalho apresenta o desenvolvimento de ferramentas para resolver os desafios dos operadores destes serviços, tendo como base métodos quantitativos e *machine learning*.

O trabalho começa por realizar um estudo de mercado e desenvolver uma *framework* de forma a analisar o potencial impacto para o negócio de novas *features*, que serão vendidas aos atuais operadores. De forma a conhecer melhor os clientes, este trabalho desenvolve um algoritmo de *clustering* para segmentar de uma forma dinâmica os utilizadores de trotinetas eléctricas, tendo como base variáveis recolhidas do histórico de viagens. Eventos correspondentes a abertura da *app* são analisados relativamente à sua conversão para reservas, de forma a desenvolver um algoritmo que converte estes eventos em unidades de procura, tendo como base conhecimento do negócio e a análise dos eventos passados. Esta conversão é essencial para a última ferramenta – previsão da procura – uma vez que se trata de um problema de procura censurada. Estes três algoritmos são aplicados e testados na cidade de Málaga, usando dados históricos de um dos principais operadores europeus.

Seguindo este método, foi possível priorizar as ferramentas a desenvolver, tendo como base o seu impacto para o negócio. Assim, o estudo de mercado permite desenvolver conhecimento sobre este setor, de forma a possibilitar o desenvolvimento dos algoritmos. Em particular, a segmentação de clientes permite identificar quatro grupos, tendo como base padrões de utilização, que são correspondentes a diferentes usos do serviço. Ao desenvolver a estratégia de *marketing* direcionada a cada grupo, é possível melhorar o serviço e ir de encontro às necessidades específicas de cada segmento. O algoritmo desenvolvido para criar unidades de procura permite analisar a conversão do evento de abrir a *app* para reservar uma viagem, através da segmentação destes eventos de acordo com vários fatores. A conversão em unidades de procura é um *input* necessário para a previsão de procura, resolvendo o problema da procura censurada. Desta forma, é possível captar procura não satisfeita, bem como ter a localização exata da procura satisfeita. Por fim, a previsão de procura baseada em modelos explicativos, nomeadamente *Random Forest*, identifica quais os fatores relevantes que determinam a procura em cada cidade e permite prever com maior precisão a procura diária em cada célula de uma grelha dinâmica. Esta ferramenta possibilita uma recolocação mais eficaz dos veículos tendo em vista o aumento da rentabilidade operacional.

Espera-se que esta dissertação contribua para a tomada de decisão baseada em métodos quantitativos, nomeadamente a estratégia de expansão e de *marketing*, bem como a melhoria da rentabilidade dos operadores. Adicionalmente, será a base para o desenvolvimento de futuras *features*.

Acknowledgments

My master's degree culminates in this thesis, which marks the conclusion of five years full of challenges, enriching experiences and personal growth. Such ride was only possible thanks to the ones that accompanied and paved the way. They deserve my words of appreciation.

Firstly, to the ones that accompanied this thesis development closely: my mentor, Roman Geugelin, for his continuous support, never-ending enthusiasm and, above all, trust in my work; my supervisor, Professor Vera Miguéis for her guidance, feedback and continuous support; my colleagues at Wunder Mobility, especially in the business development department, for all valuable inputs and support. I am very grateful for having had the opportunity to work with such smart and enthusiastic individuals.

To the ones who accompanied during these five years: my friends, both inside and outside university, for the learnings, inspiration and for always being there. They have always been a source of support when things would get a bit discouraging.

To the ones who made it all possible: my family. Firstly, my parents, for all they have done that shaped me into who I am today. A special thanks to my two brothers, who have been always a source of guidance and motivation over the years.

Contents

1	Introduction.....	1
1.1	Mobility market	1
1.1.1	Shared mobility	2
1.1.2	Shared micro-mobility	3
1.2	Company – Wunder Mobility	4
1.2.1	Fleet business unit.....	5
1.2.2	Project development context.....	5
1.3	Project objectives	5
1.4	Methodology and schedule	6
1.5	Dissertation structure	7
2	Literature review.....	8
2.1	Vehicle sharing mobility.....	8
2.2	Customer clustering	8
2.3	Fleet relocation strategy	9
2.3.1	Station-based systems.....	10
2.3.2	Free-floating systems.....	11
2.3.3	Hybrid systems	11
2.4	Demand forecast.....	11
2.4.1	City-level	12
2.4.2	Cluster-level.....	12
2.4.3	Station-level	12
2.5	Censored demand.....	12
2.6	Research gap.....	13
3	Problem definition.....	14
3.1	Context.....	14
3.2	Problem formulation	14
3.2.1	Market intelligence study	14
3.2.2	Customers analytics – segmentation	15
3.2.3	Build demand proxy from historical rides and app data	15
3.2.4	Forecast daily demand per area-level.....	15
3.3	Business relevance	15
3.3.1	Operators profitability.....	16
3.3.2	Operators competitive advantage	16
3.3.3	Data driven expansion strategy	16
3.3.4	Business insights	17
3.4	Available data.....	17
3.4.1	Bookings dataset	17
3.4.2	Open app events dataset.....	18
3.5	Project context limitations.....	18
3.5.1	Customers segmentation	18
3.5.2	Demand proxy and forecast.....	18
4	Proposed solution.....	20
4.1	Market intelligence study.....	20
4.1.1	Customers groups and demand factors.....	20
4.1.2	European cities – market attractiveness	22
4.1.3	Current industry evaluation	23
4.1.4	Feature development framework	24
4.2	Customer segmentation	25
4.2.1	Selecting clustering algorithm	25
4.2.2	Data preparation and pre-processing	26

4.2.3	Clustering.....	27
4.3	Demand proxy.....	28
4.3.1	Motivation	28
4.3.2	Methodology	28
4.4	Demand forecast.....	30
4.4.1	Prediction engineering	30
4.4.2	Feature engineering.....	31
4.4.3	Machine learning modelling	34
5	Results	38
5.1	Market intelligence study.....	38
5.2	Customer segmentation	39
5.2.1	Cluster results analysis	39
5.2.2	Business strategy	40
5.2.3	Marketing mix	41
5.3	Demand proxy.....	42
5.4	Demand forecast.....	43
5.4.1	Feature engineering and selection.....	43
5.4.2	Grid cells demand forecast	45
6	Conclusion and future work.....	47
	Bibliography.....	49
	Appendix A: Market assessment - variables	53
	Appendix B: Market intelligence study – results.....	55
	Appendix C: Ideas assessment results	57
	Appendix D: Customer clustering – results visualisation	58
	Appendix E: Demand forecast – exploratory analysis.....	60

Acronyms and Symbols

MAPE	Mean absolute percentual error
MPE	Mean percentual error
KNN	K-Nearest Neighbors
RF	Random Forest
DT	Decision Tree
SaaS	Software as a service

List of Figures

Figure 1.1: Wunder mobility business operations	4
Figure 1.2: Mobility as a service - value chain	5
Figure 1.3: Project overview - inputs (blue), outputs (grey) and impact (green)	6
Figure 4.1: Kick-scooter sharing service - unit economics (McKinsey, 2019)	24
Figure 4.2: Elbow method - total within-cluster sum of squares over number of clusters (K)	27
Figure 4.3: Davies-Bouldin index over number of clusters (k)	28
Figure 4.4: Demand forecast methodology - 3 main steps (Koehrsen, 2018)	30
Figure 4.5: Grid cell example for Málaga - open app events aggregated by cell	31
Figure 4.6: Day of the week average demand (20th January - 13th April)	32
Figure 4.7: Weekly demand units over 12 weeks (20th January - 13th April)	33
Figure 4.8: Daily demand units (20th January - 13th April)	33
Figure 4.9: Example of hold-out cross validation (Bergmeir and Benítez, 2012)	34
Figure 4.10: Nested cross validation (Bergmeir and Benítez, 2012)	35
Figure 4.11: Day forward-chaining nested cross-validation (Bergmeir and Benítez, 2012)	36
Figure 5.1: Cluster results - parallel coordinates plot with average cluster values	39
Figure 5.2: Conversion from open app events to demand units	42
Figure 5.3: Conversion from demand units to bookings	42
Figure 5.4: Conversion from demand units to bookings per grid cell with green circles corresponding to deployment points	43
Figure 5.5: Variable importance (mean decrease in accuracy) of most relevant features	44
Figure D.1: Spatial visualisation of clustering results - first view	58
Figure D.2: Spatial visualisation of clustering results - second view	59
Figure E.1: Box plot - daily precipitation (in)	60
Figure E.2: Box plot – daily temperature (°F)	60
Figure E.3: Box plot - daily speed wind (mph)	61
Figure E.4: Box plot – daily humidity (%)	61

List of Tables

Table 1.1: New mobility service trends (Viswanath <i>et al</i> , 2015).....	2
Table 1.2: Urban mobility evolution (Viswanath <i>et al</i> , 2015).....	2
Table 1.3: Mobility industry - Pestel analysis	3
Table 2.1: Vehicle sharing systems segmentation.....	9
Table 3.1: Business relevance summary - operators struggles, value-added and Wunder revenue stream	17
Table 4.1: Constant factors that affect demand	21
Table 4.2: Dynamic factors that affect demand.....	22
Table 4.3: Spearman's rho correlation matrix.....	26
Table 4.4: Demand units algorithm example	29
Table 4.5: Explanatory variables incorporated in the tested model	31
Table 4.6: Explanatory variables that can be incorporated once historical data is longer	32
Table 5.1: Generated ideas, value-added and monetization sorted by priority	38
Table 5.2: Customers segments: average metrics.....	40
Table 5.3: Results of global demand forecast for Málaga.....	44
Table 5.4: Description of most relevant features.....	45
Table 5.5: Results of cell demand forecast for Málaga	45
Table 5.6: Cell results for the two cells with highest and lowest demand	46
Table A.1: Market assessment - variables	53
Table A.2: Variables weight for each customer segment.....	54
Table B.1: Market attractiveness results – first part.....	55
Table B.2: Market attractiveness results - second part.....	55
Table B.3: Assumptions for revenue model	55
Table B.4: Pricing assumptions for revenue model.....	55
Table B.5: Revenue model results - first part.....	56
Table B.6: Revenue model results - second part	56
Table C.1: Ideas explanation, value-added, score and monetization.....	57
Table C.2: Ideas assessment - detailed cost and feasibility scores	57
Table E.1: Statistical analysis of weather explanatory variables.....	60

1 Introduction

The mobility landscape is being completely reshaped, and urban mobility poses a massive challenge to metropolitan authorities and businesses, as well as great opportunities. The global demand for passenger mobility in urbanized areas is expected to double by 2050. Meanwhile, the number of individual journeys taken daily has grown massively since 2015, thereby putting increased pressure on existing urban mobility systems (Audenhove *et al*, 2018).

The mobility industry has not been spared by the recent spate of technological advancements and innovations driven by the fourth industrial revolution. These major technological developments, including big data, artificial intelligence, Internet of Things and the emergence of new, compact forms of energy, have thrown up a range of new mobility options. Regulation is another important driver of sustainable innovation as it is generally geared towards creating the required framework for the sound introduction of new mobility solutions, ensuring those will positively contribute to reaching the optimal system. Finally, customers' expectations for fast, reliable, convenient and individualized mobility solutions are rising as fast as the mix of transport modes and services offered to them, and this trend is likely to continue. People's mobility habits are evolving dramatically, and mobility behaviours are being transformed (Audenhove *et al*, 2018).

1.1 Mobility market

In this context, a wide range of mobility services offer new kinds of transportation alternatives and money is pouring into the sector. Table 1.1 summarizes main mobility service trends, segmenting into individual- and group-based mobility. It highlights how new mobility services are replacing traditional solutions through concepts such as on-demand mobility, peer-to-peer platforms and shared mobility. These referred new mobility services are technology-enabled.

Table 1.1: New mobility service trends (Viswanath *et al*, 2015)

	Traditional mobility solutions	New mobility services	
Individual-based mobility	Private car ownership	Car sharing: peer to peer	A peer-to-peer platform where individuals can rent out their private vehicles when they are not in use
	Taxi	E-hailing	Process of ordering a car or taxi via on-demand app. App matches rider with driver and handles payment
	Rental cars	Car sharing: fleet operator	On-demand short-term car rentals with the vehicle owned and managed by a fleet operator
Group-based mobility	Car pooling	Shared e-hailing	Allows riders going in the same direction to share the car, thereby splitting the fare and lowering the cost
	Public transit	On-demand private shuttles	App and technology enabled shuttle service. Cheaper than a taxi but more convenient than public transit
		Private buses	Shared and Wi-Fi-enabled commuter buses available to the public or to employees of select companies. Used to free riders from driving to work

New and improved mobility services are making transportation even more multimodal, on-demand and shared, increasing consumer choice and convenience. The industry evolution is described in the following Table 1.2, comparing the present and future of urban mobility.

 Table 1.2: Urban mobility evolution (Viswanath *et al*, 2015)

The present and future of urban mobility

From . . .	Toward . . .
Individual car ownership as dominant form of transport	Individual car ownership as one form of multimodal, on-demand, and shared transport
Limited consumer choice and few service levels	More consumer choice and many service levels
Government-funded public transit	Public and private transit operate in parallel
Unconnected, suboptimal, transportation systems	On-demand, connected systems that use data to unlock efficiencies

1.1.1 Shared mobility

Technology is disrupting the mobility industry as a result of four main factors: autonomous driving, connectivity, electrification and sharing. In particular, vehicle sharing concepts are increasingly spreading around the world, providing on-demand mobility for customers while tackling the problem of crowded streets and lack of parking spots in metropolitan areas (Degele *et al*, 2018). Nevertheless, the main purpose behind such mobility solution is to reduce costs.

The shared mobility market in three core regions - China, Europe and the United States - accounted for almost \$54 Billion in 2016, and it should continue to experience impressive

annual growth rates in the future. Under the most positive scenario, which involves strong customer demand for self-driving taxis or shuttles, in low-density locations and in cities that take steps to enable them, the market could see 28 percent annual growth from 2015 to 2030. Even the least aggressive scenario points to steady growth based on convenience and economics; it projects 15 percent annual expansion, even if customers do not readily adopt robo-taxis and cities do not support them (McKinsey, 2019).

The following table summarizes the main trends and conditions for shared mobility operators.

Table 1.3: Mobility industry - Pestel analysis

Group	Factors
<u>Political</u>	- Governments and cities are looking for new mobility solutions to solve congestion issues and address environmental targets - Legal support and incentives for eco-friendly mobility solutions
<u>Economic</u>	- Young generations are looking for more affordable mobility solutions and cannot afford car ownership
<u>Social</u>	- Reducing car ownership - Shift towards a service-based society - People are looking for on-demand solutions
<u>Technological</u>	- Increasing mobile and internet penetration
<u>Environmental</u>	- Concern with environmental issues
<u>Legal</u>	- Lack of regulation for this kind of vehicles - Temporary prohibition and general regulation

1.1.2 Shared micro-mobility

Shared micro-mobility, a segment within the sharing mobility market, has become especially trendy over the last few years. It encompasses all shared electric bicycles, electric scooters (e-scooters) and kick-scooters, either in station-based or free-floating schemes. This market is estimated to account for \$100-150 billion by 2030 just in Europe (McKinsey, 2019).

A recent relevant trend within this segment are the kick-scooters, which started in the USA in 2017 and then expanded worldwide. Investors have poured millions of dollars into kick-scooters operators in 2018, with companies like Lime and Bird raising around \$500 and \$400 million, respectively (Dickey, 2019). On the second half of 2018, these two main American operators expanded into Europe and the market has been growing, with the number of players increasing quite fast (Hughes, 2018).

Operators provide a service on a pay per ride basis and the user can search for available vehicles through their app, after signing in. Once they find a vehicle in a desired location, users can book a ride, unlock the vehicle through a QR code and then ride to whenever they like (within permitted operating area). The price consists of a set-up fee plus a rate per minute. As of April 2019, most European operators charge 1€ for unlocking the vehicle (upfront fee) plus 0,15€ per minute.

The outlook for the micro-mobility is rather uncertain. For this market potential to become a reality, cities need to support shared micro-mobility proactively. They could, for example, boost the micro-mobility business model further to resolve traffic pain points and congestion problems. Actions might include banning cars (but not scooters) from congested or polluted

districts or creating incentives for the use of micro-mobility for short trips by significantly increasing prices for car-based shared mobility. Cities could also install intermodal hubs to make the interchange between micro-mobility and public transport more convenient. However, micro-mobility players must tread carefully, since some cities today have been hesitant to adopt the service. Issues include customers who abandon old or damaged scooters on the street; safety concerns, which still play an important role; and the low entry barriers, which means that competitors could simply invest a bit more money to steal a player's entire customer base. In a nutshell, this micro-mobility phenomenon has the potential to disrupt the industry and their success depends largely on how cities react to the service. Consequently, in addition to building their business, it is expected that micro-mobility players will have to take proactive roles in lobbying for and shaping the industry in key urban areas (Hannon *et al*, 2019).

Besides the critical regulatory issue, with cities trying to find ways to smoothly manage this new mobility agents, industry consolidation is expected to happen soon. As there are not enough major differentiator between services provided by different operators, who are struggling with very low margins, consolidation might be the only solution for them to reach profitability. In addition, given tough unit economics, operators are expected to drive costs down by leveraging economies of scale and experience (Dickey, 2019).

1.2 Company - Wunder Mobility

Wunder Mobility GmbH is a mobility tech company headquartered in Hamburg, Germany, providing Software as a service (SaaS).

It is one key player in the European mobility market by developing software that enables on-demand mobility solutions. As of April 2019, Wunder operates across the globe in more than 50 countries spreaded across all continents. It has three business units corresponding to three different products: Carpool, Shuttle and Fleet. The carpool technology connects passengers with drivers travelling in the same direction with smart route mapping algorithms. The main current markets are Rio de Janeiro, Manila and Delhi. The shuttle product enables clients to launch their own branded shuttle service. Lastly, Fleet provides software and hardware solutions to help companies launch their own free-floating sharing business. It targets the micro-mobility segment, previously described, including various vehicles such as cars, bikes, scooters and kick-scooters. The following Figure 1.1 summarizes Wunder business operations.

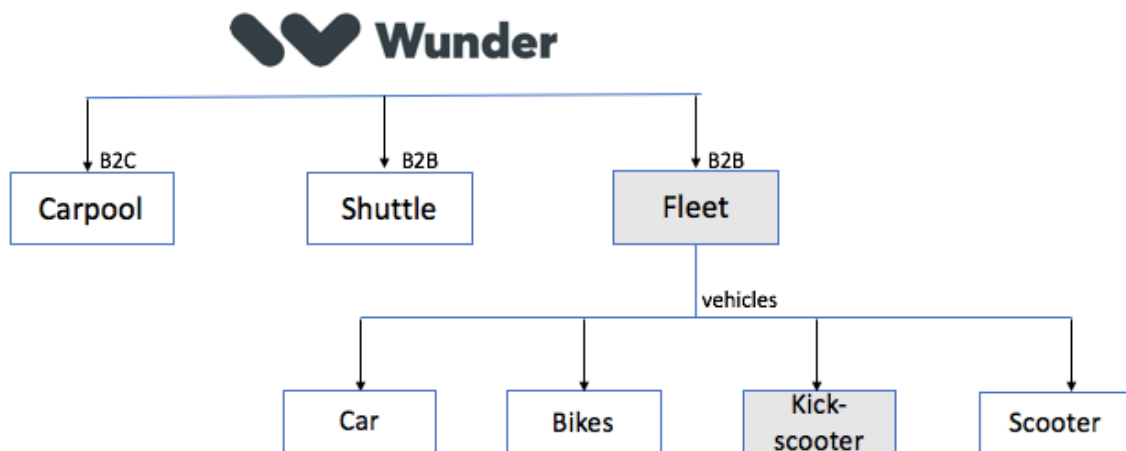


Figure 1.1: Wunder mobility business operations

1.2.1 Fleet business unit

Within the fleet segment, Wunder positions itself between hardware providers or manufacturers and operators. Clients comprise corporations, start-ups and cities or governments. Figure 1.2 illustrates the value chain of this industry.

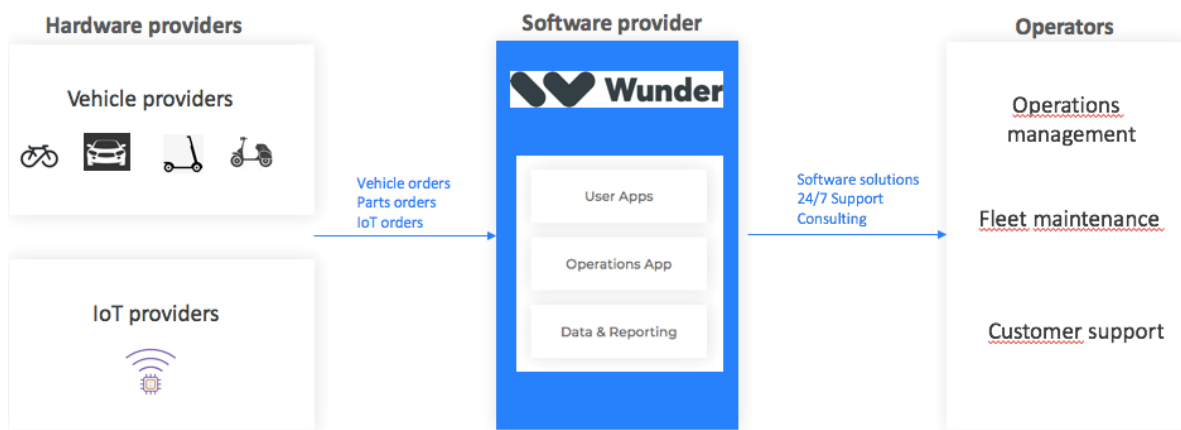


Figure 1.2: Mobility as a service (MaaS) - value chain

Wunder main value proposition is enabling clients to quickly launch their mobility solutions and scale them by providing the software solution and required expertise. Wunder offers full stack software, including customisable user app, operator dashboard and required hardware. It charges new clients an upfront fee and then a monthly fee proportional to the number of operating vehicles (fleet). From the operators' perspective, the associated cost is significantly lower than building an in-house team to develop the required software. In addition, Wunder has already expertise and experience launching such services; thus, is able to advise clients on operations, marketing and strategy.

1.2.2 Project development context

This project took place within the business development department, which is responsible to develop and implement growth opportunities, having around 15 employees. This entails acquiring new clients as well as increasing the value-added to them by providing additional services. This department is divided into three divisions according to the business units and products previously mentioned (carpool, shuttle and fleet). This work was developed within the fleet division, given that it was exclusively focused on this business unit and product. The business expertise and industry knowledge from the team members were key to the project. Moreover, this work was developed in close collaboration with the product department that manages new product features development and implementation. In addition, the engineering team support was needed to get the data required and understand it, as well as key account managers with whom clients' main hurdles were analysed, which should be addressed by this project as a way to increase value-added.

One should highlight that this work adhered to the highest ethical and confidentiality standards, by anonymizing all the data used.

1.3 Project objectives

This project focuses on business analytics for kick-scooters sharing systems, as an attempt to develop new revenue streams for Wunder, namely new product features and consulting services, as well as retaining customers and differentiate Wunder from competitors. The main

outcome is the demand forecast algorithm, as a way to optimize vehicle relocation and gain business insights, enabling data-driven decision making across many areas, such as business expansion, revenue forecast, marketing, etc. This algorithm predicts future daily demand per area-level on a dynamic grid based on machine learning models, with the following inputs: historical daily demand and additional data. App data analytics and customer clustering features are also relevant outcomes of this project. The former depicts customer journey from opening app events to booking a ride and builds demand units from app data, whereas the latter presents a dynamic clustering algorithm and outlines a marketing strategy to address each customer segment based on the marketing mix.

Figure 1.3 summarizes project outputs (grey), required inputs (blue) and project impact (green).

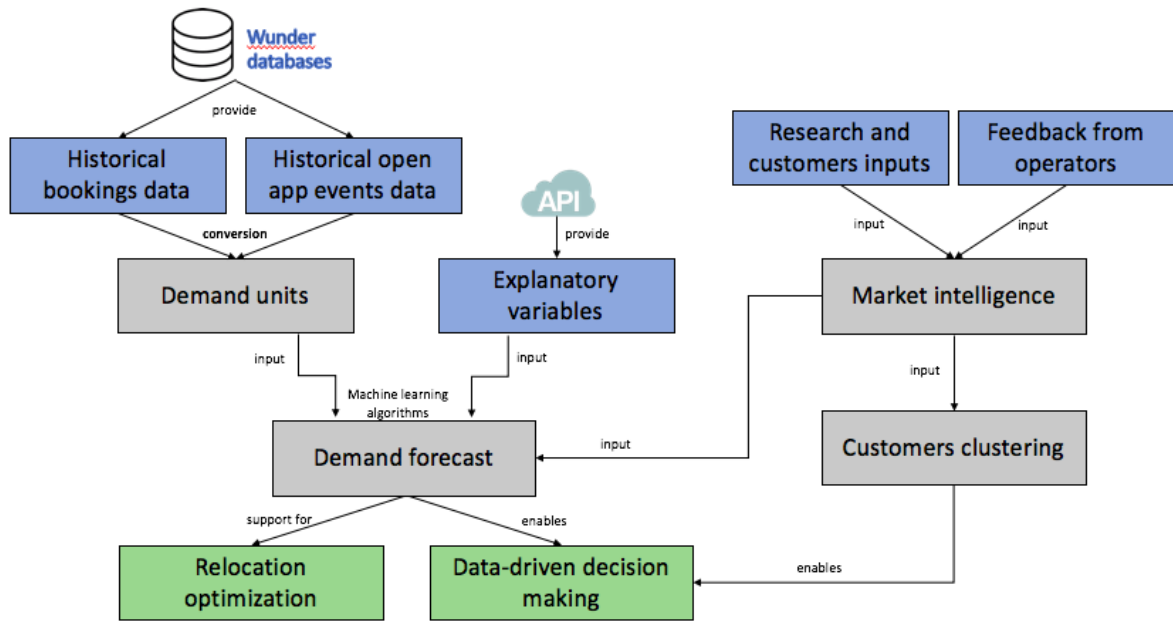


Figure 1.3: Project overview - inputs (blue), outputs (grey) and impact (green)

1.4 Methodology and schedule

This project has four main phases with one corresponding outcome for each: market intelligence study, demand proxy, daily demand forecast models and customer clustering. Firstly, the market intelligence study gives an overview over the global sharing kick-scooter market with a focus on Europe as a rapidly growing segment with special regulation and characteristics, as opposed to the USA, where this kick-scooters trend has begun. It also comprises a revenue forecast model for the main European cities by analysing critical demand drivers. This study seeks to understand how business analytics can yield the biggest impact, by identifying main operators' hurdles and thus maximizing value-added. The main identified struggle consists of very low margins due to high operating costs and low vehicle utilization. Therefore, given their rapid expansion both within current cities and to new cities, operators need to optimize vehicle relocation in order to guarantee that supply meets demand. In this context, a forecast demand model is a key input.

However, given the fact that historical rides are censored demand and many players operate in every city, each operator cannot accurately determine historical demand, as rides are not the best proxy. This led to the second step of this project: determining demand units by using historical app data and bookings data. In addition, this enables app analytics based on the conversion from open app events to bookings, which is a valuable input for app design (UX/UI) improvements.

After building an algorithm to infer demand units from app data and after gathering additional data required, all inputs necessary to forecast demand are collected. By dividing a city into a dynamic grid to aggregate historical data, daily demand for each square is predicted based on machine learning models. This feature also includes interactive visualisation tools.

Lastly, a customer segmentation based on k-means clustering was developed, which allows operators to dynamically segment customers at any point in time. This feature not only yield business insights but also enables operators to develop a targeted marketing strategy aiming at enhancing consumer experience and retain them. In this context, a marketing strategy for each customer segment in one city is also presented.

1.5 Dissertation structure

The remainder of this dissertation is structured as follows: relevant literature is summarized in Section 2, comprising a review on sharing mobility business models in an attempt to contextualize the project, as well as vehicle relocation strategies, demand forecast and customer segmentation for similar mobility sharing models, namely scooters, bikes and cars. Censored demand is also reviewed. Section 3 thoroughly describes the approached problem, including its context, business relevance, problem formulation, used data, restrictions and limitations. Section 4 explains the proposed solution, comprising the market overview analysis, customer clustering, app data analytics and demand forecast model. The approach and methodology are thoroughly presented, as well as developed models and relevant analysis. The results obtained are analysed and interpreted in Section 5. Lastly, main conclusions and future work are presented in Section 6.

2 Literature review

The literature on kick-scooters is very scarce, given how recent this trend is. In fact, to the best of author's knowledge, there is no study on kick-scooter sharing systems with regard to market intelligence, customer segmentation, demand forecast nor vehicle relocation algorithm. However, one can find some relevant research on e-scooters from which is possible to gain insights, given how similar both products and services (on-demand basis) are. In addition, there are even more research on bike-sharing systems, which are also relatively similar. On car-sharing one can find much more research, which might be insightful up to a limited extent, given the considerable difference between both services.

The literature review is structured as follows: overview on mobility sharing business models, trends and customers, followed by a review on customer clustering, especially for mobility sharing systems. Secondly, research on vehicles relocation strategies for e-scooters, bikes and car sharing models is presented. Thorough demand forecast research for such mobility solutions is also carried out, followed by a review of censored demand. To conclude, the research gap is identified, and the proposed scientific contribution of this work is presented.

2.1 Vehicle sharing mobility

Cohen and Kietzmann (2014) focus on business models for the mobility sharing economy, such as on-demand bike and car-sharing systems, evaluating their value propositions, supply chain, customer interface and financial model. It also explores the current context contribution to the development of such business models. This emerging mobility sharing economy is particularly interesting in the context of cities that struggle with population growth and increasing density (Cohen and Kietzmann, 2014). Such mobility solutions can contribute to reduce city congestion and pollution. It also highlights the need of cities and governments to cooperate with such sharing economy business models for them to work. Yang (2010) studies the relation between bike-sharing systems and government institutions, highlighting the required cooperation and incentives alignment for such initiatives to be successful. Situations in both China and Taiwan are explored.

The potential demand and costs of developing an electric vehicle sharing service with several vehicle types is examined by Galatoulas *et al* (2018). It studies how to choose the best pick-up and drop-off points, taking into account several factors, namely access to other means of transportation, population density, proximity to centers of economic activity, among others.

2.2 Customer clustering

To better understand the market where they operate, all companies in general, and companies in the mobility sharing sector in particular, need to study consumer needs and should leverage the gathered data for this purpose. In fact, many of the most interesting and practical applications of analytics are in the field of marketing (France and Ghose, 2019). As examples of potential applications of marketing analytics are: visualisation, clustering and customer segmentation. In this context, segmentation is a core activity for marketers, being required for targeting and positioning, which is crucial for marketing strategy design. From an analytics perspective, segmentation refers to the process of grouping or splitting items using a range of segmentation criteria or bases (France and Ghose, 2019). The following business areas can be leveraged by customer segmentation: branding strategy, customer acquisition, customer insight and retention, digital marketing, product development, product positioning and strategy.

Kim *et al* (2006) propose a framework for analysing customer value and segmenting customers based on their value, followed by a strategy outline customized for each customer segment. As core concept, Kim *et al* (2006) argue that is essential to build refined strategies for customers

based on their value. Therefore, the segmentation is mainly focused on a customer lifetime value model, considering the past contribution, potential value and churn probability of each client at the same time. This segmentation enables more targeted and personalized marketing.

Arunachalam *et al* (2018) evaluates the performance of different data clustering approaches for searching profitable customer segments in the UK hospitality industry using ordinal data. The research design process consists of: data exploration and pre-processing, multi-dimension scaling and visualisation, principal component analysis, cluster analysis, profiling of segments and performance comparison. Four clustering approaches are compared: hierarchical clustering, K-Medoids, fuzzy clustering and Self-Organising Maps. This comparison is based on two cluster validity indices: compactness and separation. While the former measures the closeness of objects within each cluster, the latter evaluates how well separated clusters are. The findings suggest that the Fuzzy clustering and Self-Organising Maps techniques are more efficient in revealing hidden structure in the data set.

Degele *et al* (2018) and Morency *et al* (2011) provide insights on sharing mobility services customer for e-scooter in Germany and car-sharing in Canada, respectively. Degele *et al* (2018) segment e-scooters customers based on k-means clustering technique using real data from german operators, namely demographic (age and gender) and behavioural variables. Each customer cluster is described as regards to usage frequency and patterns (average time between rides, average distance and most common usage day) and customer age. It also evaluates the impact of customer segmentation on business development, namely on the marketing mix (product, price, place and promotion).

As future work, Degele *et al* (2018) acknowledge the importance of a geographic analysis regarding the spatial distribution of rentals, taking into account the surrounding infrastructure, which could give insights on the demand per area. It is also mentioned the optimization of business area as a way to increase revenue.

Morency *et al* (2011) carry a similar study for car-sharing users, by focusing on the systematic analysis of members behaviour in terms of transactions and distance travelled. Data mining techniques are used to classify members according to temporal units expressing their behaviours.

2.3 Fleet relocation strategy

Related literature for vehicle sharing systems can be segmented based on the characteristics summarized in Table 2.1. The detailed explanation is presented after.

Table 2.1: Vehicle sharing systems segmentation

Relocation responsibility		System types			Relocation frequency	
User	Operator	Free-floating	Station-based	Hybrid	Static	Dynamic

Related literature for bike and car-sharing studies can be divided into two streams: user and operator-focused. The former is based on users' reactions and the relocation process is carried out by users, whereas the operator-based approach is based on the interventions made by an operator. Weikl and Bogenberger (2013) study relocation strategies and algorithms for free-floating carsharing systems. They also evaluate the pros and cons of user-focused against operator-focused relocation. As the whole system has to be profitable for the provider, additional costs and benefits have to be compared. If the relocation costs compensate the additional earnings, a relocation strategy may be applied (Weikl and Bogenberger, 2013).

There are three main system types regardless of the vehicle: free-floating, station-based and hybrid. On the one hand, free-floating systems (also referred as dock-less) offer a higher degree

of flexibility to customers as vehicles can be parked anywhere inside the operating area and no stations exist (Albi *et al*, 2018). This model increases flexibility on the consumer side but also complexity in terms of balancing supply and demand on the provider side (Vogel *et al*, 2011). On the other hand, in station-based systems, vehicles are rented at predefined stations and also returned to a station after usage. The positioning of such stations at public transport stops or points of interest guarantees a better availability of bikes at highly frequented locations. This makes the system more attractive to customers, as the search for vehicles can be time-consuming even if smartphone apps indicate the position of bikes on a map. Hybrid systems are a combination of free-floating and station-based, combining the convenience for the user of the former with the attractiveness of the latter (Albi *et al*, 2018).

Regarding relocation timings, there are two different modes: one is during the night when the usage rate of the system is negligible; the other is during the day when the status of the system is rapidly changing (Raviv *et al*, 2013). For bike sharing systems, the former is referred as the static bicycle repositioning problem (SBRP) and the latter as the dynamic bicycle repositioning problem (DBRP). Some operators use static repositioning, some dynamic, and some use a combination of the two (Raviv *et al*, 2013).

Furthermore, problem formulations can either assume stochastic or deterministic demand. The former tries to model the demand or number of arrivals based on a distribution (mainly Poisson), whereas the latter develops an explanatory model to predict the demand.

Wuarrington and Ruchti (2019) present a generic model suitable for different shared mobility systems, both free-floating and station-based and also for different types of vehicles, namely bikes, scooters and cars. The model is based on a two-stage stochastic approximation for dynamic rebalancing, assuming stochastic demand, where the number of clients arrivals follows a Poisson distribution. For kick-scooters or “micro” scooters, as referred in the paper, they distinguish themselves from bikes in that, owing to their small size, large numbers can in principle be repositioned by a single manned vehicle, i.e., no upper bound on the rebalancing vehicles capacity is necessary (Wuarrington and Ruchti, 2019).

Specific literature for each type of system is presented.

2.3.1 Station-based systems

Research on relocation strategies has been mostly focused on station-based systems. Jorge *et al* (2014) develop a mathematical model to optimize relocation operations that maximize the profitability of the carsharing service. The model is applied to a Lisbon case study.

For bike sharing systems, several papers focused on vehicles relocation: Aeschbach *et al* (2015), Legros (2019) and Dell *et al* (2018). For London, Aeschbach *et al* (2015) evaluate several methods to balance such systems by engaging customers. In fact, it is shown that by appropriately sending control signals to customers requesting them to slightly change their intended journeys, bike sharing systems can be balanced without using staffed trucks (Aeschbach *et al*, 2015). Using a Markov decision approach, Legros (2019) presents a decision-support tool, which help the operator to decide at any point in time how many bikes should be added or removed at each station. Assuming stochastic demand, Dell *et al* (2018) approach the rebalancing problem of a station-based bike system, aiming at determining minimum cost routes for a fleet of homogeneous vehicles in order to redistribute bikes among stations. Faghih-imani *et al* (2017) try to identify factors that drive bike-sharing systems demand. By collecting station-level occupancy data from two Spanish cities, they analyse station level customer arrivals and departures. A model is developed to estimate the influence of bicycle infrastructure, socio-demographic characteristics and land-use properties on customer arrivals and departures. The number of rebalancing bikes is estimated through a regression model. This takes into account points of interest density in each area analysed.

Ho and Szeto (2014) present an iterated tabu search heuristic to solve a static bike repositioning problem by selecting a subset of stations to visit, sequencing them and determining the pick-up/drop-off quantities under various operational constraints. The goal is to minimize total penalty costs incurred at all stations.

A framework to solve the on-demand dynamic bike sharing repositioning problem based on four core models is proposed by Regue and Recker (2014): a demand forecasting model, a station inventory model, a redistribution needs model and a vehicle-routing model.

2.3.2 Free-floating systems

Research on free-floating systems relocation strategies is not as vast as for station-based, given that such systems are more recent and not so widespread. However, Albi *et al* (2018) conclude that operators could increase the rent outs by making more free-standing bikes available. For carsharing, Weikl and Bogenberger (2013) present a new integrated two-step model for optimal vehicle positioning and relocation. This new approach consists of an offline demand prediction, followed by an optimization algorithm that finds optimal relocation strategies if necessary (Weikl and Bogenberger, 2013).

For Munich free-floating bike sharing systems, Reiss and Klaus Bogenberger (2017) study how to optimize vehicles relocation based on a detailed GPS-Data analysis. They developed a demand model in order to forecast the upcoming demand at a certain time and place, which incorporates different factors like season, weather conditions, time of the day, holidays and weekend. This model reveals optimal fleet distributions for different zones in the operating area and for different time slots.

Caggiani *et al* (2018) propose a new comprehensive dynamic bike redistribution methodology that starts from the prediction of the number and position of bikes over a system operating area and ends with a relocation decision support system. The relocation process is activated at constant gap times in order to carry out dynamic bike redistribution. The method consists on a spatio-temporal clustering, followed by a non-linear autoregressive neural network.

2.3.3 Hybrid systems

Research on hybrid systems is very limited, given that their application is not yet so widespread. Albi *et al* (2018) carry out a usage pattern and demand analysis of such bike systems. They conclude that a reposition policy could increase the booking number of free-standing bikes and thus customer satisfaction as well as system's profitability.

2.4 Demand forecast

Several authors have studied how to predict bike-sharing demand on three spatial granularity levels: city, cluster and station.

Regarding forecasting models, Regue and Recker (2014) show that machine learning explanatory methods such as neural networks and bayesian network yield better results than time series techniques, for similar problems. Studies that test for weather and holidays showed improvements of more than 50% in predictive power when those factors are included (Regue and Recker, 2014). In fact, Regue and Recker (2014) conclude that fewer trips are made in rainy, humid, windy and cold conditions. In terms of future work, Regue and Recker (2014) mention that the forecasting model should incorporate more independent variables such as large events (e.g., basketball games, concerts, etc.), transit arrivals and departures in areas nearby bike-sharing stations or real time distribution of people in the city.

2.4.1 City-level

Several researchers have tested multiple machine learning algorithms such as Ridge Regression, Adaboost Regression, Support Vector Regression, Random Forest Tree and Gradient Boosting Regression Tree and showed that the first two outperformed the others (Lin *et al*, 2018). Although predicting the total city-level rentals from all bike-sharing stations simplifies the problem greatly, it does not contribute to solving the bike rebalancing problem among stations (Lin *et al*, 2018).

2.4.2 Cluster-level

Vogel *et al* (2011) use data mining to analyse free-floating car-sharing in Berlin and forecast demand on a given area based on a regression model, which incorporates several factors related with the presence of stores, ATM, airport, etc in the neighbourhood.

Wagner *et al* (2016) develop a novel method that provides decision support to free-floating car-sharing operators. Using an extensive set of customer data in a zero-inflated regression model, they explain spatial variation in carsharing activity through environmental data, namely points of interest (such as movie theatres and airports), as well as demographic data (such as education level, population density and age distribution). To validate the model, the outcome indicators are used to predict the number of rentals before an expansion of the business area and this forecast is compared against the actual demand post-expansion. The developed approach succeeds in identifying areas with high carsharing activity and can be easily adapted to other cities, enabling data-driven expansion strategy.

2.4.3 Station-level

Station-level bike-sharing demand prediction is more challenging (Lin *et al*, 2018) and has been a popular research topic. In fact, Chang *et al* (2019) and Kang *et al* (2017) apply a wide range of machine learning techniques to predict station-level demand. Hulot *et al* (2018) have a different approach by using Markov chain based demand. Chen *et al* (2017) apply machine learning techniques to predict both rental and return demand for every station in New York.

Lozano *et al* (2018) highlight in their future research the importance to include the new dock-less (free-floating) bike sharing systems into their model, suggesting grouping the deployed fleet of bikes to create clusters, which would act as virtual stations in the city.

2.5 Censored demand

Most literature around relocation strategy neglects the fact of censored demand. If a station is out of stock, the true demand might be higher and forecasting models might underestimate the actual demand (Albi *et al*, 2018). To highlight the impact of censored demand, Albi *et al* (2018) concluded that when ignoring censored demand, the real demand is underestimated, whereas the service level is overestimated.

In fact, censored demand is a common problem to many industries and sectors. For example, Mersereau (2014) and Jain *et al* (2014) investigate the impact of censored sales historical data on demand prediction in retail. Stefanescu (2016) propose a class of models for multi-product multi-period aggregate demand forecasting, taking the impact of censored demand into account. The models are applied to two different industries: entertainment and airline.

For the hybrid bike sharing system in Munich, Albi *et al* (2018) include the effects of censoring demand in a service level analysis, when analysing demand patterns. If demand is not corrected for lost sales, the service levels in the high demand periods (or lower supplied areas) are overestimated (Albi *et al*, 2018). Consequently, the provider is not aware of the additional

customers who could be served with more bikes (Albi *et al*, 2018), which shows the importance of taking censored demand into account.

2.6 Research gap

There is relevant research on analytics support for car and bike sharing systems. However, the literature applicable for kick-scooters is very limited. As a result, this master thesis aims at contributing to the scientific knowledge of both customer clustering, demand forecast and relocation strategy for mobility sharing systems, especially kick-scooters. This work is based on an operator based, hybrid and static relocation context. In other words, operators must decide at the end of the day how to relocate the kick-scooters for the next day, with a combination of free-floating and station-based vehicles.

The proposed demand forecast model will enable data-driven decision making as regards not only to vehicle relocation but also business expansion strategy. This model presents an innovative way to deal with censored demand, which is neglected in most literature around demand forecast and relocation strategies.

3 Problem definition

3.1 Context

Kick-scooters sharing systems are a very recent mobility offer in Europe that is rapidly expanding and growing fuelled by considerable amount of investment raised by the main market players. Given the scarce available data and lack of resources, namely time and human resources, European operators are not leveraging analytics to increase profitability. On the one hand, they are missing business insights on what drives demand, customer types and behaviour, which are crucial for business strategy. On the other hand, they do not have a data-driven strategy to optimize operations, especially fleet relocation, which has to be deployed every single day.

Given the intense competitive landscape and the growing number of operators, making the right strategic decisions on where to invest as well as optimize operations and marketing campaigns to achieve profitability are critical factors for operators' success. Such factors will decide whether companies thrive in this market.

Regarding demand drivers, a few factors like weather and city infrastructure are evident. Moreover, many conclusions can be drawn from similar mobility offers, like bicycle and scooter sharing systems. However, only a data-driven approach allows to quantify the factors relevance and impact on consumer demand. In addition, with the growing sharing mobility offers, including bicycle, scooter and car sharing, both direct and indirect competition have to be taken into account when predicting demand.

Regarding operations, daily fleet charging and relocation is responsibility of operators in Europe. They currently subcontract this service to a third provider, which is responsible for collecting all vehicles at the end of day to check their status, charge them and run maintenance, if necessary. They deploy them at the beginning of the next day following operators' instructions. As of information collected in April 2019, kick-scooters are deployed in groups of 5 to 15 units in most cities. According to the feedback received by two European operators, they do not have a clear relocations strategy, which is developed manually on a weekly basis following a trial and error approach based on historical rides heatmaps. They do not have any kind of demand forecast algorithm either.

3.2 Problem formulation

Within the global scope of gain business insights to enable data-driven decision making and optimizing operations, especially vehicles daily relocation, there are four specific problems thoroughly defined below.

3.2.1 Market intelligence study

This first part characterizes the sharing kick-scooters market, by giving an overview into service, customers, players and market trends, followed by an industry analysis into barriers of entry and profitability. This study intends to provide general knowledge, especially for the European market, enabling further analysis and a market attractiveness assessment of main European cities. It is crucial to understand what the analytics focus should be in order to maximize its business impact.

Besides a generic market overview, the main research questions are:

- Which factors influence demand?
- Which European cities are more attractive to launch such mobility systems?
- What main struggles do operators face and how to prioritize feature development?

3.2.2 Customers analytics - segmentation

Following the market intelligence study, a deep dive analysis on customers is carried out using app and bookings data. The main research questions are:

- How to segment customers and characterize each group?
- What are the special needs of each segment?
- How to redefine strategy based on customer segments?
- How to design the Marketing Mix (Price, Promotion, Place and Product) for each customer segment?

3.2.3 Build demand proxy from historical rides and app data

In an attempt to increase customer retention and as a required input for business-driven demand forecast, the following research questions are proposed:

- How to define the best demand proxy from historical rides and app data?
- What does the funnel from open apps to booking rides look like?
- How can this data be used to optimize the app UX/UI and increase conversion from open app events to bookings?

3.2.4 Forecast daily demand per area-level

Given the operators daily relocation context, they are interested in having a daily demand forecast per area level based on a dynamic city grid, which divides a city area into cells whose dimensions can be defined by the user and thus can be easily adapted to every city. One example of such grid can be consulted in Figure 4.5. One should note that the operators are not interested in dynamic relocation throughout the day, as the additional revenue would not pay off the relocation cost, given the lower ride price of kick-scooters in comparison with car or bike sharing systems.

Based on the literature review, this relocation problem could be described as a static daily relocation of a hybrid system, which is a combination of station-based and free-floating systems. There are a certain number of docks, where kick-scooters can be rented and returned (station-based system). However, they can also be parked and rented anywhere in the operating area (free-floating system).

As a key input for this relocation problem, daily demand forecast pretends to answer the following research questions:

- How to deal with censored demand?
- To what extent and how can demand be forecasted with a good level of accuracy?
- What is the best model and technique to forecast demand?
- Which factors influence demand and what is their impact?

3.3 Business relevance

As a software provider company, Wunder must try to solve customers' biggest pains and work closely with them in order to provide the best service possible. This task is assigned to the business development department. Wunder success is dependent on its clients' success (operators). Therefore, by providing services that contribute to their success, it increases the value-added and revenue. Moreover, by better understanding the end-consumer and market, Wunder can provide a better service and truly differentiate itself from competitors as a market expert. For instance, it can launch new product features based on the gained insights.

In addition, it can start to offer consulting services to clients by leveraging industry expertise. One should note that given the industry age and its tremendous growth, there is not much available information on consumers.

As additional product features, a customer segmentation tool, app funnel analysis and demand forecast can represent key future revenue streams for Wunder. Furthermore, the developed algorithms can be adapted to other vehicles within fleet or even to other business units.

According to industry reports, artificial intelligence is the key to drive profitability in the micro-mobility sector. Main applications are fleet management (predictive maintenance), customer experience enhancement and targeted marketing, demand forecast and fleet relocation optimization. This thesis was the pioneer artificial intelligence project at Wunder, which is eager to expand such services in the near future.

The problem relevance for operators can be segmented into the following categories.

3.3.1 Operators profitability

A relocation optimization strategy would enable operators not only to grow their revenue, but also their profitability by increasing vehicle utilization (number of rides per vehicle per day). In fact, reaching profitability is the main struggle for current operators, mainly due to high operation costs (Azevedo and Wilhelm, 2019). Unit economics reinforce the idea that vehicle utilization (number of rides per vehicle per day) is a key industry profitability driver, since this kind of business revenue models only make sense with high utilization (Wilhelm, 2018). On the operation level, kick-scooters must be recharged and relocated in specific areas, where they are usually deployed in groups. Operators urge to have a smarter relocation strategy to make sure that kick-scooters are deployed in the best stations as a way to increase rides and thus profitability.

As an example, the operator used as case study had an average vehicle utilization of 2,2 rides per vehicle per day between January and March 2019, which is extremely low. According to industry benchmarks, this metric should be between 5 and 6 for operators to reach profitability (McKinsey, 2019).

It is also important to highlight the potential scalability of the model, which can be applied to every city and potentially contribute to optimize other sharing mobility systems.

3.3.2 Operators competitive advantage

Given the intense competition among operators and the high number of new players in the market, fleet rebalancing can be a differentiator, ensuring competitive advantage over competitors. By deploying the vehicles at the right place, operators can ensure customer retention and this way avoid losing customers to competitors. In fact, consumer surveys showed that most users only have two kick-scooters apps installed and are prone to uninstall one provider app when they cannot find a vehicle nearby after a few attempts. Moreover, by engaging with customers through targeted marketing powered by analytics, operators are more likely to retain customers and increase revenue.

3.3.3 Data driven expansion strategy

By having a dynamic demand forecast based on explanatory models, operators can understand what factors affect demand and quantify their impact. Consequently, companies can better design their expansion strategy, since they are able to better assess market attractiveness based on the identified factors and their relevance.

3.3.4 Business insights

Explanatory models yield business insights, which may help operators to better understand customers and identify most attractive markets. In fact, based on interviews conducted with operators, one can conclude that they are eager to understand what drives demand, which is crucial to better understand the market and be able to meet customer needs.

Moreover, this thesis contributes to a more accurate revenue forecast, with impacts on budgeting, resource allocation and financial management. Furthermore, this dynamic demand forecast model can provide insights for other similar mobility sharing solutions.

The main business advantages of every research topic are summarized in Table 3.1, comprising the corresponding operators struggle, value-added and potential revenue stream.

Table 3.1: Business relevance summary - operators struggles, value-added and Wunder revenue stream

<u>Research question</u>	Market intelligence	Customers segmentation	Demand proxy	Demand forecast
Operators struggle	Lack of market knowledge	Lack of customers knowledge and absence of targeted marketing	Lack of app analytics	Relocation algorithm & business insights
Value for operators	Strategy insights for business expansion	Data-driven marketing & business insights	App funnel analytics	Increase in vehicle utilization & business knowledge
Wunder - revenue stream	- Consulting services - Operations support	- Consulting services - Revenue stream from analytics	- Additional product features	- Additional product features

3.4 Available data

This project has two available data sources - historical bookings and open app events - from one main European operator present in several cities across the continent and rapidly expanding, both within current operation areas and also to new cities. Market share data regarding this operator is not available. One should note that although this operator is a relevant player in every city it operates, it is usually not the biggest player.

The used data sets are from 01/01/2019 to 27/04/2019, given that their operations only started by the end of 2018.

3.4.1 Bookings dataset

This dataset has the following variables:

- Reservation_Id: booking unique identifier (id);
- Car_Id: vehicle unique identifier (id);
- customer_Id: customer unique identifier (id);
- drivenDistance: travelled distance (km);
- startTime, endTime: date and time of ride start and finish (DD/MM/YYYY HH:MM);
- startLat, startLon: Latitude and longitude of starting point;
- endLat, endLon: Latitude and longitude of ending point;
- startCity: name of the city (string);
- Cost: price paid for the ride (€);

- `usageTime`: usage total time (min).

3.4.2 Open app events dataset

This dataset has the following variables:

- `Tracking_Id`: booking unique identifier (id);
- `Content`: string with latitude, longitude and `customerId`;
- `Timestamp`: date and time of open app event (DD/MM/YYYY HH:MM).

It is important to highlight that non logged in users have the variable *Content* empty.

3.5 Project context limitations

There are various inherent context limitations and constraints on the work developed, namely with available data and business context.

Given how recent this industry is, there isn't much available reliable data sources and information is scarce. Moreover, since most Wunder clients operate kick-scooters since December 2018, the available dataset is very short as of April 2019. This general context is a clear limitation for this project and specific limitations for every mentioned problem are presented.

3.5.1 Customers segmentation

General Data Protection Regulation (GDPR) prevents this work from including demographic variables in customer clustering. For instance, variables such as gender, age and nationality cannot be used as clustering variables. It impedes, for example, to identify tourists through user's mobile number phone country code. Consequently, customers segmentation has to be solely based on usage cases from rides data, such as frequency, ride duration, time of day, day of the week, etc.

3.5.2 Demand proxy and forecast

Available data

The main forecast constraint is related with the available data. Given how recent kick-scooters sharing models are, there is a general growth trend in most cities. The market is not yet mature and so rides generally increase over time as it becomes more widespread in operating cities and due to fleet expansion by operators.

Regarding historical bookings data, one should note that the starting points are dependent on available vehicle location, which depends both on operators' daily fleet deployment and end points of users' rides. Thus, is a censored demand context, given that demand might be higher than actual rides, in case potential customers could not find an available vehicle. Moreover, operators used promo codes as a marketing tool to engage new customers, resulting in many single-ride users. This effect has also to be considered.

As regards to app data, it is clearly not directly demand, as people might open the app just for curiosity or just to check where vehicles are. It is also influenced by vehicle relocation, as users might open the app when they see a vehicle either to spontaneously rent it or just for curiosity.

Given that Wunder does not serve all operators in the market, it is only possible to have access to partial data for a city rather than global rides data. Consequently, in addition to indirect competition (bike and scooter sharing systems), direct competition (other kick-scooters operators) have to be considered in any developed solution. For instance, the demand in a

certain area of a city may be fulfilled by a competitor and cannot be incorporated in the developed model.

Taking into account the dataset length, several limitations have to be pointed out as regards to the demand forecast model. As data is only available from December 2018 onwards, one cannot have yearly data. Therefore, effects like season and monthly seasonality cannot be yet incorporated in the model. Season can considerably affect demand not only due to weather (which the model incorporates) but also people free time, number of tourists, citizens' mood, holidays effect, etc. This constraint is further developed in the methodology section.

Moreover, one needs to consider the adoption rate of such technology, which involves a new way of mobility. With such disruptive business models, demand tends to grow over time until maturity is reached. The exploratory analysis evaluates this effect. The proposed model considers these data limitations and tries to overcome them.

Operations

Operations specificities and unique events have also to be considered, namely downtime and vehicles availability (due to breakdowns or hardware issues). It can occur server downtimes as well as operational downtime. These might lead to outliers in the dataset, which should be cleaned in the data pre-processing phase.

Regulation

Regulation plays a very important role in this business, given how disruptive for the daily transit of cities this mobility service is. Cities have been trying to regulate it to ensure citizens safety and avoid accidents. Therefore, city councils have taken measures such as: regulate business areas, limit the number of vehicles, introduce city-permit fees, or even completely ban such systems. All city-specific regulation initiatives that may have affected demand, such as forbidden areas or business area expansion, have to be considered.

4 Proposed solution

This chapter presents the proposed methodology and solution for each research question. Main results and corresponding interpretation are discussed in the following chapter 5.

Apart from the market intelligence study (Section 4.1), which is focused on Europe, the proposed solution is tested on the city of Málaga, using data from one main European operator.

4.1 Market intelligence study

This market intelligence study is not only a valuable output of this project but also a key input for the rest of the work developed. Firstly, factors that determine global demand are listed based on market expertise from company employees and feedback from operators and customers. Secondly, a market attractiveness framework is developed to assess which European cities are more promising as regards to kick-scooters sharing services. Thirdly, an industry evaluation regarding profitability and evolution is carried out aiming at identifying key operators pain points. Lastly, a feature development framework is explored, as a way to decide which product features and services should be prioritized.

4.1.1 Customers groups and demand factors

By analysing customers behaviour in several European cities, three main usage cases are identified: joy riders, commuters and tourists. The first group rides for fun and joy, whereas commuters use the service for running errands and commuting. Tourists use the service as an amusing way to move around the city and also to try it out.

Based on market expertise gained from research and feedback from operators, all relevant factors that affect daily market demand are listed. These can be segmented into constant and dynamic. The former factors remain relatively constant over time and can be further segmented into city and area-specific, whereas dynamic factors change over time and are mainly city-wide. Table 4.1 lists constant factors, whereas Table 4.2 enumerates dynamic factors. A link between the factors which influence the demand and each customer segment previously described is established in these two tables.

Table 4.1: Constant factors that affect demand

				Customer segments		
Type	Factor	Granularity	Explanation	Joy riders	Commuters	Tourists
Structural	Infrastructure	Area	Kick-scooters usage requires a proper pavement and dedicated lanes	x	x	x
Structural	Regulation	Area	Cities have adopted regulation measures such as area regulation and even prohibition	x	x	x
Market potential	Foot traffic	Area	High foot traffic is likely to generate more rides (higher market potential)	x	x	x
Market potential	Local purchasing power	City	Given the constant pricing in Europe, high purchasing power regions are more likely to ride more	x	x	
Market potential	Local tech savviness	City	Since the service involves using an app, the potential market is higher for tech savvies populations	x	x	
Market potential	City size	City	Bigger cities involve higher travel distances and are less walkable		x	x
Market potential	Population density	City	Highly populated cities tend to have congestion problems and urge to new mobility solutions		x	
Market potential	Commuting expenses	City	The more expensive other transportation options are, the more likely people search for cheaper options		x	
Market potential	Traffic congestion level	City	People are more prone to search for commuting alternatives in highly congested cities		x	
Market potential	Pollution level	City	People are more prone to search for greener commuting alternatives in highly polluted cities		x	
Market potential	Commuting habits	City	People commuting habits (car, public transportation, bicycle and walking) affect potential		x	
Environment	Neighbourhood attractiveness	Area	The presence of POI's (transportation options, leisure spots, tourist attractions, etc) influence potential demand	x	x	x

Table 4.2: Dynamic factors that affect demand

Type	Factor	Granularity	Explanation	Customer segments		
				Joy riders	Commuters	Tourists
Public Transportation	Presence of bus, train and metro stops	Area	Kick-scooters might provide the last mile solution for public transportation	x	x	x
Competition	Presence of bike, cars and scooters sharing systems	Area	Other sharing mobility services, especially bike and scooters, are direct competitors for kick-scooters	x	x	x
Weather	Rain / Temperature	City	People are more likely to use to use the service in warmer sunny days	x	x	x
Seasonality	Season	City	People tend to have more free time in summer, for example	x	x	x
Seasonality	Day of the week	City	People's free time varies over the day of the week	x	x	
Seasonality	Time of the day	City	People's free time varies over the day	x	x	x
Tourism	Tourism level	City	The presence of tourists increases the potential market, especially given they are a frequent customer type			x

One should highlight that the competitive landscape obviously affects the demand of a specific market player. However, this analysis focuses on total market demand rather than on one market player's demand.

4.1.2 European cities - market attractiveness

Before estimating each city market size, a model is developed to assess cities' attractiveness for each customer segment previously introduced. This model considers the relevant factors for each customer segment, as presented in Table 4.1 and Table 4.2. The attractiveness is measured *per capita* and thus does not take into account the potential market size. As inputs, the user must specify the breakdown of customers per segment and also the weight of each factor for every customer group. The model output is a normalized score between -1 and 1 corresponding to the weighted average of the attractiveness score of each customer segment.

After listing all relevant factors for each customer segment, it is necessary to gather data that represent each variable as accurate as possible. Each factor might have one or more variables associated. For instance, the weather factor has rainy days per year and sunny hours per year as variables. Given the lack of credible data for a few variables, one needs to find other data sources that represent the variable to the maximum extent possible. All variables have to be normalized to allow building the score model.

The developed model not only allows to evaluate the market attractiveness of each city, but also gives insights on which customer(s) segment(s) companies should target in each market, which is very valuable for their expansion strategy.

The detailed list of considered factors and corresponding proxies can be consulted in Appendix A, more specifically Table A.1. Results for a specific inputs combination are presented in Section 5.1.

Based on the market attractiveness model, one can estimate each city market size, also for each customer segment previously introduced. This model considers the potential market and then estimates the adoption rate based on the score model, as the following formula shows.

$$\text{Market size (\# rides/day)} = \text{Potential market (\# people)} \times \text{adoption rate (rides/person/day)}$$

Potential market for joy riders and commuters segments is the metropolitan area population, whereas for tourists is the yearly tourist arrivals. It was only considered the population segment between 18 and 50 years old, given that driver's license are required in most operating countries and people over 50 are very unlikely to use the service.

One should highlight that the model involves two main assumptions. Firstly, regulatory implications are assumed to only impact costs and thus are not considered. Secondly, demand is considered the unique limiting factor. In other words, the model assumes that supply will be enough to meet demand.

As required inputs, the user must specify the average global adoption rate (share of potential market) for each customer segment, from which the model calculates the adoption rate in every city based on its previously computed attractiveness score. The model assumes that commuters use the service from Monday to Friday (5 days per week), whereas joy riders only on weekends (2 days per week).

This market size (number of rides per day) can be converted to monetary value by assuming an average revenue per ride based on a given pricing (set-up fee plus price per minute) and average ride duration (minutes). One should highlight that this forecast is developed for 2019. To predict for future years, one has to consider adoption rate and growth, which are not included in this model.

To consult thorough computation logic and results, the reader is directed to Appendix B, specifically Table B.3 and Table B.4. Main results and conclusions are discussed in Section 5.1.

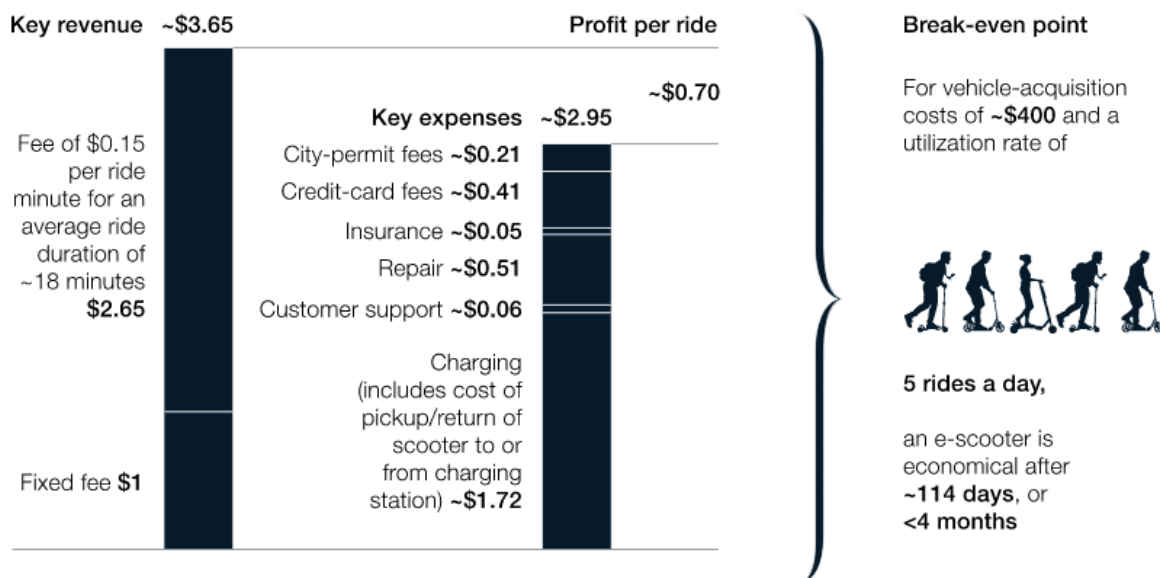
4.1.3 Current industry evaluation

After raising considerable amounts of investment, main kick-scooters operators have been focused on growing and scaling to become global players with relevant market share. However, concerns regarding profitability started to surge in 2019 and they are looking for ways to increase profit margins aiming at making their businesses more sustainable.

The kick-scooter sharing market has low entry barriers, given the low investment required, which means that competitors could simply invest a bit more money to steal a player's entire customer base. Consequently, there is an intense and growing competition, especially in Europe. This affects operators' profitability, which is extremely low in main European cities, as of April 2019. Figure 4.1 shows the unit economics for the American market, where the business looks profitable for a vehicle utilization of 5 rides per day (McKinsey, 2019). However, the main European operators have a vehicle utilization around 2, as of April 2019, which is clearly unprofitable.

An e-scooter is economical after four months.

Revenue-and-expense estimate, per e-scooter ride, \$



McKinsey&Company | Source: Expert interviews; McKinsey analysis

Figure 4.1: Kick-scooter sharing service - unit economics (McKinsey, 2019)

This is the biggest pain of European operators, who urge to make the business more profitable by increasing revenue and decreasing costs. On the revenue side, operators point out targeted marketing campaigns and fleet rebalancing as key measures to generate more revenue. On the cost side, fleet maintenance management and economies of scale are expected to drive costs down (McKinsey, 2019).

Furthermore, one key trend is regulation, as already explored. Operators may have to adapt their offering to meet local laws and needs (McKinsey, 2019). For example, ensuring customer safety by limiting speed in certain areas and finding ways to enforce customers to park the vehicle in a proper place.

4.1.4 Feature development framework

Having identified operators pain points and struggles, it is possible to start the process of feature development. The first stage is called ideation, which consists of trying to come up with product features or services that meet their needs.

To evaluate which product features should be prioritized after ideation, several factors have to be considered. First of all, one should clearly identify the associated value-added, which can be either reducing costs or increase revenue. Secondly, the cost and feasibility of each idea have to be evaluated, namely:

Cost

- Development: cost to develop the product or service;
- Testing: cost to test the product or service before starting to sell it;
- Selling: incremental cost to market the product or service and sell it;

Feasibility

- Market potential: market size and its growth potential;

- **Product/Service:** to what extent is the product or service within the company's core competencies and whether the company has the required expertise;
- **Competition:** direct competitors and potential competitors;
- **Clients:** whether current clients are willing to pay for this product or service and if they will be interested in the future;
- **Synergies and cannibalization:** potential synergies and cannibalization with other services and other business units.

Lastly, the monetization strategy must be defined. In this context, it can be either an additional fee per vehicle, fee per app user or a service (one-time) fee.

A score based on a 1-5 scale (5 being the best) is used to compare ideas and select the one(s) which should be prioritized. Main generated ideas and corresponding evaluation are presented in Section 5.1.

4.2 Customer segmentation

The customer segmentation algorithm methodology, which aims at providing business insights and enabling targeted marketing campaigns, is thoroughly described in this section. One should recall that demographic variables cannot be incorporated due to GDPR regulation, as previously explained.

Market segmentation is the process used to divide heterogeneous markets into smaller markets that can be reached more efficiently and effectively with offerings that match their unique needs and wants (Kim *et al*, 2006). Consumers within the segment should be similar to one another in one or more characteristics and, particularly, in terms of their needs and wants. An effective segmentation process requires three steps (Kim *et al*, 2006):

1. **Segmentation:** choose segmentation variables and identify distinct group of customers who differ in their needs and wants;
2. **Targeting:** Evaluate the segments and select one or more to target;
3. **Positioning:** Establish and communicate a distinct positioning for each market segment.

The main bases of customer segmentation are: geographic, demographic, behavioural and psychographic variables (Kim *et al*, 2006). In this context and due to GDPR regulation, one cannot include neither demographic nor psychographic variables. In addition, since this aims to individually cluster segments of every city, geographic variables are not appropriate. Therefore, this customer segmentation focuses on behavioural aspects, by considering how customers use the service, as regards to rides patterns, such as time duration and day of the week, as well as usage frequency.

4.2.1 Selecting clustering algorithm

Cluster analysis is the formal study of methods and algorithms for grouping objects according to measured or perceived intrinsic characteristics or similarity. Cluster analysis does not use category labels that tag objects with prior identifiers, i.e., class labels. The absence of category information distinguishes data clustering (unsupervised learning) from classification or discriminant analysis (supervised learning) (A. K. Jain, 2010).

On their clustering-based customer segmentation work, A. K. Jain (2010) suggests choosing the most appropriate clustering technique before preparing and analysing the data to determine clustering parameters.

Given performance expectations and business context, partitioning techniques are considered more appropriate than hierarchical clustering. Consequently, K-means is the selected algorithm,

as it is one of the most widely used clustering techniques (Morency *et al*, 2011) and given its simplicity and computing speed, which are crucial factors in this problem context.

The main steps of K-means algorithm are as follows (A. K. Jain, 2010):

- Select K initial clusters centres; repeat steps 2 and 3 until cluster membership stabilizes;
- Generate a new partition by assigning each pattern to its closest cluster center;
- Compute new cluster centers.

The K-means algorithm requires three user-specified parameters: number of clusters (K), cluster initialization, and distance metric. The most critical choice is K (A. K. Jain, 2010).

4.2.2 Data preparation and pre-processing

After defining the analysis goal and clustering technique, one needs to prepare the data, which involves using historical bookings dataset grouped by customer identification number. Firstly, data quality is assessed to ensure data integrity and robust results. The variable ‘Driven distance’ has many clear outliers, corresponding to unfeasible distances, which were confirmed by the operators; thus, this variable is excluded. For the considered time period, the total number of customers is 21.539, who generated 74.088 rides.

Having in mind the clustering purpose previously described, the following clustering variables are initially considered:

- Total number of rides per customer – ‘Number_Rides’;
- Total revenue per customer (€) – ‘Revenue’;
- Average ride duration (min) – ‘Time’;
- Average days between usage – ‘Avg_days_bet’;
- Percentage of rides on weekend – ‘Weekend’;

One should compute the correlation between variables to avoid clustering distortion. In fact, if two or more clustering variables have a strong correlation, they distort the clustering as they exhibit a higher emphasis than other factors (Degele *et al*, 2018). Spearman’s rho correlation is thus computed, which reflects the intensity of monotone relationship between two variables in an ordinal scale (Degele *et al*, 2018). Table 4.3 shows a relatively high correlation between ‘Number_Rides’ and ‘Revenue’, which makes business sense. The former is eliminated, as average revenue per ride is by definition correlated with average ride time, given that the ride price is proportional to its duration.

Table 4.3: Spearman's rho correlation matrix

	Number_Rides	Revenue	Time	Avg_days_bet	Weekend
Number_Rides	1.00	0.77	-0.10	0.68	-0.02
Revenue	0.77	1.00	0.35	0.47	0.01
Time	-0.10	0.35	1.00	-0.11	0.06
Avg_days_bet	0.68	0.47	-0.11	1.00	-0.06
Weekend	-0.02	0.01	0.06	-0.06	1.00

To assess how clusterable the dataset is, one can compute the Hopkins statistic as a measure of cluster validity (Banerjee and Davé, 2004). This statistic tests the spatial randomness of a dataset, by calculating the probability that a given dataset is generated by a uniform data distribution (Banerjee and Davé, 2004). The null hypothesis is rejected if the test statistic is

below 0,5 and it should be close to 0 for the dataset to be considered very clusterable. This statistic was computed in a random sample of 1000 customers and the result $H = 0,002873$ proves that the dataset is very clusterable.

The four defined clustering variables originate from different categories and differ in their scale. Consequently, to avoid distortion, the clustering variables have to be normalized by converting them into double values between 0 and 1 (Degele *et al*, 2018). After running the clustering algorithm, values are denormalized to allow their analysis and interpretation.

4.2.3 Clustering

Regarding K-means parameters, the common Euclidean distance is used, and the initial partition is randomly generated. Therefore, the only K-means parameter that needs to be defined is the total number of clusters (K). The most appropriate value for this parameter can be estimated through the elbow method, which calculates the total within-cluster sum of squares for each K (number of clusters). By plotting the results, as depicted in Figure 4.2, the optimal K can be identified as the elbow point, where the marginal gain of explained variance drops, resulting in an angle in the graph (A. K. Jain, 2010).

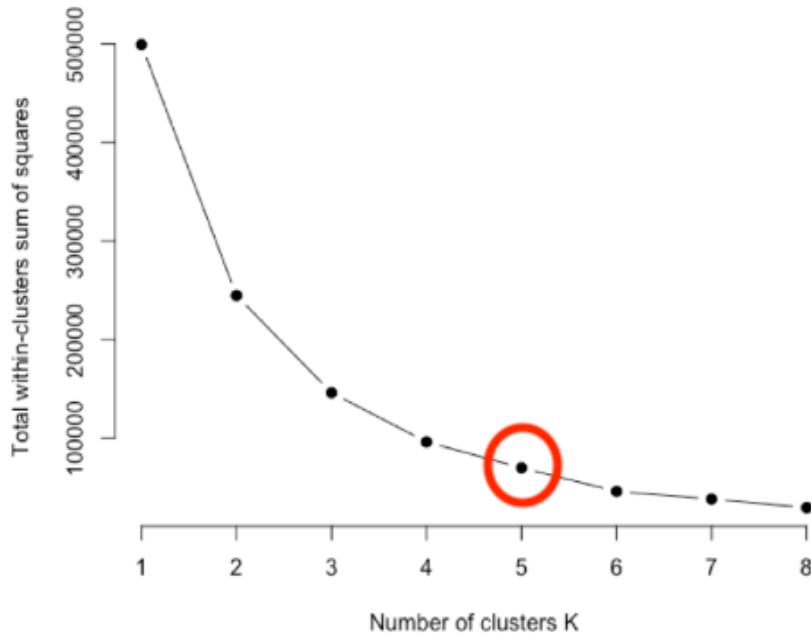


Figure 4.2: Elbow method - total within-cluster sum of squares over number of clusters (K)

Given this elbow curve, 5 looks the optimal K value. However, it is not clear and therefore another method based on the Davies-Bouldin index is used. This index measures how good the clustering scheme is, by taking into account both the distances between clusters (separation), which ideally are as large as possible, and distances within clusters (compactness), which ideally are as low as possible. This index is defined as the ratio between compactness and separation (Günter and Bunke, 2003). Therefore, the lower the index, the better the separation of clusters and the compactness inside clusters. Based on the Davies-Bouldin index, one can also estimate the optimal number of clusters, by calculating the index for several K values and choose the minimum (Günter and Bunke, 2003).

Figure 4.3 depicts the Davies-Bouldin index for every number of clusters (K) between 2 and 8.

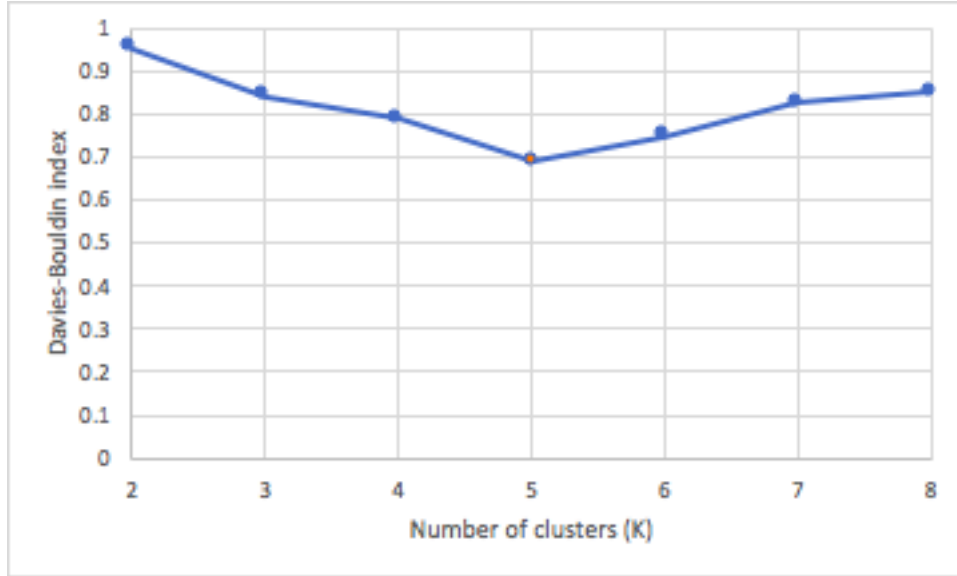


Figure 4.3: Davies-Bouldin index over number of clusters (k)

This method yields the same result as the elbow curve and thus the chosen K value is 5. Corresponding results, interpretation and marketing strategy are presented in Section 5.2.

4.3 Demand proxy

This section explores the demand proxy algorithm, starting by explaining the motivation behind it, followed by the methodology description. Results, interpretation and business validation is thoroughly discussed in Section 5.3.2.

One should note that this dataset was previously characterized in Section 3.4.2.

4.3.1 Motivation

This subsection aims at building an algorithm to convert open app events into demand units, having in mind two different goals. On the one hand, analysing the conversion from open app events to bookings, which can help to improve the app UX/UI. On the other hand, it is a required input for the demand forecasting algorithm, as neither solely open app events nor historical rides are accurate demand proxies due to three main reasons. Firstly, bookings locations depend on vehicles availability, which makes historical rides biased towards past deployment points. Secondly, as the available data is from a single operator and many players operate in every city, it may not be fully representative of total demand. For instance, if a competitor deployed vehicles in a certain area where the considered operator did not, this potential demand is completely ignored. Thirdly, for met demand (converted to bookings), the corresponding first open app events offers the exact demand location, since they are not dependent on vehicles position. For example, a user might open the app where they want to book a ride, but then walks to the nearest vehicle to actually book it. The former location is the exact demand point, while the latter is only an approximation.

4.3.2 Methodology

The methodology comprises the following steps:

1. Data pre-processing and cleaning;
2. Exploratory analysis;
3. Required business insights gathering;

4. Algorithm development;

Firstly, to transform the raw dataset, it is necessary to eliminate all rows with the variable 'content' blank, since they correspond to open app events from unlogged users. These events are for registration purposes and thus not representative of demand. Secondly, the following variables have to be extracted from the 'content' field: latitude, longitude and customerId. Following these two steps, the dataset is ready to be analysed.

A simple analysis shows that the global conversion from open app events to bookings is around 18%, which seems extremely low. No comparable benchmark is available to better assess this conversion value. Therefore, it is important to understand this low conversion by identifying potential causes. A simple analysis finds that users often open the app multiple times within minutes or even within the same minute. In fact, around 19% of open app events correspond to duplicates within the same minute. In other words, a user opens the app a second time within one minute. To segment these events, two distinct causes are suggested depending on user location in both consecutive events:

- Clear app crash: when a specific user opens the app a second time within a minute in the same day and location (latitude, longitude);
- Potential app crash: when a specific user opens the app a second time within a minute in the same day but in different locations (latitude, longitude);

The first situation is a clear app crash, since the user opens again in the same location, probably because the app just crashed. On the other hand, the second case is very likely to happen because a user opens the app again trying to find a closer vehicle. In fact, the duplicate sometimes happens in another location that is not from a walkable distance, which might be due to the fact that the user is in a bus or car looking for the nearest vehicle. Following business sense, these duplicates are eliminated, since they do not correspond to another demand unit. In case of potential app crashes, where the duplicate takes place in a different location, it is assumed that the first open app event corresponds to the actual demand point, since it is the first place where the user searches for a vehicle.

After deleting these duplicates, one can still observe multiple open app events from the same user within a short time frame. A further exploratory analysis along with operators' feedback and customer insights yields the reason behind these consecutive events: vehicle search from customers who open the app again after some time to search for vehicles. Based on business insights from operators, a time frame of 30 minutes is defined, where multiple opens from the same user within this time period are also eliminated as a way to avoid multiple counting of single demand units. It is once again assumed that the first open app event within this time frame corresponds to the actual demand point, for the same reasons. For instance, in the following example of Table 4.4, only the first and fourth rows are kept.

Table 4.4: Demand units algorithm example

row	userId	Day	Time
1	99	20/02/2019	10:01
2	99	20/02/2019	10:04
3	99	20/02/2019	10:11
4	99	20/02/2019	10:36
5	99	20/02/2019	10:44

After this second step, demand units are obtained and ready to be used for demand forecast. Results from this algorithm are interpreted and discussed in section 5.3.

4.4 Demand forecast

After obtaining demands units, this subsection explains the methodology behind the demand forecast algorithm. This is based on three main steps: prediction engineering, feature engineering and machine learning modeling, as Figure 4.4 depicts. Prediction engineering states the business needs and translates them into a machine learning problem, defining the general approach. In the second step - feature engineering - one defines the explanatory variables and extract predictor variables (features) from raw data. It also comprises dataset exploratory analysis, namely time series characteristics and explanatory variables. Lastly, machine learning modeling consists of training a machine learning model on the defined features and select them, tune for the business need and validate predictions before deploying to new data (Koehrsen, 2018).

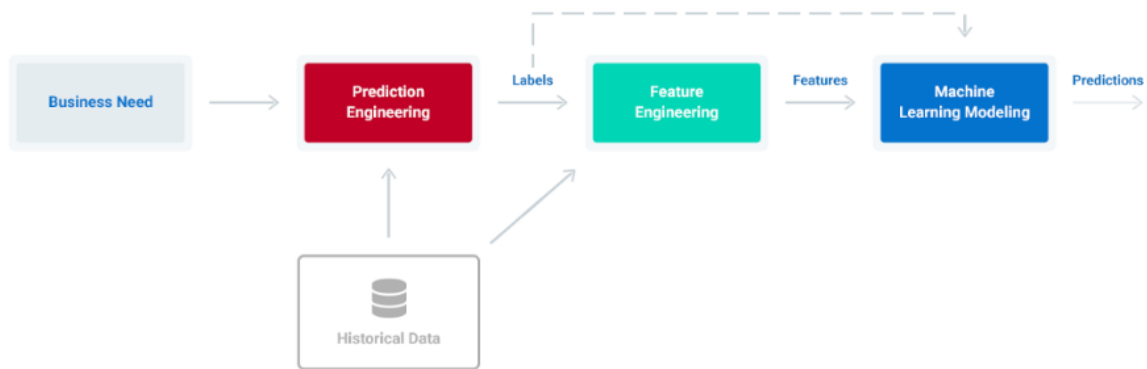


Figure 4.4: Demand forecast methodology - 3 main steps (Koehrsen, 2018)

4.4.1 Prediction engineering

In this business context, explanatory models are considered to be more appropriate than both time series models and historical averages, which are too simple. Although time series models can capture seasonality, trend, levels and even incorporate regressors, they are not as good as explanatory models in discovering relationships between variables and may require assumptions regarding external factors, which should be clearly avoided in this specific context. Explanatory models are more robust and can yield not only better results but also valuable business insights by incorporating many explanatory variables.

As demand units are single points, one needs to find a way to aggregate data geographically. The proposed solution constructs a grid over the operating area, which is defined as a matrix of $a \times b$ tiles (parameters are defined by the user). Then, all points are geographically aggregated on this grid, which allows to have historical demand per grid cell (sum of all demand units within each cell). All geographical operations are based on spatial points dataframes and corresponding methods in R software. Such geographical operations code can be used for aggregating all types of geospatial data, which is also useful for the company in the future. An example of a grid for Málaga can be seen on Figure 4.5, where open app events are aggregated and displayed according to a heatmap based on a color code.

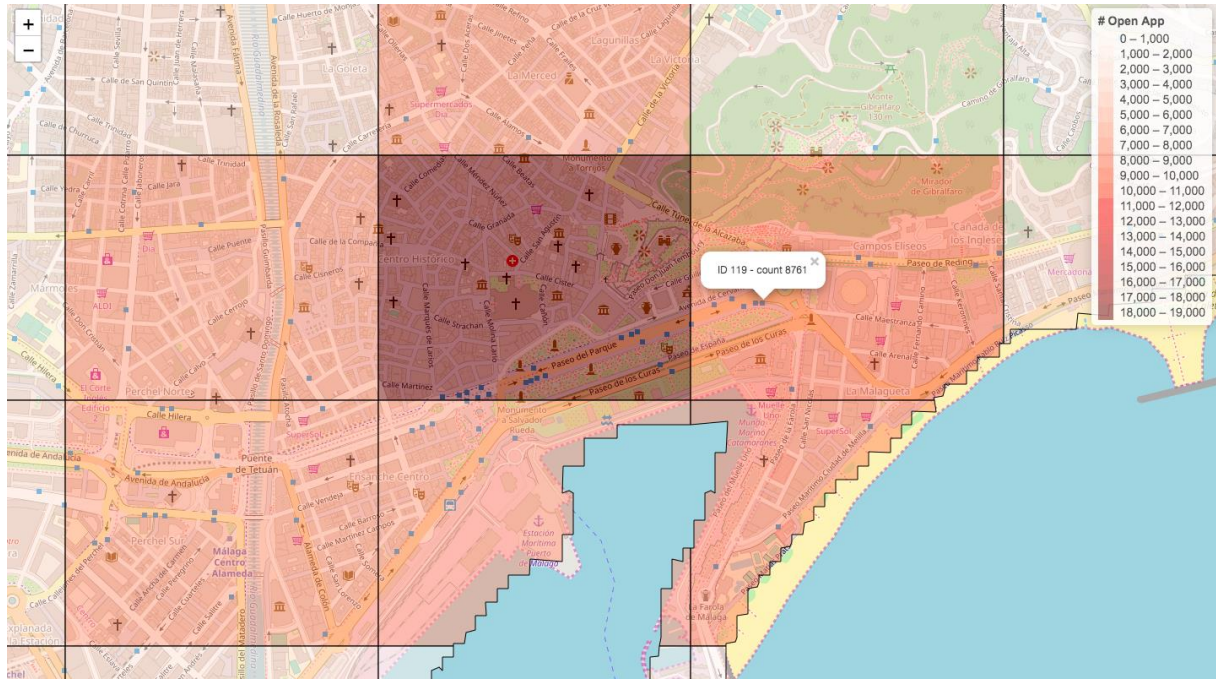


Figure 4.5: Grid cell example for Málaga - open app events aggregated by cell

The machine learning problem can be defined as daily demand forecast for each grid cell. It is a time series forecast which predicts the demand for the next day (one step ahead).

4.4.2 Feature engineering

After the problem definition based on business needs, one has to define explanatory variables to incorporate in the model. Business insights previously explored are the basis of this definition. In addition, other features extracted from historical demand are also incorporated in the model.

Regarding explanatory variables and taking into account the methodology suggested in Section 4.1, Table 4.5 lists all variables that are incorporated in the first algorithm version. Additional explanatory variables, which are very likely to influence daily demand but cannot be yet incorporated due to the dataset time period, are presented in Table 4.6.

Table 4.5: Explanatory variables incorporated in the tested model

Variable	Type	Granularity	Units	Frequency	Source
Day of the week	Categorical	City-level	-	Daily	-
Daily Precipitation	Numeric	City-level	in	Daily	www.wunderground.com
Maximum temperature	Numeric	City-level	°F	Daily	www.wunderground.com
Maximum wind speed	Numeric	City-level	mph	Daily	www.wunderground.com
Maximum humidity	Numeric	City-level	%	Daily	www.wunderground.com

Table 4.6: Explanatory variables that can be incorporated once historical data is longer

Variable	Type	Granularity	Units	Frequency	Source
Competitive intensity	Categorical	City/Area-level	-	Weekly/monthly	Operators input
Tourist arrivals	Numeric	City-level	-	Weekly	To be defined
Public holidays	Binary	City-level	-	Daily	To be defined
Single events	Binary	Area-level	-	Daily	To be defined

Following variables definition, the dataset is explored by analysing time series characteristics (level, trend and seasonality), outliers and explanatory variables distribution.

Exploratory analysis

Although the demand forecast is on grid cell level, global data from Málaga is analysed in this exploratory assessment. The time period is adjusted to eliminate the level effect due to fleet expansion, which happened in the beginning of January. Consequently, the dataset is from 20th January to 13th April 2019 (12 weeks).

- Seasonality

Different seasonality patterns can be evaluated, namely time, day of the week, weekly, monthly and season. Given the goal of forecasting daily demand, time of the day seasonality is not relevant. In addition, due to the short time frame, one cannot evaluate neither weekly, monthly nor season effects. However, the developed algorithm is able to incorporate such effects once the dataset time period allows.

Therefore, at this point in time, only day of the week seasonality can be analysed. Figure 4.6 presents the average demand for each day of the week and shows a significant increase on weekends, where people have more free time.

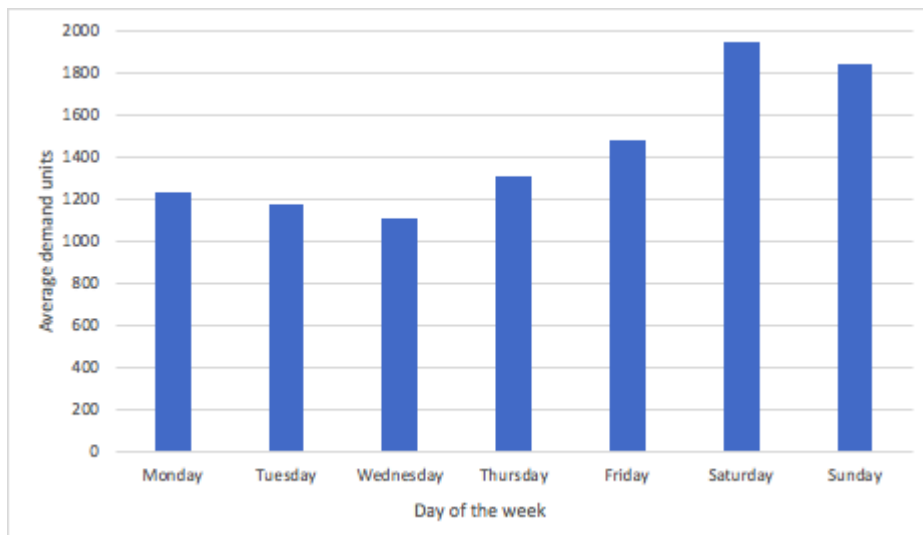


Figure 4.6: Average demand by day of the week (20th January - 13th April)

- Trend and level

Global data from Málaga is aggregated on a weekly basis to avoid the day of the week seasonality. Figure 4.7 depicts a negative trend in the last three weeks of the considered time period due to an increase in competitive intensity, according to the corresponding operator.

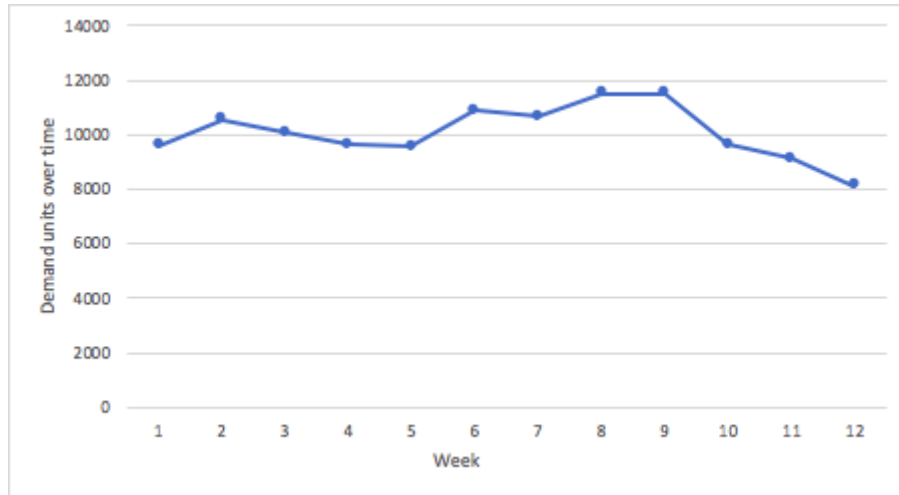


Figure 4.7: Weekly demand units over 12 weeks (20th January - 13th April)

- Outliers

Figure 4.8 shows daily global demand for Málaga, where one cannot identify any clear outlier. For forecasting purposes, outliers are automatically evaluated per grid cell based on a 90% confidence interval using the average for each day of the week, given the relevant seasonality identified. Outliers are replaced by the corresponding day of the week average over the train data set. As an example, if the daily demand of a certain grid cell on 25th February (Monday) is considered an outlier, this value is replaced by the average daily demand on Mondays of this cell. One should note that for testing purposes, the forecasted results are compared with the original dataset (without outliers' correction).

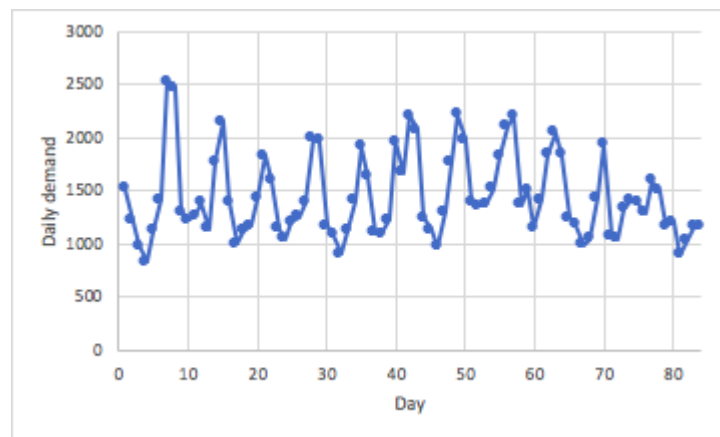


Figure 4.8: Daily demand units (20th January - 13th April)

- Explanatory variables - univariate

By analysing explanatory variables distribution, which can be found in Appendix E, one should highlight two aspects. Firstly, the variable precipitation is zero in 88% of days, since it has barely rained in Málaga over the considered time period; thus, is very likely that the models cannot incorporate successfully this variable. Secondly, maximum temperature has a low amplitude and corresponds to warm weather; thus, this variable effect for Málaga is likely to be low.

In addition to explanatory variables, other features can be extracted from historical demand. This process of generating additional features is called feature engineering. Taking into account time series forecast, the following features are considered:

- Yesterday value
- Last day of the week value
- Rolling statistic of last x days (x between 1 and 7)
- Mean
- Median
- Minimum and maximum
- Variance
- Change between previous measurements
- Change over a time window
- Relative growth over a time window

A few machine learning algorithms, namely KNN, require data pre-processing. Therefore, all numerical variables are scaled and categorical variables, namely day of the week, are converted into dummy variables (one binary variable for each level). By adding these features, a total of 38 explanatory variables are considered. To assess which variables are relevant, it is necessary to run machine learning algorithms, which is described next.

4.4.3 Machine learning modelling

This last subsection comprises the following steps: data split for training and testing, machine learning algorithms description, feature selection and evaluation metrics.

The dataset is split into two: one for training and validation plus the second one for testing. If any parameters need to be tuned, one splits the training set into a training subset and a validation set. The model is trained on the training subset and the parameters that minimize error on the validation set are chosen. Finally, the model is trained on the full training set using the chosen parameters, and the error on the test set is recorded (Bergmeir and Benítez, 2012). Figure 4.9 illustrates a hold-out cross-validation, where the data is split into a training set and a testing set. The training set is split again to tune hyperparameters - one part is used to train the model (training subset) and the other part (validation set) is used to validate the model. By following this methodology, data from the test dataset is never used to optimize parameters.

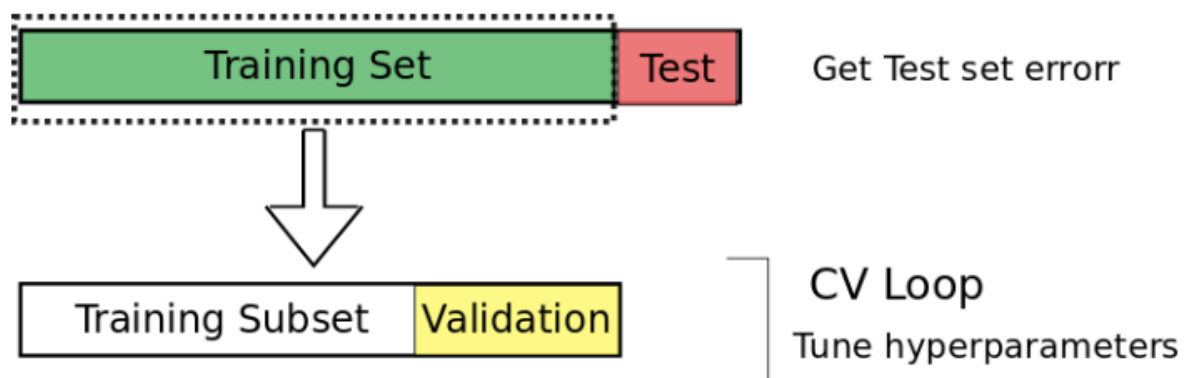


Figure 4.9: Example of hold-out cross validation (Bergmeir and Benítez, 2012)

Since the problem deals with time series data, traditional cross-validation should be avoided for two reasons: temporal dependencies and arbitrary choice of test set. In order to accurately simulate the real-world forecasting environment, in which we stand in the present and forecast the future, the forecaster must withhold all data about events that occur chronologically after the events. So, rather than use k -fold cross validation, one should utilize hold-out cross

validation where a subset of the data (split temporally) is reserved for validating the model performance (Bergmeir and Benítez, 2012). In other words, the testing or validation dataset comes chronologically after the training set. In order to arbitrarily choose the test set, one can use Nested Cross-validation. It comprises an outer loop for error estimation and an inner loop for parameter tuning (Figure 4.10). The inner loop splits the training set into a training subset and a validation set. The model is trained on the former and the parameters that minimize error on the latter are chosen. An outer loop is added, which splits the dataset into multiple different training and test sets and the error on each split is averaged in order to compute a robust estimate of model error. A nested cross-validation procedure provides an almost unbiased estimate of the true error (Varma and Simon, 2006).

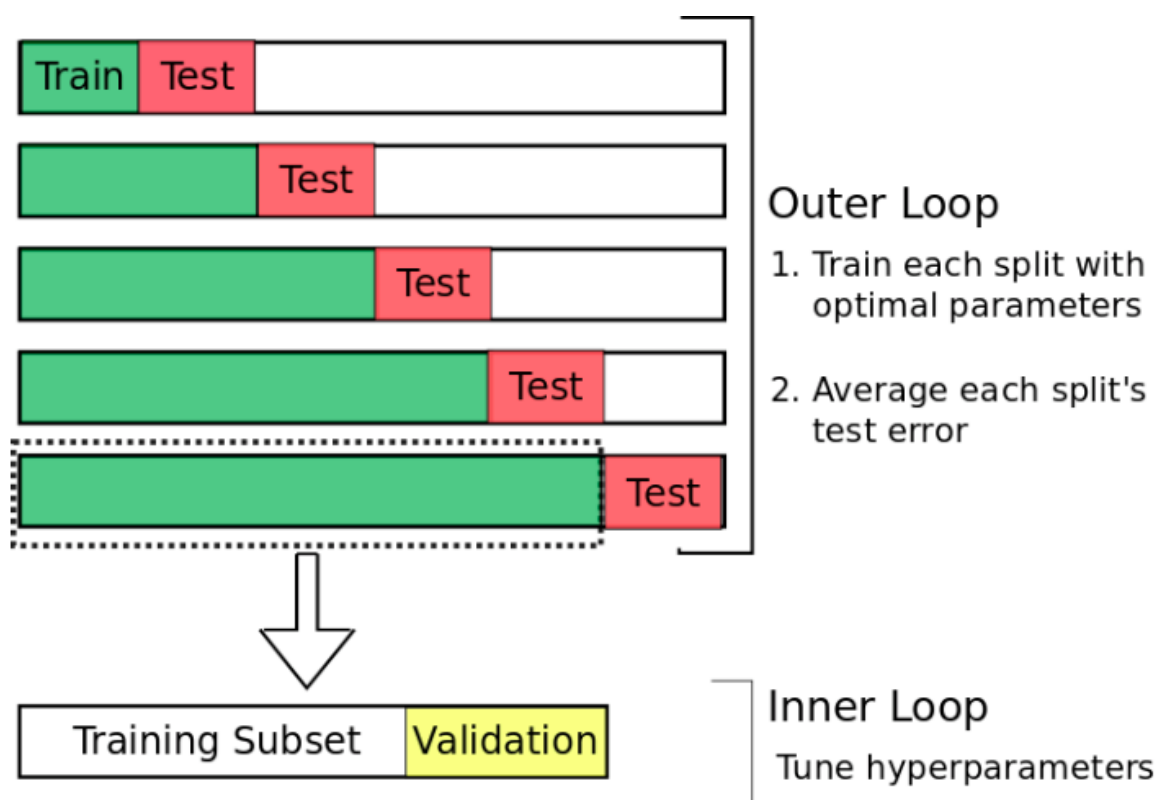


Figure 4.10: Nested cross validation (Bergmeir and Benítez, 2012)

For this single time series, the chosen method for nested cross validation is Day Forward-Chaining, which is based on forward-chaining, also referred in the literature as rolling-origin evaluation (Tashman, 2000) and rolling-origin-recalibration evaluation (Bergmeir and Benítez, 2012). Using this method, each day is successively considered as the test set and assign all previous data into the training set. This method produces many different train/test splits and the error on each split is averaged in order to compute a robust estimate of the error. Figure 4.11 illustrates this methodology.

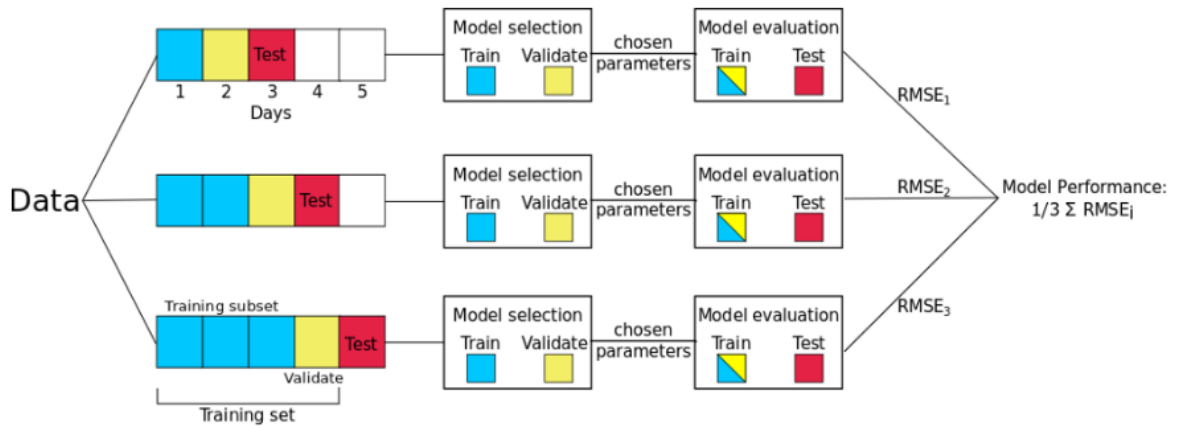


Figure 4.11: Day forward-chaining nested cross-validation (Bergmeir and Benítez, 2012)

Machine learning methods

As of today, given the historical data available, only the following two algorithms are tested: Random Forest and K-Nearest Neighbors (KNN). Techniques such as Support Vector Machines and Neural Networks are not tested, as they require more data in order not to overfit. These should also be implemented in the future. Decision Tree is not implemented because Random Forests are more robust. However, decision trees are also introduced.

- **Random Forest**

This technique derives from Decision trees (DT), which builds regression models in the form of a tree structure. As main advantages, decision trees are easy to use and to interpret. In addition, nominal variables can be treated, and the technique is independent from the variable scale. However, the performance of decision trees is usually below the standard levels and they are considered as not robust (Lin *et al*, 2018). Random Forest was developed to overcome the weaknesses revealed by DT allowing to improve the accuracy of DT and bypassing problems such as the high sensitivity to small variations in data. It is based on an ensemble of trees followed by the calculation of the mean value of the predictions obtained at the final node of each tree and avoids the lack of robustness demonstrated by a single decision tree (Lin *et al*, 2018). In this approach, each tree is grown by using a subset of independent variables randomly selected. As is the case for DT, Random Forest (RF) is considered a simple method. Regarding parameters tuning, the number of trees to grow was pre-defined as 1000 whereas the number of variables randomly sampled at each split was fine-tuned based on a range from 2 to 35 with split 1. The 'randomForest' R package was used.

- **K-Nearest Neighbors (KNN)**

K-nearest neighbours is a simple algorithm, where predictions for a new data point are made by searching through the entire training set for the K most similar instance (neighbours) and summarizing the output variable for those K instances (Lin *et al*, 2018). For regression problems, this might be the mean output variable. Since the determination of similarity between data points is based on distances, numerical variables have to be scaled and dummy variables must replace categorical variables. The only parameter is the number of neighbours to consider (K). The 'KNN' package in R was employed.

Feature selection

Since many features are generated in the feature engineering phase, one needs to select which features are indeed relevant. In fact, when data has high dimensionality, models usually choke for two reasons. Firstly, training time increases exponentially with number of features. Secondly, models have increasing risk of overfitting with increasing number of features

(Asaithambi, 2018). Feature selection helps with these problems by reducing the dimensions without much loss of the total information. It also helps to make sense of the features and its importance.

For this purpose, a global model for demand forecast in Málaga based on the random forest algorithm is tested. Variable importance is used to select which features are relevant. Two variable importance measures can be used: mean decrease in accuracy ('%incMSE') and mean decrease in node impurity ('IncNodePurity'). The first measure is computed from permuting out-of-bag data, i.e., for each tree, the prediction error on the out-of-bag portion of the data is recorded. For regression, mean squared error (MSE) is the prediction error metric. Then the same is done after permuting each predictor variable. The difference between the two are then averaged over all trees and normalized by the standard deviation of the differences. The second measure – mean decrease in node impurity - is the total decrease in node impurities from splitting on the variable, averaged over all trees. For regression, it is measured by residual sum of squares. The first measure is used as it is the most robust and informative measure (Asaithambi, 2018).

Once features are selected, one individual model for each cell grid is developed, based on the two techniques previously introduced. To reduce computation times, grid cells with very low demand, defined as less than 4 average demand units per day, are excluded from the forecast model. After running regression algorithms, one needs to choose the right metrics to evaluate and compare them. As accuracy metric, mean absolute percentage error (MAPE) provides a relative measure. To measure bias, mean percentage error (MPE) indicates whether there is a tendency to under-forecast (systematic under-estimating of future values where $MPE > 0$) or to over-forecast (systematic over-estimating of future values where $MPE < 0$). Computing time is also calculated.

A naive forecast incorporating day of the week seasonality is considered an appropriate baseline for this problem. In this simple method, the forecast is equal to the last value for the same day of the week. For instance, the forecast for the next Monday is the last Monday actual value. A simple naive method, where the forecast for tomorrow is today's demand is also evaluated. Moreover, moving averages are also used as benchmarks.

Forecast models results, interpretation and conclusions are presented in Section 5.4.

5 Results

This section presents the results obtained following the methodology described in the previous section.

5.1 Market intelligence study

For the assumed inputs, namely the variable weights presented on Table A.2, the market attractiveness study results in Madrid, Barcelona and Athens as the three most attractive European cities to launch a kick-scooter sharing business. These cities stand out given their very high potential of tourists users. In addition, local weather and commuting habits are also relevant factors, given higher temperatures, low number of rainy days and low usage of bike, contrary to northern Europe cities. Detailed results can be consulted in Appendix B – Table B.1 and B.2.

In terms of market size, for the assumed inputs presented on Appendix B – Table B.3 and Table B.4, London and Paris clearly outperform other cities given their very high population size and thus potential market. The model estimates an average daily revenue around 14.000€ and 12.000€ for London and Paris, respectively.

Concerning the product features and services development, Table 5.1 presents the generated ideas and corresponding evaluation, sorted in ascending order by priority. Detailed explanation and global score of each idea can be consulted in Appendix C – Table C.1. For the detailed score model as described in the methodology, the user is directed to Appendix C – Table C.2.

Table 5.1: Generated ideas, value-added and monetization sorted by priority

Priority	Idea	Value-added	Monetization
1	Fleet rebalancing/relocation	Cost and revenue	Fee per vehicle
2	App analytics	Revenue	Fee per app user
3	Customer analytics (segmentation)	Revenue	Fee per app user
4	Fleet maintenance management	Cost	Fee per vehicle
5	Charging systems improvement	Cost	Fee per vehicle

The fleet rebalancing feature clearly stands out, despite its high development complexity. In fact, this algorithm requires a two-step approach: demand forecast and then fleet daily deployment optimization, which are both by itself complex. However, this idea has a very high market potential, as operators are struggling with profitability and are eager to optimize vehicle deployment in order to increase vehicle utilization. In addition, this feature is a strong differentiator for Wunder because competitors do not have it, as of May 2019. There are also many potential synergies with future algorithms for the other vehicle types and also Wunder business units.

Customer segmentation and app analytics have also high potential. Both are relatively simple to develop and would have a significant impact on operators' revenue. Both ideas should be monetized through a fee per app user. App analytics has a slightly higher potential than customer analytics, given that competitively is more attractive, as there are marketing agencies specialized in customer analytics and also given potential synergies for Wunder, since other products also use an app and can leverage this algorithm.

Fleet maintenance management and charging systems improvement features are the least attractive, mainly since both are related with hardware and thus are out of Wunder expertise. In addition, there are other companies specialized in such services. Moreover, these features are hardly leveraged by the other business units, as they are specific to kick-scooters. Therefore, these ideas have low scores in terms of feasibility, specifically service, competition and synergies.

5.2 Customer segmentation

Following the methodology described in Section 4.2, k-means clustering results for Málaga are presented in Figure 5.1, which depicts the average metric for each cluster in a parallel coordinates plot. Results are also summarized in Table 5.2, where average metric values are presented, with clusters sorted by usage frequency in descending order. One should note that in addition to clustering variables (first four rows), other relevant metrics are also calculated for each cluster. The share of customers and share of revenue measure each segment business relevance, whereas the period of the day when most rides happen helps to better understand the usage of each customer segment. The reader is directed to Appendix D to visualise the clusters.

As cluster numbers do not have any meaning, one designation is assigned to each cluster according to their usage frequency (average number of rides per customer), aiming at facilitating results interpretation. Power users refers to the group with the highest average rides per customer, followed by frequent and then occasional users. Lastly, try it users have the lowest average rides per customer and are segmented into weekends and weekdays.

5.2.1 Cluster results analysis

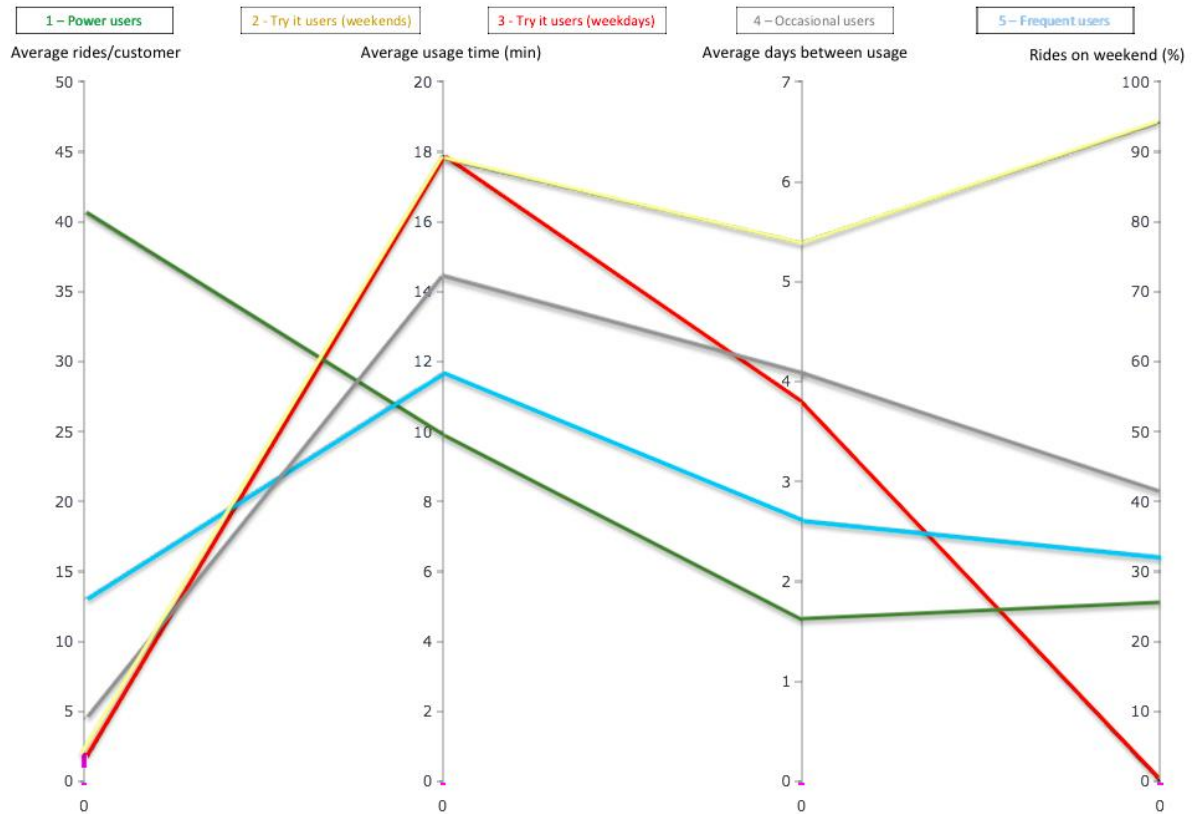


Figure 5.1: Cluster results - parallel coordinates plot with average cluster values

Table 5.2: Customers segments: average metrics

<u>Cluster</u>	1-Power users	5 - Frequent users	4 - Occasional users	3- Try it (weekdays)	2- Try it (weekends)
Average number of rides/customer	40,6	13,0	4,4	1,3	1,4
Average usage time (minutes)	9,9	11,7	14,5	17,9	1,4
Average days between usage	1,62	2,62	4,1	3,81*	5,39*
Share of rides on weekend	25,5%	32,2%	41,5%	0%	94,2%
Share of customers	0,6%	7,1%	32,6%	30,4%	29,3%
Share of revenue	6,4%	24,9%	43,3%	12,4%	12,8%
Mode - day period	Evening (>18h)	Afternoon (15-18h)	Afternoon (15-18h)	Afternoon (15-18h)	Afternoon (15-18h)
Share of customers: single ride	-	-	-	66,4%	59,2%

* ‘Average days between usage’ for ‘try it’ users excludes first time users, who only booked one ride

One can identify a few patterns in clustering results. As the number of rides per customer decreases, the average ride duration increases, as well as percentage of rides on weekends, except for the “try it” segment (weekdays). As less-frequent users tend to use the service mainly for joy, they use it longer and are more likely to ride on weekends, when they have more free time. Furthermore, the mode time period when most rides occur is the afternoon (15h-18h) for every segment, except for power users, whose mode period is the evening (later than 18h). This customer segment tends to use the service for commuting or running errands rather than joy; thus, are likely to commute back home after work using this service.

Based on these results and merging “try it” users, one can identify four main customer segments: power, frequent, occasional and “try it” users (sorted by usage frequency in descending order). The first group uses the service quite often, mainly on weekdays and it is the only group whose mode time period is the evening. They probably use the service for commuting and running errands rather than fun or joy. Due to their usage type, they have the lowest average ride duration. The second segment - frequent users - are likely to use this service for commuting as well as joy. Around one third of their rides take place on weekends. Regarding the third group, they have used the service very occasionally and more on weekends. The main purpose is very likely to be joy riding. Lastly, “try it” users tend to have long rides and a significant percentage are single-ride users.

5.2.2 Business strategy

Taking into account the share of customers and revenue of each cluster, one can evaluate how important each customer segment is for the business. Occasional users represent almost half of users and account for one third of the total revenue over this period; thus, they are a very important customer segment despite their low usage frequency. As their main ride purpose is joy, the operator should focus on increase ride frequency and thus monetization through marketing campaigns. Although they only represent 7% of customers, frequent users account for roughly one fourth of the total revenue due to their high usage frequency. Companies must focus on retaining these customers and further increase monetization. “Try it” users account

also roughly for one fourth of the total revenue but represent 60% of total users. On the one hand, this share number is likely to go down over time, as the market expands and “try it” users convert to occasional or frequent users. On the other hand, it is very likely that many customers within these “try it” clusters are tourists who only use the service one, two or three times during their holidays. A considerable percentage of these customers might become occasional users if properly targeted by operators through, for example, push messages, as an incentive to book another ride. Lastly, despite their very high usage frequency, power users are revenue-wise irrelevant due to the fact that they only account for 0,6% of total customers (129 clients). However, each customer of this segment is very valuable and is very likely to be a brand ambassador. Hence, operators should focus on building loyalty to retain them.

5.2.3 Marketing mix

Following customer insights gained through clustering as regards to usages and purposes, one can define targeted marketing strategies for each segment in order to meet their needs and wants. Marketing mix defines the strategy based on 4P's (Price, Product, Placement and Price).

Product

- For try-it and occasional users, given their joy purpose, the app could offer information on leisure activities and points of interest nearby, such as bars, cafés and museums;
- For power and frequent users, the app could display relevant information on traffic jam and useful commuting information.

Price

- A subscription model for power users could make sense given their usage frequency, as they might be prompted to buy their own vehicle rather than renting it. This subscription model could have a limited ride time, given that this customer group tends to have shorter rides. In addition, this subscription could be only valid for rides after 18h, when other users are not so likely to book rides, whereas power users are;
- Dynamic pricing might be considered, especially due to the high concentration of rides in the afternoon.

Place

- Additional research is required, namely incorporating geospatial analytics, to identify patterns;

Promotion

- For try-it and occasional users, the app could offer discounts on leisure activities and touristic places nearby, such as museums, bars and cafés. Advertising could represent an additional revenue stream for operators;
- For power and frequent users, operators could offer quantity discounts based on packages (for example 60 minutes) to increase usage frequency and simultaneously retain customers;
- Send discount through push notification when a “try it” user opens the app again and does not convert to booking, as a way to avoid losing them.

By following the clustering methodology described in Section 4.2, operators can at any point in time segment customers in a city and gain business insights. As a result, they can not only develop a targeted marketing strategy for each segment, but also launch specific marketing campaigns for each customer group. This analytics-enabled marketing activities are crucial to engage customers and retain them; thus, increasing customer lifetime value and profitability.

5.3 Demand proxy

Following the methodology and classification thoroughly described in Section 4.3, Figure 5.2 segments the open app events according to their causes.

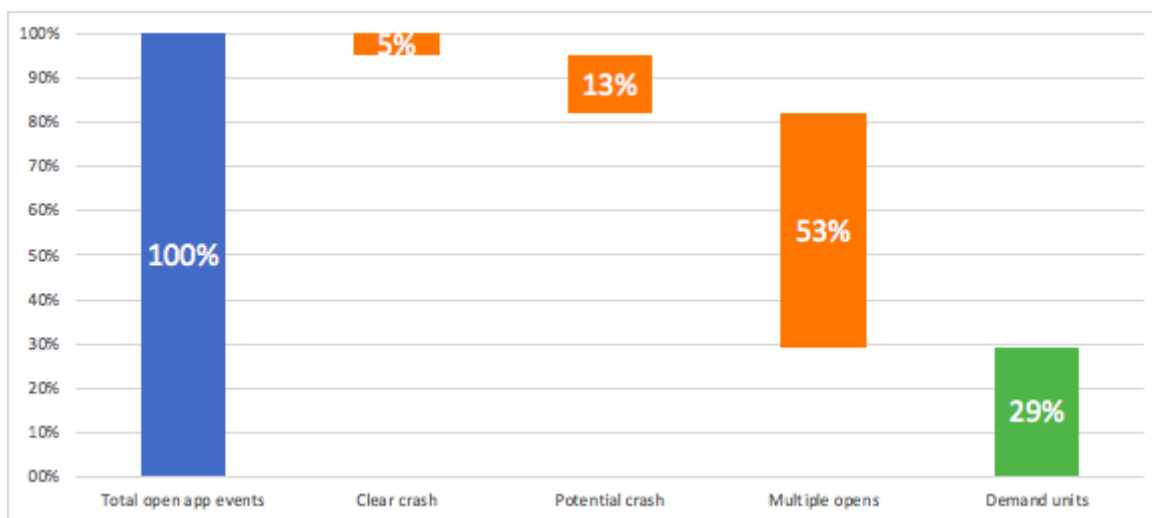


Figure 5.2: Conversion from open app events to demand units

Both clear and potential app crashes highlight the importance of app optimization as a way to improve customer experience. Since the app is recent, there is still significant room for improvement. In fact, according to consumer surveys, the app often crashes or lags when the map is open, since it is heavy.

Considering the percentage of multiple opens, one can conclude that more than half of open app events are within a 30-minute time frame from the same user, even after eliminating actual and potential app crashes. This can be either due to more app crashes, to refresh the map (which often lags according to feedback from users) or to actually search for a closer vehicle. This last option reinforces the importance of optimizing vehicle relocation and confirms that even for met demand, bookings data might not be the most accurate demand point.

Having demand units, one can analyse its conversion to actual bookings, which Figure 5.3 depicts.

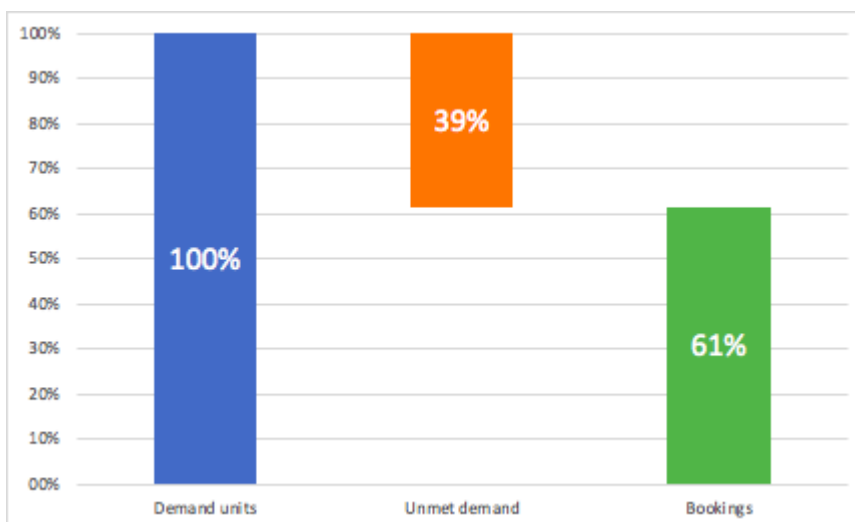


Figure 5.3: Conversion from demand units to bookings

The conversion from demand units to historical rides around 61% sounds more reasonable. Moreover, it highlights the potential impact of a relocation algorithm along with fleet expansion since up to 39% of potential demand is potentially not being met by the current vehicle availability. This demand could be met by competitors, which shows the importance of vehicle relocation optimization as competitive advantage.

Figure 5.4 shows the conversion from demand units to actual bookings per grid cell based on a heatmap. Green circles correspond to deployment points on a random day. As the figure depicts, conversion is much higher on cells where vehicles are deployed, which obviously makes sense. These kind of visualisation tools are also included in the product feature for the operators to optimize their operations.

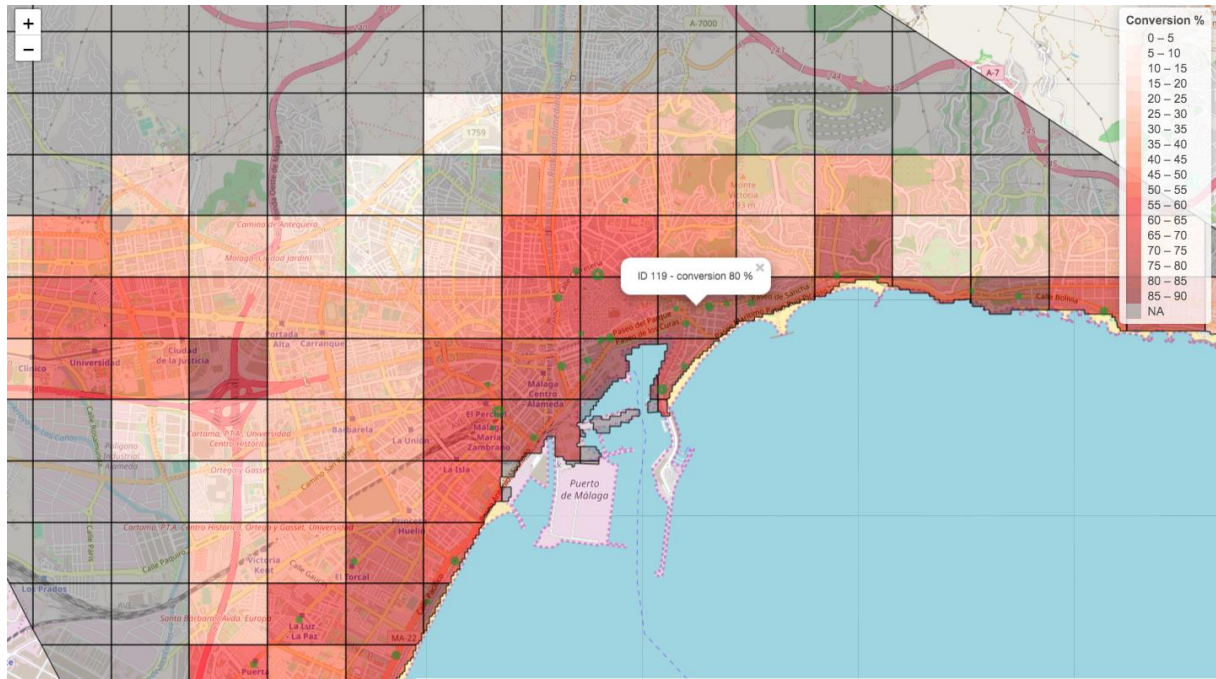


Figure 5.4: Conversion from demand units to bookings per grid cell with green circles corresponding to deployment points

By using demand units, the model can not only capture unmet demand but also better characterize met demand by having the right location point. Despite being the best possible approximation to the actual demand, this approach is not yet perfect and may not accurately represent the demand, since users might open the app before they want to book it, just to check where available vehicles are. Potential mitigation strategies for this limitation are pointed out in the future work section.

5.4 Demand forecast

For testing purposes on Málaga, a grid cell of 700m * 500m is used to geographically aggregate historical demand units.

5.4.1 Feature engineering and selection

As previously described, global demand is forecasted for Málaga using Random Forest to apply feature engineering and selection. 80% of the dataset is used for training plus validation and 20% for testing. A minimum length of 35 days is used for the training and validation. Results are presented in Table 5.3.

Despite the limited historical data (84 days), random forest clearly outperforms the other methods. The relatively high mean percentage error indicates that the algorithm is over-forecasting, since it is not able to capture the demand decrease of the last 2 weeks (see Figure 4.7). Explanatory variables to capture this drop are missing, namely competitive intensity. If these additional variables were incorporated, it is very likely that the algorithm would be able to better forecast and thus reduce MPE.

Table 5.3: Results of global demand forecast for Málaga

Algorithm	MAPE	MPE	Time
Random Forest	12,8%	-10,2%	1 min 5 sec
Total Naive forecast	17,6%	-1,7%	-
Seasonality naive	26,2%	-14%	-
Moving avg (K=2)	19,7%	-2,6%	-
Moving avg (K=3)	18,9%	-3,1%	-
Moving avg (K=4)	18,0%	-4,6%	-

Based on the variable importance of random forests, Figure 5.5 presents the relevant features, defined as having a mean decrease in accuracy over 4%. Orange bars correspond to binary variables, whereas numeric variables have blue bars. Table 5.4 describes these features. The binary variable weekend has the highest mean decrease in accuracy, followed by features resulting from historical demand. In fact, apart from the day of the week, the most relevant features are derived from historical demand. Rain and maximum temperature are relevant features, but their impact is the lowest from the defined relevant features.

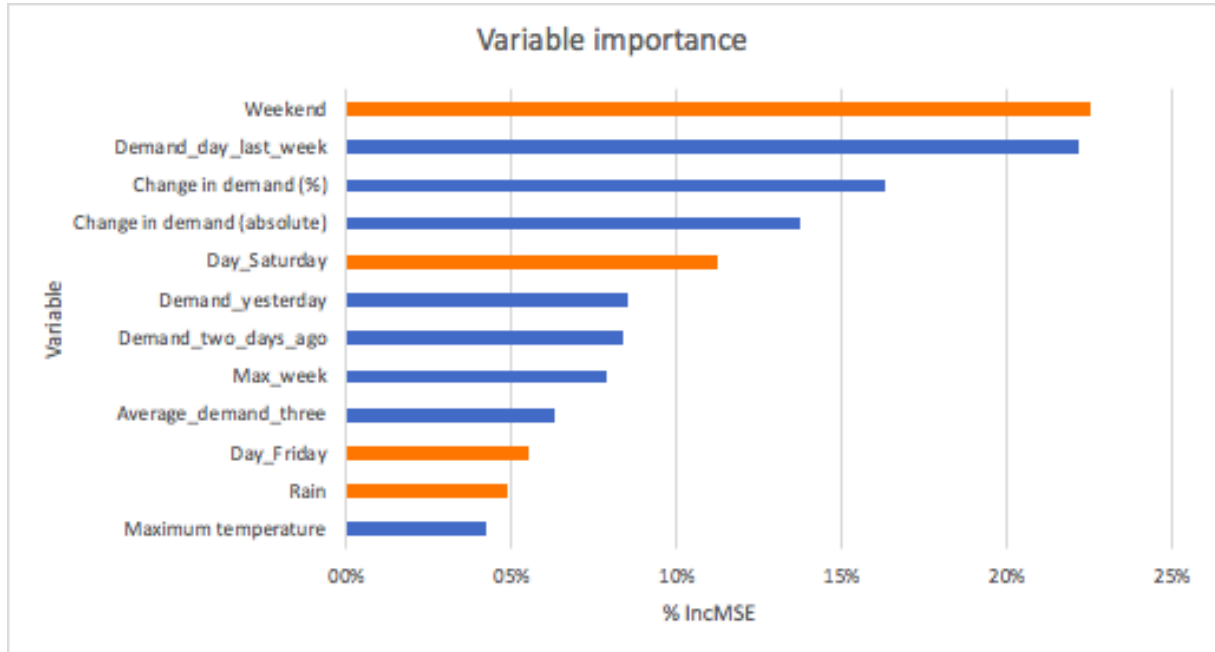


Figure 5.5: Variable importance (mean decrease in accuracy) of most relevant features

Table 5.4: Description of most relevant features

Variable	Type	Description or computing formula
Weekend	Binary	1 if Saturday or Sunday, 0 otherwise
Demand_day_last_week	Numeric	Demand of same day of last week
Change in demand (%)	Numeric	Demand (i-1) / Demand (i-2)
Change in demand (absolute)	Numeric	Demand (i-1) - Demand (i-2)
Day_Saturday	Binary	1 if Saturday, 0 otherwise
Demand_Yesterday	Numeric	Demand of the day before
Demand_two_days_ago	Numeric	Demand two days ago
Max_week	Numeric	Maximum demand of last 7 days
Average_demand_three	Numeric	Average of last 3 days demand
Day_Friday	Binary	1 if Friday, 0 otherwise
Rain	Binary	1 if Precipitation > 0; 0 otherwise
Maximum temperature	Numeric	Maximum daily temperature (°F)

One should highlight two aspects of feature selection results. Firstly, ‘change demand (%)’ and ‘change demand (absolute)’ are obviously correlated. Secondly, a few features like ‘change in demand’ and ‘demand of the day before’ are especially relevant in this context, as they are able to partially capture the negative trend of the last two weeks. They would not be so relevant in other contexts, with an absence of strong trend effects. This example shows the main advantage of using such explanatory models, as they are able to select which features are relevant for every case, making the developed model more robust and scalable.

One should highlight that these factors are the ones relevant for the specific case of Málaga during this short time period. Results are very likely to be substantially different in a few months and especially for other cities. For instance, rain and temperature are very likely to be more relevant features in cities with another climate. In fact, as an example, the feedback from operators confirms that rain is a critical demand factor for Vienna, where it has rained more often.

5.4.2 Grid cells demand forecast

After selecting relevant features based on the global demand forecast for Málaga, one can test KNN and Random Forest algorithms to forecast grid cells demand, following the methodology described in Section 4.4. Only the identified relevant features are used for this forecast. As evaluation metrics, global MAPE and MPE are considered, which correspond to the weighted average of cell metrics, using demand share as the weight. Consequently, cells with higher demand are given a greater significance.

The same benchmark algorithms are used. Total naive forecast simply uses last day actual demand, whereas seasonality naive forecast takes into account the day of the week seasonality. For instance, if we are forecasting for Friday, the total naive forecast is the day before (Thursday) actual demand, while seasonality naive forecast corresponds to previous Friday actual demand. Results are presented in Table 5.5.

Table 5.5: Results of cell demand forecast for Málaga

Algorithm	Global MAPE	Global MPE	Time
Random Forest	26,3%	-9,8%	41 min 38 sec
KNN	30,6%	-11,3%	6 min 12 sec
Total Naive forecast	35,8%	-9,1%	-
Seasonality naive forecast	41,7%	-21,5%	-
Moving average (K=2)	34,6%	-10,3%	-
Moving average (K=3)	32,6%	-10,9%	-
Moving average (K=4)	32,8%	-12,9%	-

Both machine learning models outperform other simpler techniques. As expected, Random Forest yields better results than KNN and takes more time to run, due to its higher complexity.

However, the global accuracy is not high and both algorithms over-forecast. In fact, all forecast methods tested over-forecast, due to the negative trend of last weeks. Given the strong day of the week seasonality, as previously explored, the seasonality naive forecast should outperform the total naive forecast. However, for this specific case, due to the significant negative trend of last weeks, the seasonality naive forecast is too slow to adjust and significantly over-forecasts (MPE = -21,5%), whereas total naive forecast is faster to capture the negative trend and thus outperforms.

To analyse the dispersion between cells, Table 5.6 summarizes MAPE for the two cells with highest demand as well as for the two lowest demand cells. 'First' refers to the cell with highest demand and 'Last' refers to the cell with lowest. The same forecast algorithms of Table 5.5 are considered.

Table 5.6: Cell results for the two cells with highest and lowest demand

Algorithm	MAPE			
	First	Second	Penultimate	Last
Random Forest	21,1%	21,4%	55,8%	58,0%
KNN	27,6%	27,5%	49,2%	54,2%
Total Naive forecast	31,8%	30,9%	45,3%	43,1%
Seasonality naive forecast	34,8%	35,9%	69,1%	66,2%
Moving average (K=2)	31,6%	31,4%	51,2%	50,5%
Moving average (K=3)	29,6%	29,9%	48,2%	49,7%
Moving average (K=4)	28,8%	30,2%	48,8%	48,6%

Results show that all methods forecast more accurately cells with highest demand, which are more constant and therefore easier to predict. Factors not included in the model may more intensively impact the demand in the cells presenting lower demand and consequently machine learning algorithms are prone to overfitting, which justifies why Random Forest has the worst accuracy for the penultimate and last cells.

From an operations perspective, once this algorithm is taken into production, these computing times seem feasible, as operators run it once every morning. In addition, these computing times are likely to go down if faster processors are used.

Results are very likely to improve over time and especially when more explanatory variables are incorporated, namely the ones already previously mentioned in Table 4.6. Furthermore, this model should also be tested in other cities with different characteristics and for a longer time period, as conclusions from this specific case of Málaga cannot be generalized, given its very particular context, as regards to weather characteristics and limited time period with a strong negative trend on the test period. These questions are further explored in the following section.

6 Conclusion and future work

The kick-scooters surge seems to keep thriving, as an eco-friendly last mile solution, given the rapid expansion of operators across the globe and especially in Europe. However, from the operators' perspective, the business is far from profitable and so they rely on software partners, namely Wunder, to help them redefine their strategy and optimize operations by leveraging analytics and enabling data-driven decision making.

In this project, a market overview and industry profitability assessment outline the most attractive European cities and define demand drivers. This study shows that southern European cities with proper infrastructures are promising markets. Following this global perspective, data from one main operator in Málaga (Spain) is used to test several analytical approaches that can improve kick-scooters operators activities and decision making.

By developing a clustering approach based on k-means, a solution is developed to cluster customers from any city at any time, aiming to support decision-making. For the case study of Málaga, five customer segments are characterized regarding their rides' patterns, frequency and monetary value. Specific marketing strategies based on the marketing mix are also proposed. As future work, results from other cities, namely metropolis, should be benchmarked against Málaga results to assess to what extent customer usage are city-specific. Moreover, if compliant with GDPR regulation, customer data could be integrated in the developed customer clustering, providing additional insights. For instance, demographic variables such as age, gender and income would be important for marketing purposes. Furthermore, through the mobile phone country number, it would be possible to derive which users are (foreign) tourists. These insights would contribute not only to marketing purposes but also for business expansion. As future work, the following additional aspects should be considered: incorporate geospatial variables, conduct customer surveys as well as focus groups to validate cluster results and gain additional insights. Furthermore, on the technical side, additional cluster techniques such as hierarchical methods or fuzzy clustering could also be tested.

This project also introduces a demand proxy which allows to have demand units for the forecast model and yields business insights. It segments open app events according to their cause, namely clear app crash, potential crash, multiple opens and demand units. The low global conversion from open app events to demand units highlights the considerable potential of optimizing vehicle relocation. As future work, the algorithm could be further developed by using clicks data and UX/UI inputs. Moreover, a root cause analysis should be carried out to better understand multiple opens and potential crashes. This can be achieved through customer surveys and focus groups.

A forecast model based on demand units is also proposed, providing a valuable feature for operators, who can monitor where demand is not being met by looking at conversion on area-level based on the develop grid. This algorithm is dynamic and thus scalable to every city where operators are present. The demand forecast, although very limited by the available data, outperforms naive methods. The model accuracy is expected to significantly improve once the data set is longer and by incorporating mentioned explanatory variables that could not be included due to the short data set available.

As future work for the demand forecast, several actions have to be pointed out. Firstly, the developed model should be tested in other cities, namely metropolis. This would test both its robustness and scalability; thus, potentially leading to improvements and adjustments. Secondly, all previously mentioned explanatory variables that could not be incorporated due to the limited available data set should be added to the model once the historical data allows. Such variables are environmental data, special events (holidays, festivals, football matches, etc), monthly seasonality and number of tourist arrivals.

By incorporating environmental data, namely points of interest, one can further derive what drives demand, which enables data-driven strategic decisions. For example, it can help operators better designing their deployment strategies in new cities, where historical data is non-existent, as well as cities with very limited data. This would be useful for an aggregated demand forecast model, where one should include specific cells explanatory variables. It would also allow to better understand the impact of other explanatory variables in specific cell grids. For instance, a higher temperature might have a bigger impact on a cell grid near the beach with many popular cafés. For each point of interest, the following features would be required: name, category (restaurant, bar, museum, bus station, etc), relevance proxy (for instance the number of Google reviews) and location (latitude and longitude). Despite the high potential, there are three inherent challenges. Firstly, getting the right data, especially a measure for points of interest relevance. Secondly, models can easily get too complex and computing time might increase exponentially. Thirdly, some points of interest might not only impact the cell grid where they are located but also surrounding cells, especially if located near the border. A way to mitigate this effect, following Wagner *et al* (2016) approach, could be to define a radius of influence for POIs.

Moreover, the competitive landscape variable is critical to capture positive or negative demand trends. These steps are not only very likely to considerably increase forecast accuracy, but also would allow the evaluation of how important these demand drivers are. Moreover, social demographic data should also be incorporated into the model, namely income level and population density, which are likely to influence demand. When historical data is long enough to incorporate all effects mentioned, powerful machine learning like neural networks and support vector machines are very likely to excel in finding patterns in the dataset and outperform other techniques. Furthermore, regression techniques such as Ridge, Lasso and Elastic net should also be tested, especially because they facilitate feature selection.

Having a more robust demand forecast, a feature to optimize daily relocation should be implemented. Given the operations context described, this algorithm should solve the operator-based hybrid system rebalancing, based on the area grid already implemented and using demand forecast as input. This algorithm should take into account the vehicle flow during the day in an attempt to maximize vehicle utilization, given that relocation throughout the day is not profitable. In other words, the method should evaluate the potential of deploying a vehicle in a certain square grid in terms of future daily bookings by taking into account historical routes flow starting in each cell. Furthermore, the developed demand forecast feature can be easily adjusted to forecast demand in any day period (e.g.: morning, afternoon, evening, etc).

This relocation optimization could be either based on linear optimization or simulation models using heuristics, such as Tabu Search. This feature would increase operators' vehicle utilization and thus profitability. An A/B testing should be carried out to estimate the impact of such an algorithm as regards to an increase in revenue or vehicle utilization, as a way to validate the model.

Bibliography

- Aeschbach, Philipp, Xiaojing Zhang, Angelos Georghiou, and John Lygeros. 2015. "Balancing Bike Sharing Systems through Customer Cooperation – A Case Study on London 's Barclays Cycle Hire." *2015 54th IEEE Conference on Decision and Control (CDC)*, no. Cdc: 4722–27. <https://doi.org/10.1109/CDC.2015.7402955>.
- Albi, Szymon, Pirmin Fontaine, and Stefan Minner. 2018. "Performance Analysis of a Hybrid Bike Sharing System: A Service- Level-Based Approach under Censored Demand Observations." *Elsevier* 116 (June): 59–69. <https://doi.org/10.1016/j.tre.2018.05.011>.
- Alex Wilhelm. 2018. "The Revenue, Costs, And Margins Behind Bird's Scooters." 2018. https://news.crunchbase.com/news/the-revenue-costs-and-margins-behind-birds-scooters/#Gross_Margins.
- Arunachalam, Deepak, and Niraj Kumar. 2018. "Benefit-Based Consumer Segmentation and Performance Evaluation of Clustering Approaches: An Evidence of Data-Driven Decision-Making." *Expert Systems With Applications* 111: 11–34. <https://doi.org/10.1016/j.eswa.2018.03.007>.
- Asaithambi, Sudharsan. 2018. "Why, How and When to Apply Feature Selection – Towards Data Science." *Medium*. <https://towardsdatascience.com/why-how-and-when-to-apply-feature-selection-e9c69adf2>.
- Audenhove, Francois, Andrew Smith, Guillaume Rominger, and Aurelia Bettati. 2018. "The Future of Mobility 3.0, Reinventing Mobility in the Era of Disruption and Creativity." *Aurthur D Little Insights*, no. March: 50. http://www.adlittle.com/futuremobilitylab/assets/file/180330_Arthur_D.Little_&_UITP_Future_of_Mobility_3_study.pdf.
- Banerjee, Amit, and Rajesh N. Davé. 2004. "Validating Clusters Using the Hopkins Statistic." *IEEE International Conference on Fuzzy Systems* 1: 149–53. <https://doi.org/10.1109/FUZZY.2004.1375706>.
- Bergmeir, Christoph, and José M. Benítez. 2012. "On the Use of Cross-Validation for Time Series Predictor Evaluation." *Information Sciences* 191 (May): 192–213. <https://doi.org/10.1016/j.ins.2011.12.028>.
- Caggiani, Leonardo, Rosalia Camporeale, Michele Ottomanelli, and Wai Yuen. 2018. "A Modeling Framework for the Dynamic Management of Free- Fl Oating Bike-Sharing Systems." *Transportation Research Part C* 87 (April 2017): 159–82. <https://doi.org/10.1016/j.trc.2018.01.001>.
- Chang, Pei-chann, Xiao-yong Lu, Jheng-long Wu Yahui, and Xu Min. 2019. "Bike Sharing Demand Prediction Using Artificial Immune System and Artificial Neural Network." *Soft Computing* 23 (2): 613–26. <https://doi.org/10.1007/s00500-017-2909-8>.
- Chen, Po-chuan, He-yen Hsieh, Xanno Kharis Sigalingging, Yan-ru Chen, Jenq-shiou Leu, and Senior Member. 2017. "Prediction of Station Level Demand in a Bike Sharing System Using Recurrent Neural Networks." *Institute of Electrical and Electronics Engineers*, no. 3.
- Cohen, Boyd, and Jan Kietzmann. 2014. "Ride On ! Mobility Business Models for the Sharing Economy." *SAGE*. <https://doi.org/10.1177/1086026614546199>.
- Degele, Jutta, Anna Gorr, Katja Haas, Dimitri Kormann, Sascha Krauss, Paulina Lipinski, Muhamett Tenbih, Christine Koppenhoefer, Jan Fauser, and Dieter Hertweck. 2018. "Identifying E-Scooter Sharing Customer Segments Using Clustering." *2018 IEEE International Conference on Engineering, Technology and Innovation*.

- Dell, Mauro, Manuel Iori, Stefano Novellani, and Anand Subramanian. 2018. "The Bike Sharing Rebalancing Problem with Stochastic Demands." *Transportation Research Part B* 118: 362–80. <https://doi.org/10.1016/j.trb.2018.10.015>.
- Dickey, Megan Rose. 2019. "The Electric Scooter Wars of 2018 | TechCrunch." *TechCrunch*, 2019. https://techcrunch.com/2018/12/23/the-electric-scooter-wars-of-2018/?guce_referrer_us=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_cs=dinR-ubdBAjbIZOol9qbLw&guccounter=2&guce_referrer_us=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS91cmw_cT1odHRwcovL3RIY2hjcjVuY2guY29.
- Faghih-imani, Ahmadreza, Robert Hampshire, Lavanya Marla, and Naveen Eluru. 2017. "An Empirical Analysis of Bike Sharing Usage and Rebalancing : Evidence from Barcelona and Seville." *Transportation Research Part A* 97: 177–91. <https://doi.org/10.1016/j.tra.2016.12.007>.
- France, Stephen L, and Sanjoy Ghose. 2019. "Marketing Analytics : Methods , Practice , Implementation , and Links to Other Fields." *Expert Systems With Applications* 119: 456–75. <https://doi.org/10.1016/j.eswa.2018.11.002>.
- Galatoulas, Nikolaos-fivos, Konstantinos N Genikomsakis, and Christos S Ioakimidis. 2018. "Analysis of Potential Demand and Costs for the Business Development of an Electric Vehicle Sharing Service." *Sustainable Cities and Society* 42 (May): 148–61. <https://doi.org/10.1016/j.scs.2018.07.008>.
- Günter, Simon, and Horst Bunke. 2003. "Validation Indices for Graph Clustering." *Elsevier* 24 (8): 1107–13.
- Hannon, Eric, Stefan Knapfer, Sebastian Stern, Ben Sumers, and Jan Tijs Nijssen. 2019. "An Integrated Perspective on the Future of Mobility, Part 3: Setting the Direction Toward Seamless Mobility." *McKinsey & Company*, no. January. https://www.mckinsey.com/~media/mckinsey/business_functions/sustainability_and_resource_productivity/our_insights/the_road_to_seamless_urban_mobility/an-integrated-perspective-on-the-future-of-mobility-part-3-vf.ashx.
- Ho, Sin C, and W Y Szeto. 2014. "Solving a Static Repositioning Problem in Bike-Sharing Systems Using Iterated Tabu Search." *TRANSPORTATION RESEARCH PART E* 69: 180–98. <https://doi.org/10.1016/j.tre.2014.05.017>.
- Hughes, Matthew. 2018. "Those Ridiculous Bird Scooters Are Now in Europe." *The next Web*, 2018. <https://thenextweb.com/eu/2018/08/01/those-ridiculous-bird-scooters-are-now-in-europe/>.
- Hulot, Pierre, Daniel Aloise, and Sanjay Dominik Jena. 2018. "Towards Station-Level Demand Prediction for Effective Rebalancing in Bike-Sharing Systems." *Research Council of Canada*, 378–86.
- Jain, Aditya, Nils Rudi, and Tong Wang. 2014. "Demand Estimation and Ordering Under Censoring : Stock-Out Timing Is (Almost) All You Need." *Inform*s 5463: 1–17.
- Jain, Anil K. 2010. "Data Clustering: 50 Years beyond K-Means." *Pattern Recognition Letters* 31 (8): 651–66. <https://doi.org/10.1016/j.patrec.2009.09.011>.
- Jorge, Diana, Gon Correia, and Cynthia Barnhart. 2014. "Comparing Optimal Relocation Operations with Simulated Relocation Policies in One-Way Carsharing Systems." *Institute of Electrical and Electronics Engineers*.
- Kang, Zilu, Yuting Zuo, Zhibin Huang, and Feng Zhou. 2017. "Research on the Forecast of Shared Bicycle Rental Demand Based on Spark Machine Learning Framework." *International Symposium on Distributed Computing and Applications to Business*,

- Engineering and Science*, 219–22. <https://doi.org/10.1109/DCABES.2017.55>.
- Kim, Su-yeon, Tae-soo Jung, Eui-ho Suh, and Hyun-seok Hwang. 2006. “Customer Segmentation and Strategy Development Based on Customer Lifetime Value : A Case Study.” *Elsevier* 31: 101–7. <https://doi.org/10.1016/j.eswa.2005.09.004>.
- Koehrsen, Will. 2018. “How to Create Value with Machine Learning – Towards Data Science.” *Medium*. <https://towardsdatascience.com/how-to-create-value-with-machine-learning-eb09585b332e>.
- Legros, Benjamin. 2019. “Dynamic Repositioning Strategy in a Bike-Sharing System ; How to Prioritize and How to Rebalance a Bike Station.” *European Journal of Operational Research* 272 (2): 740–53. <https://doi.org/10.1016/j.ejor.2018.06.051>.
- Lin, Lei, Zhengbing He, and Srinivas Peeta. 2018. “Predicting Station-Level Hourly Demand in a Large-Scale Bike- Sharing Network : A Graph Convolutional Neural Network Approach ☆.” *Transportation Research Part C* 97 (October): 258–76. <https://doi.org/10.1016/j.trc.2018.10.011>.
- Lozano, Á, Juan F De Paz, Gabriel Villarrubia, Daniel H De La, Iglesia Id, and Javier Bajo Id. 2018. “Multi-Agent System for Demand Prediction and Trip Visualization in Bike Sharing Systems.” *Applied Sciences*. <https://doi.org/10.3390/app8010067>.
- Mary Ann Azevedo, and Alex Wilhelm. 2019. “Bird Squeezes The Brakes, Cuts Staff To Control Costs.” <https://news.crunchbase.com/news/bird-squeezes-the-brakes-cuts-staff-to-control-costs/>.
- McKinsey. 2019. “Micromobility’s 15,000-Mile Checkup,” no. Exhibit 1: 1–4. <https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/micromobilitys-15000-mile-checkup>.
- Mersereau, Adam J. 2014. “Demand Estimation from Censored Observations with Inventory Record Inaccuracy.” *University of Florida*, no. 1976: 1–27.
- Morency, Catherine, Station Centre-ville, Station Centre-ville, and Bruno Agard. 2011. “Typology of Carsharing Members.”
- Raviv, Tal, Michal Tzur, and Iris A Forma. 2013. “Static Repositioning in a Bike-Sharing System: Models and Solution Approaches.” *EURO J Transp Logist*, 187–229. <https://doi.org/10.1007/s13676-012-0017-6>.
- Regue, Robert, and Will Recker. 2014. “Proactive Vehicle Routing with Inferred Demand to Solve the Bikesharing Rebalancing Problem.” *Transportation Research Part E* 72: 192–209. <https://doi.org/10.1016/j.tre.2014.10.005>.
- Reiss, Svenja, and Klaus Bogenberger. 2017. “ScienceDirect ScienceDirect A Relocation Strategy for Munich ’ s Bike Sharing System : A Relocation Strategy for Munich ’ s Bike Sharing System : Combining an Operator-Based and a User-Based Scheme Combining an Operator-Based and a User-Based Scheme.” *Transportation Research Procedia* 22: 105–14. <https://doi.org/10.1016/j.trpro.2017.03.016>.
- Stefanescu, Catalina. 2016. “Multivariate Customer Demand : Modeling and Estimation from Censored Sales.” *Operations Research*, 1–35.
- Viswanath, Prashanth, and Manu Mathew. 2015. “Urban Mobility at a Tipping Point.” *McKinsey Center for Business and Environment September*, no. September: 1–9. <http://www.mckinsey.com/business-functions/sustainability-and-resource-productivity/our-insights/urban-mobility-at-a-tipping-point>.
- Vogel, Patrick, Torsten Greiser, and Dirk Christian. 2011. “Understanding Bike-Sharing Systems Using Data Mining : Exploring Activity Patterns.” *Procedia - Social and*

- Behavioral Sciences* 20: 514–23. <https://doi.org/10.1016/j.sbspro.2011.08.058>.
- Wagner, Sebastian, Tobias Brandt, and Dirk Neumann. 2016. “In Free Float : Developing Business Analytics Support for Carsharing Providers \$.” *Elsevier* 59: 4–14. <https://doi.org/10.1016/j.omega.2015.02.011>.
- Weikl, Simone, and Klaus Bogenberger. 2013. “Relocation Strategies and Algorithms for Free-Floating.” *IEEE Intelligent Transportation Systems Magazine* 5 (October): 100–111. <https://doi.org/10.1109/MITS.2013.2267810>.
- Wuarrington, Joseph, and Dominik Ruchti. 2019. “Two-Stage Stochastic Approximation for Dynamic Rebalancing of Shared Mobility Systems.” *Automatic Control Lab - Zurich*, 1–37.
- Yang, Chi-jen. 2010. “Technological Forecasting & Social Change Launching Strategy for Electric Vehicles : Lessons from China and Taiwan.” *Technological Forecasting & Social Change* 77 (5): 831–34. <https://doi.org/10.1016/j.techfore.2010.01.010>.

Appendix A: Market assessment - variables

Table A.1: Market assessment - variables

Factor	Variable	Granularity	Source
Purchasing power	Monthly net income	City	https://www.numbeo.com
Weather	Hours sunshine/year	City	https://www.currentresults.com
Weather	Rainy days/year	City	https://www.currentresults.com
City size/population	Total metropolitan area population	City	https://population.un.org/
City size/population	Percentage of people between 18 and 50 years old	Country	https://population.un.org/
Population density	Population density (people/km ²)	City	https://population.un.org/
Transportation expense	Share of income spent on transport) (times the monthly net income)	Country	https://ec.europa.eu/eurostat
Traffic congestion level	Traffic congestion level (score)	City	https://www.tomtom.com
Pollution level	Pollution level (score)	City	https://www.tomtom.com
Commuting habits	Share of people using car to commute	Country	https://www.citylab.com
Commuting habits	Share of people walking to commute	Country	https://www.citylab.com
Commuting habits	Share of people using bicycle to commute	Country	https://www.citylab.com
Commuting habits	Share of people using public transportation to commute	Country	https://www.citylab.com
Tech savviness	Tech savviness ranking (score)	Country	https://ec.europa.eu
Tourism activity	Tourist arrivals (2018)	City	https://ec.europa.eu
Tourism activity	Yearly tourism growth (%)	City	https://ec.europa.eu

Table A.2: Variables weight for each customer segment

		Weights		
Factor	Variable	Joy riders (50%)	Commuters (30%)	Tourists (20%)
Purchasing power	Monthly net income	30%	-	-
Weather	Hours sunshine/year	30%	10%	30%
Weather	Rainy days/year	25%	10%	25%
City size/population	Total metropolitan area population			40%
City size/population	Percentage of people between 18 and 50 years old	Potential market	Potential market	
Population density	Population density (people/km2)		15%	
Transportation expense	Share of income spent on transport) (times the monthly net income)		15%	
Traffic congestion level	Traffic congestion level (score)		5%	
Pollution level	Pollution level (score)		5%	
Commuting habits	Share of people using bicycle to commute		10%	
Commuting habits	Share of people walking to commute		5%	
Commuting habits	Share of people using public transportation to commute		10%	
Commuting habits	Share of people using car to commute		10%	
Tech savviness	Tech savviness ranking (score)	15%	5%	
Tourism activity	Tourist arrivals (2018)			Potential market
Tourism activity	Yearly tourism growth (%)			5%

Appendix B: Market intelligence study - results

Table B.1: Market attractiveness results - first part

	Scandinavia		UK/Ireland		Benelux		German speaking				
	Stockholm	Copenhagen	London	Dublin	Brussels	Amsterdam	Berlin	Hamburg	Frankfurt	Munich	Vienna
Global score	0.08	-0.04	0.18	-0.25	-0.51	-0.23	-0.10	-0.24	-0.19	-0.14	0.03
Daily commuting	-0.08	-0.57	0.12	0.02	0.01	-0.68	-0.18	-0.26	-0.20	-0.18	0.04
Tourists	-0.31	-0.42	0.83	-0.91	-1.13	-0.47	-0.16	-0.55	-0.59	-0.46	-0.15
Joy riding	0.34	0.42	-0.04	-0.16	-0.57	0.13	-0.03	-0.10	-0.03	0.02	0.10

Table B.2: Market attractiveness results - second part

	Southwestern Europe						Central/Eastern Europe				
	Paris	Madrid	Barcelona	Lisbon	Rome	Milan	Prague	Budapest	Sofia	Warsaw	Athens
Global score	0.64	0.61	0.42	0.25	-0.04	-0.30	-0.19	-0.22	-0.59	0.57	0.64
Daily commuting	0.27	0.61	0.18	0.20	0.12	-0.07	-0.23	-0.13	-0.32	0.66	0.27
Tourists	1.19	0.88	0.74	0.52	0.07	-0.38	0.14	-0.10	-0.60	1.03	1.19
Joy riding	0.64	0.50	0.44	0.17	-0.17	-0.39	-0.30	-0.31	-0.75	0.33	0.64

Table B.3: Assumptions for revenue model

	Average adoption rate	Average trip length(min)	Daily usage	Weekly usage	Comments
Commuting	0.01%	12.00	2	5	Commuters will use the service twice a day only during working days
Tourism	0.01%	10.00	1	7	Tourists will use it once per day and 7 days/week
Joy riding	0.05%	10.00	1	2	Joy riding will be mainly during weekend

Table B.4: Pricing assumptions for revenue model

Pricing	€
Set-up fee	1
Price per minute	0.15

(current pricing as of April 2019)

Table B.5: Revenue model results - first part

	Scandinavia		UK/Ireland		Benelux		German speaking				
	Stockholm	Copenh agen	London	Dublin	Brussel s	Amster dam	Berlin	Hamburg	Frankfu rt	Munich	Vienna
Average daily revenue	€1,267	€1,129	€13,971	€1,085	€985	€2,122	€2,626	€1,170	€1,108	€1,476	€2,308
Daily commuting	€336	€199	€2,543	€231	€400	€252	€683	€403	€344	€378	€483
Tourists	€496	€535	€9,479	€679	€360	€1,442	€1,244	€355	€406	€692	€1,370
Joy riding	€435	€396	€1,949	€175	€225	€427	€699	€412	€358	€406	€456

Table B.6: Revenue model results - second part

	Southwestern Europe						Central/Eastern Europe				
	Paris	Madrid	Barcelona	Lisbon	Rome	Milan	Prague	Budapest	Sofia	Warsaw	Athens
Average daily revenue	€12,058	€5,979	€5,860	€2,750	€5,182	€2,988	€2,348	€1,979	€590	€1,192	€4,360
Daily commuting	€3,078	€1,376	€1,472	€527	€800	€741	€421	€455	€243	€442	€1,024
Tourists	€7,620	€3,017	€3,163	€1,649	€3,686	€1,741	€1,638	€1,140	€159	€452	€2,603
Joy riding	€1,360	€1,586	€1,225	€574	€696	€505	€289	€384	€187	€298	€732

Appendix C: Ideas assessment results

Table C.1: Ideas explanation, value-added, score and monetization

Priority	Idea	Explanation	Value-added	Score (1-5)	Monetization
1	Fleet rebalancing algorithm	A machine learning algorithm to optimize fleet relocation would reduce its costs and drive more rides	Revenue and cost	4	Fee per vehicle
2	App analytics	App analytics would help optimizing UX/UI and thus increase revenue	Revenue	3,875	Fee per app user
3	Customer analytics	Customer analytics would drive business insights and help targeted marketing campaigns aimed at increasing revenue	Revenue	3,75	Fee per app user
4	Fleet maintenance management	Preventive maintenance powered by analytics would reduce breakdowns and save costs	Cost	3,25	Fee per vehicle
5	Charging systems improvement	Facilitating vehicle charging systems by developing new more efficient ways to do it would allow cost reductions	Cost	2,75	Fee per vehicle

Table C.2: Ideas assessment - detailed cost and feasibility scores

	Cost			Feasibility				
Priority	Development	Testing	Sales/MK	Market potential	Product / Service	Competition	Clients	Synergies /Cannibalization
1	1	4	4	5	3	5	5	5
2	4	4	3	4	4	4	4	4
3	4	5	5	4	4	2	4	2
4	3	3	4	4	2	3	4	3
5	2	2	4	5	1	2	5	1

Note: How to correctly interpret these scores?

- Costs: the lower estimated costs are, the higher the score
- Feasibility: the higher the feasibility is, the higher the score

Appendix D: Customer clustering - results visualisation

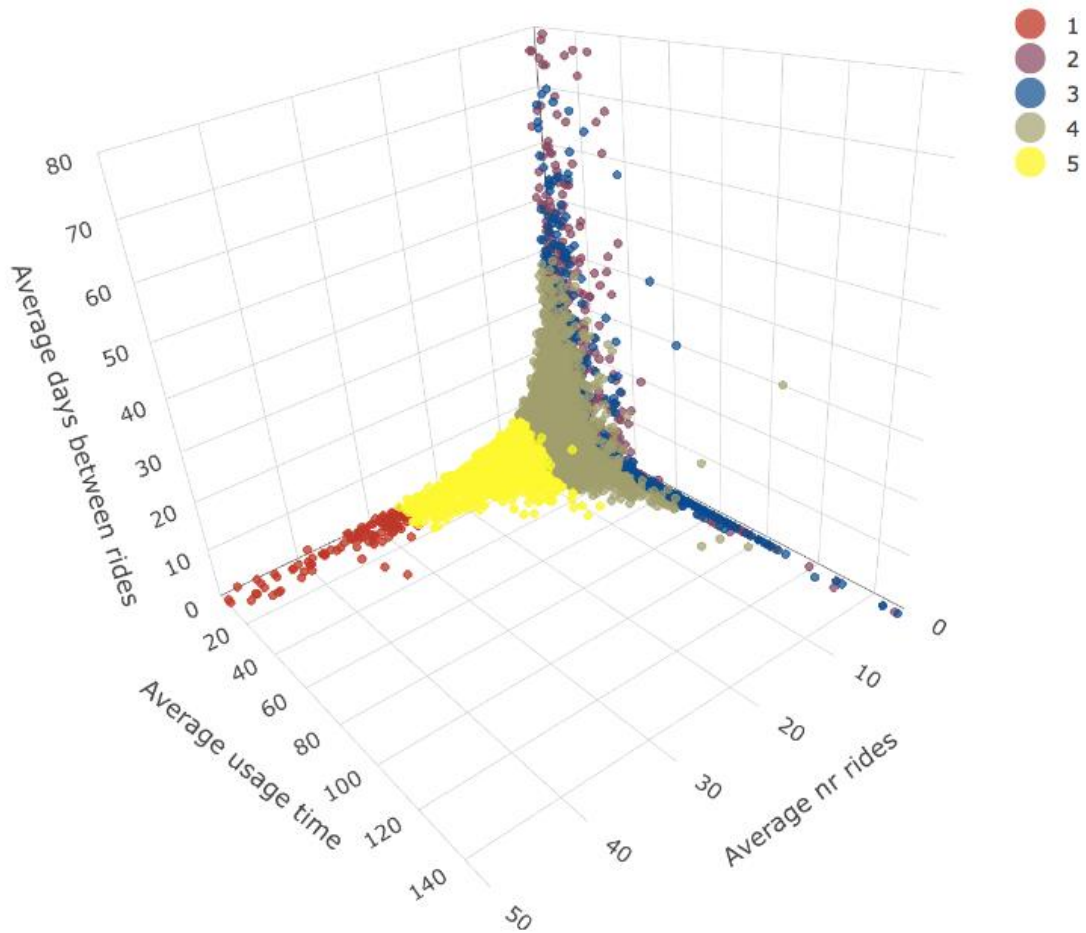


Figure D.1: Spatial visualisation of clustering results - first view

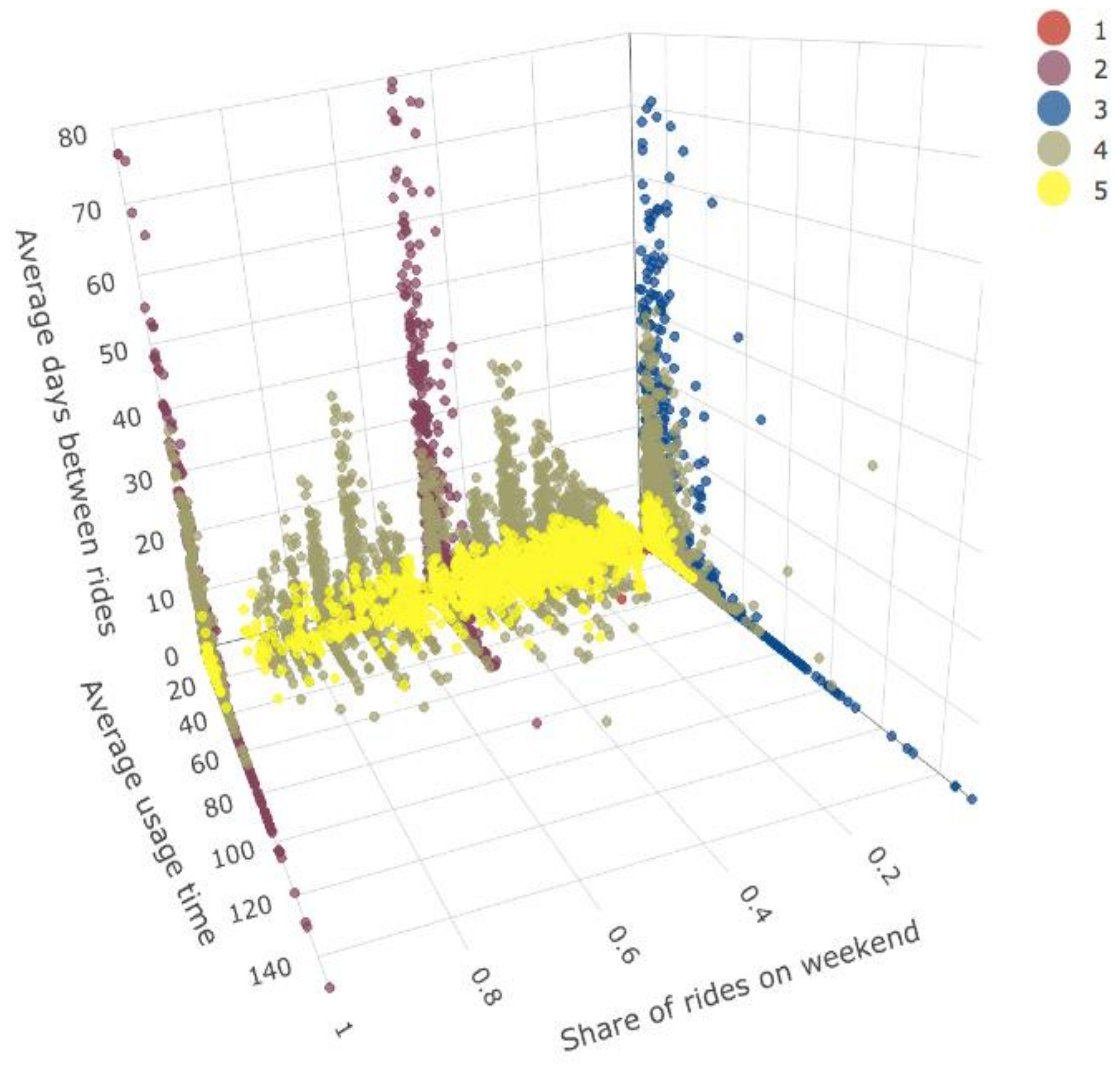


Figure D.2: Spatial visualisation of clustering results - second view

Appendix E: Demand forecast - exploratory analysis

Table E.1: Statistical analysis of weather explanatory variables

	Precipitation (in)	Maximum temperature (°F)	Maximum wind speed (mph)	Maximum humidity (%)
Minimum	0	59	9	58
1st quartile	0	64	13,75	76
Median	0	67	16	87
Mean	0,01619	67,21	16,37	83,62
3rd quartile	0	70	29	93,25
Maximum	0,43	82	31	100

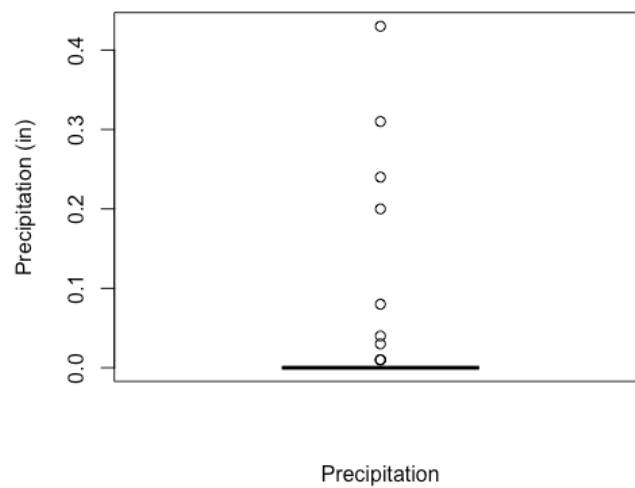


Figure E.1: Box plot - daily precipitation (in)

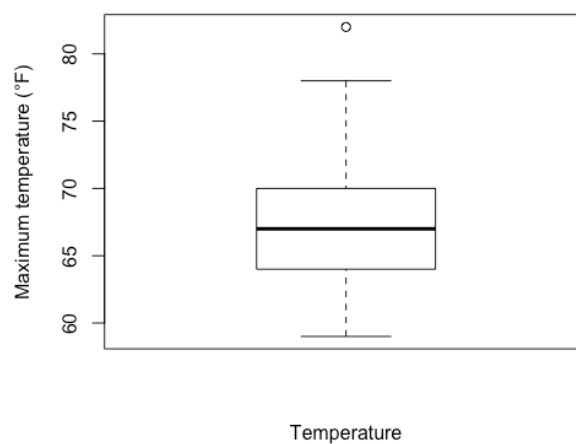


Figure E.2: Box plot – daily temperature (°F)

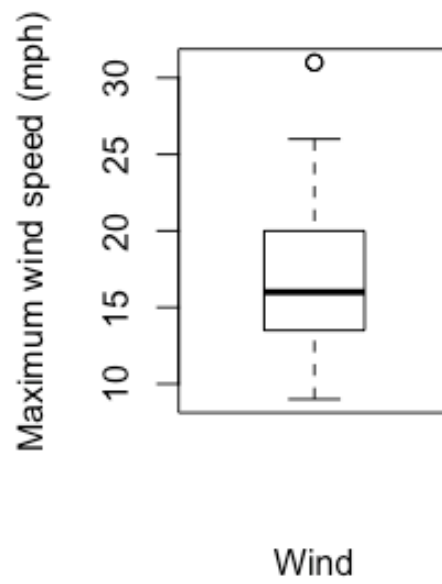


Figure E.3: Box plot - daily speed wind (mph)

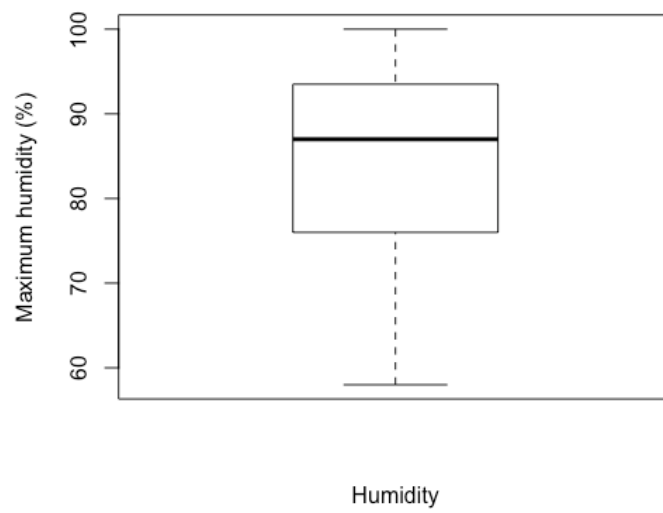


Figure E.4: Box plot – daily humidity (%)