FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Comportamentos Atípicos em Espaços Comerciais: Avaliação do Impacto de Eventos Externos

Vítor Castro

Mestrado Integrado em Engenharia Informática e Computação

Orientador: Rui Maranhão

Co-orientador: Fernando Freitas

26 de Janeiro de 2015

Comportamentos Atípicos em Espaços Comerciais: Avaliação do Impacto de Eventos Externos

Vítor Castro

Mestrado Integrado em Engenharia Informática e Computação

26 de Janeiro de 2015

Dedicado aos meus Pais:
Jerónimo Fernando Silva Castro e
Clementina Vieira Ferreira Castro

Resumo

Nos dias atuais, os retalhistas estão cada vez mais a apostar em tecnologias que permitam conhecer quantos, onde e como se movimentos os consumidores no interior dos seus espaços comerciais de forma a aumentar a sua vantagem competitiva. Um exemplo dessa tecnologia é o BIPS *(Business Intelligence Positioning System)* que, através da deteção de sinais de radiofrequência dos dispositivos móveis, consegue traçar os percursos dos consumidores.

Esta sistema já se encontra implementado, fornecendo ao gestor do espaço comercial variadas informações sobre os seus visitantes. No entanto, devido ao grande volume de dados, torna-se árdua a tarefa de análise dos dados. De forma a facilitar esta tarefa, propomos um método automático de descoberta de valores atípicos no comportamento dos consumidores. Esse método foi aplicado ao número de visitas a um centro comercial usando os dados recolhidos pelo BIPS .

De forma a avaliar se eventos externos são responsáveis pelos comportamentos atípicos num espaço comercial, procedeu-se à análise da influência da realização de eventos desportivos numa loja de comércio de produtos desportivos. Sabendo o impacto desses eventos, os gestores dos espaços comerciais podem implementar medidas preventivas, mudando o seu paradigma de gestão reativa para um paradigma de gestão proactiva.

Palavras Chave: *Data mining*, retalho, análise de dados, marketing

Abstract

Currently, retailers are investing on technologies that allow to know how many, where and how consumers walks within their commercial areas in order to increase their competitive advantage. An example of this technology is BIPS *(Business Intelligence Positioning System)* that detects radio signals from mobile devices, tracing the trajectories of customers.

This system is already implemented, providing retail managers data about their visitors. However, due to the large volume of data, it becomes hard to analyze such amount of data. In order to make this task easier, we propose an automatic method of finding outliers in consumer behavior. This method was applied to the number of visits to a shopping center using the data collected by BIPS.

In order to assess whether external events are responsible for the atypical behavior in a store, we analyze the influence of sports events in store. Knowing the impact of these events, retail managers can implement preventive measures, evolving from a reactive to a proactive management.

Keywords: Data mining, retail, data analysis, marketing

Agradecimentos

Este espaço está reservado para as pessoas que considerei mais importantes e que fizeram parte deste meu percurso académico.

Agradeço em primeiro lugar aos meus pais sem os quais não seria aquilo que sou hoje.

Agradeço ao Professor Doutor Rui Maranhão pela sua orientação durante esta dissertação e por toda a sua ajuda e disponibilidade.

Agradeço a todos os colaboradores da Movvo em especial aos engenheiros Fernando Freitas, Tiago Rodrigues e João Teixeira por toda a ajuda e disponibilidade que me deram ao longo desta dissertação.

Por fim agradeço a todos os meus amigos por terem estado presentes e por me terem apoiado.

A todos, o meu Muito Obrigado.

Vítor Castro

"Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito.
Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes"

Marthin Luther King

Conteúdo

Introdução 1
1.1 Contexto/Enquadramento 2
1.2 Motivação e Objetivos 2
1.3 Definição do Problema 4
1.4 Questões de Pesquisa 4
1.5 Estrutura da Dissertação 5
Estado da Arte 6
2.1 BIPS 6
2.2 Data Mining 15
2.3 Conclusão 24
Tecnologias Utilizadas 25
3.1 JSON 25
3.2 R 26
3.3 Conclusões 27
Deteção de números de visitas atípicos num centro comercial 28
4.1 Compreensão de Negócio 28
4.2 Compreensão dos Dados 29
4.3 Preparação dos Dados 43
4.4 Modelação 44
4.5 Avaliação 49
4.6 Desenvolvimento 51
Avaliação do impacto de eventos numa loja de desporto 56
5.1 Compreensão de Negócio 56
5.2 Compreensão dos Dados 58
5.3 Preparação dos Dados 65
5.4 Modelação 65
5.5 Avaliação 70
5.6 Desenvolvimento 73
Conclusões e Trabalho Futuro 74
6.1 Conclusões Finais 74
6.2 Satisfação dos Objetivos 75
6.3 Trabalho Futuro 76

Referências 77

Lista de Figuras

2.1 Logótipo do BIPS 8
2.2 Arquitetura do Sistema BIPS 9
2.3 Relatório do BIPS - Parte 1 10
2.4 Relatório Final do BIPS - Parte 2 12
2.5 Exemplo de *outliers* num conjunto bidimensional 19
2.6 Ciclo CRISP-DM 21
2.7 Representação das fases e tarefas do modelo CRISP-DM 22
2.8 Descrição de um diagrama de caixas de bigodes 24

4.1 Serie Temporal - Número de visitas ao longo do tempo 34
4.2 Gráfico de barras da média das visitas por dia da semana 35
4.3 Diagrama de caixas de bigodes – Número de visitas por dia da semana 36
4.4 Gráfico de barras das média de visitas por mês 37
4.5 Diagrama de caixas de bigodes – Número de visitas por mês 38
4.6 Gráfico de barras da média das visitas por estado do tempo 39
4.7 Diagrama de caixas de bigodes – Número de visitas por estado do tempo 40
4.8 Teste de Tukey HSD – Variável Meses 43
4.9 Teste de Tukey HSD –Variável *Dia da Semana* 45
4.10 Teste de Tukey HSD –Variável *Clima* 47
4.11 Abordagem 1 –Percentagem de *outliers* calculados 49
4.12 Abordagem 2 – Percentagem de *outliers* calculados 50
4.13 Abordagem 3 – Explicação Gráfica 50
4.14 Abordagem 3 – Percentagem de *outliers* calculados 51
4.15 Abordagem 3 – Percentagem de *outliers* removidos 52
4.16 Abordagem 3 – Percentagem de *outliers* calculados com novo valor de exclusão . 52
4.17 Abordagem 3 – Percentagem de *outliers* removidos com novo valor de exclusão 53
4.18 Diagrama de arquitetura 56
4.19 Modelo de Dados da BD de eventos 58

5.1 Média das Visitas por Zona por Dia 62
5.2 Média do Tempo de Permanência por Zona por Dia 63
5.3 Número Médio de Visitas e Tempo Médio de Permanência por Zona por Dia . . . 64
5.4 Gráfico de barras da média das visitas por dia da semanas 65
5.5 Gráfico de barras da média do tempo médio de permanência por dia da semana . 66
5.6 Gráfico de barras da média das visitas por mês 67
5.7 Total de visitas à loja de desporto ao longo dos meses 69
5.8 Análise do Número Médio de Visitas por Dia nas Zonas 70
5.9 Análise do Tempo Médio de Permanência por Dia nas Zonas 70

5.10 Análise das Primeiras Visitas à Zona Calçado Running 72
5.11 Análise das Primeiras Visitas à Zona Têxtil Equipamento Running 72

Lista de Tabelas

2.1 Comparação entre BIPS e seus concorrentes 16

4.1 Descrição do conjunto de dados relativo ao número de visitas por dia 33
4.2 Resumo estatístico da variável Temperatura Média 34
4.3 Resumo da variável Tempo 35
4.4 Resumo da variável Visitas 36
4.5 Descrição das novas variáveis 37
4.6 Média das visitas por dia da semana 38
4.7 Total das visitas por mês e média por mês 39
4.8 Média das visitas por estado do tempo 40
4.9 Tabela da Análise de Variância - Variável Meses 41
4.10 Tabela da Análise de Variância - Variável *Weekday* 44
4.11 Tabela da Análise de Variância - Variável *Clima* 46

5.1 Descrição do conjunto de dados de deteções da loja de desporto 61
5.2 Descrição do conjunto de dados da loja de desporto após operações 62
5.3 Número médio e tempo médio de visitas por zona 63
5.4 Resumo da variável Data 65
5.5 Média das visitas por dia da semana 65
5.6 Tempo médio de permanência por dia da semana (min.) 66
5.7 Média das visitas por mês 67
5.8 Tabela de eventos próximos dos dias *outliers* 69
5.9 Tabela Estatística do Número Médio de Visitas às Zonas 71
5.10 Tabela Estatística do Tempo Médio de Permanência nas Zonas (min.) 71
5.11 Percentagem da Média das Primeiras Visitas às Zonas 74

Abreviaturas

BIPS	*Business Intelligence Positioning System*
RTLS	*Real Time Locating System*
GPS	*Global Positioning System*
Wi-Fi	*Wireless Fidelity*
IMSI	*International Mobile Subscriber Identity*
MAC	*Media Access Control*
GSM	*Global System for Mobile Communications*
IDE	*Integrated Development Enviornment*
API	*Application Programming Interface*
BD	Base de Dados
HTTP	*Hypertext Transfer Protocol*

Capítulo 1

Introdução

O comportamento do consumidor é cada vez mais objeto de estudo por parte dos retalhistas. Estes estão cada vez mais convictos que o conhecimento sobre o comportamento dos utentes dos seus espaços comerciais pode trazer vantagem competitiva sobre os seus concorrentes. Este conhecimento servirá de base para o estreitamento da relação com o cliente, de forma a prever as suas necessidades e a melhorar a experiência de compra. Estes fatores contribuem para o aumento da taxa de retenção do cliente: uma preocupação dos retalhistas de pequenas e grandes dimensões [Mil02].

Em negócios de pequena dimensão, como por exemplo lojas de rua, torna-se relativamente simples conseguir estes objetivos enquanto em grandes superfícies comerciais, a dificuldade em atingir estes objetivos aumenta consideravelmente. É o caso dos centros comerciais e dos grandes retalhistas especializados. Nestes espaços existem gestores que têm a responsabilidade de tomar decisões que originem uma maior perceção de valor por parte dos clientes e também no aumento da margem bruta de lucro através das vendas e da redução de custos [cJK11]. Torna-se fundamental para estes gestores tirar proveito de informação relevante dos visitantes destes espaços comerciais.

Por este motivo, temos assistido a um maior investimento em tecnologias e processos que permitam obter o máximo de informação possível sobre os consumidores. Esta informação consiste num grande conjunto de dados armazenados, a partir de diversas fontes, em bases de dados. Devido ao grande volume de dados, torna-se insuficiente apenas a aquisição de dados. Estes devem ser traduzidos em conhecimento através de técnicas automatizadas. Recorre-se a ferramentas de tratamento de dados e produzem-se relatórios que devem ser fornecidos aos retalhistas para facilitar o seu trabalho diário.

1.1 Contexto/Enquadramento

Esta dissertação enquadra-se na área do retalho, mais concretamente num centro comercial e numa loja de desporto, e também na área de marketing associada a estes dois negócios.

Este tema proposta pela empresa Movvo - empresa criada em maio de 2009 e ligada a atividades de consultoria e programação informática - à Faculdade de Engenharia da Universidade do Porto no âmbito de dissertações de mestrado e foi orientada pelo Professor Rui Maranhão da Faculdade de Engenharia da Universidade do Porto e pelo Engenheiro Fernando Freitas da empresa mencionada.

A Movvo foi criada por três investigadores universitários que têm como principal objetivo estabelecer uma ponte entre o mundo académico e o empresarial. A empresa destaca-se principalmente nas áreas de Segurança e Computação, Redes de Pesquisa Operacional e *Marketing Research*. Esta empresa investiu no desenvolvimento da sua própria tecnologia, o BIPS, e baseando-se no sucesso deste conseguiu um grande reconhecimento no mercado ao desenvolver vários produtos baseados nesta tecnologia.

O BIPS é uma tecnologia de localização de alta precisão de dispositivos móveis, através da deteção dos sinais de radiofrequência que estes emitem em tempo real, permitindo-o seguir de forma anónima o sinal ao longo do tempo. A partir desta tecnologia, é possível saber, por exemplo, quantas pessoas estão num determinado espaço, quais os percursos utilizados, as zonas mais visitadas, o tempo médio de permanência e as lojas visitadas. O presente projeto irá ser desenvolvido tendo como base os dados recolhidos por esta tecnologia que vai ser descrito na secção 2.1 deste documento.

1.2 Motivação e Objetivos

O sucesso de qualquer espaço comercial passa pela compreensão do uso que é feito pelos utilizadores e intervindo na sua alteração em caso de necessidade. Até recentemente, os retalhistas limitava-se ao uso de contadores nos acessos para saber o número de visitantes dentro do espaço comercial. Apesar de ser possível saber o número de visitas por dia, por hora e por acesso, não era possível saber o comportamento dos visitantes no interior do espaço comercial, como por exemplo, o tempo médio em compras, o tempo médio em passeio ou os caminhos percorridos. Outros métodos de compreensão do comportamento do consumidor, baseiam-se em pesquisas como entrevistas ou estudos de marketing, que tendem a entender o processo cognitivo do ato de compra, em vez dos padrões de consumo num contexto real de compra. Além disso, estes estudos são geralmente dispendiosos e baseiam-se em amostras pequenas da população.

É neste contexto que surge a tecnologia BIPS, que se baseia na deteção de dispositivos móveis ao longo do tempo, permitindo saber em tempo real, não só o números de visitantes, mas também a sua localização durante a visita, o tempo passado em cada zona e o percurso efetuado desde o acesso ao espaço comercial. Estas informações são valiosas para o suporte a decisões, como a melhoria da

qualidade do serviço, a redução de tempos de espera nas filas ou otimização do *layout* do espaço comercial. Além disso, é importante para outras tarefas a nível operacional ou de marketing, tais como: a otimização do escalonamento de colaboradores para determinado períodos; a amplificação da eficiência das tarefas de marketing, como por exemplo, a adaptação do espaço em função do comportamento e dos hábitos dos visitantes ou simplesmente a colocação de painéis de publicidade em locais relevantes; ou a avaliação do sucesso de campanhas comerciais de lojas, de modo a concluir se esta teve sucesso ou não só na própria loja, atraindo um maior número de visitantes no espaço comercial e aumento de compras não planeadas.

No entanto, uma tecnologia como o BIPS, por si só é insuficiente, uma vez que os dados de captação não se traduzem em conhecimento. É necessária uma análise dos dados armazenados, através de técnicas de *data mining*, para extrair conhecimentos de suporte à decisão. Atualmente, a Movvo possui um produto – o Retail Movves – que fornece um conjunto de métricas, num *dashboard online*, indicando o número de visitantes, tempos médios de permanência, percursos agrupados por diferentes períodos de tempo e por diferentes zonas lógicas configuradas para um determinado espaço comercial.

O produto apresenta as métricas calculadas, no entanto ainda depende da avaliação de um analista que, observando os dados, deverá indicar se o comportamento está dentro da normalidade e, no caso de haver mudanças face ao comportamento padrão por parte dos consumidores, deverá encontrar os motivos para a alteração do comportamento. Além disso, o conceito de "normalidade" pode ser difícil de ser avaliado pelo analista, uma vez que deverá ter em consideração todo o historial de visitas ao espaço comercial, considerando o contexto (dia da semana, época do ano, clima, época de promoções ou eventos dentro e fora do espaço comercial).

Esta dissertação tem como objetivo a avaliação do impacto de eventos externos, quer sejam campanhas ou eventos patrocinados, através de análise do histórico do comportamento do centro comercial, relacionando-o com eventos externos. Em primeiro lugar, irá ser criado um método de identificação de valores atípicos (*outliers*) nas visitas a um determinado centro comercial, indicando-se os eventos externos realizados nesses dias; e em segundo lugar, analisar-se-á se eventos desportivos patrocinados por uma loja de comércio de produtos desportivos, incitou alterações de comportamento nessa loja. Para cumprir estes objetivos irão ser utilizadas técnicas de *data mining*. Estes objetivos de *data mining* sustentam o objetivo de negócio relacionado com a melhoria dos processos de análise, tornando mais ágil a avaliação das alterações do comportamento, indicando a possível causa dessa alteração. Assim, o utilizador final irá facilmente ter o conhecimento de todas as situações anómalas relativas ao número de visitantes no espaço comercial com menor esforço e tempo. Isto enriquecerá o BIPS com funcionalidades inovadoras que se traduzirão num aumento de valor para o produto e para os clientes.

Existem tecnologias concorrentes do BIPS e com produtos semelhantes que apresentam os dados agregados em aplicações ou relatórios gerados automaticamente através de técnicas de análise de dados, no entanto deixam a análise e interpretação dos dados a cargo dos utilizadores do produto. A Movvo pretende com a análise de dados agregados facilitar o trabalho dos gestores de espaços comerciais, diferenciando o seu produto dos demais concorrentes e mantendo-se na vanguarda do seu mercado de atuação.

1.3 Definição do Problema

O principal problema é o facto de ser necessário um esforço do gestor do espaço comercial para interpretar os resultados do Retail Movves para conseguirem tomar decisões que se mostrem rentáveis para o espaço comercial da sua responsabilidade. Pretende-se especificar, no relatório final produzido, anomalias ou comportamentos inesperados nos dados processados relativos ao número de visitas. Assim, requer-se a criação de um conjunto de caraterísticas que façam parte de imediato do BIPS resultando assim num aumento de valor do produto.

Pretende-se também tentar justificar as situações anormais detetadas nos dados processados com eventos para tentar demonstrar o impacto que estes possam ter no número de visitas numa loja ou num centro comercial.

1.4 Questões de Pesquisa

Com vista o desenvolvimento da aplicação a que se refere esta dissertação, foi encontrada a necessidade de procurar respostas a algumas questões essenciais inerentes ao projeto.

Tendo em conta que para a resolução do mesmo seria necessário o recurso a técnicas de *data mining*, considerou-se fulcral pesquisar informação relativamente a este ponto. Procurou-se perceber em que consistia *data mining* e que possíveis tipos de algoritmos seriam úteis para a aplicação final. Concluiu-se, que tendo em conta os problemas referidos, seria necessário pesquisar sobre *outliers* e formas de deteção destes para tentar encontrar situações anómalas no número de visitas. Assim surgem questões como o que é um *outlier*? Como será possível fazer uma deteção de *outliers*?

Uma vez que o sistema a desenvolver será incorporado no BIPS (secção 2.1), foi necessário estudar o mesmo. Foi necessária a compreensão do seu funcionamento e das métricas captadas pois estas serão fundamentais para o projeto atual para dar resposta a perguntas como qual a razão que levou à criação do BIPS? De que modo se torna útil para o gestor de uma superfície comercial?

Sentiu-se também a necessidade de saber quais as aplicações existentes no mercado dentro das áreas mencionadas. Fez-se portanto um estudo relativo às aplicações existentes no mercado e uma análise com o objetivo de descobrir se alguma conseguia atualmente responder ao problema definido.

1.5 Estrutura da Dissertação

A presente dissertação encontra-se estruturada da seguinte forma:

- Capítulo 1 - Introdução
- Capítulo 2 -Estado da Arte em que se estuda os produtos concorrentes do BIPS e ainda técnicas de *data mining* necessárias para ajudar a resolver os problemas encontrados.
- Capítulo 3 - Abordagem das tecnologias que irão ser utilizadas para o desenvolvimento da aplicação.
- Capítulo 4 - Detalhe da implementação da ferramenta de deteção de *outliers* para um centro comercial.
- Capítulo 5 - Apresentação da avaliação do impacto de eventos externos numa loja de desporto.
- Capítulo 6 - Apresentação das conclusões finais.

Capítulo 2

Estado da Arte

Neste capítulo é descrita a tecnologia BIPS, a partir da qual são obtidos os dados de deteção, que servem de suporte para as métricas indicadoras de desempenho de um espaço comercial. É também apresentada uma lista dos principais concorrentes no mercado, alguns com princípios tecnológicos diferentes, explorando as vantagens e as desvantagens face à tecnologia da Movvo. No capítulo ainda são descritos os métodos de análise de dados, ou seja, as técnicas de *data mining* e conceitos que se consideram adequados para a compreensão do projeto e para o cumprimento dos objetivos expostos.

2.1 BIPS

O BIPS é a tecnologia que irá servir de base para o desenvolvimento do projeto. Vai ser descrito o seu modo de funcionamento, a sua arquitetura, as questões de privacidade e segurança, o produto e finalmente uma análise da concorrência.

2.1.1 Descrição Geral

O BIPS, principal tecnologia da empresa Movvo, é uma tecnologia inovadora que deteta sinais de radiofrequências (GSM, WI-FI, ou *Bluetooth*) de dispositivos móveis em tempo real - RTLS - e permite seguir de forma anónima esse sinal ao longo do tempo num determinado espaço. A tecnologia encontra-se patenteada como "*Tagless radio frequently based self correcting distributed real time location system*" [DCC13].

O BIPS destaca-se pela sua eficiência em seguir o sinal de um dispositivo móvel em espaços interiores assim como em espaços exteriores. Muitos dos sistemas RTLS existentes são conhecidos pela sua eficiência em seguir um sinal em espaços exteriores mas não o conseguem fazer em espaços interiores devido à ausência de sinal ou devido às várias interferências existentes.

Figura 2.1: Logótipo do BIPS

O GPS - o sistema mais popular de localização no exterior - torna-se ineficaz em espaços interiores pois o sinal nestes espaços degrada-se bastante devido, por exemplo, a barreiras físicas ou a perturbações magnéticas [Arb11].

Cada dispositivo móvel, como um *tablet* ou um telemóvel, emite um sinal de radiofrequência com uma potência e identificação única, por exemplo, o MAC *address* ou o IMSI. O BIPS captura esse sinal através de antenas construídas para este fim. São necessárias pelo menos três antenas para a localização de um dispositivo num determinado local, enquanto para deduzir a presença no local basta apenas uma antena. É através destas antenas que é feita uma aproximação da distância e do posicionamento do dispositivo que emite o sinal.

Visto que o sinal emitido por um dispositivo pode ser afetado por vários fatores externos, tais como o ruído eletromagnético, o número de pessoas e até mesmo a temperatura ambiente, o BIPS tem a capacidade de se autocalibrar. É importante referir que a privacidade do visitante, em todas as fases, está assegurada como se poderá verificar neste capítulo.

2.1.2 Arquitetura do Sistema

O BIPS é constituído fundamentalmente por três tipos de componentes: dois instalados no local de monitorização que são responsáveis pela captura e agregação dos dados e um terceiro componente localizado remotamente num *data center* para processamento dos resultados como ilustra a figura 2.2. O "*bNode*" é um nó composto por hardware específico que inclui uma antena para deteção de radiofrequências que enviam a informação das deteções sem qualquer processamento. O "*bServer*" é um servidor local que armazena e garante a privacidade e segurança dos dados. Este componente é responsável pela agregação dos dados diários que recebe de todos os "*bNodes*" a ele conectados, pelo cálculo das posições e pelo envio da informação para a base de dados. O servidor de resultados é responsável pela agregação dos resultados que poderão ser consultados pelo utilizador final numa interface gráfica.

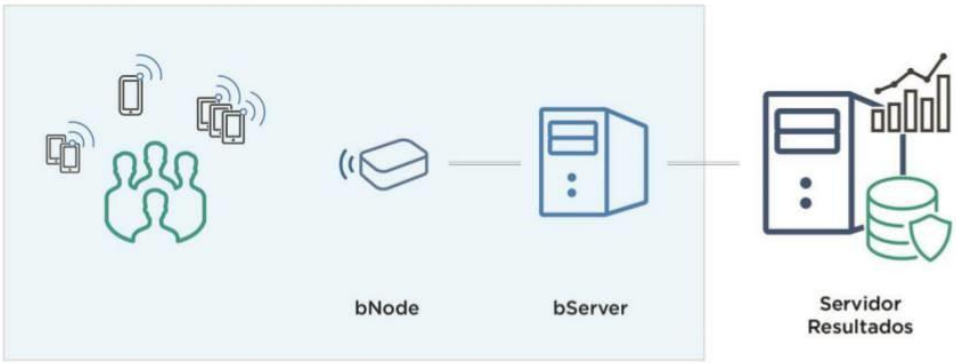


Figura 2.2: Arquitetura do Sistema BIPS

2.1.3 Segurança e Privacidade

O BIPS cumpre, em todas as fases, todos os requisitos de segurança de informação estabelecida no padrão ISO 27001. Isto garante um modelo adequado de implementação e operação do BIPS garantindo assim que a informação captada é tratada com elevados padrões de gestão e de proteção ao nível da segurança da informação que está em todas as fases protegida [iso05].

É captada pelas antenas do BIPS o MAC *address* ou o IMSI do dispositivo móvel. Este é codificado em tempo real antes de ser armazenado no sistema com recurso a uma função *hash*. Assim torna-se impossível reconhecer o dispositivo. Caso o visitante volte a deslocar-se ao espaço comercial, num dia diferente, será atribuída uma nova referência de modo a que o rasto seja perdido e não seja possível reconhecer o dispositivo em questão nem seguir-lhe os seus passos em dias diferentes.

Relativamente à comunicação entre componentes, é necessária uma chave privada e todas as mensagens trocadas entre o BIPS são encriptadas com recurso ao algoritmo *Advanced Encryption Standard* - AES [PT01]. O único ataque conhecido a este algoritmo demora um tempo computacional impraticável a tentar comprometer o sistema.

2.1.4 Retail Movves

Como referido na secção relativa à arquitetura do BIPS, são necessários três componentes para o correto funcionamento. Torna-se importante fazer a distinção de tecnologia e do produto. A tecnologia resume-se à capacidade do sistema detetar os dispositivos móveis através do componente "*bNode*"e do respetivo armazenamento dessa informação captada no componente "*bServer*".

Como foi referido, os dados de deteção das antenas, contendo a identificação devidamente cifrada e a localização dos dispositivos, são guardados na base de dados sem qualquer tratamento. De forma periódica, existe um processo que faz a consulta à base de dados para transformar os dados de deteção em métricas: número de visitas ao espaço e às zonas configuradas no sistema, percursos efetuados, revisitas efetuadas no mesmo dia, entre outras. Estas métricas são disponibilizadas num *dashboard* - o Retail Movves. Estas métricas podem ser consultadas, por

escolha do utilizador, em diferentes períodos temporais (hora, dia, semana ou mês) e apresentadas de diferentes formatos.

De forma a fornecer mais detalhes sobre as métricas e o tipo de análises disponibilizadas no *dashboard*, procede-se em seguida à análise de um dos componentes relevantes do *dashboard* – um relatório semanal com algumas métricas. Este relatório final parcial encontra-se dividido em duas imagens (Figura 2.3 e Figura 2.4) e segue-se então uma descrição dos pontos mais relevantes do mesmo.

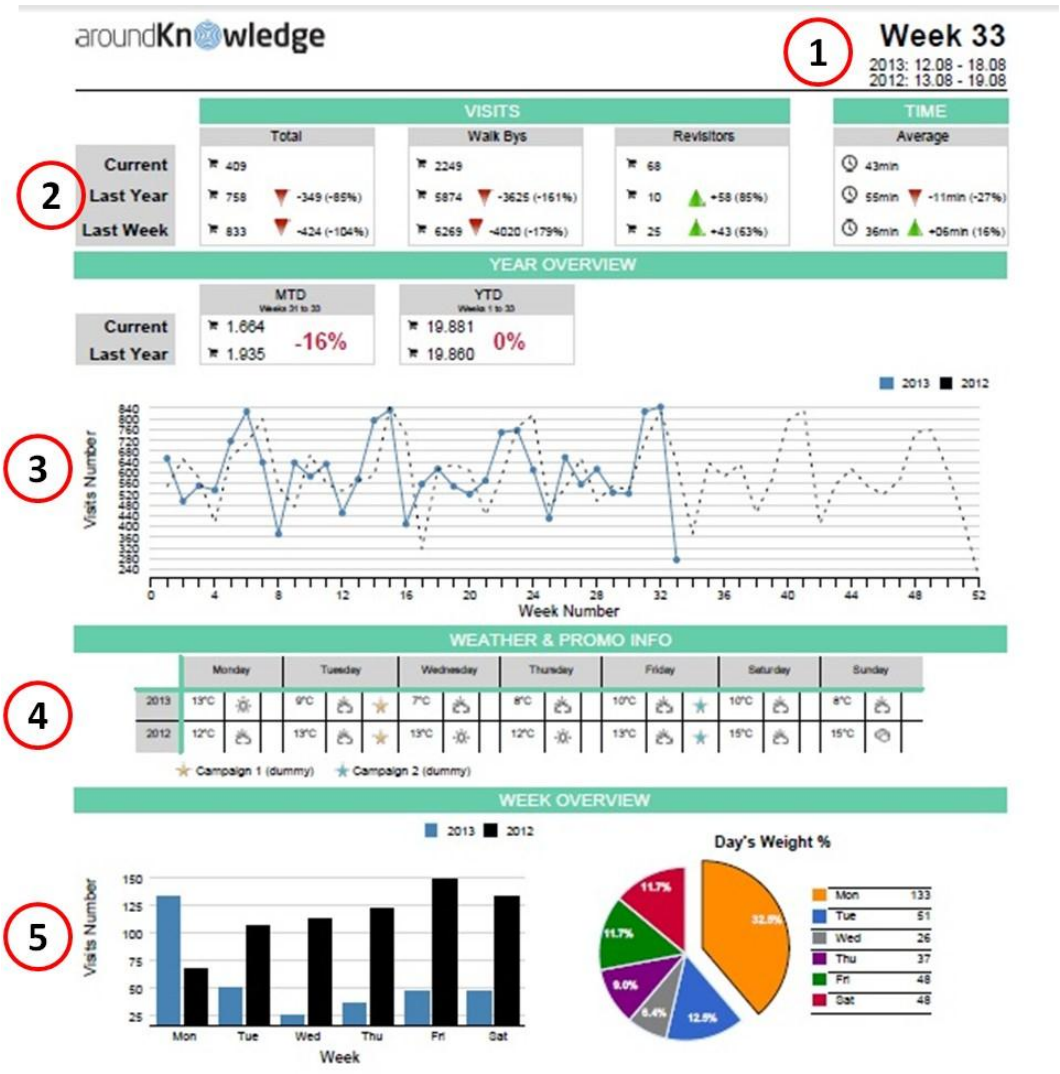


Figura 2.3: Relatório do BIPS - Parte 1

No ponto 1 encontramos o período ao qual se refere a análise. Este relatório corresponde à semana 33 do ano 2013 entre os dias 12 e 18 de agosto. Todos os relatórios são confrontados com o período homólogo no ano anterior que corresponde à mesma semana de 2012. No exemplo, a semana 33 do ano 2012 corresponde ao período entre os dias 13 e 19 de agosto.

No ponto 2, são apresentados os números de visitas, os números de pessoas que foram detetados fora do espaço (mas que não entraram) e os números de pessoas que regressaram no mesmo dia durante essa semana (33 de 2013), à semana do período homólogo do ano anterior (33 de 2012) e à semana anterior (32 de 2013). É ainda mostrado o tempo médio de visita. A acompanhar os valores da semana do período homólogo do ano anterior e da semana anterior estão as variações em valor absoluto e em valor percentual.

No ponto 3 é mostrada a análise gráfica anual das visitas e efetua-se a comparação com o ano anterior. A linha a cheio representa o ano em análise e mostra apenas os valores até à semana a que se refere o relatório (semana 33) e a linha a tracejado representa o número de visitas por semana do ano anterior razão pela qual a linha tracejado continua atá à semana 52.

No ponto 4 encontra-se uma comparação meteorológica dos dois anos para a semana em questão. Esta informação é importante para ajudar a relacionar o número de visitas com a meteorologia.

No ponto 5 observa-se uma análise gráfica semanal. No gráfico de barras é possível uma comparação do número de visitas por dia da semana que está a ser analisada e da semana do ano anterior correspondente e no gráfico circular temos a informação percentual e numérica do número de visitas por dia da semana em análise.

No ponto 6 (figura 2.4) observa-se análises tabular e gráfica das visitas por hora para a semana em análise. O detalhe permite dar ao utilizador final uma perceção das horas mais movimentadas por dia.

No ponto 7 está representado em formato de tabela e gráfico circular as visitas por zona do centro comercial para a semana em questão. As zonas são configuradas no *dashboard* e não existe um número, nem área limites para a sua definição. Este relatório mostra um espaço onde foram definidas apenas três zonas.

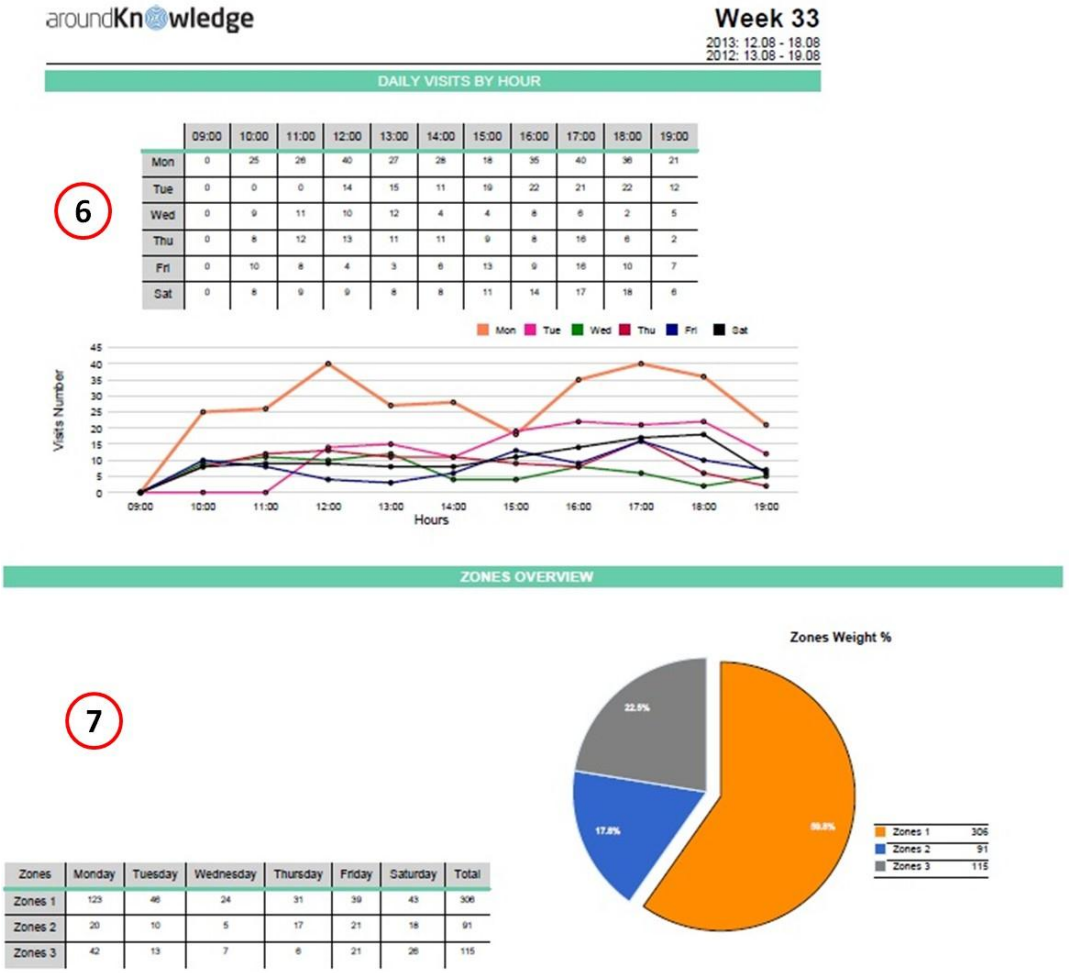


Figura 2.4: Relatório Final do BIPS - Parte 2

É importante referir que por questões de confidencialidade, os valores disponibilizados neste relatório final são fictícios.

2.1.5 Análise de Concorrência

Existem atualmente no mercado ferramentas que podem ser consideradas similares ao BIPS. Procede-se agora a uma pequena análise dos seguintes produtos concorrentes:

- PeCo
- Footfall
- Euclid
- Path Intelligence
- Nomi

2.1.5.1 PeCo

O PeCo, ou People Counting, é uma ferramenta que inclui técnicas de análise de imagens em sistemas de vigilância e permite a análise e o processamento de informações contidas nas imagens de vídeo e possibilita a obtenção de um conjunto de dados úteis para a eficiência operacional e funciona como um suporte inteligente ao negócio [webf].

O PeCo possibilita várias informações tais como o percurso dos visitantes, informações que possibilitam uma análise de tendências das pessoas, dias e horários com mais/menos movimento [webf]. Permite ainda algo que o BIPS não permite que é uma estimativa da idade das pessoas e ainda saber o género dos visitantes graças à análise de vídeo. Cada câmara tem incorporado um sensor inteligente que analisa as imagens obtidas pela respetiva câmara.

Este sistema trabalha com câmaras específicas, caso se pretenda instalar equipamentos para não interferir com o sistema de vigilância ou no caso de não se pretender instalar novos equipamentos. É também possível incorporar o PeCo no sistema de vigilância funcionando assim ambos em simultâneo [webf].

Foi realizado um estudo, por iniciativa da Movvo, com o intuito de comparar o desempenho do BIPS com o do PeCo colocando os dois sistemas a funcionar em paralelo. Relativamente à contagem de pessoas o PeCo contabilizou um número de pessoas superior ao dobro do número de pessoas contabilizado pelo BIPS. Concluiu-se que o BIPS apresentou mais valores mais próximos da realidade e a razão apontada para tamanha diferença de valores foi o facto de o PeCo contar pessoas repetidas quando uma pessoa estava no raio de alcance de mais do que uma câmara.

2.1.5.2 Footfall

O Footfall é uma ferramenta que tem como objetivo ajudar gestores de negócio com tomadas de decisão que possibilitem um aumento dos lucros através da análise do que se passa nos locais de venda [webb]. Esta aplicação permite a contagem de pessoas, permite uma análise do comportamento das pessoas nas filas de espera e, tal como o BIPS e o PeCo, a elaboração de um relatório com a análise dos dados [webb].

O Footfall funciona com base em detetores de infravermelhos posicionados nas entradas/saídas das lojas ou dos centros comerciais e também nas caixas de pagamentos, quando se trata de lojas para compreensão do comportamento das pessoas na fila de espera. Relativamente às filas de espera, o sistema possibilita o cálculo, em tempo real, das extensões médias das filas, os tempos de espera e os tempos de transação [webb].

Relativamente à contagem do número de pessoas, os sensores nas entradas/saídas permitem de facto que se tenha uma noção do número de visitantes. Contudo este sistema conta pessoas repetidas e não permite a total perceção se um visitante está a entrar ou a sair da loja. Não é

possível também perceber o comportamento do visitante visto que os sensores não permitem seguir o rasto da pessoa no interior da loja.

Conclui-se que o ponto forte do Footfall é a informação recolhida relativa às filas de espera das lojas pois é possível perceber o tamanho da fila de espera e o tempo que uma pessoa passa na fila e também o tempo que uma pessoa demora a ser atendida logo pode-se considerar uma aplicação mais útil para lojas do que propriamente para as zonas exteriores às lojas como é o caso do centro comercial.

2.1.5.3 Euclid

O Euclid é uma ferramenta que pretende demonstrar o comportamento das pessoas num determinado espaço. O Euclid permite seguir o sinal de um dispositivo móvel com tecnologia WI-FI [weba]. Tal como o BIPS, esta ferramenta possibilita a contagem do número de pessoas, determinar o tempo que cada pessoa passa numa determinada zona, saber o caminho que percorrem entre outras métricas.

A privacidade da pessoa também está assegurada com esta ferramenta pois os dispositivos móveis emitem um pequeno *"ping"* quando procuram por redes WI-FI para uma conexão. Estes *"pings"* incluem o endereço MAC do dispositivo mas o Euclid está preparado para codificar este endereço recorrendo a um algoritmo de *hash* unidirecional [weba].

Para além do relatório final, adaptado ao cliente, com as informações consideradas essenciais para o negócio, existe uma funcionalidade que possibilita o envio das informações mais importantes para o correio eletrónico ou para o telemóvel do responsável da área de negócio sendo esta uma característica útil e que não está presente nos sistemas concorrentes.

Esta ferramenta é considerada concorrente direta do BIPS pois assenta em tecnologias muito semelhantes mas o BIPS destaca-se pela capacidade de detetar, para além de sinais WI-FI, também sinais Bluetooth e GSM enquanto o Euclid, se limita apenas a sinais WI-FI.

2.1.5.4 Path Intelligence

O Path Intelligence é uma ferramenta considerada concorrente direta do BIPS, estando efetivamente no mesmopatamar.

Esta tecnologia permite detetar anonimamente e localizar todos os dispositivos móveis existentes num determinado espaço e possibilita a análise do movimento destes dispositivos nesse mesmo espaço [webe]. Graças a esta ferramenta é possível saber o percurso dos visitantes, o tempo que eles ficam no espaço ou numa determinada zona desse espaço, o percurso que fazem, o número de visitas e entre muitas outras métricas.

O resultado final é um relatório com informações poderosas para os gestores da área de negócio que de outra forma não conseguiriam obter. Tal como o BIPS este sistema deteta sinais de Bluetooth, WI-FI e GSM.

Ao nível da tecnologia, não parece existir grandes diferenças entre os dois sistemas daí o Path Intelligence ser considerado um concorrente direto do BIPS. A diferença está no produto, ou seja, nos relatórios com as análises efetuadas por cada uma das empresas.

2.1.5.5 Nomi

O Nomi é uma plataforma construída com o intuito de melhorar os vários aspetos que possam contribuir para experiencia do cliente final numa loja [webd].

Esta ferramenta permite medir o comportamento e o perfil dos clientes no interior da loja. Permite também identificar quais os investimentos e movimentações de marketing necessárias e como melhorar o processo de conquista de novos clientes e retenção dos atuais clientes com recurso a uma aplicação web que permite em tempo real ter acesso a informações úteis para o negocio com base no comportamento dos clientes no interior da loja [webd].

Esta aplicação, tal como o Euclid, apenas capta sinais WI-FI. É uma aplicação que está desenhada para lojas para tendo em conta as métricas que consegue captar e as informações que apresenta ao utilizador final, pode também ser adaptada para centros comerciais sendo por isso considerada uma aplicação concorrente do BIPS.

2.1.5.6 Conclusão

Após este estudo pode-se fazer um resumo que evidência as principais diferenças entre as tecnologias e apresenta-se estas mesmas na tabela 2.1. É possível definir dois principais grupos de concorrentes: aqueles que possuem tecnologias diferentes (PeCo e Footfall) e aqueles com tecnologias semelhantes (Euclid, Path Intelligence e Nomi). Deste grupo apenas o Path Intelligence consegue captar os três tipos de sinais que capta o BIPS (WI-FI, Bluetooth e GSM) sendo portanto um concorrente direto. Outra conclusão é o fato do Nomi e do Footfall estarem mais adaptados para funcionarem em lojas apesar do Nomi ser adaptável também a centros comerciais.

Assim conclui-se também que o BIPS destaca-se de alguns concorrentes visto que abrange um maior número de tecnologias de deteção e permite compreender comportamentos de visitantes no interior do centro comercial e conta com elevado grau de fiabilidade o número de visitantes. Apresenta-se na tabela seguinte um estudo em que se compara o BIPS com as aplicações concorrentes enumeradas:

Sistema	Descrição	Vantagens	Desvantagens	Tecnologias
BIPS	- Permite seguir o sinal de um dispositivo móvel ao longo do tempo.	- Permite saber o caminho percorrido por um visitante e o tempo demorado. -Permite cobrir a totalidade de uma área.	- Não conta visitantes sem dispositivos móveis	- Bluetooth - WI-FI - GSM
PeCo	- Análise de Vídeo.	- Permite estimar o género dos visitantes e a idade.	- Leva a erros na contagem das pessoas. - Difícil saber o caminho, percorrido mesmo em circuito fechado	- Vídeo
Footfall	- Contagem de pessoas através dos infravermelhos situados nos acessos.	- Permite o estudo relativo a filas de espera.	- Conta pessoas,repetidas. - Não recolhe informação relativamente ao que se passa no interior.	-Infravermelhos
Euclid	- Produto semelhante ao BIPS sendo considerado concorrente direto.	- Permite saber o caminho percorrido por um visitante e o tempo demorado.	-Cobertura limitada ao WI-FI	- WI-FI
Path Intelligence	- Produto semelhante ao BIPS sendo considerado concorrente direto.	- Permite saber o caminho percorrido por um visitante e o tempo demorado.	- Não conta visitantes sem dispositivos móveis	- Bluetooth - WI-FI - GSM
Nomi	- Produto semelhante ao BIPS sendo considerado concorrente direto.	- Permite saber o caminho percorrido por um visitante e o tempo demorado.	-Cobertura limitada ao WI-FI	-WI-FI

Tabela 2.1: Comparação entre BIPS e seus concorrentes

2.2 Data Mining

O crescimento exponencial das bases de dados das empresas levou à necessidade urgente de novas técnicas para tornar toda a informação existente nos dados em conhecimento útil para as empresas [LL98].

Para responder a esta necessidade, muitos investigadores com especialização em estatística, bases de dados, aprendizagem automática, redes neuronais, reconhecimento de padrões e muitas outras, trabalharam neste tipo de problema [Joh97] Todos estes esforços levaram ao aparecimento de uma nova área de pesquisa: *Data Mining* ou descoberta de conhecimento [LL98].

Data mining é precisamente a exploração e análise de grandes quantidades de dados de forma automática ou semiautomática com o intuito de descobrir informações relevantes e potencialmente úteis como situações anómalas, padrões, regras de conhecimento para resolver um determinado problema de uma empresa num espaço de tempo mínimo [LL98, Sch02] .

Data mining classifica-se tipicamente em duas categorias: descritiva e preditiva. A análise descritiva concentra-se em encontrar padrões de conhecimento sem uma ideia pré-determinada sobre o que pode ser esse padrão, enquanto a análise preditiva foca-se no uso de campos da base de dados para prever valores ou situações futuras de outras variáveis de interesse [She07].

Contudo apenas encontrar informação não é suficiente. O grande objetivo das empresas é fazer uso desta informação encontrada e torna-la em valor e apoio para possíveis tomadas de decisão lucrativas [LL98].

2.2.1. Deteção de *Outliers*

Uma vez que se pretende detetar valores atípicos, ou *outliers*, no número de visitas de superificies comerciais, torna-se necessário compreender o que significa este termo e que tipo de técnicas é necessário analisar.

Outliers, ou também referidos como deteção de anomalias, exceções, observações discordantes, refere-se ao problema de detetar padrões em dados que não demonstram o comportamento esperado [CBK09]. Informalmente, um *outlier* é qualquer valor de dados que parece estar fora do lugar em relação ao restante dos dados [Kno02]. *Outlier* pode também ser definido como uma observação (ou subconjunto de observações), que parece ser inconsistente com os restantes elementos do conjunto de dados [Ord96]. A definição evidente de um *outlier* seria uma observação que se desvia tanto a partir de outras observações e consequentemente concluindo que os valores dessa observação foram gerados de uma forma diferente e que não seguem o mesmo modelo estatístico dos restantes dados, sendo então considerados valores discrepantes ou *outliers* [Haw80]. Um *outlier* pode ser um ou vários dos seguintes pontos:

- Um valor extremo ou relativamente extremo
- Um contaminante, isto é, uma observação de outra distribuição (possivelmente desconhecida)
- Um valor de dados legítimo, mas surpreendente / inesperado
- Um valor de dados que foi medido ou registado incorretamente [Kno02]

Consideremos o atributo único idade. Um homem adulto inserido num grupo de crianças é considerado um *outlier*. No entanto, *outliers* não precisam ser valores extremos. Por exemplo, se uma pessoa de estatura baixa entrasse numa sala cheia de jogadores de basquetebol, então ela seria um *outlier*, pois a sua altura seria claramente diferente das alturas dos restantes indivíduos [Kno02] uma vez que os jogadores desta modalidade são, por norma, de estatura alta.

Um *outlier* é uma observação "suspeita"que merece um estudo mais cuidadoso para determinar [Kno02]:

- Se é realmente ou não uma observação válida
- E o motivo dessa observação ser tão diferente do restante conjunto de dados.

Um *outlier* pode estar a fornecer informações que o restante conjunto de dados não consegue fornecer devido ao facto de surgir a partir de uma combinação de circunstâncias inesperadas. O interesse de um valor anómalo pode ser vital para a compreensão de um profissional relativamente ao que se passa no seu negócio. Esta conclusão surgirá após uma análise desse *outlier* e das causas que o provocaram [Ric06].

Contudo um *outlier* não é sempre necessariamente sinónimo de um resultado prático útil ou que permita a aprendizagem do ser humano. Um *outlier* pode ser também uma observação inválida derivada de vários fatores, como por exemplo, uma introdução incorreta de um valor no sistema por parte do utilizador ou uma medição errada. Neste caso o *outlier* pode ser eliminado (implica a remoção dos dados desse *outlier* do conjunto de dados), corrigido (implica refinar o modelo subjacente) ou até mesmo ignorado ficando esta decisão sempre ao critério de um profissional da área de negócio [Kno02].

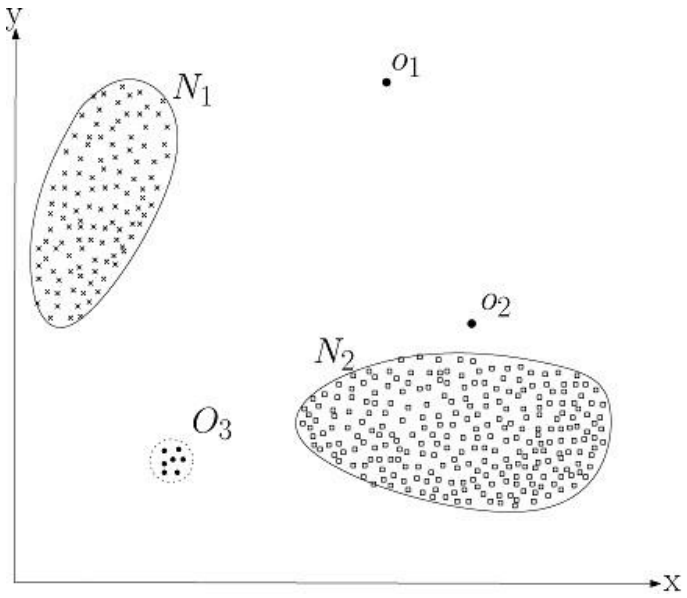


Figura 2.5: Exemplo de *outliers* num conjunto bidimensional

Na figura 2.5, pode-se visualizar um exemplo de *outliers* num conjunto de dados [CBK09].

Os dados na figura 2.5, têm duas regiões de valores esperados, N1 e N2, visto que a maior parte dos dados observados encontram-se nessas duas regiões. Os pontos que estão mais afastados dessas duas regiões, os pontos O1, O2 e O3, são considerados *outliers* ou anomalias [CBK09]. As regiões N1, N2 e O3 podem ser considerados *clusters*, aglomerados ou grupos. Um *outlier* pode pertencer simultaneamente a vários *clusters*, ou grupos, dentro de um conjunto de dados [Kno02]. Estes são considerados *outliers* multivariados. *Outliers* multivariados são classificados como: *outliers* brutos ou estruturais [Whi92]

Outliers brutos são aquelas observações que são irregulares para um ou mais atributos individuais. *Outliers* estruturais são irregulares em relação à estrutura de covariância dos dados não irregulares [Kno02].

Imaginando, como exemplo, uma ordenação linear de N valores de dados, então o valor mais pequeno e o valor maior consideram-se valores extremos. Os restantes valores, menores ou maiores, e maiores que o valor mais pequeno do conjunto e menor que o maior valor do conjunto, são exemplos de valores relativamente extremos [Kno02]. Se um valor extremo ou relativamente extremo é ou não um *outlier*, depende exclusivamente se o ponto é ou não um valor inesperado para o modelo de distribuição dos dados [Kno02]. O termo "extremo"precisa ser definido em termos de

contexto da área de negócio, já que muitas noções podem ser associadas [Kno02]. Os dados não-anómalos são considerados dados básicos. Dados básicos representam a grande maioria dos dados num determinado conjunto de dados [Kno02].

O método tradicional de transformar dados em conhecimento dependendo da análise e interpretação manual [FPsS96] pode ser melhorado e tornado mais eficiente aliando o conceito de *outliers* a uma técnica de *data mining*. Contudo a associação de *outliers* a técnicas de *data mining* leva a alguns problemas tais como:

- A identificação eficiente de *outliers*
- A quantidade de informações adicionais que um algoritmo de deteção de *outliers* pode proporcionar [Kno02].

As técnicas de deteção de *outliers* existentes atualmente permitem descrever o espaço de dados em que ocorrerem e fornecem informação sobre a relação entre os vários de *outliers*. Um tipo de algoritmos bastante utilizado na deteção de *outliers*, são os algoritmos baseados na distância em que os objetos que estão a uma distância considerável de qualquer outro *cluster*, ou seja, que não têm muitos vizinhos, são considerados *outliers* [Rog10]. Algoritmos baseados na distância resumem-se a uma função de distância em que se utiliza como métrica para identificar *outliers* em grandes conjuntos de dados, uma distribuição de probabilidade desconhecida. Portanto, *outliers* são identificados com base na densidade da vizinhança mais próxima [Pei06].

Os algoritmos de deteção de *outliers* com base na distância, mostram que a deteção destes pode ser feita de forma eficaz para grandes conjuntos de dados em que se obtém uma complexidade espacial de $O\ (K\ N^2)$ no pior dos casos em que K é a dimensão dos conjuntos de dados e N representa o número de objetos [Rog10]. Estes algoritmos funcionam com dois parâmetros: p e D em que p é a percentagem mínima de objetos que estão fora da vizinhança e D é a distância absoluta do seu afastamento do conjunto de dados [Rog10, Pei06] Estes dois parâmetros são definidos por tentativa e erro por parte do utilizador através de vários ensaios e tentativas visto que não existe uma forma universal para os determinar durante a pesquisa de *outliers*. Contudo o valor inicial mais utilizado de p é 0.995 [Rog10]. Existe uma outra variável – M – que representa o número máximo de objetos dentro de uma vizinhança de um *outlier* e é fundamental para definir um objeto como um *outlier* ou não *outlier*. Sempre que haja pelo menos $(M\ +1)$ vizinhos, a pesquisa é interrompida e o objeto é declarado um não-*outlier* ou valor comum. Se houver menos do que $(M\ +1)$ vizinhos, o objeto é declarado um *outlier* [Rog10]. A variável M é calculada pela forma $M=N(1-p)$ [Rog10, Pei06]. Este raciocínio mostra-se coerente com a ideia básica dos algoritmos de deteção de *outliers* baseados na distância que um conjunto de dados anómalos deve estar suficientemente longe dos restantes pontos de dados. Uma vantagem da definição *outlier* com base na distância é a não necessidade de assumir uma forma paramétrica dos dados. Contudo, é necessário encontrar os valores apropriados de D e p [KN98].

Existem outras técnicas de deteção de *outliers*. A deteção de *outliers* univariados que consiste na deteção de *outliers* para cada variável de um conjunto de dados [LJK00]. Para este fim existem

várias técnicas tais como os diagramas de caixas de bigodes em que é possível detetar *outliers* intermédios e *outliers* severos. A sua popularidade deriva da sua simplicidade e é a mais popular para deteção de *outliers* univariados [DC11]. Este tipo de gráficos vai ser abordado com mais pormenor neste capítulo.

2.2.2 Metologia de Data Mining

A implementação de *data mining* requer uma utilização adequada da combinação de ferramentas e recursos humanos. Foi então criada uma metodologia que integra de forma adequada e bem gerida técnicas e recursos humanos num processo para implementação de *data mining* [Rog10].

2.2.2.1 CRISP-DM

CRoss Industry Standard Process for Data Mining (CRISP-DM) [CCK+00] é uma metodologia geral e um modelo de processo que abrange todas as fases de *data mining* que podem ser aplicadas em vários tipos de atividades [She07]. Este modelo de processo é considerado um padrão na área de *data mining* e está associado à NCR Corporation e SPSS Inc [Rog10] e é ilustrado na figura 2.6, .

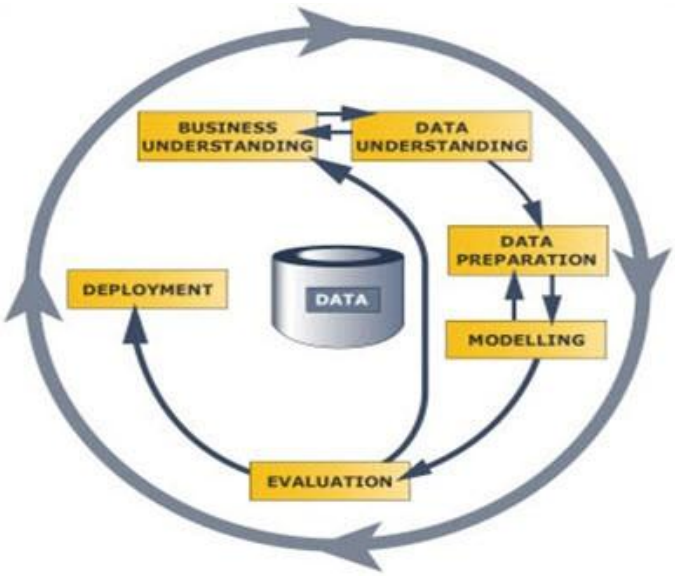


Figura 2.6: Ciclo CRISP-DM

1. Compreensão do negócio

(a) Compreender os objetivos do projeto e requisitos; Definição dos problemas de data mining

2. Compreensão dos dados

(a) Coleta inicial dos dados familiarização; Identificar problemas de qualidade de dados; Resultados óbvios iniciais

3. Preparação dos dados

(a) Registo e seleção dos dados; Limpeza dos dados; O resultado desta fase é um conjunto de dados limpos que podem ser utilizados no processo de *data mining*

4. Modelação

(a) Seleção e aplicação de várias técnicas de modelação, com parâmetros apropriados; Seleção de técnicas e ferramentas de *data mining*

5. Avaliação

(a) Determinar se os resultados correspondem com os objetivos do negócio; Identificar questões de negócio que deveriam ter sido abordadas anteriormente

6. Desenvolvimento

(a) Colocar os modelos resultantes em prática

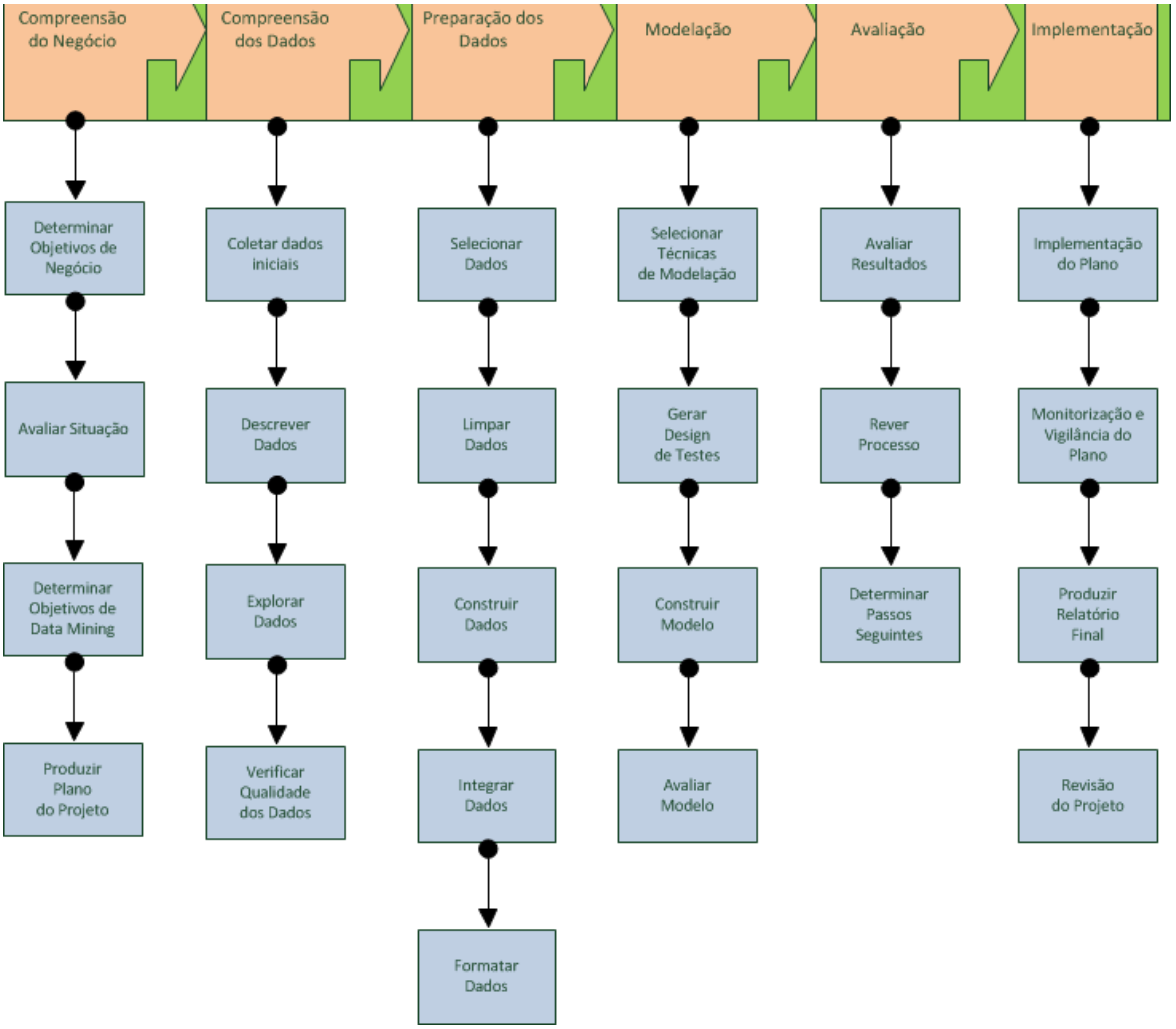


Figura 2.7: Representação das fases e tarefas do modelo CRISP-DM

Esta metodologia tem a particularidade de ser cíclica, o que permite que as fases de desenvolvimento do projeto sejam também cíclicas, ou seja, se após a avaliação dos resultados obtidos se concluir que a modelação feita não foi adequada ou que os dados utilizados não foram bem preparados, pode-se recomeçar o processo.

A cada fase estão associadas várias tarefas como ilustrado na figura 2.7 que funcionam como guia para o cumprimento eficiente de cada uma das fases.

É importante referir que é aplicada uma maior percentagem de esforço à tarefa de formular corretamente o problema de *data mining* e é gasta uma maior percentagem de tempo (60% a 70%) na tarefa de preparação dos dados [CHS+98].

Conclui-se que o modelo CRISP-DM fornece uma importante orientação para quem estiver a desenvolver um projeto de *data mining* pois descreve claramente as fases e tarefas fundamentais para o desenvolvimento do projeto com sucesso sendo que o resultado de cada fase será utilizado numa fase seguinte deste modelo. Definida a metodologia a utilizar, torna-se agora importante enquadrar as várias fases do CRISP-DM no desenvolvimento de projeto. Esta fase é descrita ao longo dos capítulos 4 e 5

2.2.3 Métodos Estatísticos

Data mining é uma técnica que abrange várias áreas importantes sendo uma delas a estatística. Esta razão leva ao estudo de alguns métodos estatísticos que serão utilizados ao longo desta dissertação.

2.2.3.1 Diagramas de Caixas de Bigodes

Estes diagramas são muito utilizados no ramo da estatística e são particularmente úteis para representar graficamente e analisar uma variável separada do conjunto de dados [GI91, SJDG11]. Este tipo de diagramas são ferramentas versáteis de exploração e análise de dados e são utilizados para resumir dados univariados [GI91]. Os diagramas de caixas de bigodes recorrem a métricas estatísticas como mediana, primeiro e segundo quartis e ainda às observações máximas e mínimas de valores não *outliers* [GI91]. Um diagrama de caixas de bigodes é construído da seguinte forma:

A caixa é um retângulo que se estende desde o quartil 1 até ao quartil 3 e que representa 50% dos dados. Este retângulo é denominado também por intervalo interquartil (IQR). Este é um intervalo robusto para interpretação porque a região central de 50% não é afetada por *outliers* [DC11]. No interior da caixa existe uma linha horizontal que representa a mediana dos dados. Existem duas linhas verticais (superior e inferior à caixa) que representam os valores dos dados extremos que não são *outliers* (bigodes). Estes intervalos são calculados pelas seguintes fórmulas [DC11, GI91]:

- Q3 + K (Q3-Q1) - Limite superior
- Q1 - K (Q3-Q1) - Limite inferior

A variável *K* permite a definição do tamanho das linhas verticais que se estendem desde a caixa. Usualmente a esta constante atribuem-se valores entre 1.5 e 3.0 [DC11]. O valor 1.5 tende a considerar várias observações como *outliers* e o valor 3.0 pode falhar a considerar muitas observações como *outliers* [DC11]. Todos os valores do conjunto de dados que são superiores e inferiores aos resultados destas fórmulas, são considerados *outliers*.

Uma representação visual de um diagrama de caixas de bigodes, e respetiva explicação, está presente na figura 2.8.

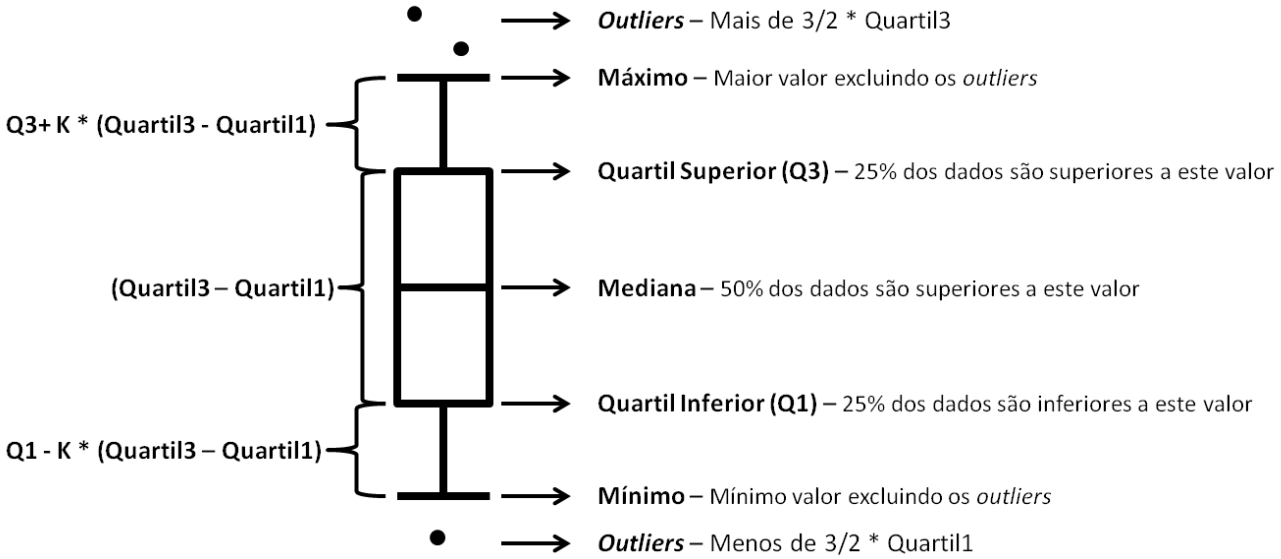


Figura 2.8: Descrição de um diagrama de caixas de bigodes

Este método gráfico para identificação de *outliers* é extremamente apelativo não só pela sua simplicidade mas também porque os diagramas de caixas de bigodes não usam poucas observações extremas na construção do intervalo interquartis o que faz com que estes gráficos não sofram do efeito de "*masking*". *Masking* é fenómeno pelo qual a presença de alguns *outliers* faz com que outros sejam difíceis de detetar [DC11]. Este tipo de gráficos serão fundamentais para o desenvolvimento do presente projeto.

2.2.3.2 Análise de Variância

O teste ANOVA - *Analysis of Variance* - assenta num grupo de teste de hipóteses e é um processo de analisar as diferenças nas médias entre vários grupos [CA13]. O principal objetivo deste teste paramétrico é comparar médias de mais de duas populações e então testar se uma dada variável qualitativa chamada *factor* tem algum efeito numa variável quantitativa. É chamado teste paramétrico visto que a sua hipótese é sobre os parâmetros da população, nomeadamente a média e o desvio padrão. A análise de variância clássica é chamada *F-test* [GS13].

A análise de variância tem como objetivo testar H_0: $\mu_1 = \mu_2 = ... = \mu_K$ face a H_1: $\mu_i \neq \mu_j$ para qualquer i ≠ j, em que μ_i é a média da população [GS13, CA13].

Caso se aceite H_0, conclui-se que as médias populacionais são iguais. Caso se rejeite H_0 então pode-se concluir que as médias populacionais são diferentes, ou seja, pelo menos uma das médias é diferente das restantes.

Na análise de variância, a total variação (soma total dos desvios quadrados da média) das amostras são particionadas em duas partes: soma dos quadrados dentro de grupos (SSW) e soma dos quadrados dentro de grupos (SSB). SSW é também denominado por erro ou soma residual dos quadrados. O teste estatístico *F-test* é construído através da razão das duas somas dos quadrados (SSW e SSB).

Complementando o teste ANOVA é usual utilizar o teste *Honest Significant Difference* que é um método proposto para a comparação de médias de populações distribuídas normalmente com o objetivo de determinar quais os grupos da amostra que apresentaram diferenças.

Enquanto o teste ANOVA permite provar se existem diferenças nas médias, o que acontece quando se rejeita H_0, o teste de Tukey permite determinar quais essas diferenças. Este teste é feito comparando-se a diferença absoluta (em módulo) entre as várias médias pareadas duas a duas, a um valor δ , previamente calculado. Este valor calcula-se através da fórmula $\delta = q\sqrt{\frac{QME}{n}}$ em que q representa a amplitude total, QME representa o quadrado médio do resíduo e n representa o número de observações. Serão consideradas significativas ao nível de significância pré determinado de α aquelas diferenças entre médias cujo valor absoluto for maior que o δ calculado.

2.3 Conclusão

Nesta fase conclui-se que as técnicas *data mining* estudadas comprovam que é possível o desenvolvimento deste projeto. Usando técnicas de deteção de *outlier* será possível encontrar nos dados analisados situações anómalas. Estas situações poderão ser encontradas através da comparação semanal, mensal ou anual com as médias das visitas obtidas nos períodos homólogos.

É possível concluir que a metodologia escolhida será também de extrema utilidade neste projeto. Para além de ter as várias fases devidamente definidas, considera-se que é um apoio fundamental para um projeto desta área e o facto de ser uma metodologia cíclica, permite que caso os resultados obtidos não sejam relevantes, se possa voltar à fase inicial e recomeçar o processo.

Conclui-se também que este tipo de técnicas é perfeitamente adaptável e bastante importante para resolver o problema especificado. Tendo uma noção mais concreta do número de visitas do espaço comercial é possível uma gestão mais eficiente por parte dos gestores do espaço. Os métodos estatísticos apresentados serão de grande utilidade para deteção de *outlier* e também para perceber o impacto que variáveis do tipo *factor* terão em variáveis quantitativas.

Este projeto está dependente da grande quantidade de dados armazenada relativamente ao que se passa no interior do centro comercial. Após um estudo das métricas captadas pelo sistema BIPS, concluiu-se que estes dados serão extremamente úteis para o desenvolvimento do projeto. Métricas, como tempo médio de visita, serão calculadas através de algoritmos desenvolvidos que são executados com tempo computacional baixo.

Capítulo 3

Tecnologias Utilizadas

Neste capítulo irão ser descritas as tecnologias a utilizar para o desenvolvimento do projeto. No fim é feita uma conclusão na qual se apresenta a justificação para a escolha destas tecnologias.

3.1 JSON

Foi necessário recorrer a uma linguagem que facilitasse a troca de informaçao entre aplicações. É neste contexto que surge o JSON.

O JSON (Notação de Objetos JavaScript, do inglês JavaScript Object Notation) [webc] é um formato livre e leve para trocas de dados utilizado normalmente na troca de mensagens entre uma aplicação do tipo cliente-servidor. É um formato que é simples para humanos lerem e escrever e também é fácil para máquinas processarem e gerarem.

É baseado num subconjunto da linguagem de programação JavaScript, padrão ECMA-262 3ª Edição – Dezembro de 1999 [CB14]. JSON é um formato de texto que é utilizado por programadores de várias linguagens como por exemplo: JAVA, C, C++, Python e muitas outras.

Estas propriedades fazem do JSON uma linguagem de troca de dados simples e ideal. Um ficheiro do formato JSON pode ser construído segundo duas estruturas universais:

1. Uma coleção de pares chave/valor. Esta coleção pode ser conhecida noutras linguagens como objeto, registo, estrutura, dicionário, tabela ou *hash*
2. Uma lista ordenada de valores. Na maioria das linguagens esta lista é percebida como um vetor, lista, sequência ou *array*.

Em JSON, uma coleção começa sempre com ""(chaveta esquerda) e termina sempre com (chaveta direita). Cada nome é seguido por ":"(dois pontos) e os pares chave/valor são separados por ","(vírgula). Uma lista ordenada começa com "["(parêntesis reto esquerdo) e termina com "]"(parêntesis reto direito). Os valores são separados por ","(vírgula). Um valor pode ser uma *string* entre aspas, ou um número, ou um booleano, ou nulo, ou uma coleção ou uma lista ordenada.

3.2 R

Foi necessário recorrer a uma linguagem de programação utilizada na área da estatística e capaz de processar grandes quantidades de dados com baixo tempo computacional. A linguagem escolhida foi o R.

O R [webg] é um ambiente de desenvolvimento de software gratuito para cálculos estatísticos e para gráficos. É uma linguagem de programação bastante utilizada na área da estatística para análises da área e também para análises de dados. Foi desenvolvido inicialmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da universidade de Auckland, Nova Zelândia, e foi sendo desenvolvido através da colaboração de vários programadores do mundo.

O R é uma parte oficial do projeto GNU da Fundação Livre de Software (do inglês, Free Software Foundation's) e a Fundação R tem objetivos semelhantes a outras fundações de código aberto como Fundação Apache ou Fundação GNOME.

Alguns objetivos da Fundação R são o suporte de desenvolvimento contínuo do R, a exploração de novas metodologias e o ensino e treino da computação estatísticas. O R é disponibilizado atualmente gratuitamente sendo que o seu código fonte está também disponível pela licença GNU GPL (GNU General Public License).

O desenvolvimento contínuo do R levou à criação de uma comunidade ativa que trabalha na evolução e melhoria constante desta linguagem de programação sendo esta comunidade uma das razões que fez tornar o R tão popular e tão bem sucedido.

Esta comunidade foi responsável pela criação de novos pacotes (ou bibliotecas) que permitem alargar as funcionalidades desta linguagem de programação tornando-a assim mais flexível e pronta para vários e distintos objetivos de programação sendo que existe um número alargado de pacotes disponíveis gratuitamente.

Os pacotes disponíveis para a linguagem R são de grande utilidade para este projeto em concreto. Sendo necessário trabalhar com *json*, instalou-se o package "*rjson*"que permite a conversão de objetos R em objetos JSON e vice-versa [CB14].

Foi escolhido para o desenvolvimento deste projeto um IDE chamado RStudio. Este IDE fornece um conjunto de ferramentas integradas que ajudam a aumentar a produtividade com a linguagem R. Fornece ainda uma interface organizada e uma consola que suporta a execução direta de código, o acesso facilitado a documentação bem como ferramentas para gráficos e depuração [webh]. O RStudio é atualmente o principal ambiente de desenvolvimento integrado para a linguagem R.

3.3 Conclusões

A tecnologia JSON foi imposta pela empresa Movvo havendo apenas liberdade na opção de escolha pela linguagem R.

A escolha teria de residir por uma linguagem que fosse perfeitamente compatível com JSON e que fosse indicada para programação estatística. Estas foram duas das razões que levaram à escolhida da linguagem R pois com recurso a packages específicos existentes, desenvolvidos pela comunidade que atualmente tem a preocupação de melhorar constantemente o R, considerou-se uma ferramenta adequada para o trabalho. Outra razão fundamental para a escolha desta linguagem estatística é o facto de ser atualmente a linguagem de programação gratuita mais utilizada em análise estatística e em análise de dados [top].

Capítulo 4

Deteção de Números de Visitas Atípicos num Centro Comercial

Neste capítulo irá ser explorada a descoberta de valores de visitas a um determinado centro comercial que apresentam um grande afastamento das restantes. Irão ser criados os critérios para a definição de períodos com visitas fora do "normal" e apurada a existência de eventos nesses dias. Esta análise deverá ser incluída como parte integrante do Retail Movves. Seguiu-se a metodologia de CRISP-DM para a exploração do problema, e descrevem-se as várias etapas que foram referidas na secção dedicada a esta metodologia.

4.1 Compreensão de Negócio

Como já foi referido, cada vez mais os retalhistas estão cientes que o conhecimento do comportamento do consumidor dentro do seus espaços comerciais é importante para aumentar a vantagem sobre os seus concorrentes. Por este motivo, tem-se assistido a um aumento no investimento em tecnologias que permitam o acréscimo quantitativo e qualitativo de informação sobre o comportamento dos consumidores. É o caso do BIPS, a fonte origem dos dados que estão a ser analisados nesta dissertação, que deteta os movimentos dos consumidores em espaços comerciais, através do seguimento dos sinais de radiofrequência dos dispositivos móveis transportados pelos visitantes.

O BIPS encontra-se instalado em vários centros comerciais, mas vamos focar a nossa investigação num dos maiores centros comerciais portugueses, localizado na região norte de Portugal que contém atualmente 270 lojas divididas por várias zonas, como por exemplo, cinema e restauração.

Este centro comercial atrai diariamente uma elevada quantidade de visitantes. Dependendo do dia da semana, do clima e dos eventos que decorrem na região de influência do centro comercial, o número de visitas pode variar. O gestor do centro comercial já possui um conhecimento empírico sobre o número de visitas ao centro comercial através do Retail Movves, no entanto tem de fazer uma avaliação baseada na observação dos valores atuais e dos valores de histórico mostrados na interface para concluir se um determinado número de visitantes está dentro do "padrão normal". O Retail Movves, neste momento, apenas apresenta os valores para o utilizador final, não produzindo qualquer juízo matemático sobre a variação dos valores face a todo o histórico.

Com o crescente volume de dados, a que um gestor tem acesso sobre o seu negócio, as primeiras críticas dizem respeito ao fato de ser demasiada informação, geralmente dados numéricos, que não se traduzem em informação útil. Para resolver esta questão, as empresas e as organizações investem na contratação de recursos humanos para analisarem os dados.

Para colmatar esta limitação, também podem ser utilizadas técnicas de *data mining* para fornecer mais detalhes sobre os dados e reduzir o conjunto de análise ao evidenciar resultados com importância de acordo com heurísticas pré-definidas.

Ao mesmo tempo, existe um conjunto de eventos que influenciam o comportamento do consumidor dentro do centro comercial. Estes eventos podem ser internos, por exemplo, campanhas promocionais, abertura de novas lojas, eventos culturais; ou externos, tais como, concertos ou eventos desportivos. Um gestor do centro comercial está atento, não só ao que se passa dentro do centro comercial, mas também no exterior e, por isso, deve manter um histórico sobre todos os eventos que vão decorrendo.

Pretendemos com este projeto complementar o Retail Movves para que este seja capaz de identificar os valores atípicos, efetuando análises do histórico. Ao mesmo tempo, será acrescentada a funcionalidade de criação de eventos numa base de dados para que se possam cruzar as informações.

4.2 Compreensão dos Dados

O BIPS coleta atualmente um grande conjunto de informação relativamente ao que se passa no interior do centro comercial. Selecionou-se um conjunto de dados, para estudo que se descreve em seguida.

4.2.1 Análise do Número de Visitas por Dia

Foi extraído da API da Movvo um conjunto de dados relativos ao centro comercial que após algumas operações resultou no conjunto de dados apresentado na tabela 4.1 que serviu de base para o desenvolvimento do projeto e possui os atributos, e possíveis valores, apresentados na mesma tabela.

Variável	Descrição
Data	Data das visitas no formato dd-mm-aaaa.
Visitas	Número inteiro positivo representativo do número de visitas por dia.
Temperatura Média	Número inteiro positivo representativo da temperatura em Celsius.
Clima	Descrição do estado meteorológico do dia podendo assumir os seguintes valores: - *Clear* - *Clouds* - *Drizzle* - *Mist* - *Partly Cloudy* - *Rain* - *Scattered Cloud*

Tabela 4.1: Descrição do conjunto de dados relativo ao número de visitas por dia

Fez-se uma exploração deste conjunto de dados para compreender melhor as suas características e tentar obter algumas conclusões relativamente à relação do número de visitas com as restantes variáveis descritas para o período de 30-07-2013 a 30-06-2014. Durante o período de tempo referido, todas as variáveis estão corretamente instanciadas.

Métricas	Valores
Mínimo	3ºC
Máximo	25ºC
Mediana	15ºC
Média	14.5ºC

Tabela 4.2: Resumo estatístico da variável Temperatura Média

Estes valores apresentados na tabela 4.2 não permitem, numa primeira fase, extrair conhecimento relativamente ao impacto no número de visitas da variável *Temperatura Média*.

Clima	Ocorrências
Clear	161
Clouds	126
Drizzle	85
Mist	1
Partly Cloud	1
Rain	94
Scattered Clouds	1

Tabela 4.3: Resumo da variável Clima

Relativamente ao resumo da tabela 4.3, pode-se verificar de imediato que os estados do tempo *Mist*, *Partly Cloud* e *Scattered Clouds*, não são estados relevantes pois ocorreram apenas uma vez em todo o conjunto de dados. Os restantes estados são mais frequentes.

Mínimo	Máximo	Mediana	Média	Dia com menos visitas	Dia com mais visitas
20245	75460	34951	37014	25-12-2014	23-12-2014

Tabela 4.4: Resumo da variável Visitas

Segundo a tabela 4.4, o centro comercial em questão atrai em média 37014 visitas por dia.

Quanto ao número de visitas ao longo do tempo, o gráfio 4.1 permite uma boa análise. Este gráfico trata-se de uma série temporal e optou-se por construir uma serie temporal pois o número de visitas possuiu uma organização temporal entre os vários registos e torna-se relevante analisar o número de visitas recolhidas em função do tempo.

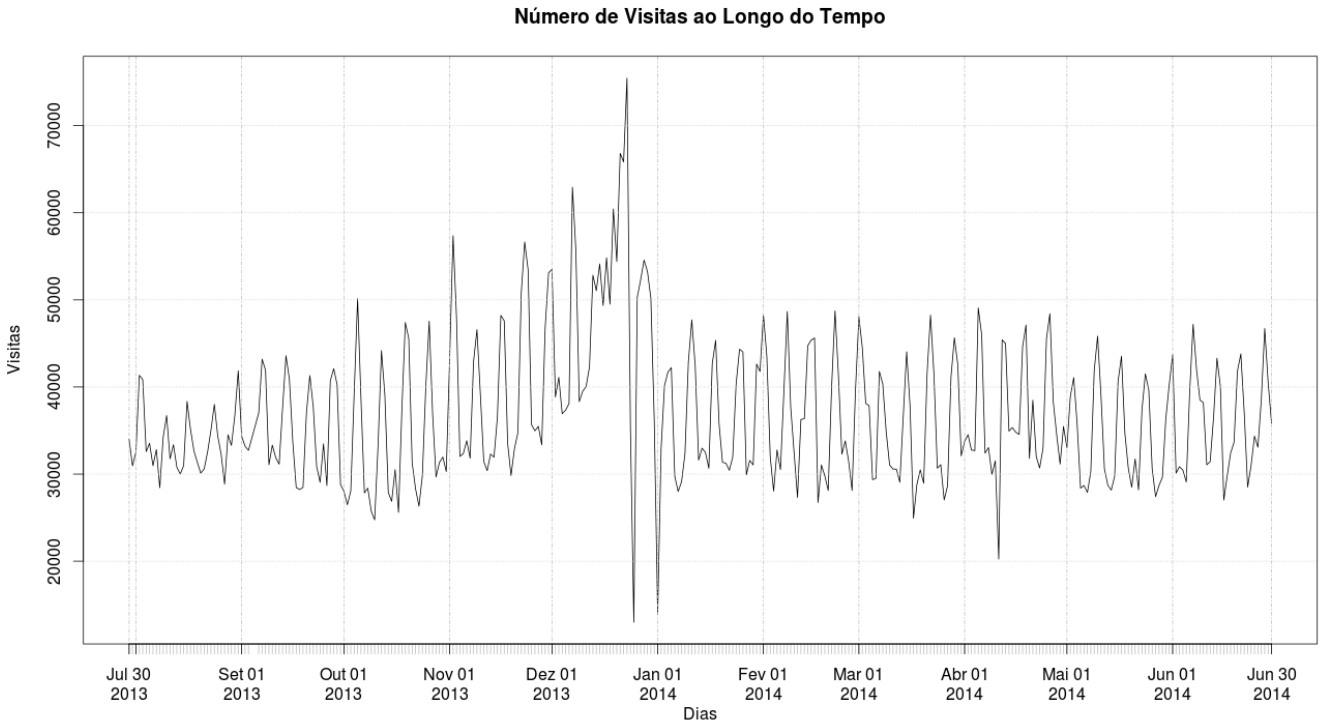


Figura 4.1: Serie Temporal - Número de visitas ao longo do tempo

Segundo a imagem 4.1, é notório um comportamento anormal e um elevado número de visitas no mês de dezembro de 2013. Este comportamento é justificado pela quadra Natalícia que atrai um elevado número de visitantes. O dia com maior número de visitas foi o dia 23 de dezembro de 2013. No dia 25 de dezembro é possível visualizar um baixo número de visitas que é justificado pelo encerramento da maioria das lojas do centro comercial, exceto os cinemas.

Apesar das conclusões úteis que se podem retirar deste estudo, considera-se que estas são ainda insuficientes. É necessário encontrar novas variáveis para retirar conhecimento com maior utilidade e também é necessário cruzar os valores das diferentes variáveis. Assim, para o presente conjunto de dados foram criadas as variáveis que se apresentam na tabela 4.5:

Novas Variáveis	Descrição
Dia da Semana	Valor categórico correspondente ao dia da semana baseada na data do dia
Mês	Valor categórico correspondente ao mês baseada na data do dia

Tabela 4.5: Descrição das novas variáveis

As variáveis descritas na tabela 4.5, são mais úteis do que a própria data pois tornam mais simples a manipulação e validação dos dados. Contudo irá ser feita também uma análise para cada uma destas variáveis para se perceber o real impacto destas e as suas relações com o número de visitas.

4.2.2 Relação Entre o Número de Visitas Diárias e o Dia da Semana

Para relacionar o número de visitas com o dia da semana, torna-se útil descobrir a média de visitas por cada dia. Assim, apresenta-se na tabela 4.6 e na figura 4.2 o número médio de visitas por dia da semana:

Dia da Semana	Média de Visitas
Domingo	41 611
Segunda-feira	33 289
Terça-feira	32 219
Quarta-feira	32 150
Quinta-feira	32 809
Sexta-feira	40 326
Sábado	46 499

Tabela 4.6: Média das visitas por dia da semana

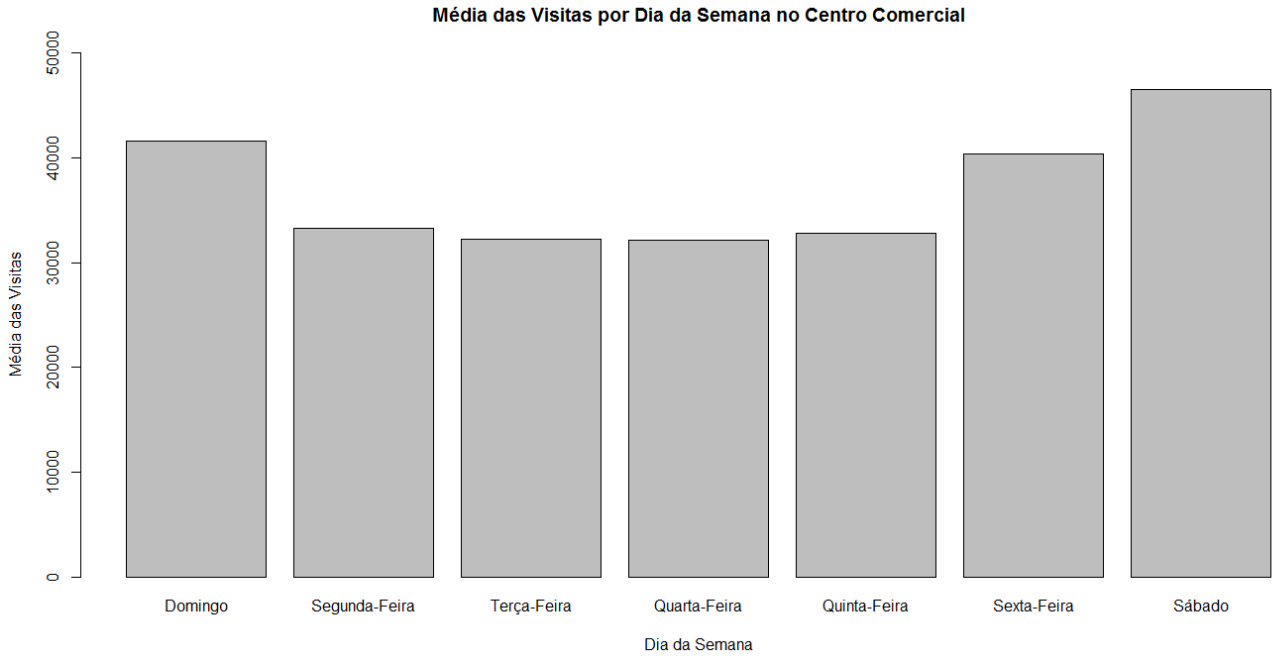


Figura 4.2: Gráfico de barras da média das visitas por dia da semana

De modo a explorar-se melhores estas variáveis, criou-se um diagrama de caixas de bigodes, figura 4.3, que relaciona o número de visitas com o dia da semana.

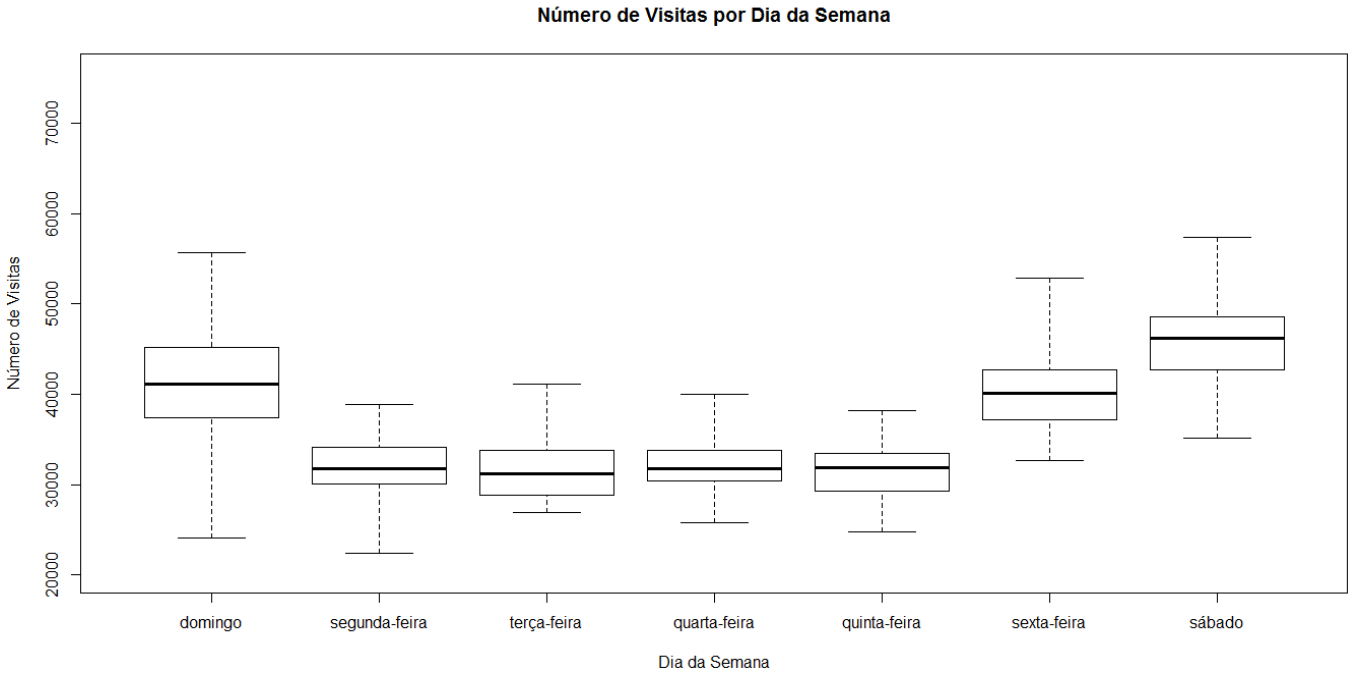


Figura 4.3: Diagrama de caixas de bigodes – Número de visitas por dia da semana

Através dos elementos anteriores, é possível observar que o número de visitas difere claramente consoante o dia da semana. Estes dois elementos permitem uma distinção do número de visitas em dois grupos em função do dia da semana devido à proximidade dos valores:

- Grupo 1 - segunda-feira, terça-feira, quarta-feira e quinta-feira
- Grupo 2 - sexta-feira, sábado e domingo

4.2.3 Relação Entre o Número de Visitas Diárias e o Mês do Ano

Realizou-se uma análise muito semelhante à anterior. Contudo torna-se essencial referir que não existe informação do número de visitas para todos os meses do ano. Assim, apresenta-se na tabela 4.7 o número total de visitas por mês e a média das visitas por dia de cada mês e na figura 4.4 o número de visitas por mês do ano:

Mês	Total de Visitas	Média das Visitas
Julho	64950	32475
Agosto	1046525	33758.87
Setembro	980778	35027.79
Outubro	1038667	33505.39
Novembro	1197290	39909.67
Dezembro	1491332	49711.07
Janeiro	1062379	36633.76
Fevereiro	1020157	36434.18
Março	1107449	35724.16
Abril	1097142	36571.4
Maio	1057404	34109.81
Junho	1086101	36203.37

Tabela 4.7: Total das visitas por mês e média por mês

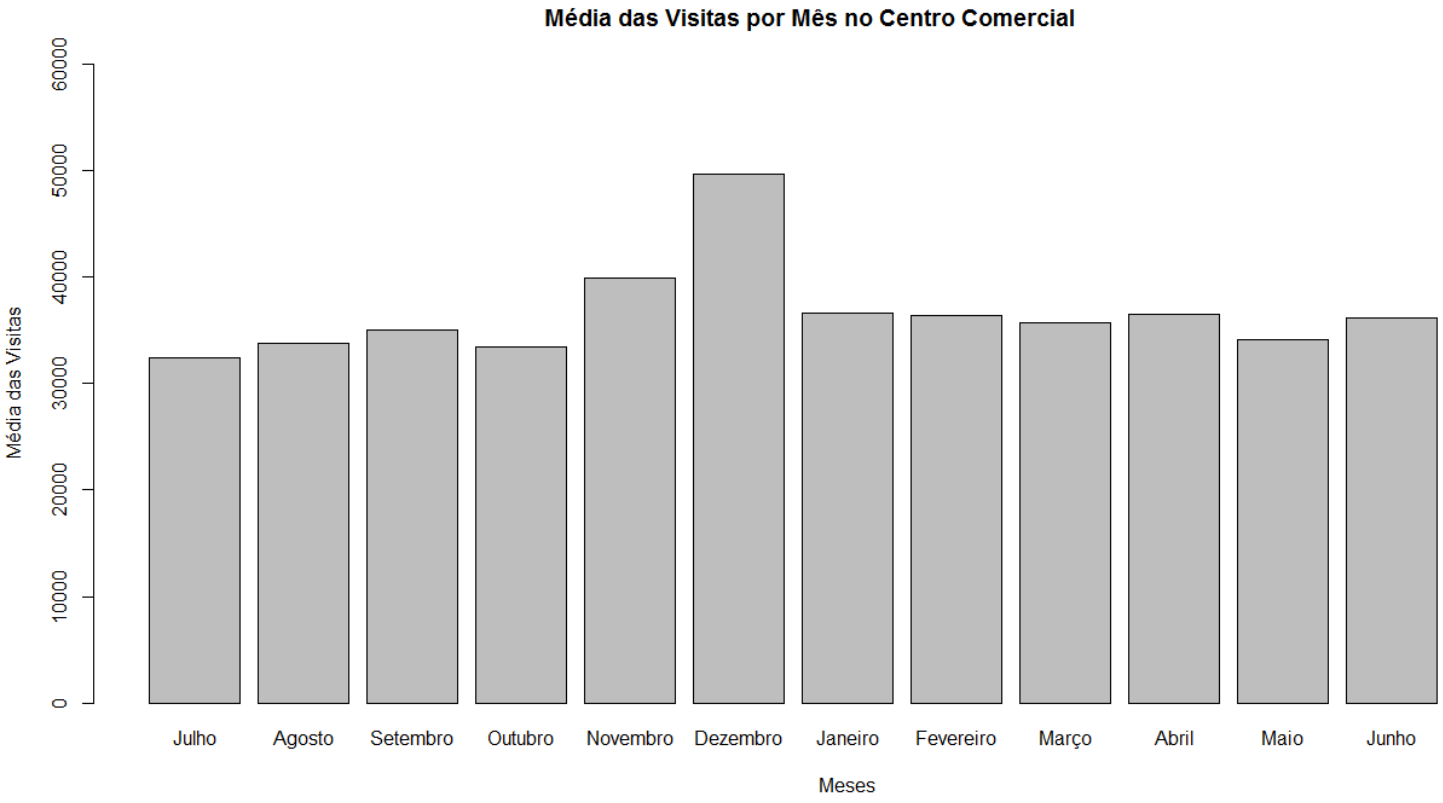


Figura 4.4: Gráfico de barras das média de visitas por mês

De modo a explorar-se melhores estas variáveis, criou-se um diagrama de caixas de bigodes, figura 4.5, que relaciona o número de visitas com o mês do ano.

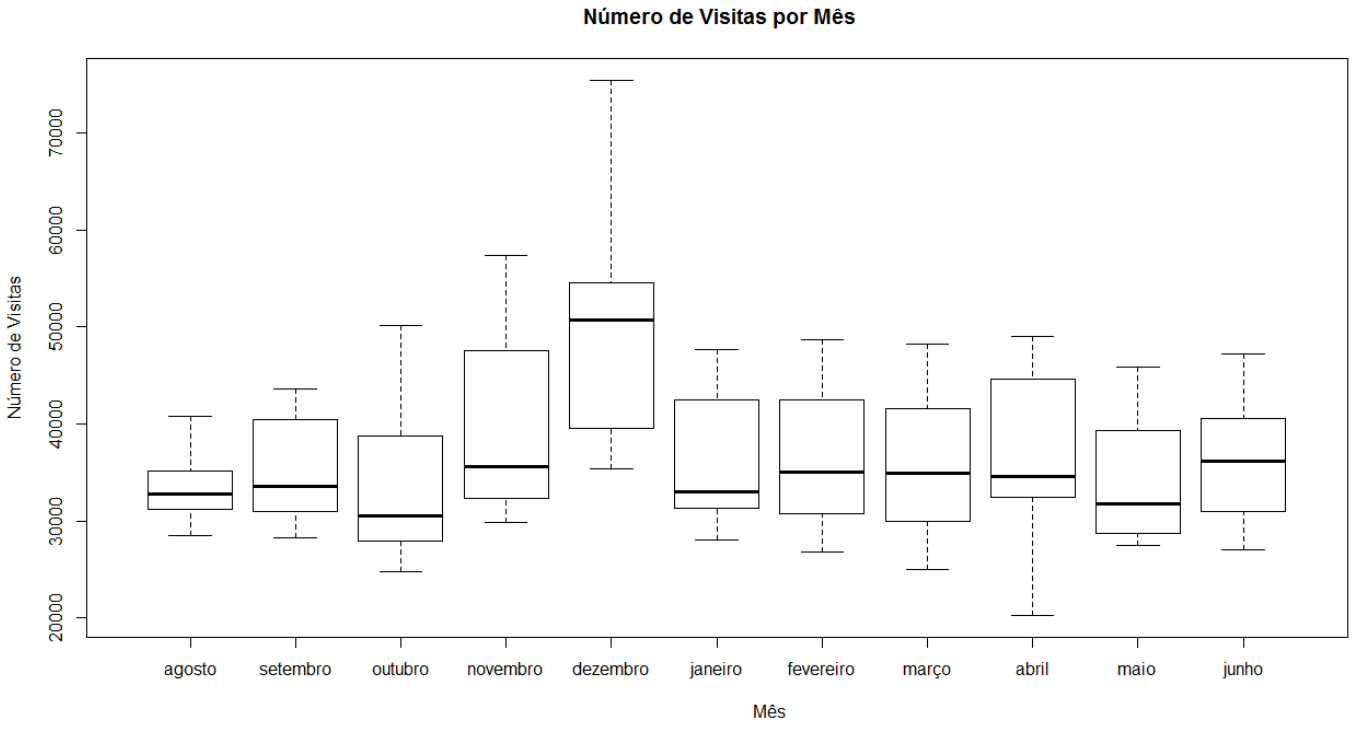


Figura 4.5: Diagrama de caixas de bigodes – Número de visitas por mês

É possível observar, por fim, que existe um número elevado de visitas durante o mês de dezembro o que é justificado pela quadra Natalícia.

Para além deste ponto, torna-se evidente que a diferença comportamental entre os restantes meses não é clara o que numa primeira fase não permite concluir o impacto que esta variável tem no número de visitas.

4.2.4 Relação entre o Número de Visitas e o Estado do Tempo

Para este caso é necessário também saber o número médio de visitas para cada estado do tempo. Esta informação está presente na tabela 4.8 e na figura 4.6.

Clima	Média das Visitas
Clear	33 933
Clouds	34 292
Drizzle	37 176
Mist	38 347
Partly Cloud	50 143
Rain	34 515
Scattered Clouds	39 284

Tabela 4.8: Média das visitas por estado do tempo

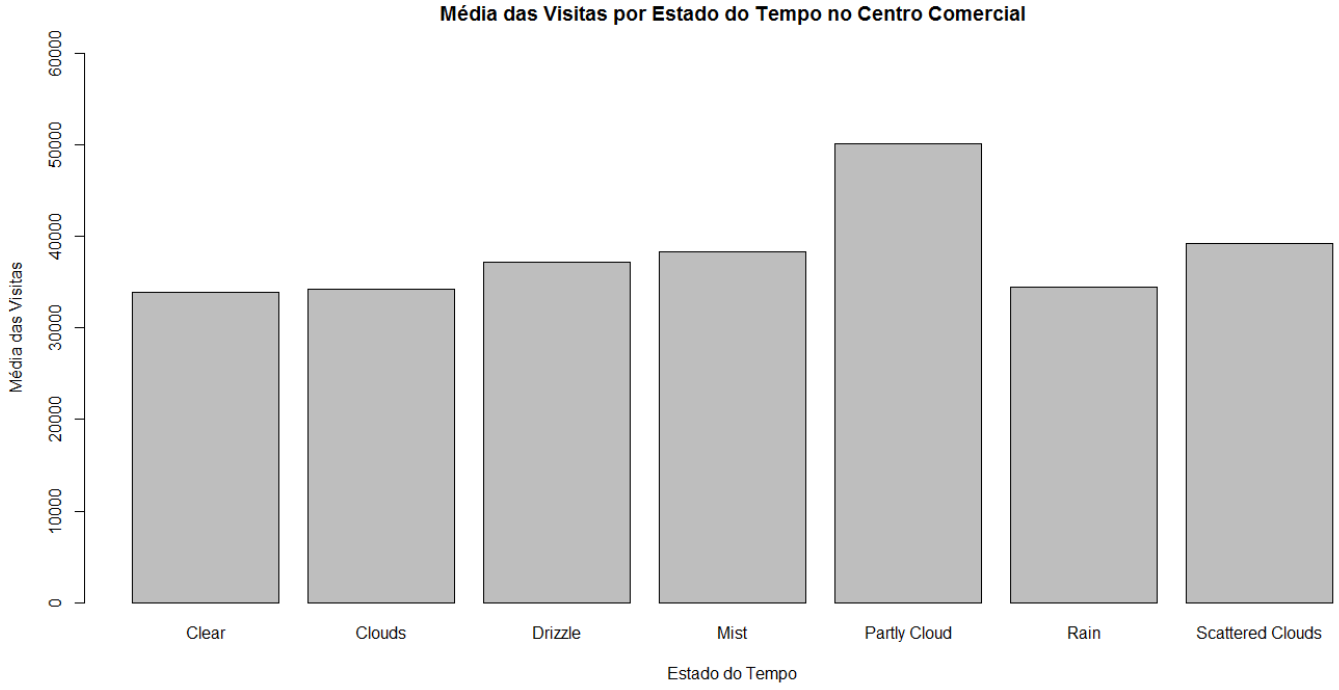


Figura 4.6: Gráfico de barras da média das visitas por estado do tempo

De modo a enriquecer esta análise, criou-se um diagrama de caixas de bigodes, figura 4.7, que relaciona o número de visitas com estado do tempo.

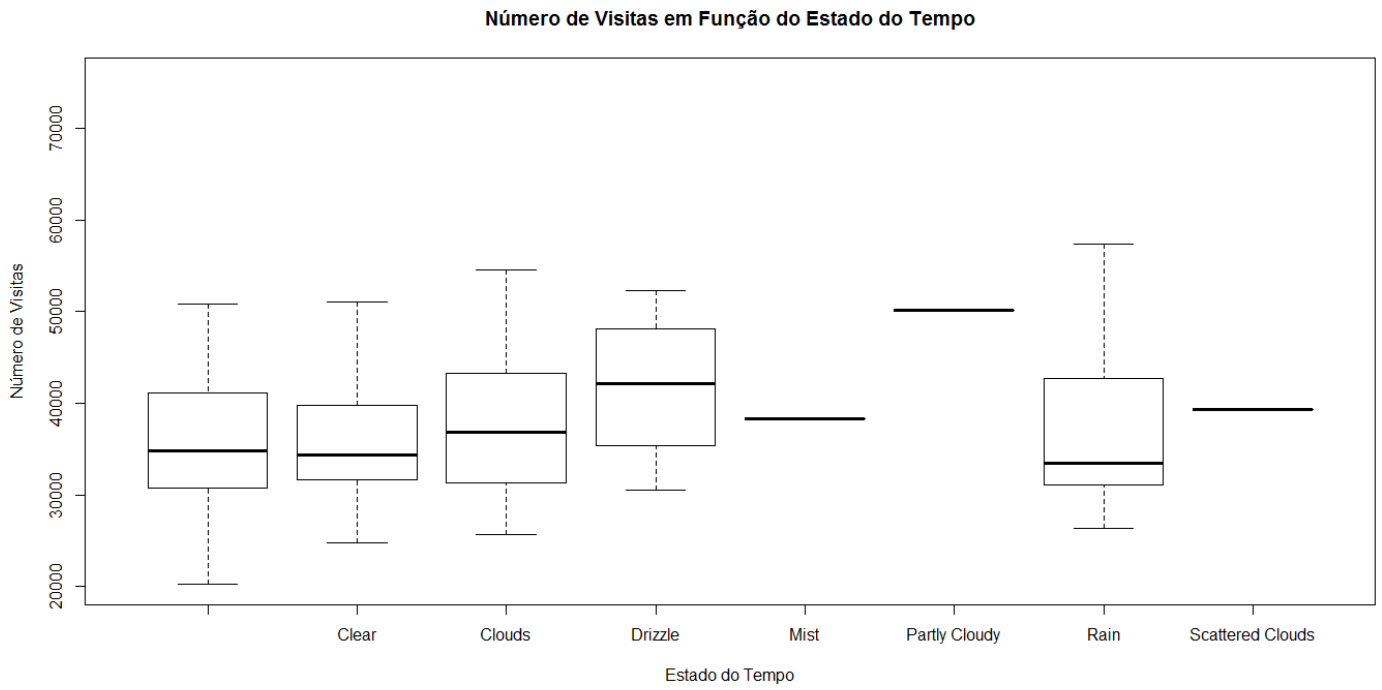


Figura 4.7: Diagrama de caixas de bigodes – Número de visitas por estado do tempo

Com base nos elementos apresentados, é notório, como já referido, que há três estados meteorológicos – *mist, partly cloud e scattered clouds*- que contêm apenas uma observação pelo que se considera que estes não são relevantes. Relativamente aos restantes estados meteorológicos não se permite ainda tirar conclusões com elevado grau de confiança relativamente ao seu impacto no número de visitas.

4.2.5 Análise da Influência das Variáveis no Número de Visitas

Como já referido, existem variáveis que têm influência no número de visitas e outras que aparentemente podem não ter influência alguma. Contudo vai-se proceder à avaliação da influência através de testes estatísticos.

Pretende-se portanto recorrer a testes de hipóteses para concluir estatisticamente quais as variáveis que têm efetivamente, ou não, impacto no número de visitantes do centro comercial e qual o grau desse impacto.

Para tal é necessário o recurso ao teste ANOVA - *Analysis of Variance* - e ao teste de *Tukey*.

Vai-se aplicar estes testes às seguintes variáveis:

- Meses
- Dia da Semana
- Estado do Tempo

4.2.5.1 Variável Meses

Sejam *i* e *j* as combinações entre os meses do ano e μ a média das visitas.

Formulação das hipóteses estatísticas:

H_0: $\forall_{i,j}\, \mu i = \mu_j$
H_1: $\exists_{i,j}\, \mu_i \neq \mu_j$

Aplicando o teste ANOVA para esta variável obtém-se a tabela 4.9

	Graus de Liberdade	Soma dos Quadrados	Quadrados das Médias	F	Pr(>F)
Meses	11	6.263e+09	569330321	11.97	<2e-16
Resíduos	350	1.665e+10	47568404		

Tabela 4.9: Tabela da Análise de Variância - Variável Meses

O valor de F diz-nos o quão longe estamos da hipótese da variável *Meses* não ser relevante para o número de visitas.

Calculando-se o valor crítico para um nível de 5% com os graus de liberdade 11 e 350, obtemos: $F^{krit}_{11.350}(5\%) = 1.8160$. Uma vez que F=11.97 e F> $F^{krit}_{11.350}(5\%)$ logo rejeitamos H_0.

É possível a mesma conclusão se Pr(F) $\leq$ 5%. Uma vez que 2e-16% < 5%, reforçamos a conclusão que a variável *Meses* tem impacto no número de visitas. Assim é concluído que existem diferenças no número de visitas ao centro comercial consoante o mês do ano.

De seguida, utilizar-se-á o método de Tukey HSD para se saber quais essas diferenças. Construindo intervalos de confiança para todos os pares de médias com um grau de confiança de 95% para todos os intervalos, obtemos a representação gráfica apresentada na figura 4.8.

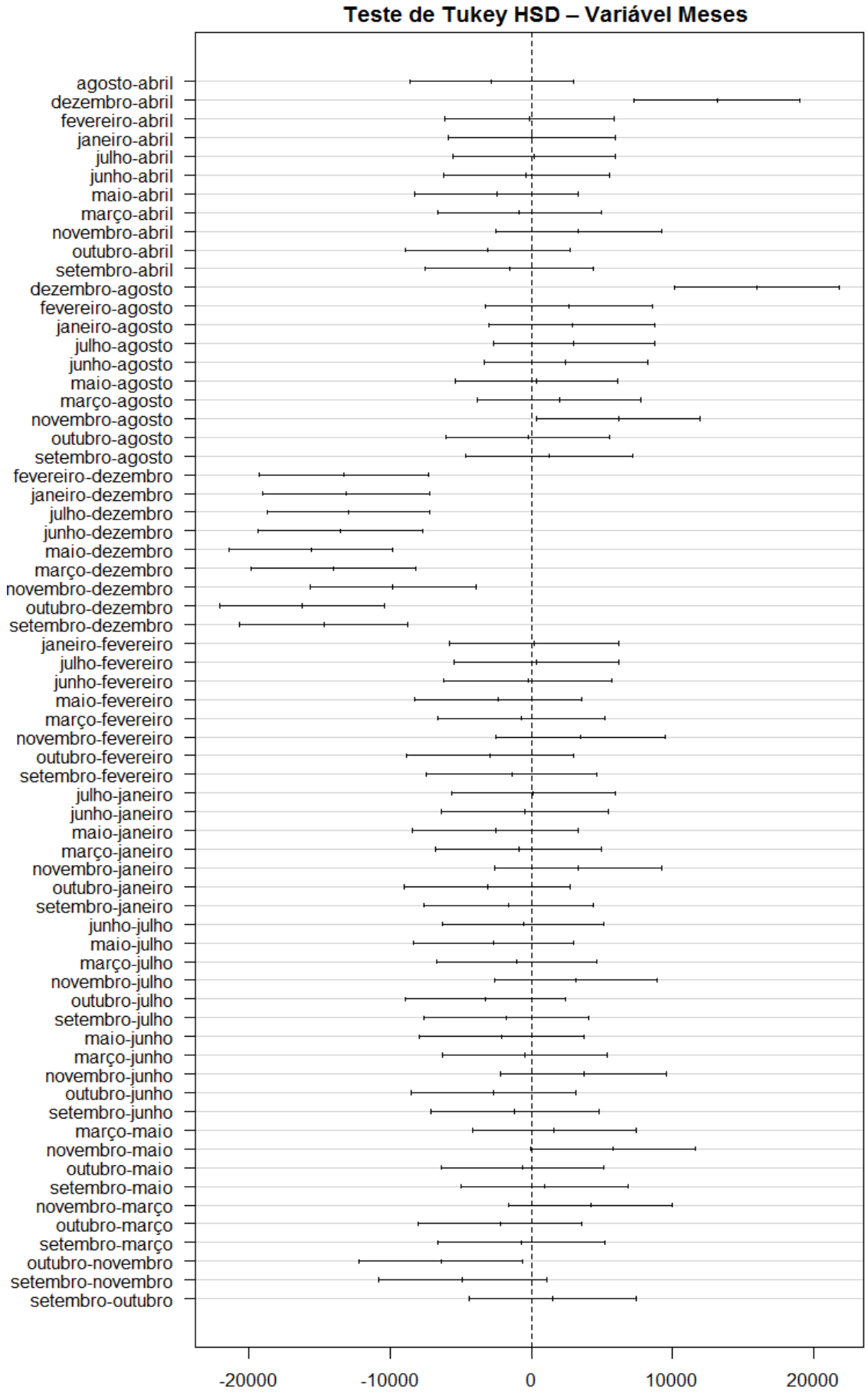

Teste de Tukey HSD – Variável Meses
agosto-abril
dezembro-abril
fevereiro-abril
janeiro-abril
julho-abril
junho-abril
maio-abril
março-abril
novembro-abril
outubro-abril
setembro-abril
dezembro-agosto
fevereiro-agosto
janeiro-agosto
julho-agosto
junho-agosto
maio-agosto
março-agosto
novembro-agosto
outubro-agosto
setembro-agosto
fevereiro-dezembro
janeiro-dezembro
julho-dezembro
junho-dezembro
maio-dezembro
março-dezembro
novembro-dezembro
outubro-dezembro
setembro-dezembro
janeiro-fevereiro
julho-fevereiro
junho-fevereiro
maio-fevereiro
março-fevereiro
novembro-fevereiro
outubro-fevereiro
setembro-fevereiro
julho-janeiro
junho-janeiro
maio-janeiro
março-janeiro
novembro-janeiro
outubro-janeiro
setembro-janeiro
junho-julho
maio-julho
março-julho
novembro-julho
outubro-julho
setembro-julho
maio-junho
março-junho
novembro-junho
outubro-junho
setembro-junho
março-maio
novembro-maio
outubro-maio
setembro-maio
novembro-março
outubro-março
setembro-março
outubro-novembro
setembro-novembro
setembro-outubro
-20000
-10000
0
10000
20000

Figura 4.8: Teste de Tukey HSD – Variável Meses

Esta representação permite concluir que o mês de dezembro comparativamente aos restantes meses do ano apresenta diferenças significativas o que é justificado pela quadra Natalícia uma vez que esta época atrai um número elevado de pessoas ao centro comercial. O mês de novembro também se destaca em alguns casos, nomeadamente quando comparado com os meses de agosto e outubro, em que a diferença significativa é clara, e também com setembro em que se verifica também uma ligeira diferença. Conclui-se que os meses novembro e dezembro são os que têm mais influência no número de visitas

4.2.5.2. Variável Dia da Semana

Sejam *i* e *j* as combinações entre os dias da semana e μ a média das visitas.

H_0: $\forall_{i,j}\ \mu i = \mu_j$
H_1: $\exists_{i,j}\ \mu_i \neq \mu_j$

Aplicando o teste ANOVA para a variável *Dia da Semana* obtém-se a tabela 4.10

	Graus de Liberdade	Soma dos Quadrados	Quadrados das Médias	F	Pr(>F)
Meses	6	1.037e+10	1.728e+09	48.89	<2e-16
Resíduos	355	3.534e+07	3.534e+07		

Tabela 4.10: Tabela da Análise de Variância - Variável *Dia da Semana*

O valor de F diz-nos o quão longe estamos da hipótese da variável *Dia da Semana* não ser relevante para o número de visitas.

Calculando-se o valor crítico para um nível de 5% com os graus de liberdade 6 e 355, obtemos: $F^{krit}_{6.355}(5\%) = 1.8160$. Uma vez que F=48.89 e F> $F^{krit}_{6.355}(5\%)$ logo rejeitamos H_0.

É possível a mesma conclusão se Pr(F) $\leq$ 5%. Uma vez que 2e-16% < 5%, reforçamos a conclusão que a variável *Dia da Semana* tem impacto no número de visitas. Assim é concluído que existem diferenças no número de visitas ao centro comercial consoante o dia da semana.

De seguida, utilizar-se-á o método de Tukey HSD para se saber quais essas diferenças. Construindo intervalos de confiança para todos os pares de médias com um grau de confiança de 95% para todos os intervalos, obtemos a representação gráfica apresentada na figura 4.9.

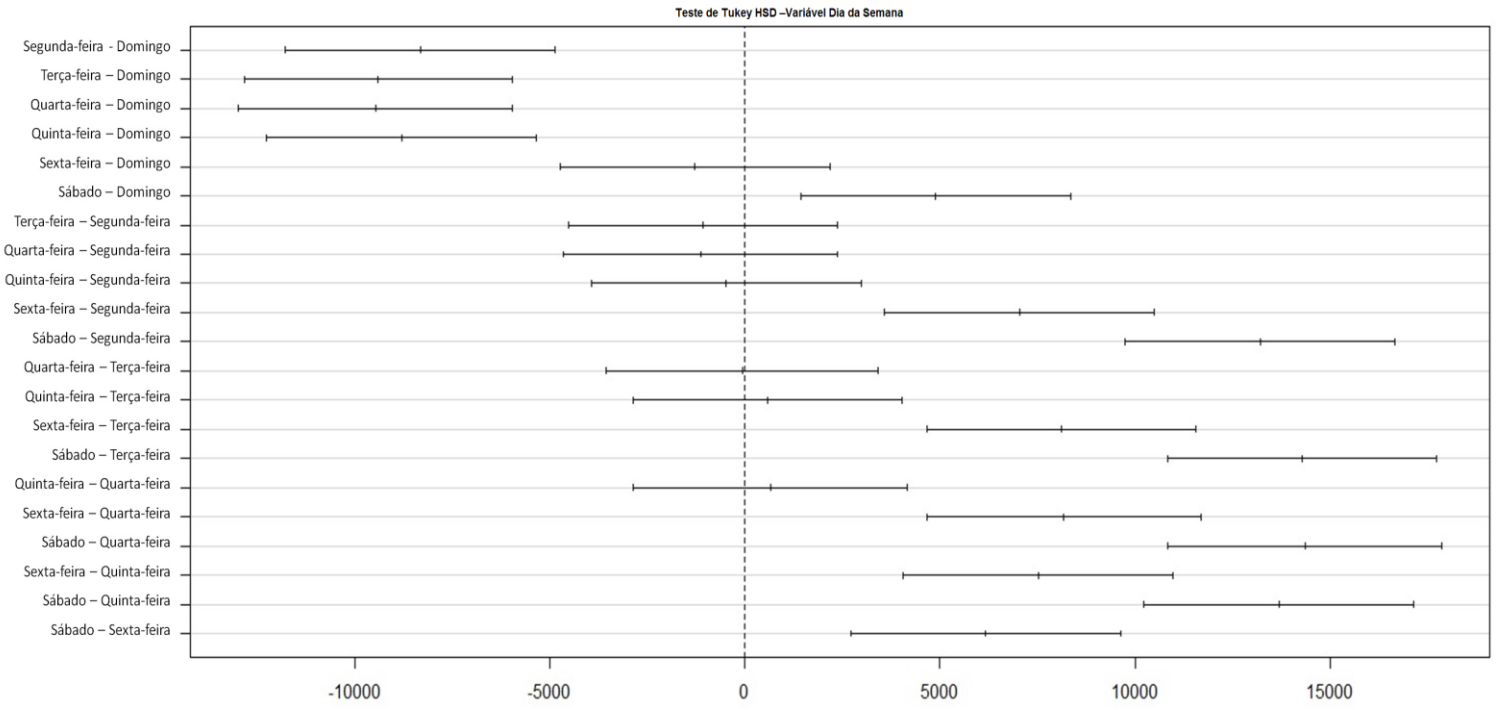


Figura 4.9: Teste de Tukey HSD –Variável *Dia da Semana*

Conclui-se assim que o domingo comparativamente a todos os dias da semana, apresenta diferenças significativas, excetuando a sexta-feira pois são dias considerados semelhantes. O sábado e a sexta-feira também apresentam algumas diferenças quando comparados com a segunda-feira, a quarta-feira e a quinta-feira. Relativamente ao dia de sábado é possível verificar ainda uma diferença significativa com a terça-feira e a sexta-feira.

Fez-se ainda um teste em que se denominou o dia da semana de todos os dias feriados por "feriado". Concluiu-se que os feriados têm impacto no número de visitas do centro comercial e concluiu-se ainda que o um dia feriado é semelhante aos domingos e às sextas-feiras.

4.2.5.3. Clima

Sejam:

Sejam i e j as combinações entre os diferentes estados do clima e μ a média das visitas.

Formulação das hipóteses estatísticas:

H_0: $\forall_{i,j}\ \mu i = \mu_j$
H_1: $\exists_{i,j}\ \mu_i \neq \mu_j$

Aplicando o teste ANOVA para a variável *Clima* obtém-se a tabela 4.11

	Graus de Liberdade	Soma dos Quadrados	Quadrados das Médias	F	Pr(>F)
Meses	7	7.14e+08	1.02e+08	1.62	12.7
Resíduos	354	2.22e+10	6.27e+07		

Tabela 4.11: Tabela da Análise de Variância - Variável *Clima*

O valor de F diz-nos o quão longe estamos da hipótese da variável *Clima* não ser relevante para o número de visitas.

Calculando-se o valor crítico para um nível de 5% com os graus de liberdade 7 e 354, obtemos: $F^{krit}_{7.354}(5\%) = 2.035471$. Uma vez que F=1.62 e F> $F^{krit}_{7.354}(5\%)$ logo rejeitamos H_0.
É possível a mesma conclusão se Pr(F) $\geq$ 5%. Uma vez que 12.7% > 5%, reforçamos a conclusão que a variável *Clima* não tem impacto no número de visitas.

De seguida, utilizar-se-á o método de Tukey HSD para comprovar esta conclusão.

Construindo intervalos de confiança para todos os pares de médias com um grau de confiança de 95% para todos os intervalos, obtemos a representação gráfica apresentada na figura 4.10:

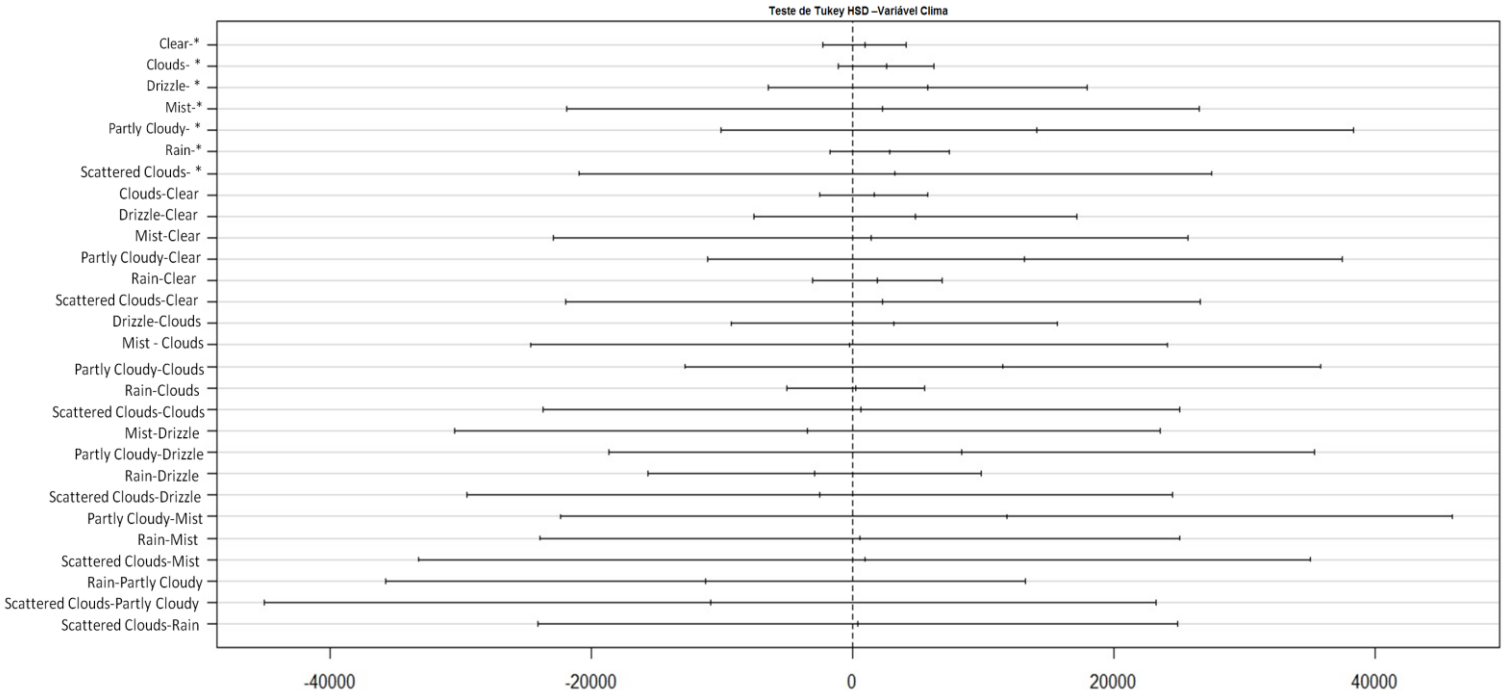


Figura 4.10: Teste de Tukey HSD –Variável *Clima*

Conclui-se assim que a variável *Clima* não tem um impacto significativo no número de visitas do centro comercial em estudo. Isto é considerado normal uma vez que Portugal é considerado um dos países mais amenos da Europa, com um clima mediterrânico, ou seja, um tipo de clima temperado, o que caracteriza regiões cuja temperatura varia regularmente ao longo do ano, mas apesar disto consegue apresentar 4 estações bem definidas. Contudo os estados meteorológicos *Rain* e *Clouds* são aqueles que mais perto estão de provocar algum impacto no número de visitas, o que é expectável do ponto de vista do senso comum.

4.3 Preparação dos Dados

Durante o estudo da secção anterior foi possível apresentar os dados iniciais, explora-los, verificar a qualidade destes e ainda apresentar dados que foram criados para simplificar o trabalho nomeadamente as variáveis *meses* e *dia da semana*.

Estes dados iniciais foram extraídos da API em formato *json* e posteriormente este ficheiro foi interpretado pelo R e guardou-se esta mesma informação numa estrutura de dados específica do R denominada por *data frames* pelo facto de ser uma estrutura que permite uma manipulação simplificada dos dados.

Concluiu-se que os atributos originais não eram suficientes para avançar com o desenvolvimento deste projeto.

Como referido, criaram-se novos atributos que se revelaram bastante uteis e foram também construídos novos dados.

Foram criadas as variáveis *"Dia da Semana"* e *"Meses"* por serem mais úteis do que a própria data e tornam mais simples a manipulação e validação dos dados.

Verificou-se ainda que existem dados que foram considerados irrelevantes. Com base nisto excluiu-se os dados das visitas diárias do mês de julho de 2013 pois, como referido, apenas tinha duas observações o que não permite tirar grandes conclusões.

Os dias 25 de dezembro de 2013 e o 01 de janeiro de 2014 foram retirados pois nestes dias as lojas do centro comercial encerraram exceção feita aos cinemas.

Assim eliminaram-se estes dias do conjunto de dados para não influenciar a análise dos mesmos.

4.4 Modelação

Os parâmetros que se pretendem estudar são o número de meses e o produto de uma constante pelo intervalo interquartil (IQR) (como já referido, termo da estatística descritiva). Para perceber o real impacto desta constante no problema decidiu-se analisar vários valores. Começou-se com um valor mínimo de 1.1 e foi-se aumentando o valor desta constante até ao especialista de negócio aceitar os resultados obtidos. Após vários estudos, concluiu-se utilizar os valores 1.5 (valor por defeito no cálculo dos gráficos de caixas de bigodes), 1.9. Este valores foram denominados de valor de avaliação e valor de exclusão.

Nessa fase, pretende-se fazer uma deteção de *outliers* univariada, estudar os parâmetros de entrada e calibra-los de modo a obter os valores otimizados para o cálculo de dias com número de visitas considerados anómalos.

Os meses sujeitos a esta análise pertencem ao intervalo de tempo de agosto de 2013 a junho de 2014.

A proposta de solução passa por escolher um mês para ser analisado - mês *pivot* - e calcular os *outliers* desse mês em função de um conjunto de meses anterior a esse *pivot* para avaliação do algoritmo.

Exemplificando:

Imaginando que se pretende calcular o número de *outliers* em novembro, cria-se os conjuntos de meses outubro, setembro-outubro, agosto-setembro-outubro e para cada um deles calcula-se o quartil 1 e o quartil 3 para o número de visitas dos dias da semana. Assim, para cada dia da semana calcula-se o intervalo $[quartil1 - valor\ Avaliaçao \times (quartil3 - quartil1)\ ;\ quartil3 + valor\ de\ Avaliação \times (quartil3 - quartil1)]$ e posteriormente determina-se os dias de novembro que estão fora deste intervalo. Desta forma consegue-se concluir o impacto que o número de visitas dos conjuntos de meses tem no mês em análise.

Foram feitas 3 abordagens ao problema:

1. Contagem de *outliers*
2. Eliminação de todos os *outliers* detetados
3. Contagem de todos os *outliers* e eliminação dos *outliers* severos.

- **Contagem de *outliers***

Esta primeira abordagem consiste em utilizar a metodologia descrita e contar os dias *outliers* dos meses *pivots*.

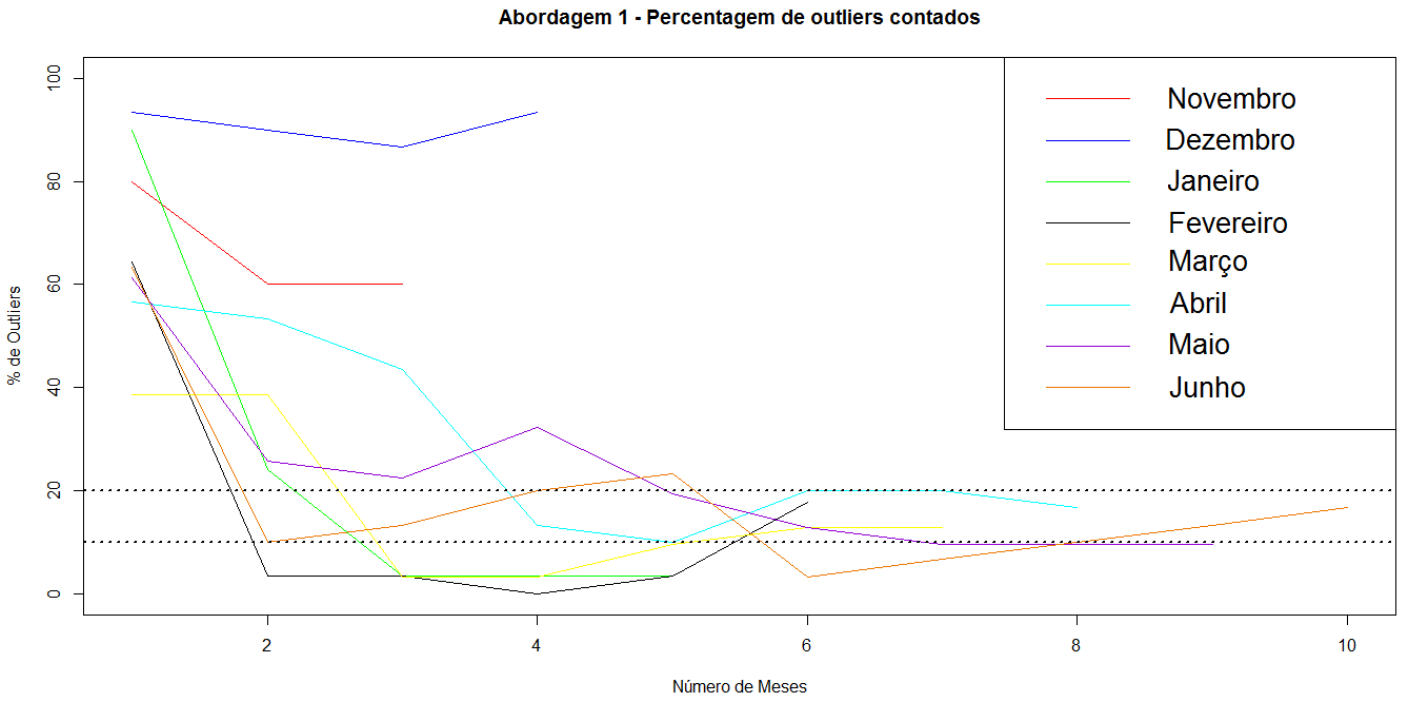


Figura 4.11: Abordagem 1 –Percentagem de *outliers* calculados

O gráfico da figura 4.11 representa a deteção de *outliers* para o mês de novembro. O mês *pivot* escolhido foi novembro pelo facto de ser o mês considerado mais constante e com um maior número de meses anteriores cujos comportamentos não fogem do comportamento considerado normal e pretende-se alcançar uma percentagem de 10% a 20% de *outliers*. A cada iteração de análise do mês *pivot*, o conjunto de meses anteriores a este, aumenta o que está presente no eixo horizontal. Após a análise de novembro e respetivos meses anteriores (outubro, setembro e agosto), o algoritmo avança para dezembro e repete o processo até ao último mês do conjunto - junho. Cada linha do gráfico representa o mês *pivot* e os vários pontos dessa linha o número de *outliers* para cada conjunto de meses anteriores.

Após este processamento conclui-se através da figura 4.11 que a percentagem de *outliers* tende a baixar e a variar menos à medida que o número de meses anteriores ao *pivot* aumenta. Esta primeira abordagem foi fundamental para ganhar uma perceção relativamente ao comportamento dos *outliers*. Uma vez que o conjunto de dados é limitado a nível histórico, visto que não existe período homólogo, existem meses *pivots* com poucos meses anteriores. É o caso de setembro, por exemplo. É uma limitação do projeto pois enriqueceria o mesmo ao recorrer às visitas do período homólogo para uma análise mais robusta.

- **Eliminação de todos os *outliers* detetados**

Nesta abordagem optou-se pela eliminação de todos os os *outliers* detetados com o intuito de não permitir que estes influenciassem os possíveis *outliers* dos meses seguintes ao *pivot*. Os resultados obtidos estão apresentados na figura 4.12

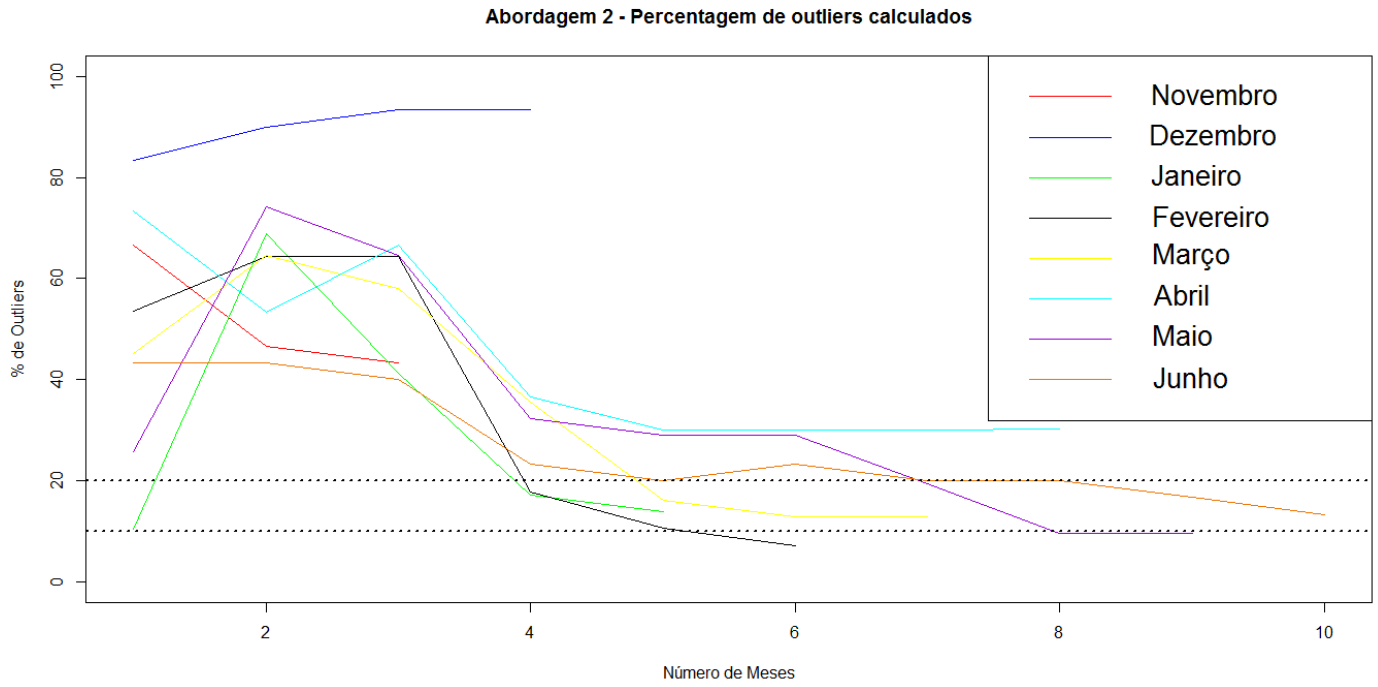


Figura 4.12: Abordagem 2 – Percentagem de *outliers* calculados

- **Eliminação de *outliers* severos**

Utilizando o princípio da abordagem anterior, de eliminar *outliers*, decidiu-se não eliminar todos os *outliers* encontrados, mas apenas os severos. Nesta abordagem utilizaram-se duas constantes: valor de avaliação (como já referido) e de exclusão. Para determinar os *outliers* severos utilizou-se um *range* de 1.9 (valor de exclusão) para aumentar o intervalo interquartil.

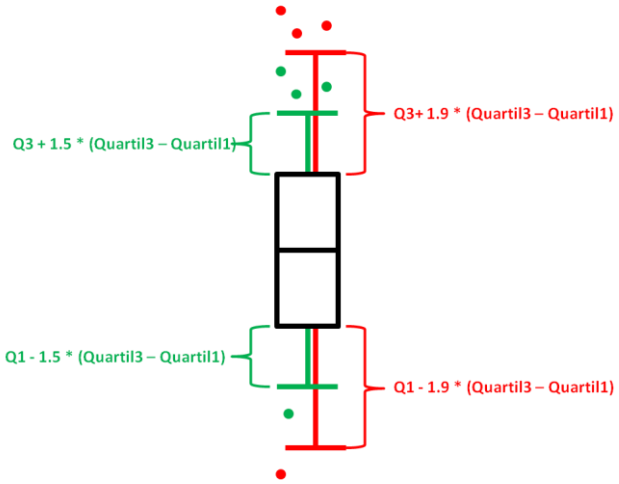


Figura 4.13: Abordagem 3 – Explicação Gráfica

A figura 4.13 ilustra este processo. Numa primeira fase são calculados todos os *outliers* com o valor de avaliação de 1.5. Segundo o exemplo apresentado na figura, o resultado seria o somatório de todos os *outliers* representados (4 de cor verde + 4 de cor vermelha = 8 *outliers*). Na segunda fase seriam eliminados todos os *outliers* severos. Estes são os valores calculados com o valor de exclusão de 1.9. Segundo o exemplo da figura, seriam eliminados todos os *outliers* de cor vermelha. Utilizando este raciocínio desenvolveu-se um algoritmo e analisou-se os seus resultados.

Assim apresenta-se na imagem 4.14 e apresenta-se ainda na figura 4.15 a percentagem de *outliers* removidos a cada iteração.

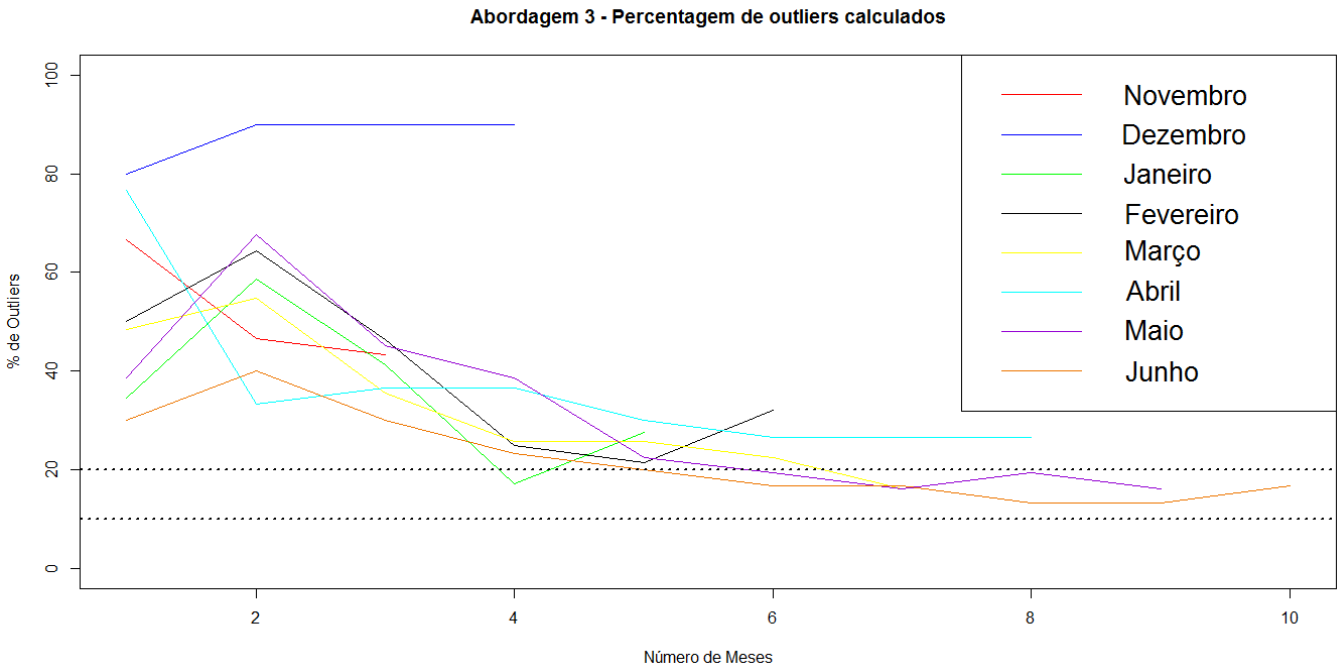


Figura 4.14: Abordagem 3 – Percentagem de *outliers* calculados

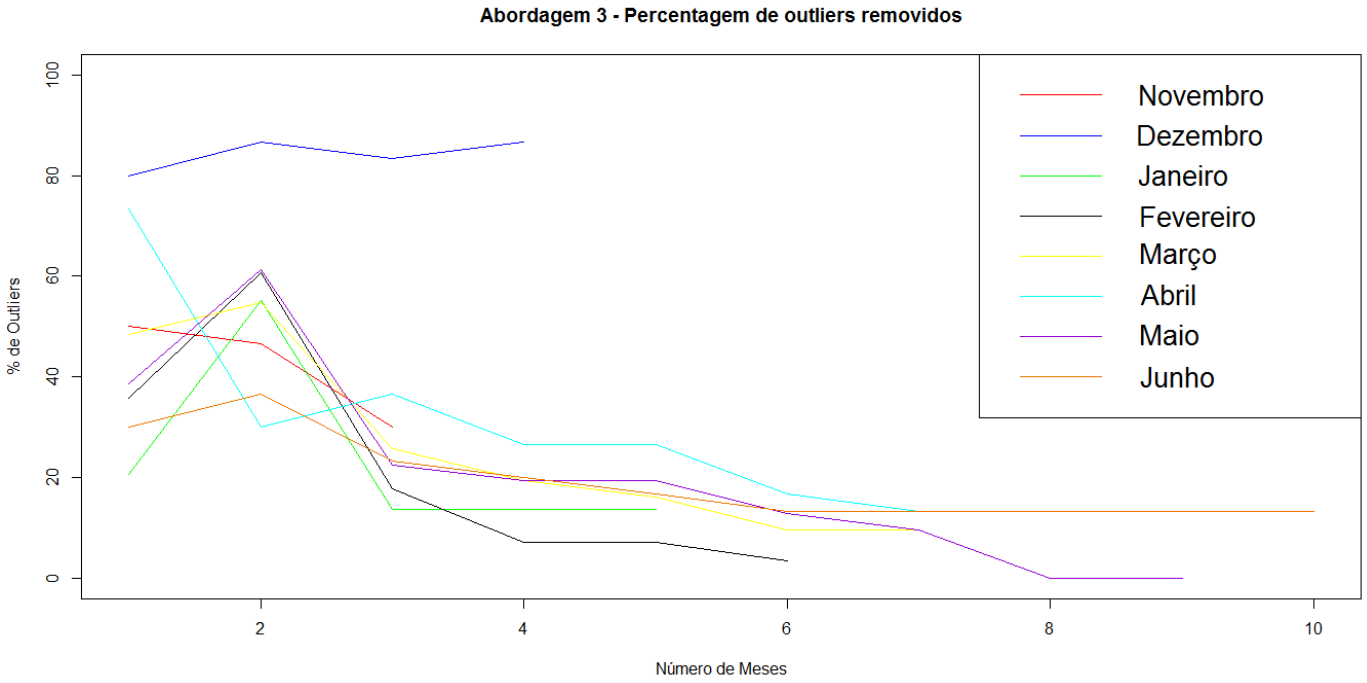


Figura 4.15: Abordagem 3 – Percentagem de *outliers* removidos

Optou-se ainda, para esta mesma abordagem, aumentar o valor de exclusão de 1.9 para 2.1. O valor de avaliação manteve-se em 1.5. O objetivo desta alteração foi estudar também o impacto do valor de exclusão nesta abordagem. Assim apresenta-se na imagem 4.16 e apresenta-se ainda na figura 4.17 a percentagem de *outliers* removidos a cada iteração.

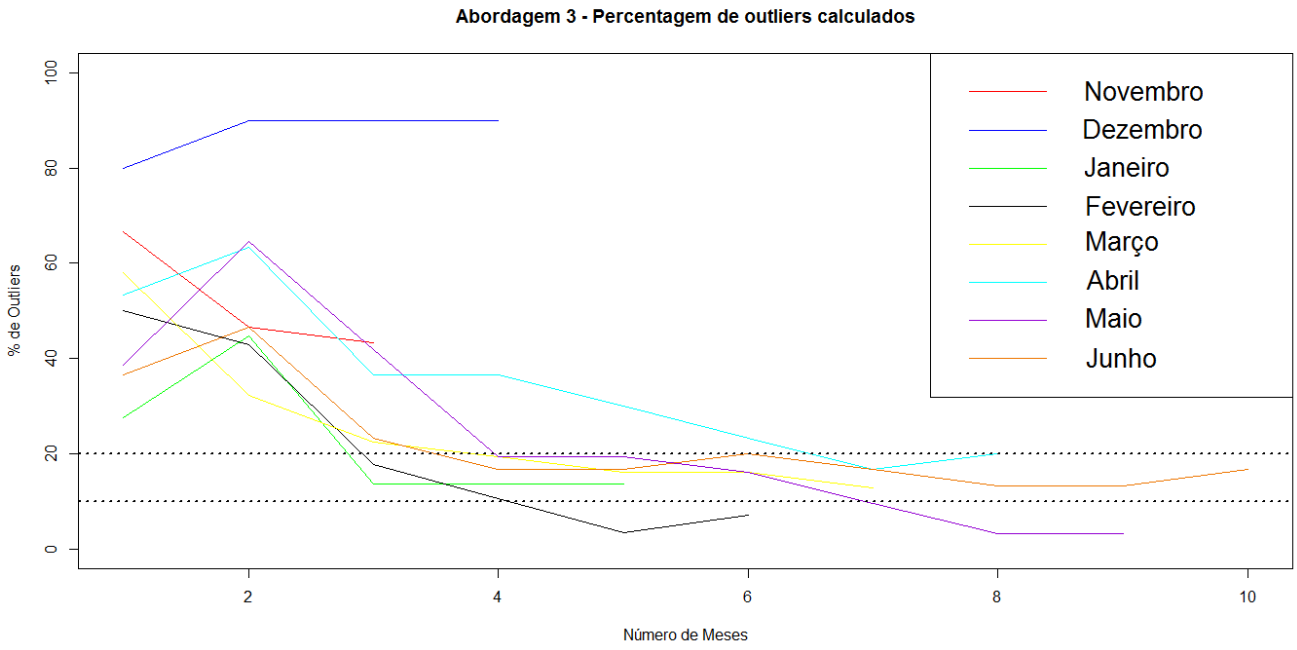


Figura 4.16: Abordagem 3 – Percentagem de *outliers* calculados com novo valor de exclusão

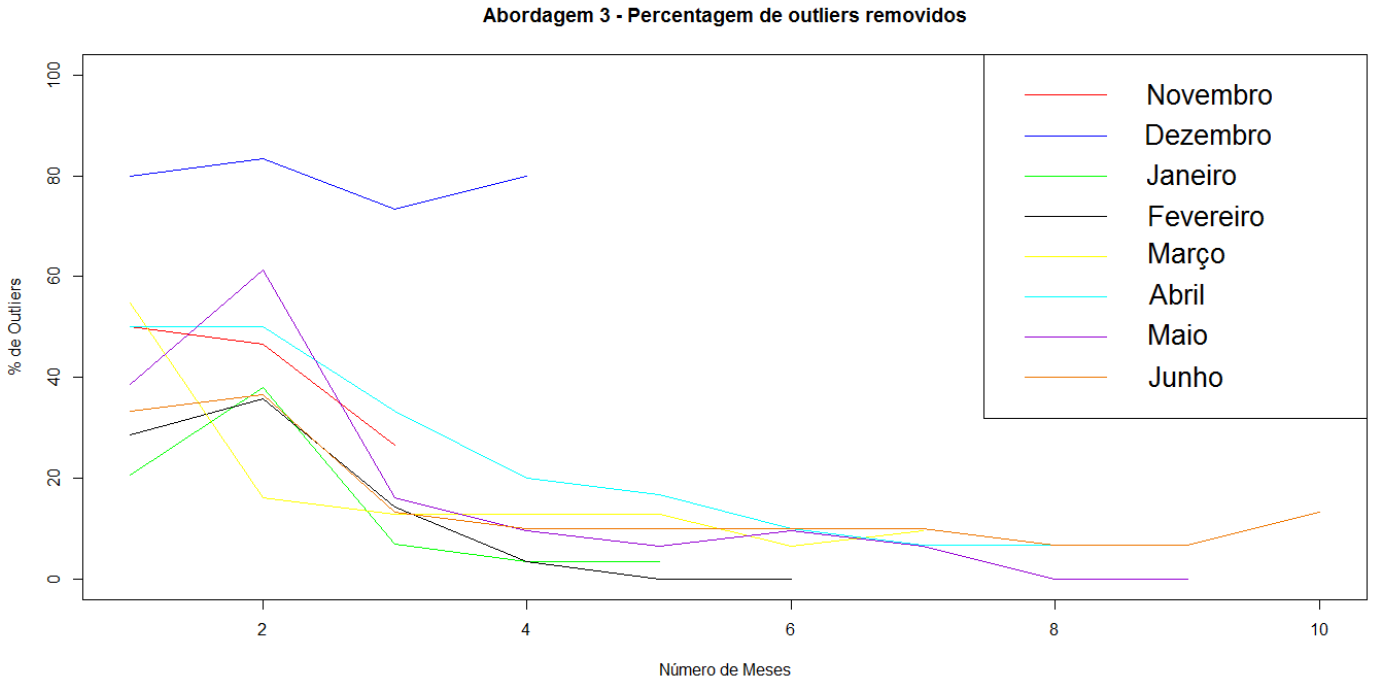


Figura 4.17: Abordagem 3 – Percentagem de *outliers* removidos com novo valor de exclusão

Nesta terceira abordagem implementou-se uma nova característica no algoritmo. No caso de haver mais de 50% de dias excluídos, o algoritmo adapta-se a essa situação e deixa de considerar esses dias *outliers*, passando-os a considerar dias normais nas iterações seguintes. Desta forma tenta-se que o algoritmo aprenda e consiga adaptar-se caso existam meses com números de visitas muito heterogéneas.

4.5 Avaliação

Nesta secção são avaliados os resultados obtidos na secção anterior. Esta fase de avaliação foi desenvolvida juntamente com um especialista da área de negócio da Movvo.

Como referido, era fundamental definir a percentagem de *outliers* pretendida. Esta decisão é puramente empírica e o intervalo definido foi [10%-20%] de *outliers* num mês. Os resultados apresentados são de uma vista mensal pois o comportamento dos dados é mais homogéneo e facilita a opinião final sobre a o melhor conjunto dos dados iniciais.

Avalia-se de seguida os resultados das três abordagens detalhadas na secção anterior.

- **Contagem de *outliers***

Como já referido, esta abordagem permitiu obter uma perceção do comportamento dos *outliers* segundo a metodologia pensada para resolver este problema. Através da figura 4.11, deduz-se que com poucos meses obtém-se um valor muito alto de dias *outliers*.

Como esperado, conclui-se que dezembro possui um grande número de valores *outliers* quando comparado com o conjunto de meses anteriores devido à quadra Natalícia que leva um número bastante alto de visitas ao centro comercial. Conclui-se ainda que após dezembro, o número de dias *outliers* tende a diminuir, pois os valores das visitas dos meses anteriores são bastante heterógenos, o que faz com que seja mais difícil detetar *outliers*. Isto leva também, ao fim de algum tempo, a que o número de *outliers* varie menos e se torne mais estável.

- **Eliminação de todos os *outliers* detetados**

Esta abordagem foi considerada insuficiente e uma abordagem a evitar. O facto de eliminar todos os *outliers* encontrados, e o facto de haver um grande número de *outliers* com poucos meses, como ilustra a figura 4.12, faz com que numa fase inicial sejam eliminados bastantes dias do conjunto de dados e o resultado seja amostras de dados, bastante pequenas. Ao terceiro mês conclui-se que existe um número de *outliers* bastante elevado comparativamente à abordagem anterior. Com o aumento do número de meses, conclui-se que o número de *outliers* estabiliza, o que apesar de parecer positivo, não é correto, pois o conjunto de dados torna-se bastante reduzido após recursivas eliminações de dias considerados *outliers*.

Nesta fase foi necessário repensar no processo. Surgiu assim a terceira abordagem.

- **Eliminação de *outliers* severos**

Mantendo-se a intenção de eliminar *outliers*, para não influenciar os meses seguintes, e com a noção de que não seria correto eliminar todos os dias *outliers*, optou-se pela estratégia de eliminar apenas os *outliers* severos.

Avaliando os resultados apresentados na figura 4.14, conclui-se que o número de *outliers* calculados diminui a partir dos dois meses. Contudo, de forma geral, o número de *outliers* está acima do intervalo pretendido. Só a partir dos sete meses é que o número de *outliers* começa a atingir o intervalo pretendido. Avaliando ainda a figura 4.15, é possível concluir que à medida que o número de meses aumenta, a percentagem de *outliers* eliminados diminui.

Apesar de se ter um modelo correto, o número excessivo de *outliers* forçou uma revisão do processo. Assim, foram alterados os valores de avaliação e de exclusão, de modo a alterar o intervalo interquartil e concluiu-se que o valor de exclusão tem mais impacto na diminuição do número de *outliers*. Seguindo este raciocínio, aumentou-se o valor de exclusão de 1.9 para 2.1 e obteve-se os resultados ilustrados na figura 4.16 e conclui-se que o número de *outliers* começa a atingir o intervalo pretendido após os três meses.

Avaliando ainda a figura 4.17 é possível concluir que à medida que o número de meses aumenta, a percentagem de *outliers* eliminados também diminui. Estes resultados vão ao encontro do pretendido e pode-se concluir através destes resultados os dados ideias iniciais sendo eles:

- Intervalo de meses: [4,6]
- Valor de avaliação: 1.5
- Valor de exclusão: 2.1

4.6 Desenvolvimento

Pretende-se nesta secção explicar o desenvolvimento desta ferramenta e como se pretende interligar esta com o BIPS.

4.6.1 Produto

O primeiro ponto necessário de compreensão é que o resultado deste projeto não será uma aplicação nova.

Durante o desenvolvimento desta ferramenta recursiva, teve-se em atenção a sua otimização. Esta situação foi tida em atenção para que a ferramenta não perdesse eficiência em função da quantidade de dados processados.

Esta ferramenta estará sempre em execução e sempre a atualizar-se para que, a qualquer momento, o gestor do centro comercial possa ver um número atualizado e correspondente à realidade do número de dias anómalos no centro comercial.

Caso o gestor do centro comercial pretenda, no futuro, será possível alterar o conjunto de dados inicial, considerado ideal. Assim será possível incluir mais meses, ou remover, do conjunto inicial de modo a corresponder às suas exigências. Esta flexibilidade já é possível atualmente no algoritmo desenvolvido.

Ainda não está definido como o resultado deste trabalho irá ser incorporado no sistema BIPS, pois a equipa da Movvo responsável pela API MiddleWare ainda está a pensar numa solução que torne o sistema já existente capaz de trabalhar com os ficheiros produzidos em R. Este trabalho está atualmente a ser planeado pelos colaboradores da empresa e estes irão executá-lo o mais brevemente possível para que o trabalho realizado esteja presente na próxima versão do produto. Contudo este trabalho é responsabilidade de outra equipa da empresa. Apesar disto existe já uma arquitetura definida e vai ser descrita no ponto seguinte.

Esta ferramenta, como já mencionado, será integrada no *dashboard* produzido pelo BIPS. Assim, acedendo ao *dashboard*, a qualquer momento, o gestor do centro comercial poderá visualizar quais os dias que foram considerados anómalos e ainda uma lista de eventos que possam justificar esse número de visitas anormal no centro comercial. Será ainda apresentada a percentagem dos dias que foram considerados *outliers*. Considera-se ainda de interesse referir que as pessoas que irão ao centro comercial sem dispositivos móveis, serão efetivamente visitas, mas não serão contados pelo sistema. Esta situação não é passível de resolução.

4.6.2 Arquitetura Geral

O sistema desenvolvido possui uma arquitetura que é ilustrada na imagem 4.18:

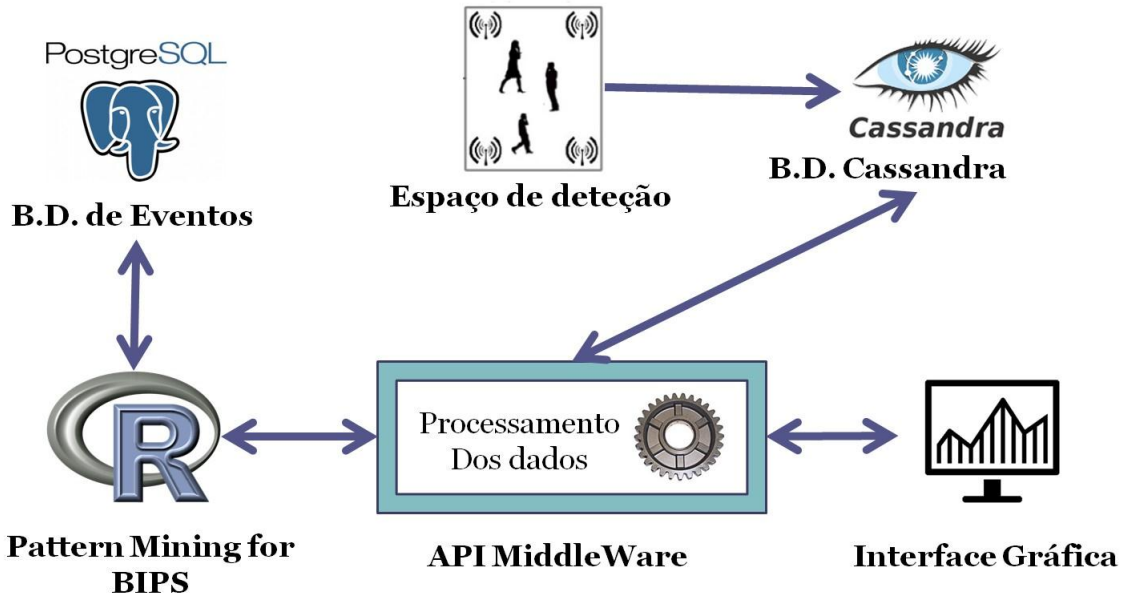


Figura 4.18: Diagrama de arquitetura

O espaço de deteção diz respeito à tecnologia BIPS. As designadas antenas (como já referido, *bNode*) captam os dados de deteção e enviam-nos para um servidor (componente *bServer*) que contém uma base de dados Cassandra. A partir deste momento os dados ficam armazenados, seguros e disponíveis. A API desenvolvida pelos colaboradores da empresa, permite o acesso aos dados já processados e aglomerados em métricas, ficando assim disponíveis para o relatório final ou também designado *dashboard* ou simplesmente interface gráfica. Com esta API aumenta-se a segurança com a centralização dos acessos à base de dados. Este é o processo atual.

No fim do desenvolvimento deste projeto, os ficheiros produzidos em R irão aceder também à API para obtenção da informação para análise. Como já referido, será necessário uma base de dados externa com eventos de diversos tipos armazenados. Haverá um acesso a uma base de dados para saber possíveis justificações para determinadas situações anómalas ocorridas em determinados intervalos de tempo.

Procede-se agora à explicação desta base de dados de eventos.

4.6.3 Base de Dados de Eventos

A criação de uma base de dados com eventos possibilita a resposta ao problema de negócio identificado. A base de dados foi implementada em PostgreSQL.

A ideia consiste em registar o maior leque possível de situações que possam, direta ou indiretamente, causar impacto no número de visitas no centro comercial. Foi feita uma entrevista junto de colaboradores de centros comerciais com o intuito de perceber quais as situações que causam impacto no normal funcionamento de um centro comercial. As conclusões foram as seguintes:

- No primeiro sábado de cada mês o cinema é grátis para crianças. Torna-se assim útil analisar se esta iniciativa atrai normalmente mais pessoas ao centro comercial, em concreto, à zona dos cinemas.

- Estreias e antestreias de filmes com grande expectativa podem atrair nesses dias, e também nos dias seguintes, mais pessoas ao centro comercial, e uma vez mais, à zona dos cinemas para verem o filme. Torna-se útil, como no caso anterior, para justificar se esses dias fazem com que haja um aumento do número de pessoas no centro comercial.

- Promoções, saldos e campanhas específicas são como se sabe sinónimo de muitas pessoas nos centros comerciais. Torna-se portanto importante registar estes eventos pois levam muita gente ao centro comercial e possibilita justificar um elevado número de pessoas na zona pertencente à loja responsável por estes eventos.

- Padrões sazonais têm uma grande utilidade. Com base nisto é possível fazer comparações por estação. É possível analisar se no Verão, por exemplo, o número de visitas ao centro comercial é menor comparativamente com o Inverno podendo assim obter-se mais um fator para justificar um aumento do número de visitas.

- Épocas festivas exigem por si só um registo na base de dados. Natal, Páscoa, Ano Novo, Dia dos Namorados, Carnaval, São João podem ter influência direta no número de visitas ao centro comercial. Segundo a opinião dos colaboradores entrevistados, no Natal e no Dia dos Namorados o número de visitas aumenta bastante enquanto no São João e no Ano Novo o número de visitas diminui drasticamente.

- O registo da inauguração e encerramento de lojas poderá ter uma grande utilidade pois permitirá uma análise ao impacto dessa ocorrência. O número de visitantes na zona dessa loja específica poderá aumentar ou diminuir e assim será possível saber o impacto que teve junto dos visitantes.

- Eventos desportivos tais como Campeonatos do Mundo e da Europa de futebol, Volta a Portugal em Bicicleta ou eventos desportivos patrocinados por lojas de desporto existentes no centro comercial em questão irão ser registados pois é admissível que os visitantes se desloquem às lojas de desporto do centro comercial provocando assim algum impacto no número de visitas ao centro comercial. Desta forma será possível analisar o impacto destes eventos nestas mesmas lojas.

- Pretende-se também registar eventos pontuais tais como por exemplo música ao vivo no centro comercial, épocas de exames ou campanhas específicas de determinadas lojas pois são eventos que poderão ter impacto no número de visitas.

O esquema da base de dados é o ilustrado na figura 4.19:

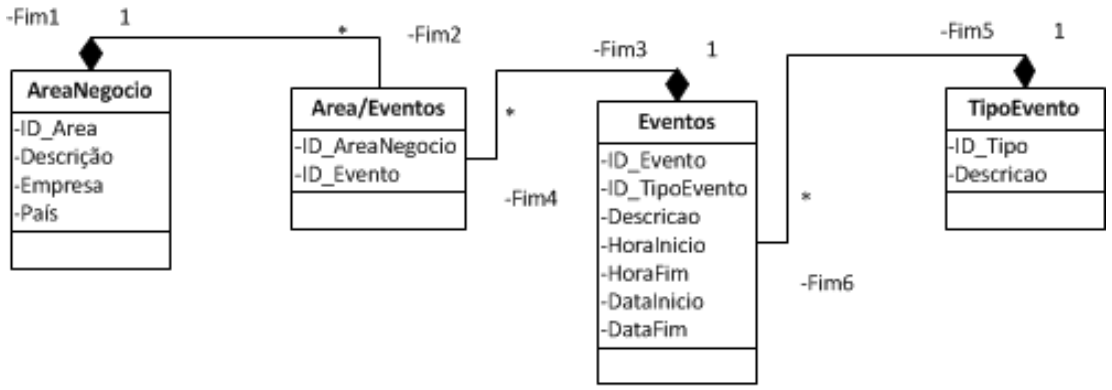


Figura 4.19: Modelo de Dados da BD de eventos

Na eventualidade de a Movvo expandir o seu negócio para outras áreas além da área do retalho, a base de dados estará preparada com a existência da tabela "AreaNegocio" e assim será possível associar eventos a áreas de negócio. A tabela "TipoEvento" permitirá agrupar os eventos por géneros visto que há eventos dos mais diversos tipos. Assim será possível agrupar eventos por cinema, festas, meteorologia entre outros.

Esta base de dados é acedida por um servidor desenvolvido durante este projeto que é invocado pelo R através de pedidos HTTP no qual se implantou código R. Este servidor recebe como parâmetros as datas atípicas anteriormente calculadas e devolve os eventos externos, registados na base de dados mencionada, que ocorreram perto dessas datas. Esta estratégia permite o funcionamento do sistema proposto nesta dissertação e faz com que este esteja disponível a qualquer instante para qualquer colaborador da empresa.

A solução consiste em utilizar a plataforma Domino, que permite criar *API Endpoints*, cuja funcionalidade principal é resolver o atrito existente com a integração dos modelos produzidos em R em sistemas de produção. *API Endpoint* é um serviço *web* que permite definir endereços ou pontos de conexão a outros serviços *web*. Assim é possível tornar o código produzido pela linguagem R acessível a aplicações externas. Fazendo o *upload* para o Domino da função em R que procura eventos para as datas recebidas como parâmetros, criou-se um novo *API Endpoint*

em que se especifica o ficheiro e a função, com o endereço da base de dados, a invocar quando esta API é utilizada. O Domino implementa o ficheiro R num servidor de baixa latência e atende a pedidos HTTP. Quando um pedido chega, o Domino retira os argumentos do pedido HTTP, passa-os parâmetros para a função R especificada, e retorna os resultados numa resposta HTTP.

Capítulo 5

Avaliação do Impacto de Eventos numa Loja de Desporto

No seguinte capítulo vai-se abordar todas as fases associadas à implementação de um caso de estudo. Seguiu-se, uma vez mais, a metodologia de CRISP-DM para a exploração do problema, seguindo-se as várias etapas já referidas anteriormente.

5.1 Compreensão de Negócio

Neste capítulo pretende-se apresentar um caso de estudo que consiste na aplicação de um caso prático do problema apresentado no capítulo anterior.

O problema de negócio associado ao presente caso de estudo, pertence a uma loja de desporto portuguesa. Esta loja de desporto está presente em várias zonas do país sendo oficialmente a maior cadeia de lojas de desporto em Portugal e tem atualmente mais de 100 lojas na Península Ibérica. A maior loja da zona norte desta cadeia de lojas de desporto está no centro comercial em estudo no capítulo anterior.

Esta loja específica, apesar de possuir um elevado número de visitas, continua com a ambição de aumentar os lucros e para tal, os gestores juntamente com a equipa de marketing, apostam em várias iniciativas com o objetivo de reforçar a visibilidade da marca e elevar a fidelidade dos clientes com a marca para, deste modo, conseguirem os seus objetivos.

Uma das principais iniciativas é o recurso a patrocínios. O patrocínio é visto como uma operação que envolve a troca de recursos materiais ou financeiros por alguns direitos de comunicação [pat11]. Deste modo, a loja de desporto em estudo patrocina vários eventos desportivos com o intuito de promover a marca junto do público em geral. Esta loja em questão não patrocina eventos financeiramente mas sim com materiais da sua loja. [pat11].

Neste sentido, a loja de desporto aliou-se à Runporto. A Runporto é uma organizadora de eventos desportivos da cidade do Porto cuja missão passa por colocar todos os habitantes da cidade do Porto a praticar a desporto. A Runporto apresenta estruturas, de apoio a eventos, personalizáveis (como por exemplo: pórticos, insufláveis, e o próprio pódio) para os seus patrocinadores indo isto ao encontro dos objetivos da loja de desporto.

Surge então o problema desta loja de desporto. Apesar de utilizarem 38% do orçamento de marketing, desta loja, em patrocínios, o gestor da loja não consegue saber efetivamente o valor de retorno nem medir o valor do impacto. Atualmente o gestor da loja recorre aos serviços de uma empresa para resolver este problema mas os serviços obtidos não satisfazem o responsável da loja de desporto uma vez que não são úteis para perceber efetivamente, se a decisão de patrocinar um determinado evento foi acertada. [pat11].

Atualmente o processo de análise resume-se a procurar ativamente em jornais, revistas, e outros meios, onde apareceu a marca durante o evento patrocinado e então fazer uma contagem de tempo por centímetro quadrado. A grande limitação é que nada garante que a pessoa que esta no espaço do evento, ou a ver na imprensa o evento patrocinado, observa realmente a publicidade à marca [pat11].

É então proposta uma solução com este caso de estudo para melhorar este processo utilizando a ideologia apresentada no capítulo anterior.

Uma vez que há registo do número de visitantes da loja de desporto, pois o BIPS está implementado nesta loja específica, torna-se possível determinar o número de *outliers* relativo ao número de visitas e ao tempo médio de visita de cada zona da loja. Após este cálculo, é fundamental relacionar as datas anómalas e então relaciona-las com eventos desportivos que estão, como referido anteriormente, armazenados numa base de dados externa de eventos. Após este processo, verificando que as primeiras zonas visitadas são zonas a que corresponde o evento em questão, podemos concluir que o patrocínio atraiu visitantes à loja. Se o visitante efetivamente comprou, ou não, algum produto, isto já não é possível determinar, pois não há acesso ao registo de vendas da loja.

Assim considera-se que existe relação entre o evento patrocinado pela loja de desporto e as zonas deste que possuem produtos necessários para o evento em questão. Pela análise do impacto das visitas pretende-se responder a um problema identificado na literatura utilizando a solução desenvolvida neste projeto.

Procede-se de seguida à análise e compreensão dos dados obtidos desta loja de desporto.

5.2 Compreensão dos Dados

Foi extraído da API da Movvo um conjunto de dados relativos à loja de desporto que serviram de base para o desenvolvimento do presente caso de estudo. Na tabela 5.1 estão descritas as variáveis que representam esse conjunto para o período de tempo de 01-05-2014 a 14-10-2014.

Variável	Descrição
dev	Identificação do dispositivo
ts	*Timestamp* de cada deteção
x	Coordenada do eixo dos X correspondente à posição do dispositivo detetado
y	Coordenada do eixo dos Y correspondente à posição do dispositivo detetado

Tabela 5.1: Descrição do conjunto de dados de deteções da loja de desporto

Para o período de tempo mencionado, explorou-se a variável *"Dev"* e concluiu-se que cerca de 200300 pessoas visitaram a loja de desporto em estudo.

Este conjunto de dados foi sujeito a algumas operações para facilitar a tarefa de exploração e compreensão dos dados. As operações foram as seguintes:

1. Obtenção das zonas. Recorrendo à API da Movvo, determinou-se as zonas da loja de desporto através das coordenadas X e Y.

2. Remoção de todas as deteções que não representam visitantes, ou seja, todos os dispositivos com mais de 4 horas de permanência na loja (empregados da loja de desporto ou por exemplo, pistolas de marcar preços que funcionam com tecnologia *WI-FI*).

3. Processamento de métricas. Através do *timestamp*, calcula-se o tempo médio de permanência do dispositivo em cada zona e determina-se ainda o número de dispositivos distintos em cada zona.

4. Decomposição do *timestamp* para criação das variáveis data, mês e dia da semana para simplificar a manipulação dos dados.

Após estas operações obteve-se o conjunto de dados apresentado na tabela 5.2.

Variável	Descrição
zone	Valor categórico representativo das zonas existentes na loja de desporto.
avg_time	Número representativo do tempo médio de visitas por zona.
count	Número inteiro positivo representativo do número de visitas por dia.
date	Data das visitas no formato dd-mm-aaaa baseado no *timestamp*.
weekday	Valor categórico representativo do dia da semana baseado na data.
month	Valor categórico representativo do mês do ano baseado na data.

Tabela 5.2: Descrição do conjunto de dados da loja de desporto após operações

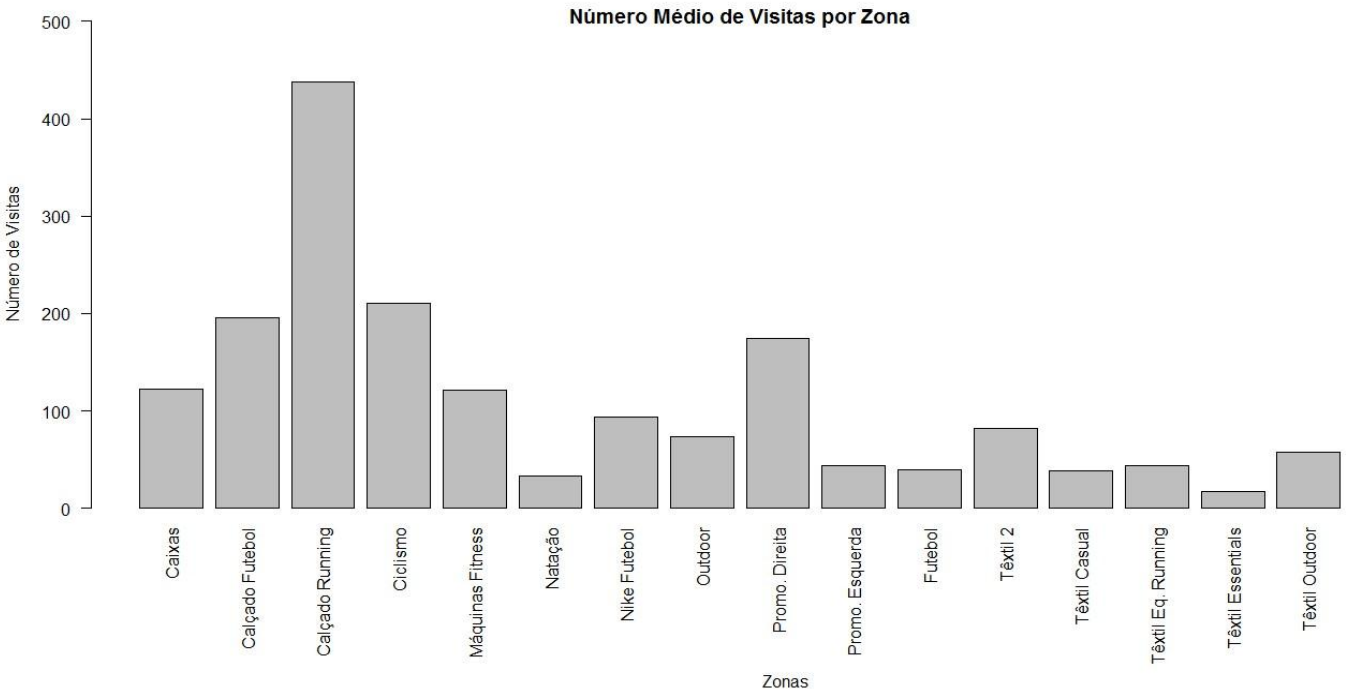
Fez-se uma exploração destes conjuntos de dados para compreender melhor as suas caraterísticas.

Relativamente à variável *zone* conclui-se que existem 16 zonas diferentes. De modo a termos informação mais relevante, determinou-se o número de visitas e o tempo médio de permanência em cada zona, por dia, e apresenta-se esta informação na tabela 5.3.

Zona	Número Médio de Visitas	Tempo Médio de Visita (min)
Têxtil Casual	438	21.28
Têxtil Equipamentos Running	196	20.47
Têxtil Outdoor	211	22.43
Têxtil Essentials	174	21.12
Têxtil 2	82	28.36
Futebol	38	47.22
Nike Futebol	121	29.38
Calçado Futebol	44	40.77
Calçado Running	122	40.77
Outdoor	73	25.48
Ciclismo	58	25.99
Natação	33	41.86
Máquinas Fitness	40	41.49
Promocional Esquerda	44	31.38
Promotional Direita	17	41.93
Caixas	94	25.45

Tabela 5.3: Número médio e tempo médio de visitas por zona

Apresenta-se nas figuras 5.1 e 5.2 a representação gráfica da informação presente na tabela anterior.



.

Figura 5.1: Média das Visitas por Zona por Dia

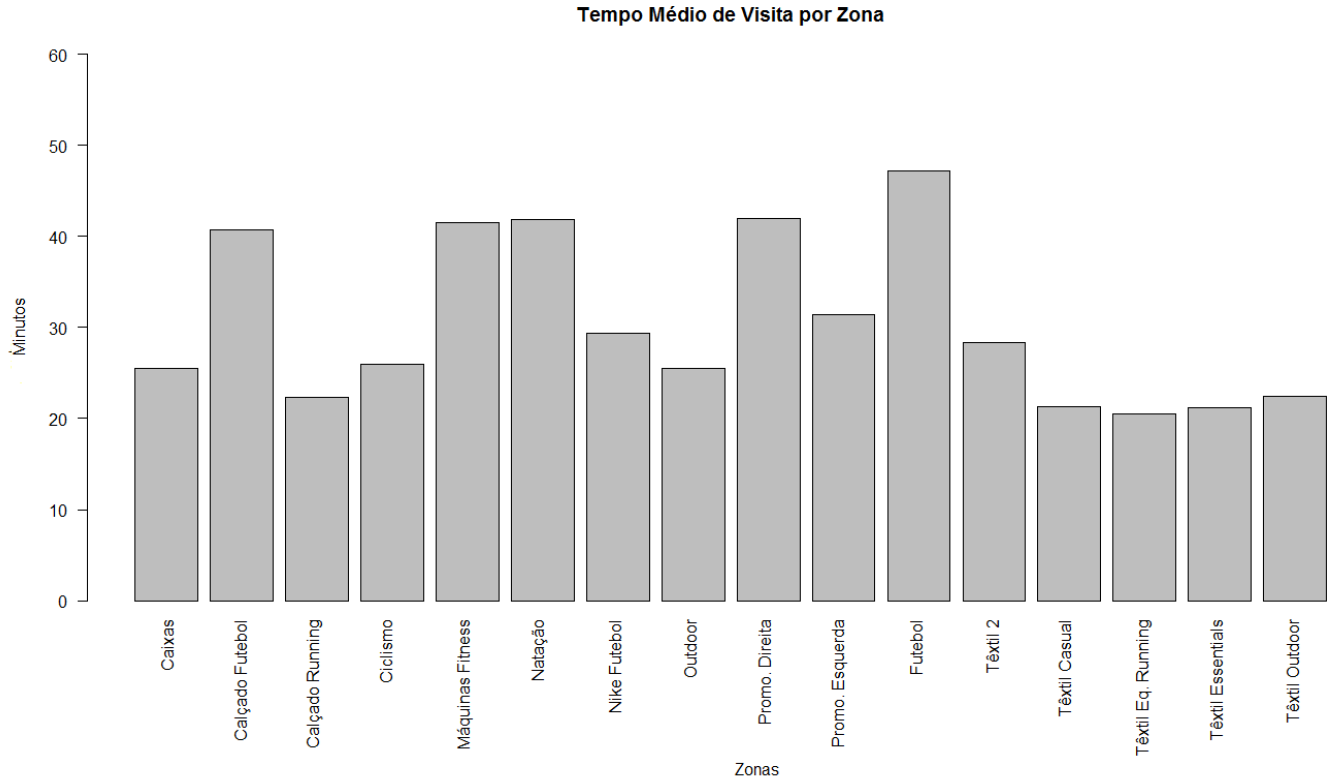


Figura 5.2: Média do Tempo de Permanência por Zona por Dia

Pode-se verificar que a zona *Têxtil Casual* é a zona que mais visitas possui. Relativamente à zona *Caixas*, esta não é uma zona com um elevado número de visitas. Esta zona corresponde ao terminal de pagamento da loja. Verifica-se que a loja possui imensas visitas mas só em média 94 pessoas por dia compram algum produto. As zonas *Promocional Esquerda* e *Promotional Direita* referem-se à frente da loja. No que respeita aos tempos médios de permanência, a zona *Futebol* é a zona onde os visitantes passam mais tempo, enquanto a zona *Têxtil Equipamentos Running* é onde passam mais tempo. Uma vez que existe informação de número de visitas e tempo médio de permanência para cada zona, decidiu-se fazer uma representação gráfica destas duas métricas num só elemento. Assim na figura 5.3 apresenta-se uma representação gráfica do número de visitas por zona aliado ao tempo ao tempo de permanência nas mesmas

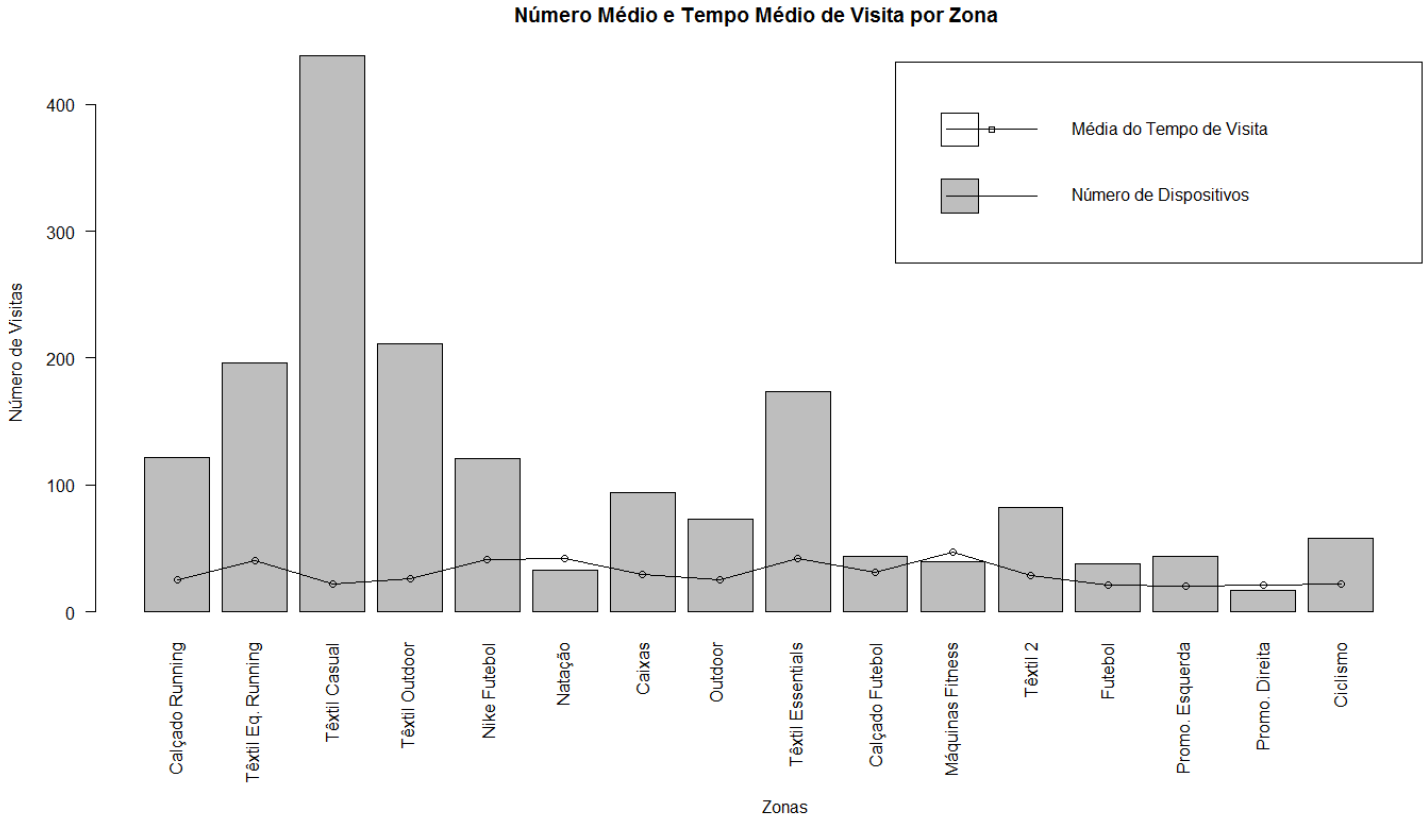


Figura 5.3: Número Médio de Visitas e Tempo Médio de Permanência por Zona por Dia

O gráfico da figura 5.3 permite retirar conclusões úteis tais como o facto de as zonas *Máquinas Fitness* e *Natação* receberem poucas visitas mas possuírem um tempo médio de visita elevado. Isto justifica-se pelo facto de serem zonas com produtos que requerem um grande período de análise antes da compra como é o caso, por exemplo, das máquinas de fitness que são produtos caros e exigem uma maior reflexão antes da compra. A zona *Têxtil Casual* possui um elevado número de visitas mas em contrapartida possui um tempo de visita bastante baixo.

Data	
Início	01-05-2014
Fim	14-10-2014

Tabela 5.4: Resumo da variável Data

Durante o período de tempo presente na tabela 5.4, todas as variáveis estão corretamente instanciadas, o que permite a exploração deste modelo de dados.

5.2.1 Relação Entre o Número de Visitas e o Tempo Médio de Visita Com o Dia da Semana

Para relacionar o número de visitas com o dia da semana, torna-se essencial descobrir a média de visitas por cada dia. Apresenta-se na tabela 5.5 e na figura 5.4 o número médio de visitas por dia da semana:

Dia da Semana	Média das Visitas
Domingo	1156
Segunda-feira	1314
Terça-feira	934
Quarta-feira	870
Quinta-feira	915
Sexta-feira	1022
Sábado	1314

Tabela 5.5: Média das visitas por dia da semana

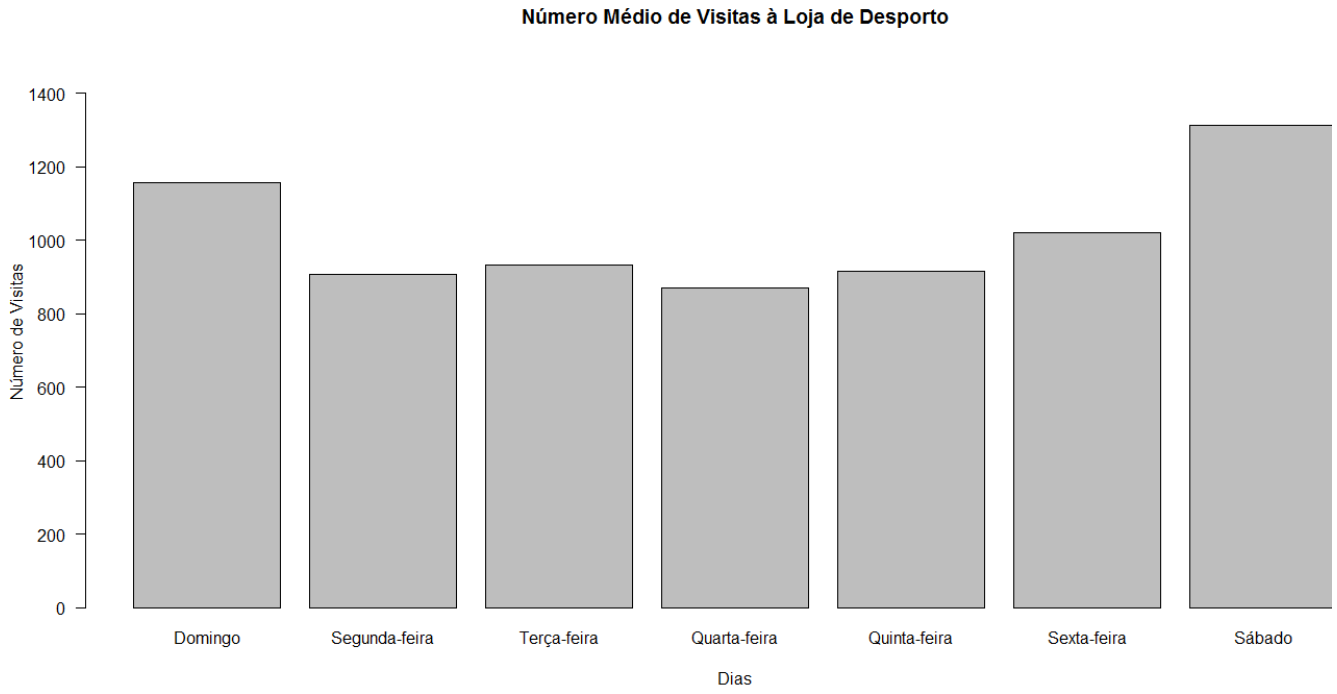


Figura 5.4: Gráfico de barras da média das visitas por dia da semanas

Através dos elementos anteriores, é possível concluir que o número de visitas difere claramente consoante o dia da semana. Estes dois elementos permitem dividir em dois grupos o número de visitas em função do dia da semana devido à proximidade dos valores:

- Grupo 1 - segunda-feira, terça-feira, quarta-feira e quinta-feira
- Grupo 2 - sexta-feira, sábado e domingo

Segue-se o mesmo procedimento para o tempo médio de visita por dia na loja e apresenta-se na tabela 5.5 e na figura 5.4 o número médio de visitas por dia da semana:

Dia da Semana	Tempo médio de permanência
Domingo	21.86
Segunda-feira	24.06
Terça-feira	25.29
Quarta-feira	25.52
Quinta-feira	23.89
Sexta-feira	23.32
Sábado	21.95

Tabela 5.6: Tempo médio de permanência por dia da semana (min.)

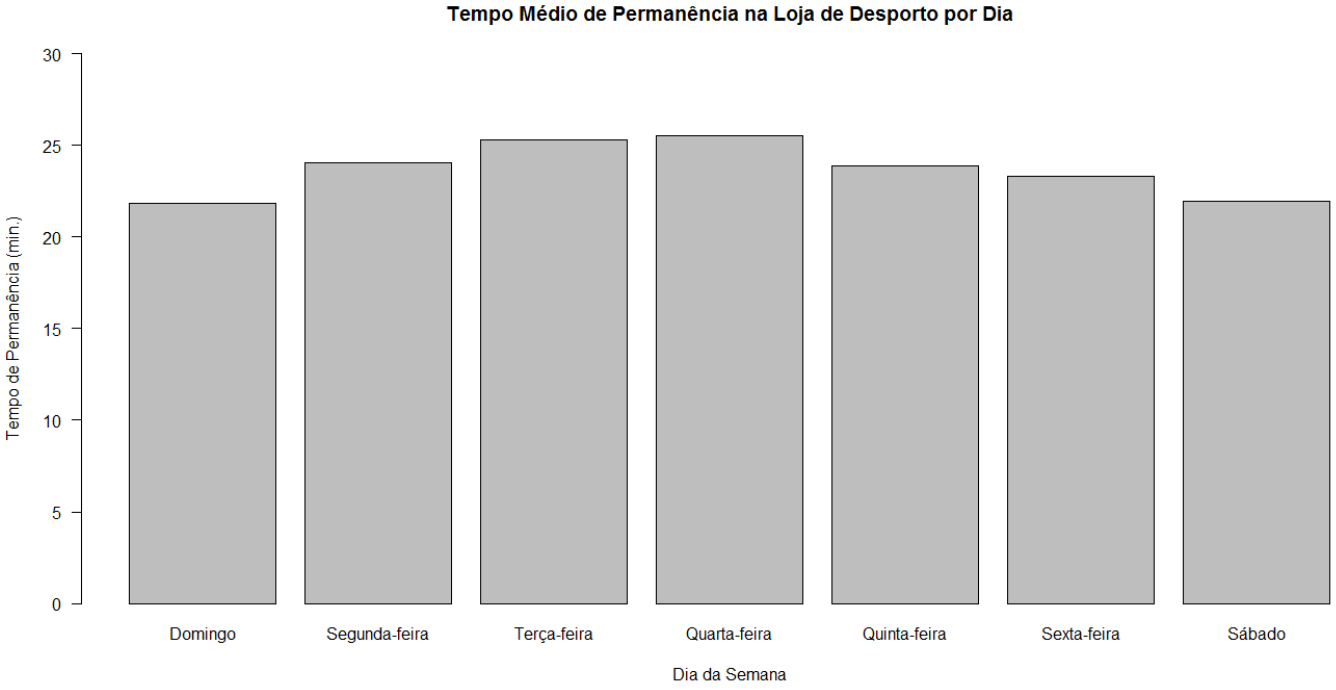


Figura 5.5: Gráfico de barras da média do tempo médio de permanência por dia da semana

5.2.2 Relação Entre o Número de Visitas Com o Mês do Ano

Para relacionar o número de visitas com o dia da semana, torna-se essencial descobrir a média de visitas por cada dia. Assim, apresenta-se na tabela 5.7 e na figura 5.6 o número médio de visitas por dia da semana:

Mês do Ano	Média de Visitas
Maio	745
Junho	839
Julho	1033
Agosto	1089
Setembro	981
Outubro	1033

Tabela 5.7: Média das visitas por mês

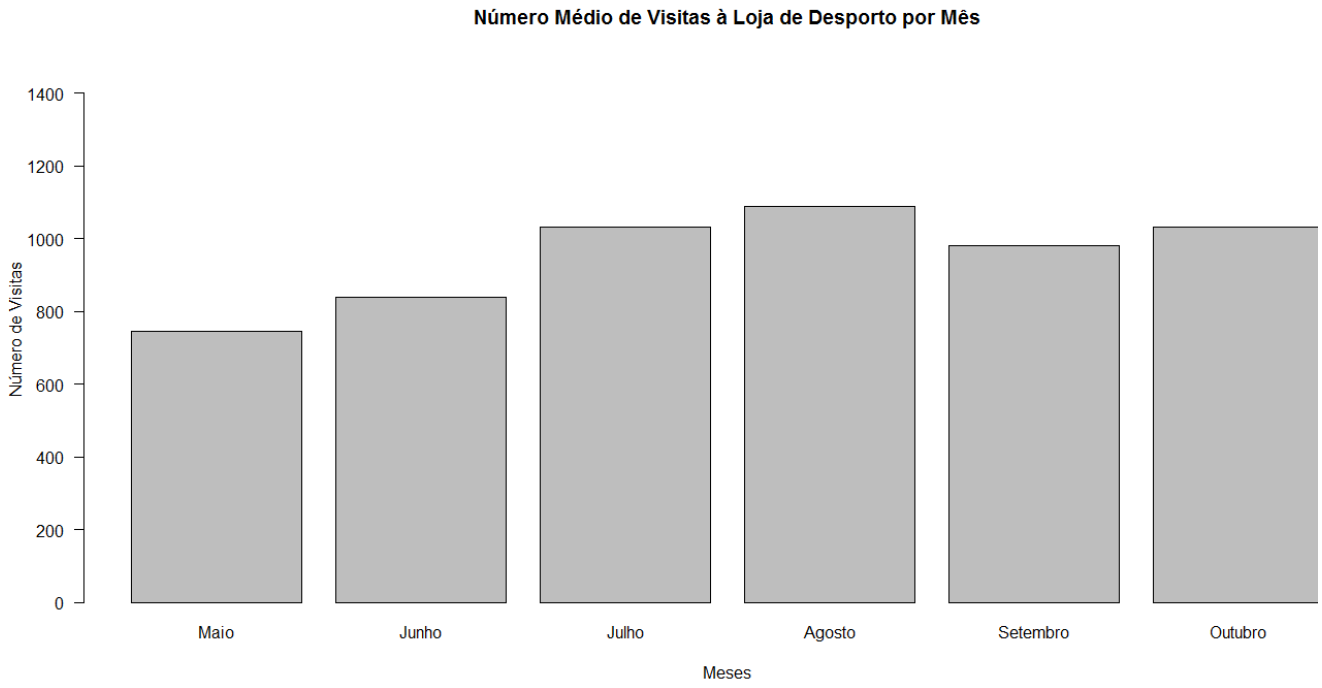


Figura 5.6: Gráfico de barras da média das visitas por mês

Através dos elementos anteriores, é possível observar que o número de visitas difere claramente consoante o dia da semana. Destaca-se o alto número de visitas durante o mês de agosto derivado às férias de verão. Não se apresenta o tempo médio de permanência por mês pois não é relevante uma vez que esta métrica varia em função dos dias e a nível do mês não é possível ter a perceção da variação dos dias.

5.3 Preparação dos Dados

Os dados iniciais utilizados foram extraídos da API para formato *json* e posteriormente interpretados pelo R e guardou-se esta mesma informação numa estrutura de dados específica do R denominada por *data frames* pelo facto de ser uma estrutura que permite uma manipulação simplificada dos dados.

Como referido na secção anterior, foram também construídos novos dados. Foram criadas as variáveis *date*, *weekday* e *months*. Estas foram criadas através da variável *"ts"* que armazena a informação temporal da deteção do dispositivo numa determinada zona. Estas foram criadas pois são mais úteis do que a própria data e tornam mais simples a manipulação e validação dos dados aliando ainda ao facto de serem valores mais percetíveis do que o *timestamp*.

A variável *zone* também foi criada com recurso às variáveis *x* e *y* que, como referido, correspondem às coordenadas do dispositivo móvel detetado. Esta transformação era essencial para ser possível relacionar um dado evento com as zonas da loja de desporto. Esta transformação passa a ignorar as variáveis *x* e *y* e só foi possível graças à implementação de uma função que interage com a API da Movvo e permite a criação das zonas. Esta variável foi armazenada como tipo *factor*.

Concluiu-se ainda que existem observações que foram consideradas irrelevantes. Como mencionado, uma das operações do desenvolvimento deste caso de estudo, foi a eliminação registos que representavam dispositivos com mais de 4 horas seguidas de deteções pois são associados a colaboradores da loja ou qualquer dispositivo existente na loja, que emita sinais *WI-FI* ou *bluetooth*, daí que não haja interesse em considerar estes registos. Após esta limpeza nos dados considera-se que o conjunto de dados torna-se mais rico pois deixa de ter registos sem importância que possam influenciar o resultado final.

5.4 Modelação

Nesta fase, pretende-se fazer uma deteção de *outliers* univariada e relaciona-los com eventos patrocinados pela loja de desporto, com o intuito de perceber se houve um impacto no número de visitas da loja ou no tempo de permanência dos visitantes nas zonas da loja. Após a análise individual aos meses todos, foram escolhidos para análise os meses de agosto e setembro sendo a justificação apresentada na secção seguinte. Assim recorrendo ao algoritmo apresentado no capítulo anterior, calcularam-se os *outliers* para o mês *pivot* setembro e os resultados obtidos foram os seguintes:

- 06-09-2014
- 13-09-2014
- 14-09-2014
- 20-09-2014
- 27-09-2014

- 28-09-2014

Foi feito um pedido ao servidor para saber os eventos associados à loja de desporto que decorreram perto destas datas e os resultados apresentam-se na tabela 5.8.

Data	Evento
14-09-2014	Meia Maratona do Porto
21-09-2014	Corrida do Homem e da Mulher
27-09-2014	The Color Run

Tabela 5.8: Tabela de eventos próximos dos dias *outliers*

Numa segunda fase analisou-se o número total de visitas à loja durante os meses de agosto e setembro. Assim apresenta-se na figura 5.7 o número de visitantes à loja.

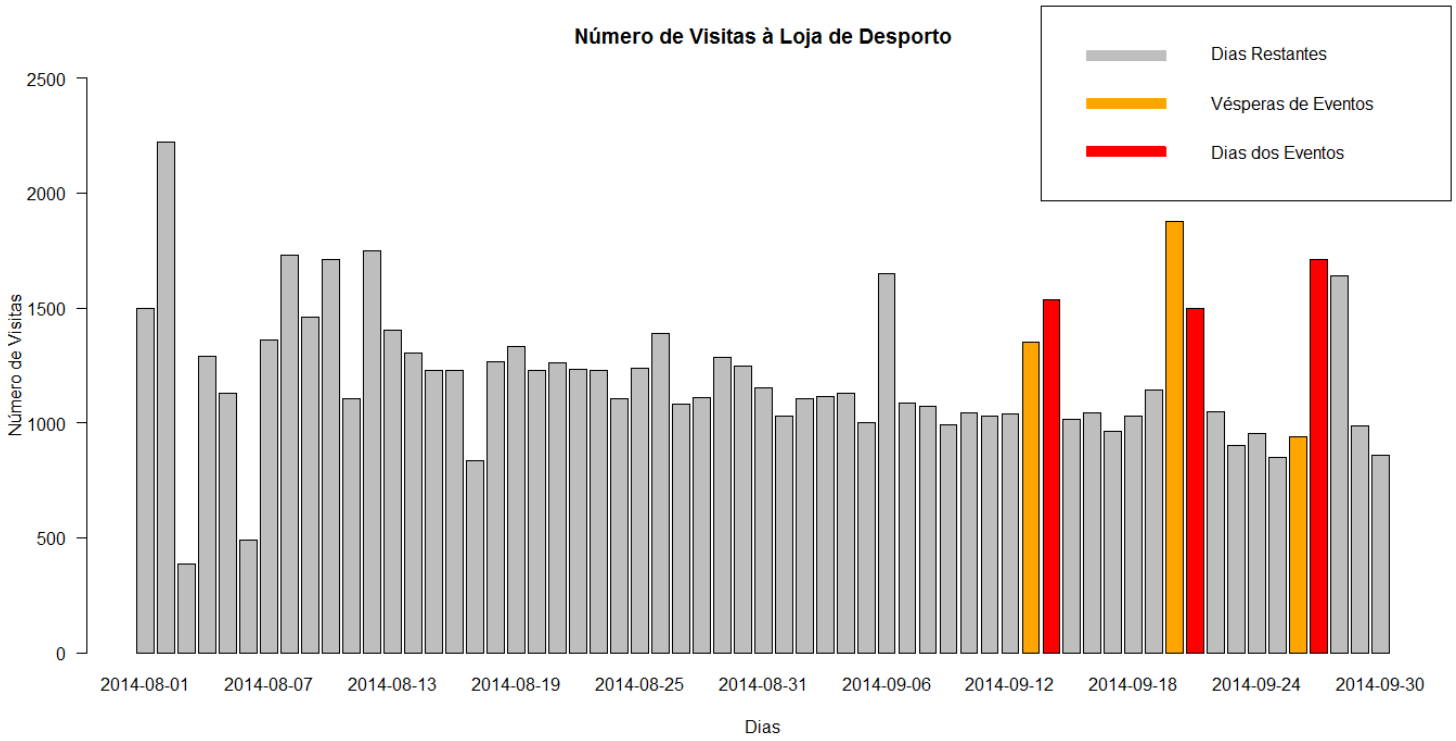


Figura 5.7: Total de visitas à loja de desporto ao longo dos meses

É possível verificar na figura 5.7 os dias correspondentes aos eventos da tabela 5.8 a cor vermelha e coloriu-se ainda a cor laranja a véspera dos eventos para ser possível a comparação relativa ao número de visitas nos dois dias. Uma vez que os eventos encontrados a apresentados na tabela 5.8 são eventos de atletismo, encontra-se portanto a necessidade de relacionar estes eventos com as zonas existentes na loja com equipamentos de corrida. As zonas são *Calçado Running* e *Têxtil Equipamentos Running* pelo que se irá apenas considerar estas duas zonas nas próximas análises.

Foi necessário estudar o número de visitas e o tempo de visita nestas duas zonas e apresentam-se nas figuras 5.8 e 5.9 os resultados.

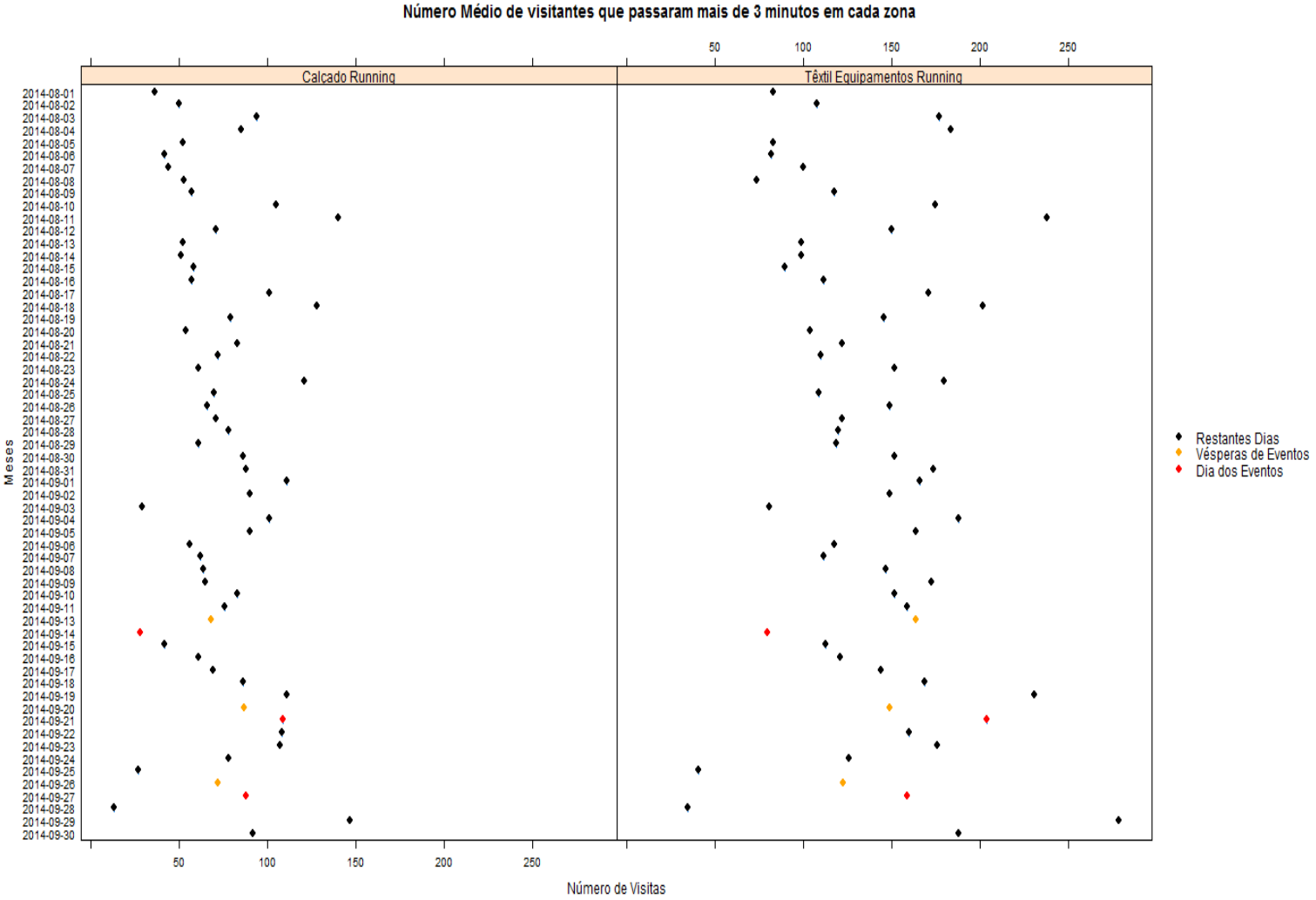


Figura 5.8: Análise do Número Médio de Visitas por Dia nas Zonas

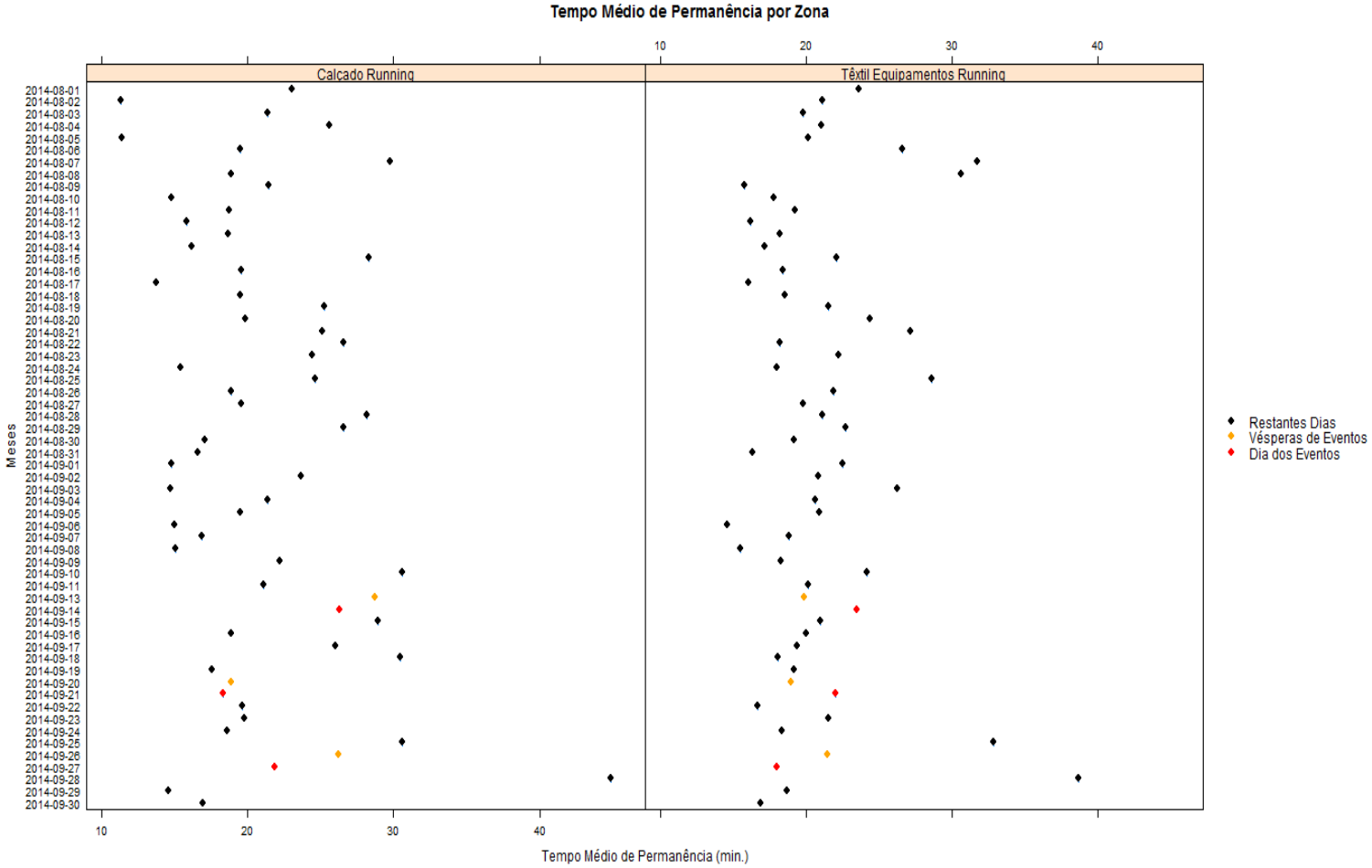


Figura 5.9: Análise do Tempo Médio de Permanência por Dia nas Zonas

Feita a análise ao número médio de visitas e ao tempo médio de permanência nas zonas, verifica-se nas figuras 5.8 e 5.9 estas métricas ao longo do tempo. Após esta análise às zonas, decidiu-se aprofundar o estudo e fez-se o estudo das primeiras zonas visitadas pelos clientes. Partiu-se do pressuposto que se um visitante for à loja de desporto influenciado por algum evento patrocinado, então ele iria de imediato às zonas com produtos adequados para esses eventos, à procura de algum produto desejado. Assim calculou-se a percentagem de clientes que visita as zonas *Calçado Running* e *Têxtil Equipamentos Running* em primeiro lugar durante os meses de agosto e setembro. Ilustra-se na figura 5.10 a percentagem de primeiras visitas à zona *Calçado Running* e na figura 5.11 a percentagem de primeiras visitas à zona *Têxtil Equipamentos Running*.

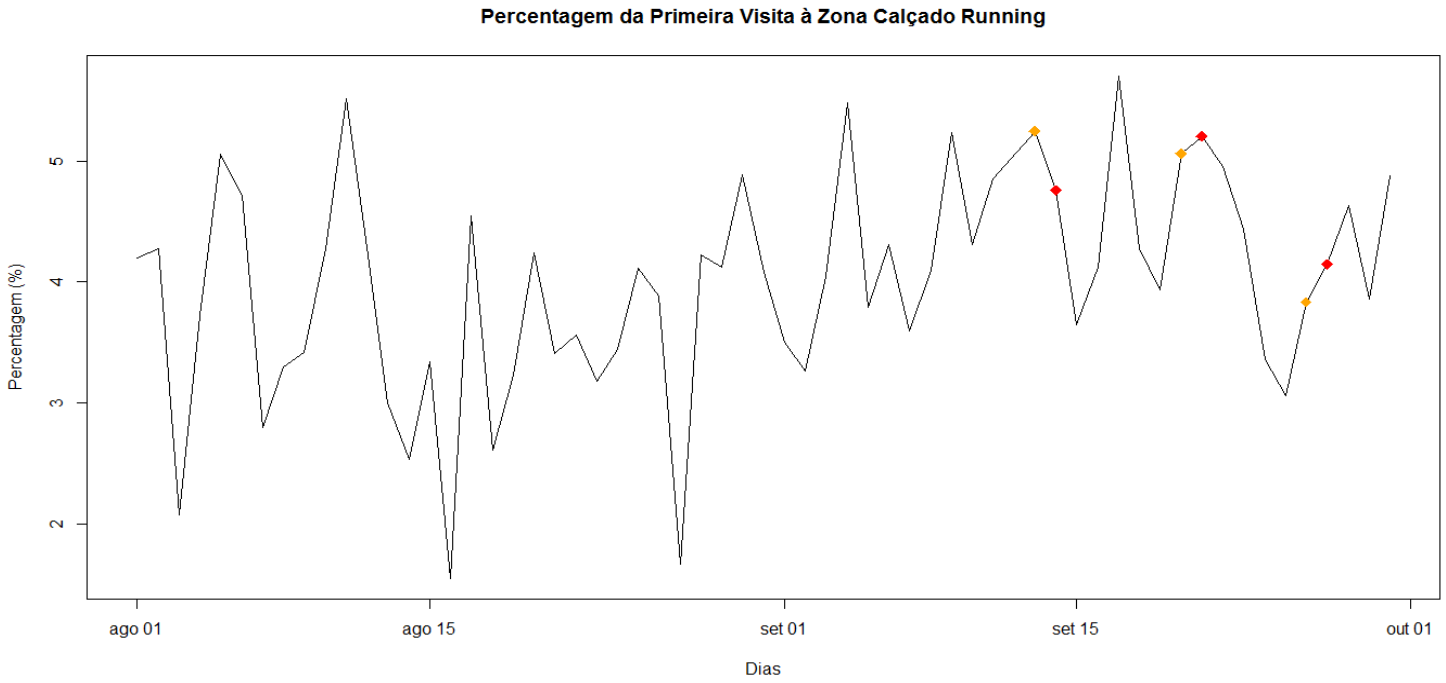


Figura 5.10: Análise das Primeiras Visitas à Zona Calçado Running

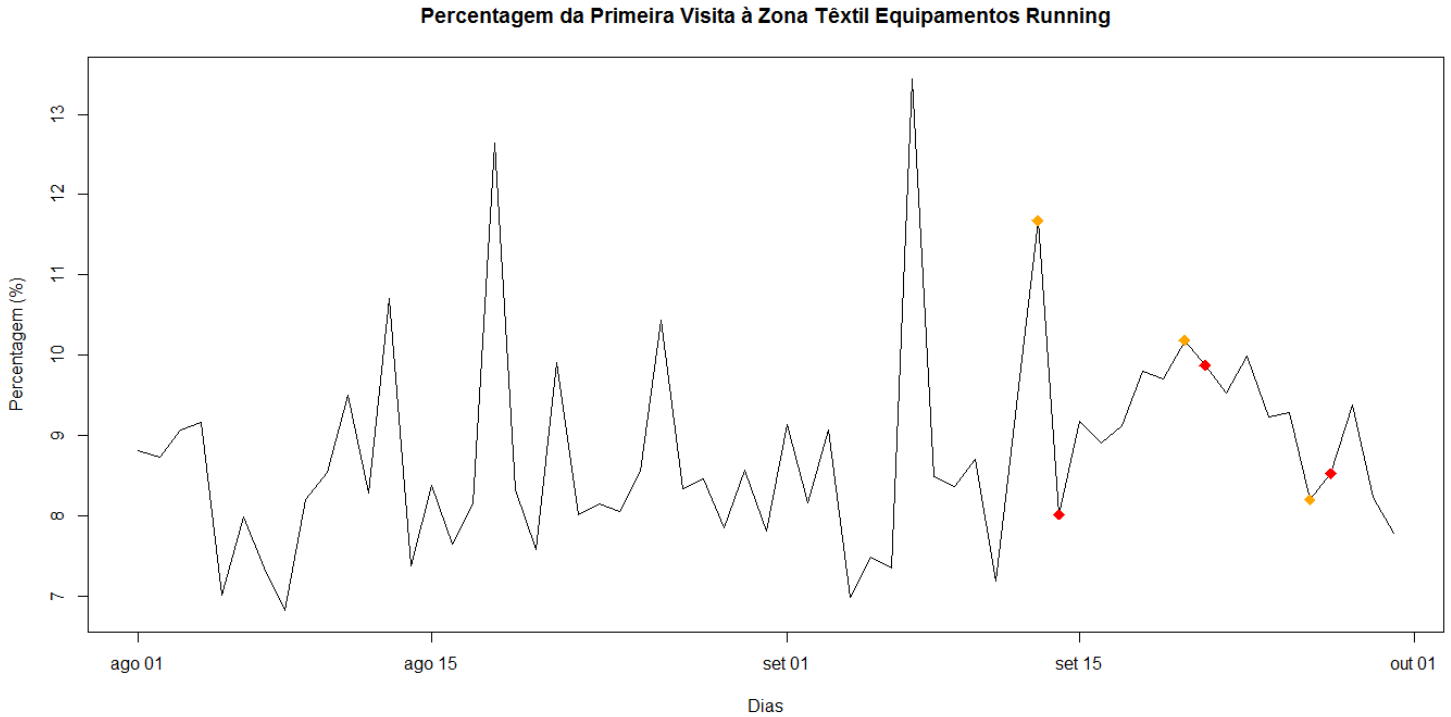


Figura 5.11: Análise das Primeiras Visitas à Zona Têxtil Equipamento Running

É possível verificar nas figuras anteriores os dias correspondentes aos eventos da tabela 5.8 a cor vermelha e a cor laranja as véspera dos eventos, para ser possível a comparação com os restantes dias.

Após este estudo segue-se na secção seguinte a avaliação do mesmo.

5.5 Avaliação

Em primeiro lugar considera-se essencial justificar a escolha dos meses de agosto e setembro.

Realizou-se uma pesquisa de eventos patrocinados pela loja de desporto para o período de tempo especificado anteriormente e não foram encontrados eventos patrocinados pela loja de desporto durante o mês de agosto. Optou-se então por realizar uma análise deste mês e também do mês de setembro devido ao facto de ser um mês com vários eventos. Analisaram-se os restantes meses mas nenhum apresentou resultados relevantes. Estes dois meses também não apresentaram resultados relevantes contudo foi considerado relevante para esta dissertação comparar dois meses consecutivos - tendo um vários eventos e outro sem qualquer evento - sendo assim possível estudar as diferenças existentes nesta serie temporal constituída por dois meses consecutivos.

Relativamente à figura 5.7 é possível verificar um valor bastante alto de visitas no dia 20- 09- 2014. Este dia é a véspera do dia em que ocorreu o evento *The Color Run*. A partir desta informação é possível tirar a conclusão que este evento aparentemente teve impacto no número de visitas à loja de desporto. Para comprovar esta situação, fez-se uma análise às zonas *Calçado Running* e *Têxtil Equipamentos Running*.

Esta análise demonstrou que a conclusão retirada não é totalmente verdade visto que nas figuras 5.8 e 5.9 é possível concluir que este dia não apresenta valores relevantes nem para o número médio de visitas nem para o tempo médio de permanência dos visitantes nas duas zonas. O mesmo se pode concluir para os restantes eventos detetados.

Para reforçar a conclsão extraída das figuras referidas, determinou-se a média e o desvio padrão das métricas mencionadas. Assim apresenta-se na tabela 5.9 a média e o desvio padrão do número médio de visitas às duas zonas e na tabela 5.10 as mesmas medidas estatísticas para a média do tempo médio de permanência nas duas zonas. Apresenta-se também na tabela 5.11 as médias das visitas e dos tempos de permanência para ambas as zonas dos dias dos eventos e das respetivas vésperas.

	Calçado Running	**Têxtil Equipamento Running**
Média	75	140
Desvio Padrão	28	46

Tabela 5.9: Tabela Estatística do Número Médio de Visitas às Zonas

	Calçado Running	Têxtil Equipamento Running
Média	21,29	21.05
Desvio Padrão	5,91	4,48

Tabela 5.10: Tabela Estatística do Tempo Médio de Permanência nas Zonas (min.)

Datas de Eventos e Vésperas	Número de Visitas		Tempo de Permanência (min.)	
	Calçado Running	Têxtil Equipamento Running	Calçado Running	Têxtil Equipamento Running
13-09-2014	68	164	28.7	19.86
14-09-2014	28	80	26.3	24.46
20-09-2014	87	149	18.9	18.9
21-09-2014	109	204	18.3	22.04
26-09-2014	72	123	26.2	21.5
27-09-2014	88	159	21.8	18.01

Tabela 5.11: Média do Número de Visitas e do Tempo Médio de Permanência nas Zonas em Dias de Eventos e Respetivas Vésperas

Com a informação apresentada nas tabelas 5.9 e 5.10, reforça-se a conclusão que os eventos não tiveram impacto no número de visitas nem no tempo de permancência nas zonas visto que estas médias são próximas das médias dos números de visitas e dos tempos de permanência, dos dias dos eventos e das respetivas vésperas, apresentadas na tabela 5.11. Os desvios padrão ajudam a comprar esta conclusão uma vez que é um valor alto o que ajuda a comprovar que os valores são bastante dispersos. A média do número de visitas do dia 21-09-2014 são superiores mas este é o dia do evento pelo que não se considera relevante.

Após concluir que os eventos patrocinados não tiveram um impacto significativo no número médio de visitas nem no tempo médio de permanência, decidiu-se estudar a seguinte abordagem: Se um visitante for à loja de desporto influenciado por um evento, então à pessoa irá à zona que contém os produtos desejados para utilizar no evento. Seguindo este raciocínio fez-se uma análise à percentagem das primeiras zonas visitadas pelos clientes. Assim nas figuras 5.10 e 5.11 é possível concluir que aparentemente houve um impacto no número de visitas na véspera do evento "Meia Maratona do Porto"nas duas zonas. Contudo, comparando com os restantes dias do intervalo de tempo, conclui-se que não é uma situação isolada pelo que não é possível assegurar que este evento tenha tido um impacto significativo na loja de desporto.

Reforçou-se esta conclusão com as médias e os desvios padrão das primeiras visitas em ambas as zonas para cada mês em ambas as zonas sendo esta informação apresentada na tabela 5.13.

Comparou-se estes elementos com as médias das primeiras visitas dos dias dos eventos e respetivas vésperas e apresenta-se estes dados na tabela 5.12.

Datas de Eventos e Vésperas	Calçado Running	Têxtil Equipamento Running
13-09-2014	5.3%	11.7%
14-09-2014	4.7%	8%
20-09-2014	5.1%	10.2%
21-09-2014	5.2%	10%
26-09-2014	3.8%	8.2%
27-09-2014	4.1%	8.5%

Tabela 5.12: Percentagem da Primeira Visita às Zonas em Dias de Eventos e Vésperas

Meses	Calçado Running	Têxtil Equipamento
Agosto	3.68%	8.53%
Setembro	4.38%	9.02%

Tabela 5.13: Percentagem da Média das Primeiras Visitas às Zonas

Comparando com as médias dos meses com as médias dos dias dos eventos, e vésperas, conclui-se novamente que as diferenças não são muito significativas. Para a zona *Calçado Running* as médias das primeiras visitas para os dias 13, 14, 20 e 21 de setembro de 2014 são efetivamente superiores às médias do mês mas a diferença não é considerada muito relevante para concluir que houve impacto. Para a zona *Têxtil Equipamentos Running* houve uma diferença no dia 13-09-2014 de quase 3% em relação à média do mês mas não foi considerado um dado suficientemente forte para justificar que o evento teve impacto no número de visitas nesta zona visto que também há outros dias com percentagens de primeiras visitas tão ou mais altas como ilustrado na figura 5.11.

Conclui-se que as médias e os desvios padrão não estão muito espaçados pelo que se demonstra que não houve impacto nas primeiras visitas às zonas da loja.

Após este estudo conclui-se assim que os eventos patrocinados não têm um impacto significativo no número de visitas à loja de desporto. Conclui-se que não há um grande impacto no número de visitas nem no tempo médio de permanência nas duas zonas em análise.

5.6 Desenvolvimento

Como já mencionado, o presente capítulo refere-se a um caso de estudo. Utilizou-se a ferramenta desenvolvida no capítulo anterior para resolver um problema concreto existente da literatura. O trabalho deste capítulo serve para demonstrar o seu valor num caso específico e não será integrado desta forma no *dashboard* produzido pelo BIPS. Esta aplicação desenvolvida recorre numa primeira fase ao algoritmo desenvolvido no capítulo anterior para calcular os dias *outliers* da loja de desporto. No passo seguinte do algoritmo implementado no presente capítulo, são extraídos da base de dados e, após processamento, armazenados num *data frame*, todos os eventos desportivos existentes na base de dados de eventos para os dias calculados no passo anterior. Por fim, tenta-se determinar o impacto do evento desportivo determinando a percentagem de pessoas cuja primeira zona visita na loja foram as zonas associadas aos eventos desportivos extraídos. Com este estudo provou-se que é possível criar uma aplicação automatizada, cujo fluxo permita estudar o impacto de eventos a vários níveis, entre os quais o estudo à primeira zona visitada, algo que até agora não existia na empresa. Era impossível também relacionar qualquer métrica com eventos uma vez que não existia também base de dados de eventos. Para realizar este processo seria necessário fazer uma pesquisa manual pelo dia pretendido e então relacionar com as métricas pretendidas.

Capítulo 6

Conclusões e Trabalho Futuro

Neste capítulo são apresentadas as considerações finais, as satisfações dos objetivos, os possíveis desenvolvimentos futuros e as lições retiradas a partir da reflexão sobre todo o trabalho desenvolvido.

6.1 Conclusões Finais

Atualmente os gestores de grandes superfícies comerciais possuem uma grande necessidade de compreensão dos comportamentos dos visitantes. Possuir conhecimento sobre necessidades e comportamentos dos clientes pode trazer vantagem competitiva sobre os concorrentes. Para tal surge a necessidade de se recorrer a ferramentas tecnológicas e é neste contexto que surge o BIPS - tecnologia da empresa Movvo – que permite detetar sinais de radiofrequência de dispositivos móveis e permite seguir esse sinal ao longo do tempo. Com base nesta tecnologia é possível uma compreensão acerca do comportamento dos visitantes num dado espaço interior, através de um relatório produzido denominado por Retail Movves. Contudo esta tecnologia ainda apresenta algumas limitações, nomeadamente no que toca ao facto de ainda não ser capaz de detetar dias atípicos de forma automática. É necessário um esforço manual do gestor do espaço comercial para realizar esta tarefa.

Concluiu-se que seria possível melhorar o BIPS para ultrapassar o problema mencionado utilizando técnicas de *data mining* para deteção de dias *outliers*. Para além da deteção de dias atípicos, definiu-se também o objetivo de relacionar estes dias com eventos externos para tentar justificar os dias *outliers* com eventos, devido ao facto de alguns eventos, terem impacto no número de visitas de um espaço comercial. O BIPS serviu de auxílio para a resolução deste problema identificado sendo que a solução desta dissertação será implementada na próxima versão da tecnologia.

Desenvolveu-se uma ferramenta capaz de detetar *outliers* num centro comercial português, localizado no norte do país, que atrai diariamente um número bastante elevado de visitantes. Apesar do gestor deste espaço comercial possuir atualmente um conhecimento empírico sobre o número de visitas ao centro comercial através do Retail Movves, necessita de fazer uma avaliação baseada na observação dos valores das visitas e concluir se um determinado dia está fora do padrão considerado normal, e precisará posteriormente, de novo esforço manual para compreender o que aconteceu nesse dia para ter existido esse número de visitas atípico.

Para resolver este problema, desenvolveu-se um algoritmo que faz uma deteção de *outliers* univariada para o número de visitas num centro comercial. Concluiu-se que o algoritmo desenvolvido para o cálculo dos dias *outliers* está dependente de um conjunto de dados ideal que foi determinado a partir do estudo de várias abordagens. Implementou-se também uma base de dados com eventos com o intuito de apresentar possíveis justificações para os dias atípicos.

Utilizou-se esta solução num caso de estudo de uma loja de desporto com o objetivo de avaliar o impacto de eventos patrocinados pela loja de desporto no retorno do número de visitas. Analisou-se o número de visitas à loja, o número de visitas e tempo de permanência nas zonas de atletismo (os eventos patrocinados são maioritariamente eventos de atletismo) e ainda a percentagem da primeira zona visitada pelos clientes. Concluiu-se que os eventos encontrados para os dias determinados como *outliers*, pelo algoritmo desenvolvido, não tiveram um impacto significativo no número de visitas à loja e mais concretamente às zonas da loja com artigos de atletismo.

6.2 Satisfação dos Objetivos

Após o trabalho desenvolvido pode-se concluir que o principal objetivo foi concluído. Era necessário um algoritmo capaz de calcular dias de visitas anómalos para que os gestores das superfícies comerciais não necessitassem de fazer esta análise manualmente. Este problema detetado foi resolvido com sucesso pelo que os gestores terão uma ferramenta que os ajudará ainda mais nas suas tomadas de decisão. Adicionalmente implementou-se uma base de dados com vários tipos de eventos, para que o gestor da superfície comercial tivesse uma pequena noção do que poderia justificar esses valores anómalos. Desta forma não será necessário pesquisar, para todos os dias atípicos, razões que possam ter influenciado esse número de visitas ao espaço comercial. Foi possível ainda elaborar um caso de estudo, no qual se utiliza este sistema para resolver um problema encontrado na literatura; a impossibilidade de uma loja de desporto conseguir determinar, se um evento patrocinado, teve ou não algum impacto na loja no que toca a visitas. Este caso de estudo permite responder ao problema encontrado e permite ainda demonstrar que é possível criar uma aplicação automatizada cujo fluxo permita estudar o impacto de eventos a vários níveis entre os quais, o estudo à primeira zona visitada, algo que até agora não existia na empresa. Era impossível também relacionar qualquer métrica com eventos externos uma vez que não existia também base de dados de eventos.

Para realizar este processo seria necessário fazer uma pesquisa manual pelo dia pretendido e então relacionar com as métricas pretendidas. Qualquer colaborador da empresa poderá aceder a esta base de dados de eventos.

O resultado desta dissertação facilitará a vida dos gestores pois terão ao seu dispor uma ferramenta de deteção de *outliers* automática e que apresenta possíveis justificações para a existência desses dias anómalos. Este processo era feito anteriormente de forma manual e nem sempre era tarefa fácil pois seria necessário ter sempre presente o padrão normal do histórico das visitas e só assim poderiam avaliar um dia como anómalo ou normal. Agora estão mais aptos para tomar decisões estratégicas para o negócio pois através das situações anómalas calculadas automaticamente conseguem preparar-se melhor para caso voltem a ocorrer no futuro. Contudo está presente a noção de que é possível melhorar o trabalho desenvolvido e de diversas formas.

6.3 Trabalho Futuro

O estado atual do projeto é passível de vários melhoramentos. Relativamente ao trabalho apresentado no capítulo 4, existe uma possível melhoria clara que é a inclusão do período homólogo na deteção de *outliers*. A empresa pretende implementar este algoritmo na próxima versão do produto pelo que decidiu-se, no início do presente projeto, trabalhar com os dados apresentados. A razão pela qual se avançou com o projeto tendo por base esta decisão, foi devido ao facto de haver bastantes clientes para os quais não existem dados do período homólogo e pretende-se que este algoritmo esteja implementado para estes clientes na próxima versão do produto de modo a haver uma solução disponível. Existindo dados do período homólogo torna-se necessário melhorar este algoritmo. Apesar disto, o resultado desta dissertação torna-se importante, pois a entrada de novos clientes para a Movvo, vai exigir um algoritmo que detete *outliers* sem período homólogo. Incluir mais variáveis no processo de cálculo de *outliers*, como por exemplo, clima ou temperatura, é outra melhoria possível e que pode levar a uma aplicação final mais eficiente.

Relativamente ao trabalho apresentado no capítulo 5, considera-se também que há espaço para melhorias. Apesar de se ter concluído que os eventos desportivos não têm impacto no número de visitas nem no tempo de permanência nas zonas, nem se ter detetado impacto nas primeiras zonas visitadas, é possível ainda analisar se as pessoas que visitaram em primeiro lugar as zonas *Calçado Running* e *Têxtil Equipamentos Running* foram posteriormente à caixa e permaneceram lá tempo suficiente para se concluir que realizaram alguma compra. Apesar não ser possível afirmar com exatidão que as pessoas efetivamente compraram, esta é uma heurística que permite uma aproximação dos valores reais. Importa realçar que no futuro, um evento patrocinado, poderá vir a ter impacto nestas zonas ou até mesmo noutras zonas da loja. Concluindo que houve um aumento na percentagem de pessoas que visitaram em primeiro lugar uma destas zonas e foram depois à caixa, é possível também estudar se houve um aumento de compras não planeadas. Este estudo é possível determinando as zonas que os clientes visitaram posteriormente e nas quais permaneceram mais de 3 minutos. São análises bastante úteis e que podem resultar num sistema de avaliação de impacto de eventos poderosa para os gestores da loja de desporto em questão.

Referências

[Arb11] Investigation of gps observations for indoor gps/ins integration, 2011. [CA13]

Su Chen e Ibrahim Ahmad. Nonparametric anova using kernel methods. P.ProQuest Dissertations Theses Global, 2013.

[CB14] Alex Couture-Beil. *rjson: JSON for R*, 2014. R package version 0.2.14. URL: http://CRAN.R-project.org/package=rjson.

[CBK09] Varun Chandola, Arindam Banerjee e Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3):15:1–15:58, July 2009. URL: http://doi.acm.org/10.1145/1541880.1541882, doi:10.1145/1541880.1541882.

[CCK+00] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer e R. Wirth. CRISP-DM 1.0: Step-by-step data mining guide. www.crisp-dm.org, 2000.

[CHS+98] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees e Alessandro Zanasi. *Dis- covering Data Mining: From Concept to Implementation*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1998.

[cJK11] In chul Jung e Young S. Kwon. Grocery customer behavior analysis using rfid-based shopping paths data, 2011.

[DC11] Yinaze Herve Dovoedo e Subhabrata Chakraborti. Contributions to outlier detection methods: Some theory and applications. P.ProQuest Dissertations Theses Global, 2011.

[DCC13] R.U. Di Cera Colazingari. Tagless radio frequency based self correcting distributed real time location system, September 5 2013. US Patent App. 13/820,433. URL: http://www.google.com/patents/US20130231131.

[FPsS96] Usama Fayyad, Gregory Piatetsky-shapiro e Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[GI91] Kenneth M. Goldberg e Boris Iglewicz. Bivariate extensions of the boxplot and distribution-free quartile-based tests. P.ProQuest Dissertations Theses Global, 1991.

[GS13] Jamie H. Gleason e Shlomo Sawilowsky. "comparative power of the anova, approximate randomization anova, and kruskal-wallis test. P.ProQuest Dissertations Theses Global, 2013.

[Haw80] Douglas M Hawkins. *Identification of outliers*, volume 11. Springer, 1980.

[iso05] ISO27001: Information Security Management System (ISMS) standard. *Online: http://www.27000.org/iso-27001.htm*, October 2005.

[Joh97] G. H. John. Enhancements to the data mining process, 1997.

[KN98] Edwin M. Knorr e Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. pages 392–403, 1998.

[Kno02] Edwin M Knorr. *Outliers and data mining: finding exceptions in data*. PhD thesis, The University of British Columbia, 2002.

[LJK00] Jorma Laurikkala, Martti Juhola e Erna Kentala. Informal identification of outliers in medical data, 2000.

[LL98] Charles X. Ling e Chenghui Li. Data Mining for Direct Marketing: Problems and So- lutions. In *Knowledge Discovery and Data Mining*, pages 73–79, 1998. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.4902.

[Mil02] Gabriel Sperandio Milan. A estratégia de retenção de clientes eo estabelecimento de relacionamentos como vantagem competitiva: um plano de ações aplicado a uma empresa de medicina de grupo. 2002.

[Ord96] Keith Ord. Outliers in statistical data: V. barnett and t. lewis, 1994, 3rd edition, (john wiley & sons, chichester), 584 pp., [uk pound]55.00, isbn 0-471-93094-6. *International Journal of Forecasting*, 12(1):175–176, 1996. URL: http://EconPapers.repec.org/RePEc:eee:intfor:v:12:y:1996:i:1:p:175-176.

[pat11] Patrocínio desportivo- estudo de caso da sport zone. Faculdade de Desporto da Univer- sidade do Porto FADEUP, 2011. URL: http://www.dart-europe.eu/full. php?id=793887.

[Pei06] Y. Pei. Discovering and ranking outliers in very large datasets. pages 112–112, 2006.

[PT01] Federal Information Processing e Announcing The. Announcing the advanced encryption standard (aes), 2001.

[Ric06] J. Rice. *Mathematical Statistics and Data Analysis*. Number p. 3 in Advanced series. Cengage Learning, 2006. URL: http://books.google.pt/books?id=EKA-yeX2GVgC.

[Rog10] J. P. Rogers. Detection of outliers in spatial-temporal data. 2010.

[Sch02] S. L. Schertel. Data mining and its potential use in textiles: A spinning mill, 2002.

[She07] Y. Shen. *A Formal Ontology for Data Mining: Principles, Design, and Evolution*. Canadian theses. Library and Archives Canada = Bibliothèque et Archives Canada, 2007. URL: http://books.google.pt/books?id=KlH8GVEcQ7cC.

[SJDG11] Ying Sun e Marc G. Hart Jeffrey D. Genton. Inference and visualization of periodic sequences. P.ProQuest Dissertations Theses Global, 2011.

[top] Top languages for analytics, data mining, data science. http://www.kdnuggets.com/2013/08/languages-for-analytics-data-mining-data-science.html. Acedido em 2014-03.

[weba] Página oficial do euclid. http://euclidanalytics.com/. Acedido em 2015-01.

[webb] Página oficial do footfall. http://www.footfall.pt/. Acedido em 2015-01.

[webc] Página oficial do json. http://www.json.com/. Acedido em 2015-01.

[webd] Página oficial do nomi. http://nomi.com/. Acedido em 2015-01.

[webe] Página oficial do path intelligence. http://www.pathintelligence.com/. Acedido em 2015-01.

[webf] Página oficial do peco. http://www.visual-tools.com/en. Acedido em 2015-01.

[webg] Página oficial do r. http://www.r-project.org/. Acedido em 2015-01.

[webh] Página oficial do rstudio. http://www.rstudio.com/. Acedido em 2015-01.

[Whi92] R. A. White. The detection and testing of multivariate outliers. 1992.