FACULDADE
DE ECONOMIA
UNIVERSIDADE
DO PORTO

Mapping the Cypriot Housing Market: Trends, Drivers and Future Forecasts

Marianna Konstantinou

Dissertation
Master's in Modelling, Data Analysis and Decision Support Systems

Supervised by:
PhD Bruno Miguel Delindro Veloso (FEP)

Co-supervised by:
PhD Maria Eduarda da Rocha Pinto Augusto da Silva (FEP)

2026

Abstract

Real estate is one of the largest investment markets globally, contributing significantly to economic growth, wealth creation and financial stability. Residential property prices affect households, businesses, and policymakers, making accurate market analysis and prediction essential for informed decision-making.

However, the residential property market of the Republic of Cyprus remains underexplored. Existing studies often treat the housing market as homogeneous, even though different property types are influenced by different pricing mechanisms. This study addresses this gap through data-driven analysis of the Cypriot residential real estate market, aiming to support real estate companies, policymakers, and the public in decision-making.

A dataset of over 21,000 residential property listings was constructed through web scraping. A location desirability index was developed using open geospatial data; properties were segmented using K-means clustering, and Decision Tree, Random Forest (RF), Extreme Gradient Boosting (XGBoost) and Gradient Boosting (GB) models were applied within each cluster. At the macroeconomic level, the Cyprus House Price Index (HPI) was analysed using econometric and time-series forecasting models.

Results revealed different price drivers across residential segments, with the RF–XGBoost ensemble explaining up to 35.9% of price variation for apartments, RF–GB explaining 34.6% for mid-size houses, and RF explaining 65.9% for luxury/large properties. For HPI forecasting, Autoregressive Integrated Moving Average (ARIMA)(1,1,2) achieved the highest accuracy (Mean Absolute Percentage Error (MAPE) = 1.37%), outperforming models incorporating macroeconomic indicators, indicating that HPI can be forecast more accurately using its own historical values.

Keywords: Residential real estate, Cyprus, time-series, machine learning, forecasting, clustering

Resumo

O setor imobiliário é um dos maiores mercados globais, contribuindo significativamente para o crescimento económico, a criação de riqueza e a estabilidade financeira. Os preços das propriedades residenciais afetam agregados familiares, empresas e decisores políticos, tornando essencial uma análise e previsão rigorosas para decisões informadas.

No entanto, o mercado imobiliário residencial de Chipre permanece pouco explorado. Os estudos existentes tratam frequentemente o mercado habitacional como homogéneo, apesar de os diferentes tipos de propriedade serem influenciados por mecanismos distintos de formação de preços. Este estudo aborda esta lacuna através de uma análise orientada por dados do mercado imobiliário residencial cipriota, com o objetivo de apoiar empresas, decisores políticos e o público na tomada de decisões.

Foi construído um conjunto de dados com mais de 21.000 anúncios de imóveis residenciais através de web scraping. Foi desenvolvido um índice de desejabilidade locacional utilizando dados geoespaciais abertos; as propriedades foram segmentadas através de clustering K-means, aplicando-se modelos de Árvore de Decisão, Random Forest (RF), Extreme Gradient Boosting (XGBoost) e Gradient Boosting (GB) em cada cluster. Ao nível macroeconómico, o Índice de Preços da Habitação (HPI) do Chipre foi analisado através de modelos econométricos e de séries temporais.

Os resultados revelaram diferentes fatores determinantes de preço entre os segmentos residenciais, com o ensemble RF–XGBoost a explicar até 35,9% da variação de preços para apartamentos, RF–GB 34,6% para casas médias e RF 65,9% para propriedades de luxo. Para a previsão do HPI, o modelo Autoregressive Integrated Moving Average (ARIMA)(1,1,2) obteve a maior precisão (Mean Absolute Percentage Error (MAPE) = 1,37%), superando os modelos com indicadores macroeconómicos, o que indica que o HPI pode ser previsto utilizando apenas os seus valores históricos.

Palavras-chave: Mercado imobiliário residencial, Chipre, séries temporais, aprendizagem automática, previsão, agrupamento

Contents

Abstract ii
Resumo iii
Glossary vi
List of Figures vii
List of Tables viii
Chapter 1 1
Introduction 1
1.1. Context and Motivation 1
1.2. Problem Definition and Objectives 3
1.3. Research Questions 3
1.4. Dissertation Structure 4
Chapter 2 5
Literature Review 5
2.1. Residential properties 5
2.2. Real estate modelling evolution 7
2.3. State of the Art 9
2.4. Research Gap and Study Motivation 12
Chapter 3 14
Dataset, Pre-processing and Descriptive Statistics 14
3.1 Web-scraping 14
3.2 Dataset Pre-processing 16
3.3 Amenity-Based Location Desirability Index 19
3.4 Descriptive statistics 22
3.5 Relationship Identification 27
Chapter 4 35
Clustering and Within-Cluster Prediction 35
4.1. Clustering and Variable Selection 35
4.2. K-means Clustering 39
4.3. Ward's Method 41
4.4. Clustering Results 42
4.5. Per-Cluster Price Prediction 45

Chapter 5 51
External Factors and Forecasting of the HPI 51
5.1. External Factors 51
5.2. Time series Preprocessing 52
5.3. Granger Causality and Cross-Correlation Analysis 57
5.4. Forecasting the HPI 58
Chapter 6 64
6.1. Main conclusions 64
6.2. Limitations and Future Work 66
Bibliography 68
Appendix 72

Glossary

Artificial Neural Network (ANN)
Augmented Dickey–Fuller Test (ADF Test)
Autoregressive Distributed Lag (ARDL)
Autoregressive Integrated Moving Average (ARIMA)
Coefficient of determination (R^2)
Cross-Correlation Function (CCF)
Decision Tree (DT)
Extreme Gradient Boosting (XGBoost)
Generalised Linear Models (GLM)
Geographical Information Systems (GIS)
Gradient Boosting Machine (GBM)
Gross Domestic Product (GDP)
House Price Index (HPI)
Mean Absolute Error (MAE)
Mean Absolute Percentage Error (MAPE)
Multi-Criteria Evaluation (MCE)
OpenStreetMap (OSM)
Pearson's correlation (PC)
Principal Component Analysis (PCA)
Random Forest (RF)
SHapley Additive exPlanations (SHAP)
Vector Autoregression (VAR)

List of Figures

Figure 1: Cyprus structure 2026 2
Figure 2: The Cyprus HPI quarterly (Analytics, 2024) 7
Figure 3: Boxplots of Numerical Variables 18
Figure 4: Bar chart of price in ranges 23
Figure 5: Bar chart of property_price_sqm in ranges 23
Figure 6: Bar chart of land_size_sqm in ranges 23
Figure 7: Boxplots of veranda_covered Figure 8: Boxplot of veranda_uncovered 24
Figure 9: Bar chart of city 25
Figure 10: Bar chart of furnished 25
Figure 11: Bar chart of property status 25
Figure 12: Bar Chart of house_type 26
Figure 13: Bar charts of binary variables 27
Figure 14: PC Heatmap 28
Figure 15: Spearman's Rank Correlation 29
Figure 16: SHAP Summary Plot, RF Model 38
Figure 17: Gower-silhouette Figure 18: Elbow Method 41
Figure 19: Dendogram of Ward's Method 42
Figure 20: K-Means Clusters (k=3): Property Size vs Price 43
Figure 21: 10 most important features Standard Urban Apartments 47
Figure 22: 10 most important features Mid-Size Properties 48
Figure 23: 10 most important features Luxury/Large Properties 48
Figure 24: Average eargings % change, GDP - seasonaly adjusted, HPI, Labour Force, Number of permits Time-series plots 53
Figure 25: ACF Plots 53
Figure 26: Lag Plots of HPI 54
Figure 27: Lag Plots of Labour Force 55
Figure 28: Lag Plots of Number of permits 55
Figure 29: ACF on the first difference 56
Figure 30: Cross Correlation with HPI 58
Figure 31: ARDL residual diagnostics 61
Figure 32: Forecast of the HPI - ARDL model 61
Figure 33: ACF residual plot and Residual distribution – ARIMA 62
Figure 34: ARIMA (1,1,2) forecast of HPI 63

List of Tables

Table 1: Distribution of amenities per website 15
Table 2: Amenities' and Disamenities' weights 20
Table 3: Descriptive statistics of the most important numerical variables 22
Table 4: Cramer's V associations 31
Table 5: Goodman–Kruskal Tau Test 32
Table 6: 10 most significant adjusted residuals and contingency tables 32
Table 7: Kruskal-Wallis Results: Price vs Categorical Variables 36
Table 8: Standard Deviation of cluster mean on scaled data 39
Table 9: Chi-square test for clustering 40
Table 10: Gower Silhouette for Ward's Methods 42
Table 11: Cluster composition per method 43
Table 12: Summary of the Clusters' characteristics using K-Means 44
Table 13: Models' Hyperparameters 46
Table 14: Accuracy matrix per cluster prediction 47
Table 15: Granger causality predictors – HPI 58
Table 16: ARDL forecast accuracy 61
Table 17: Accuracy Results of HPI Forecast 62
Table 18: Summary of the studies' limitations 74
Table 19: Dataset Description 77
Table 20: Cramer's V for Binary Variables 78
Table 21: External Indicators 79

Chapter 1

Introduction

Real estate is one of the world's largest and most resilient investment markets, attracting substantial capital despite periodic economic disruptions. As the residential property market continues to expand, accurate property valuation and price forecasting have become increasingly important for investors, real estate professionals, policymakers, and buyers. While online real estate platforms provide extensive information on property listings, they offer limited insight into the factors that determine property values. At the same time, advances in data analytics, machine learning, and geospatial technologies have created new opportunities to better understand housing markets through large-scale data analysis. This study addresses these challenges by investigating the determinants of residential property prices in Cyprus, segmenting the market into distinct property groups, and developing forecasting models for both property segments and the Cyprus House Price Index.

1.1. Context and Motivation

Cyprus, officially known as the Republic of Cyprus, is an island located in the Eastern Mediterranean Sea. The country has been an independent nation since 1960 and a member of the European Union since 2004. The official currency of Cyprus has been the Euro (€) since 2008. Approximately 37% of the island's territory remains under Turkish occupation as a result of the 1974 Turkish invasion, which is separated by the United Nations Buffer Zone as shown in Figure 1.

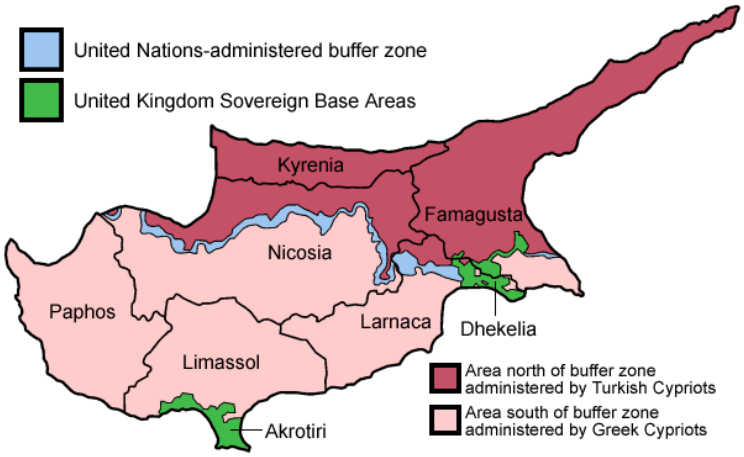


Figure 1: Cyprus structure 2026

This project focuses exclusively on the internationally recognised Republic of Cyprus, which covers the southern part of the island and has a total area of 5.896 square kilometres. According to the 2021 census, the island's total population is approximately 918.000, of whom 690.900 are Greek Cypriots (UK, 2026).

As one of Cyprus's most important economic sectors, the real estate market plays a key role in attracting both domestic and international investments. Real estate comprises residential, commercial, land, and tourism-related properties (Estates, 2025). Residential properties include housing options such as apartments, houses, and villas that can be used for permanent residence or as holiday homes (Estates, 2025). Commercial properties include offices, retail outlets, shopping centres, and industrial facilities, while tourism-related properties include hotels, resorts, and other guest accommodation options (Estates, 2025). This project focuses on residential properties and excludes the other real estate categories.

According to the article released by LiVE (2025), the Real Estate Market in Cyprus has experienced significant growth, with sales contracts up by 13% from 2024. The total value of property transactions between January and September 2025 reached €3.49 billion, indicating strong and dynamic market activity. Real estate is considered one of the most profitable sectors in Cyprus for 2025 as the country remains a popular destination for international property investors. During this period, 3.469 sales contracts were submitted by non-residents, of which 1.846 originated from non-EU countries (Kathimerini, 2025).

The expansion of the real estate market has increased the need for reliable property valuation and forecasting. Accurate predictions can help real estate companies improve pricing strategies and investment decisions, assist policymakers in monitoring housing market dynamics and designing effective policies and provide buyers, sellers and investors with better information for decision-making. These needs motivate the development of data-driven models that can identify the factors influencing residential property prices and forecast future market trends.

1.2. Problem Definition and Objectives

Choosing the right price for real estate property significantly influences attracting buyers and achieving a successful sale. Nelson (2024) mentions that a good strategy is to compare your assets with similar properties in your submarket and asset type. To the best of my knowledge, current studies primarily focus on forecasting real estate price growth in Cyprus. However, a noticeable research gap remains in investigating both external and internal factors that influence Cypriot real estate prices. Furthermore, limited research employs clustering approaches to group properties by common characteristics and forecast HPI trends.

This project aims to analyse the external and internal factors that influence real estate prices in Cyprus, develop clustering methods to group properties with similar characteristics, identify price drivers by predicting prices within each cluster, and apply time-series and machine-learning methods to forecast the house price index. Furthermore, the project provides insights to support decision-making processes for stakeholders – real estate companies, government authorities, and the public.

1.3. Research Questions

This study focuses on answering the following research questions about the Cypriot Real Estate Market:

1. What property-specific aspects and external factors influence residential real estate prices in Cyprus?

2. How can residential real estate properties be clustered, and to what extent can time series forecast the house price index?
3. What insights can be drawn from the analysis to support residential real estate companies, policymakers, and the public in decision-making?

1.4. Dissertation Structure

The subsequent chapters describe the methodology used to analyse both the external and internal determinants of real estate prices in Cyprus, apply clustering techniques to group properties based on common characteristics, and utilise time-series models to forecast the house price index.

Chapter 2 presents a literature review of prior studies on the topic and relevant techniques, and discusses the limitations identified. Chapter 3 covers data collection via Web scraping and preprocessing of raw data. It also identifies relationships between variables and presents the use of Geographical Information Systems (GIS) and Multi-Criteria Evaluation (MCE) to construct a location desirability score. Chapter 4 applies clustering techniques to group properties into segments and identifies price drivers within clusters using tree-based models. Chapter 5 examines the relationship between external factors and the HPI and applies forecasting techniques. Chapter 6 concludes the study by discussing the main findings, limitations and directions for future research.

Chapter 2

Literature Review

This chapter provides the theoretical framework for understanding the residential real estate market in Cyprus and the evolution of real estate modelling approaches in the academic literature. It begins by describing the characteristics of residential properties in Cyprus, the market structure, regional differences, investment conditions, acquisition costs, and the role of foreign investors. It places particular emphasis on the HPI as a key indicator for monitoring price dynamics over time. The chapter then reviews the evolution of real estate pricing models, highlighting how real estate analysis has shifted from simple linear models to more complex approaches. Finally, it critically examines the limitations of prior studies on housing price modelling, clustering, and prediction. Overall, this chapter connects the Cypriot residential context with the broader academic developments in real estate modelling.

2.1. Residential properties

The Republic of Cyprus is considered a highly attractive destination for residential real estate investment, thanks to its economic stability, favourable tax policies and strategic location within the European Union. The residential sector accounts for approximately 60% of the real estate market, with a market value exceeding €40 billion in 2025 (Estates, 2025). Investment returns in this sector typically range from 4% to 7% gross per year, with apartments generally outperforming houses (Estates, 2025).

Property prices have been rising steadily, driven by strong demand and limited supply, as reflected in the ongoing growth of the residential property index. Cyprus appeals to foreign investors because it has no annual property or inheritance taxes. Furthermore, investments of €300.000 or more (plus VAT) can qualify foreign buyers

for permanent residency, commonly referred to as a "Golden Visa" (Investments, 2025).

Although Cyprus is geographically small, regional differences in the real estate market are notable. Based on (Estates, 2025), Limassol has the highest property prices and rental yields, driven by demand for businesses and luxury buyers. Nicosia offers more affordable options with a stable rental market. Paphos attracts tourism-focused and expatriate buyers, and Larnaca presents value-oriented opportunities with strong growth potential and competitive rental yields.

The acquisition cost of purchasing a residence in Cyprus is approximately 8-10% of the property's value, including VAT, transfer fees, stamp duty, and legal expenses. Stamp duty in Cyprus is calculated on a sliding scale: 0.15% on the first €170,860 of the property value and 0.20% on any amount exceeding this threshold (Realty, 2025). Real estate agent commissions in Cyprus typically range between 3% and 5% of the sale price (Realty, 2025).

Newly built properties are subject to 19% VAT on the purchase price (Realty, 2025). Nevertheless, a reduced VAT rate of 5% is available for first-time buyers acquiring a primary residence, subject to specific conditions. Transfer fees constitute another significant expense and are calculated based on the property's market value at progressive rates: 3% for the first €85.000, 5% for the next €85.000, and 8% for any amount exceeding €170.000 (Realty, 2025).

Based on Estate (2024), between 2021 and 2024, foreign real estate investments in Cyprus varied by region. In Nicosia, most buyers were Greeks (1.626), Britons (1.584), and Australians (545). Limassol attracted primarily Russians, followed by UK and Israeli nationals, while Cypriots purchased 62.460 properties in the city. Paphos's top foreign buyers were Britons, Russians, and Israelis, with Britons owning 4,483 properties. In Larnaca, citizens from the UK, Lebanon, and Israel were most active. Finally, in the vacant areas of Famagusta, demand came mainly from British, Greek, and Lebanese buyers.

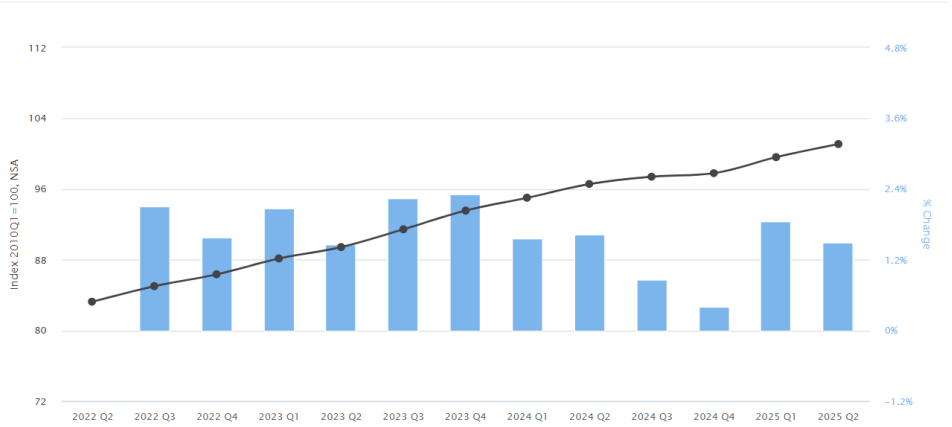


Figure 2: The Cyprus HPI quarterly (Analytics, 2024)

A crucial indicator in the Real estate sector is the HPI. According to Case et al. (1991) HPI is a time-series indicator that tracks changes in the general level of residential property prices in a specific market. It captures how prices rise or fall relative to a base period, allowing comparisons over time while controlling for differences in the house type. It is a crucial indicator used in research, policy and financial markets. For Cyprus, the HPI is published quarterly and indicates the average change in residential property prices. The data are collected from sources like the Central Bank of Cyprus and the Statistical Service of Cyprus.

Figure 2 shows the Cyprus HPI time series and its quarterly changes from 2022 to 2025. The black line represents the HPI level, while the blue bars show the quarterly percentage changes. The HPI indicate a steady upward trend throughout the 3 years. Specifically, the HPI increases from 83.25 in 2022 to 101.08 3 years later. From 2022 to 2023, the HPI increased strongly, while in early to mid-2024 it continued to rise at a noticeably slower rate. In 2025, house prices started to increase more rapidly again.

2.2. Real estate modelling evolution

Real estate pricing has long been a topic in economic research, with early studies seeking to explain real estate prices through observable structural and locational characteristics. One of the foundational contributions to real estate pricing modelling is the hedonic pricing framework introduced by Rosen (1974), which conceptualises housing prices as a function of individual attributes, such as size, location, and

neighbourhood characteristics. This approach laid the foundation for the widespread use of regression-based models in housing market analysis.

Based on the hedonic pricing theory, subsequent studies expanded the scope of real estate pricing modelling by incorporating market segmentation and macroeconomic dynamics. Goodman and Thibodeau (1998) proven that housing markets are not homogeneous and can be divided into sub-markets, while Malpezzi (2003) provided an applied review on hedonic models, noting their practical relevance to housing policy and valuation. By utilising repeat-sales indices, Case and Shiller (1989) demonstrated that house prices follow patterns rather than changing randomly, laying the foundation for research showing that housing markets experience cycles that reflect broader economic conditions.

Researchers also acknowledged the importance of spatial dependence and geographic factors in real estate pricing evaluation. Anselin (1988) used spatial econometric methods, which allowed researchers to account for spatial autocorrelation in housing prices, while later studies demonstrated that neighbour quality and spatial spillovers are significantly important in property evaluation (Dubin, 1998).

The growing availability of digital data has shifted recent research towards data-driven and computational approaches. Web scraping techniques have been identified as an important tool for collecting real estate data from online available platforms. Santos (2018) extracted real estate listings using web-scraping tools and used data mining techniques for preprocessing, followed by hierarchical clustering to group properties into homogeneous segments.

Mohan et al. (2019) studied the impact of major macroeconomic indicators on housing prices over time using a Vector Autoregression (VAR) model. The study found that the HPI itself exhibits the largest share of variability, suggesting strong persistence and influence on past price movements.

Stundziene et al. (2021) further studied the external factors influencing real estate prices in Lithuania, using macroeconomic, labour-market, business, financial, and real

estate indicators, demonstrating that real estate cycles are closely related to economic cycles. Quarterly changes in GDP, consumer spending, and household cash deposits were identified as the most influential attributes, while labour market indicators were the least informative.

Lahoz (2007), on the other hand, analysed the geographical and locational factors that affect real estate pricing using MCE and Regression Analysis. The findings indicate that factors such as crime rates, the number of foreign residents in the area, airport accessibility, income levels, and proximity to green areas and water play a critical role in real estate values.

The heterogeneity of housing markets has also motivated the use of unsupervised learning techniques to identify distinct segments. Wiersma et al. (2021) performed Principal Component Analysis (PCA) combined with K-means clustering and Ward's methods to segment German urban housing markets. Using multiple criteria to determine the optimal number of clusters and bootstrapping robustness, the study identified four distinct city clusters, ranging from major metropolitan areas with high prices to cities characterised by affordability and weaker economic fundamentals. These findings reinforce earlier evidence that real estate markets exhibit structural diversity.

Recent studies have begun exploring the integration of AI and computer vision into real estate analysis. Küp et al. (2025) investigated the use of image classification techniques to assess how picture quality affects listing accuracy and visual attractiveness.

2.3. State of the Art

Santos (2018) performed web scraping to extract real estate advertisements in Porto from online platforms. The primary objective was to repeatedly scrape data to build a historical dataset of price updates and forecast future property price updates.

Regarding forecasting performance, Santos (2018) found that the Random Forest (RF) model outperformed the Artificial Neural Network (ANN), achieving 69.55% accuracy.

Mohan et al. (2019) aimed to identify the specific external economic factors that explain the variations in the HPI using the VAR model. Data from 1999 to 2017 were collected for Amherst, New York, on five macroeconomic indicators: crude oil price, 30-year mortgage interest rate, consumer price index, Dow Jones Industrial Average, and unemployment rate. The study found that changes in HPI are explained by 92.7% of its own variance one month ahead and 74.5% 12 months ahead. Additionally, the mortgage interest rate had the greatest impact among the indicators, explaining 5% of the variance.

Stundziene et al. (2021) studied the external factors influencing housing prices in Lithuania, focusing on a range of macroeconomic, labour-market, business, financial, and real-estate indicators, as well as consumer and business expectations. Using correlation and regression analysis, causality tests, and autoregressive distributed lag (ARDL) models, the study found that 3 factors had the most significant impact on the real estate market: quarterly changes in Gross Domestic Product (GDP) growth, consumer expenditure, and households' cash and deposit levels.

Geographical determinants have been examined using GIS and MCE to assess the impact of locational factors on real estate prices. The variables were grouped into several categories depending on their content. The MCE weights were defined using regression analysis coefficients.

Wiersma et al. (2021) examined housing market segmentation in Germany using a two-step clustering approach that combined K-means with Ward's hierarchical clustering, following dimensionality reduction via PCA. The variables were grouped into 4 categories: market size and dynamics, price level and dynamics, socioeconomic indicators and dynamics and regional demographics.

The robustness of the clustering results was assessed using bootstrapping. Several criteria were used to determine the optimal number of clusters, which was ultimately

set to 4, and robustness tests showed that 88.5% of the city groupings were statistically significant.

In their study, Yılmaz and Kestel (2020) utilised the HPI to forecast house prices in Turkey and assess the predictive reliability of various statistical models. For this purpose, they used monthly HPI data from January 2010 to May 2019, alongside relevant macroeconomic and financial indicators. The study employed Generalised Linear Models (GLM), VAR models, and Univariate Time series models, including Autoregressive Moving Average, Autoregressive Integrated Moving Average (ARIMA), and exponential smoothing models (such as Holt-Winters), to generate HPI predictions.

The study concluded that the forecasting power of GLM and VAR models depends heavily on the predictive performance of the exploratory variables, thereby limiting overall predictive performance. However, they remain useful for identifying key housing drivers. In contrast, seasonal ARIMA and Holt-Winters' exponential smoothing models demonstrated higher forecasting accuracy.

Soylemez et al. (2026) conducted time-series modelling to forecast the HPI across 27 European countries using quarterly data from Q4 2013 to Q3 2024. The models employed included ARIMA, Holt Linear Trend, Additive Damped Trend and Exponential Smoothing. The results highlight the importance of country-specific forecasting approaches rather than a universal model: in 9 countries, ARIMA provided the most accurate forecasts, while in 7 countries, Holt Linear Trend or Additive Damped Trend models provided the best predictions.

Despite these results, prior studies have several limitations that motivate further research. Prior studies mainly focus on other jurisdictions, limiting the generalisability of their findings to the Cypriot residential real estate market. Several studies examine either internal property characteristics or external macroeconomic determinants in isolation, rather than integrating both perspectives into a single framework. Research on housing market segmentation remains limited, with few studies applying clustering techniques to identify distinct property groups before modelling prices. Geospatial

analyses often rely on simplified assumptions, subjective weighting schemes, or incomplete spatial variables, reducing the reliability of location-related measures.

In forecasting, many studies rely on a single modelling approach or evaluate only a limited set of techniques, which may perform poorly during structural market changes. Furthermore, several studies are constrained by limited datasets, short observation periods, insufficient historical price information, or a lack of detailed locational data, all of which reduce predictive performance.

These limitations highlight the need for a comprehensive framework that combines internal property characteristics, spatial factors, market segmentation, machine learning, and macroeconomic analysis to better explain and predict residential real estate prices and to apply forecasting techniques to the HPI. Table 18 in the Appendix summarises the literature review and the limitations identified.

2.4. Research Gap and Study Motivation

The real estate market remains one of the most financially significant sectors in Cyprus, offering a valuable opportunity to examine the factors that drive property prices. While previous studies have advanced our understanding of housing market forecasting worldwide, notable research gaps remain. As highlighted by Soylemez et al. (2026) real estate forecasting modelling varies substantially across European countries, making it difficult to generalise findings from prior studies to the Cypriot market, the focus of this study.

Previous studies also identified limited integration of external and internal determinants as a limitation. Several studies consider internal property characteristics and external macroeconomic determinants separately, leaving their combined influence on property prices relatively underexplored.

The absence of prediction within cluster-based market segments represents a significant research gap in prior studies. While most existing literature focuses on modelling housing prices at the aggregate market level, it largely ignores that

properties can naturally be grouped into clusters based on common characteristics. This raises an important question: Do different drivers affect prices in each cluster? Understanding this variation is critical, as properties in different clusters may exhibit distinct pricing dynamics. By addressing this gap, researchers can implement more tailored and robust predictive models that reflect the housing market's heterogeneity.

Building on the gaps identified in previous research, this study analyses the Cypriot real estate market, a country that has received limited attention in prior literature, to identify the property-specific characteristics and external factors that drive residential housing prices. Using cluster-based segmentation, the study addresses the lack of segment-specific prediction to test whether distinct property segments show different pricing dynamics and price drivers. At the aggregate market level, it further forecasts the HPI against key macroeconomic, labour market, housing supply and demographic indicators. The insights aim to support real estate companies, policymakers, and the public in making informed, data-driven decisions, while advancing understanding of real estate price formation in a market-specific context.

Chapter 3

Dataset, Pre-processing and Descriptive Statistics

Chapter 3 describes the methodological approach adopted to collect, prepare and explore the data underlying this study. The chapter begins with the web-scraping process used to gather residential data from three Cypriot real estate websites, followed by the dataset preprocessing process applied to handle missing values, inconsistencies, and outliers and create a structured, organised dataset. This chapter also presents the construction of a composite Amenity-Based Location Desirability Index (mce_score) built from open geospatial data to capture the influence of location on property prices. A descriptive statistical overview categorised by the different types of variables, and a relationship examination between those variables are included using correlation and association tests appropriate to each variable type.

3.1 Web-scraping

Web scraping is the automated process of collecting structured or unstructured data from the web by employing intelligent automation techniques. It addresses several limitations of manual data collection, which can be time-consuming and prone to human error. Furthermore, it outperforms alternative data-gathering methods as it is faster, more cost-effective, and highly flexible.

This study uses Python because of its user-friendly syntax and extensive ecosystem of data-collection libraries. Among Python's web scraping tools, Selenium is the most effective for dynamic websites (Bale et al., 2022). Additionally, Selenium can reduce the processing time and data consumption by 70% and 90%, respectively, compared to Beautiful Soup (Sharma & Borkar, 2024). For these reasons, Selenium was selected for this study, as it was considered more suitable.

WEBSITE	COUNT	PERCENTAGE (%)
Website A	10,467	48.4%
Website B	9,322	43.1%
Website C	1,858	8.6%

Table 1: Distribution of amenities per website

Three Cypriot real estate web pages were web-scraped for analysis using Selenium. The web pages were selected based on their popularity and whether their terms and conditions permitted web scraping.

As shown in Table 1, most data were web-scraped from Website A (48.4% of the dataset), while the fewest properties were scraped from Website C (less than 10%). The raw dataset includes 21,647 properties: 68.9% apartments and 31.0% houses. Although the division is not well balanced, the number of collected houses can still be considered a representative set of listings.

The data extracted via web scraping were structured into four main variables: page_url, title, price, description and features. Once the data were collected and organised, text-mining techniques were applied to create a comprehensive database for analysis. The text in each variable was decomposed into distinct variables, using automated techniques complemented by manual validation checks to ensure accuracy.

The individual datasets were then merged into a unified dataset, and their columns were aligned to a common schema based on content. Manual verification supported this process to ensure each column represented consistent and equivalent information across datasets.

Table 19 in the Appendix presents all variables included in the dataset after merging the four distinct datasets. The variables were categorised by content into identifier, text, numerical, categorical, and binary. More specifically, the dataset contains 88 variables: 1 identifier, 3 text, 16 numerical, 14 categorical, and 54 binary.

The column "Source(s)" in Table 19 in the Appendix demonstrates which websites contained the variable. The three websites differed in the type and depth of

information they provided for property listings. Website A offered limited information, while Websites B and C provided more extensive details to describe the property.

3.2 Dataset Pre-processing

Data pre-processing began with removing duplicate records. Next, all columns were standardised to ensure consistency, with variable names converted to a common format. Duplicate columns were then identified and removed after ensuring that overlapping information was merged.

For the numerical variables, errors were corrected, and values were converted into appropriate numerical types (integers or floats) based on the variable content. Certain variables were merged (e.g., uncovered_parking and covered_parking into parking) because they had high levels of missing data when considered separately.

Continuing with categorical variables, the original heating_type variable provided heating categories; however, the categories were not well defined, so this variable was converted into a binary variable (has_heating). The district variable was cleaned by removing invalid entries caused by variations in translation from Greek to English. The district variable was then mapped and categorised into cities to check for errors and impute missing values.

The binary variables were then assessed for sparsity and uniqueness. An 85% threshold of zero values was applied to identify overly rare variables. Variables with low positive-class occurrence provide limited learning signal and can reduce model performance due to insufficient variation. Out of the 54 binary variables, 36 variables exhibited more than 85% zero values. For those, the Mann-Whitney U test was performed to determine whether there is a statistically significant difference in price distribution between group 0 (the feature is absent) and group 1 (the feature is present) at a significance level of 5% (Mann & Whitney, 1947). Based on the results, 6 features were dropped: provision_for_storage_heating, ceiling_fans, cul-de-sac,

projectors, solar_photovoltaic_panels, and title_deed, as their p-values were >0.05, indicating no statistically significant difference in price distribution between properties with and without these features.

3.2.1. Handling Missing Values

A common issue in real-world data collection is the absence of recorded information from certain variables in the dataset, called missing values. According to Saar-Tsechansky and Provost (2007) missing values can significantly affect the performance and reliability of predictive models if not handled appropriately.

At this stage of the analysis, missing values were intentionally left untreated, as the descriptive analysis aimed to reflect the dataset's original structure and distribution. By retaining the missing values at this stage, the descriptive statistics and relationship analysis could provide a more accurate representation without the influence of imputation methods. Missing-value treatment was deferred to later stages of the study, according to the specific requirements of each modelling technique.

However, variables with an excessive proportion of missing values were considered uninformative for further analysis. Consequently, the numerical variables monthly_communal_fees, garden_area, renovation_year, and uncovered_parking, and the categorical variables cooling_type, electricity_type, and construction_type, were removed from the dataset, as each contained more than 90% missing values.

3.2.2. Outliers Treatment

In this study, outliers and errors were treated differently across variables, depending on each variable's nature and content. To identify these values, boxplots were inspected manually.

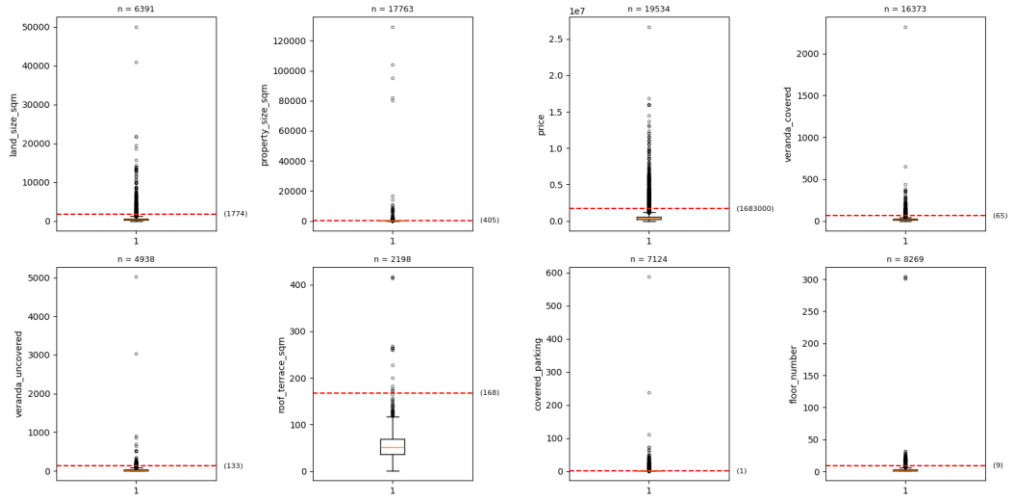


Figure 3: Boxplots of Numerical Variables

By inspecting Figure 3, we removed land_size_sqm and property_size_sqm values below 16 sqm. For land_size_sqm, observations exceeding 40,000 sqm were treated as invalid and removed, while for property_size_sqm, five extreme upper-tail outliers were removed. For price, the extreme observation of €26,670,000 and implausibly low sale prices of €550, €1,000, and €1,200 were removed.

Unrealistic observations were similarly identified for veranda_covered (2,320 sqm), veranda_uncovered (3,032 sqm, 5,025 sqm, and eight additional outliers ranging from 513–911 sqm), covered_parking (43 parking spaces), floor_number (values exceeding 300), and roof_terrace_sqm (10 observations exceeding 200 sqm), all of which were removed.

Outliers adjacent to the upper fences were retained if they represented realistic property characteristics. Rather than eliminating these observations based solely on a statistical threshold, they were retained to preserve the authentic diversity of residential properties. In total, 2,122 observations were removed during the outlier-removal process.

3.3 Amenity-Based Location Desirability Index

To capture the influence of the location of the residential property on price, this study adopted the same MCE approach proposed by Lahoz (2007). This approach follows the principle of MCE, a technique widely used in spatial decision analysis and hedonic property research, whereby multiple weighted spatial criteria are aggregated into a single composite score (Malczewski, 2004).

The evaluation criteria for this method were limited to physical and environmental accessibility factors derived from open geospatial data; socioeconomic determinants and demographic composition were excluded because publicly available data were lacking. A composite locational desirability index was constructed for the 250 individually geocoded points, producing up to 237 distinct MCE values.

Using the osmnx Python package, OpenStreetMap (OSM) data via the Overpass API were used to retrieve amenity and disamenity locations programmatically. OSM was selected as it provides a freely available, regularly updated and comprehensive database of points of interest for Cyprus. For this study, 15 amenity and disamenity categories were extracted using standard OSM key-value tags.

The accessibility analysis was conducted at the district level, where each district was first assigned to its city (e.g., Strovolos – Nicosia) and then converted to geographic coordinates using the Nominatim geocoding service in the osmnx package. This process transformed each district into a single map point, and then, via OSM, official boundary polygons were created for the five cities. Using GeoPandas, the official district points were identified. Approximately 77.8% of the properties were successfully assigned to an official district. The remaining observations could not be matched to any locality name and could not be geocoded.

Before performing any spatial calculations, all geographic data were transformed from the geographic coordinate reference system (WGS84) to the projected coordinate reference system (EPSG:3857) so distances could be calculated in meters. As this map projection is not perfectly accurate at Cyprus's latitude, the calculated

distances are slightly overestimated; however, this affects only the absolute scale of the accessibility scores rather than the relative ranking of districts.

The distance decay method was then applied by measuring the Euclidean distance between each district centroid and every amenity and disamenity. It assumes that nearby amenities have a greater positive influence on accessibility, while nearby disamenities have a greater negative impact, with both effects decreasing exponentially as distance increases.

The decay rate was calibrated separately for each category (e.g., 1,500 m for restaurants and 25,000 m for airports) to reflect their differing spatial influence on residential desirability. To improve computational efficiency, radius-based neighbour searches were performed using a k-d tree spatial index. Rather than calculating the distance between every district and amenity, the algorithm evaluated only amenities within the specified search radius, substantially reducing computational cost.

To ensure all amenities and disamenities used the same units and scale, each amenity and disamenity variable was standardised using min-max normalisation on a common scale ranging from 0 to 1, where 0 indicates the lowest accessibility observed across the five cities and 1 the highest.

CATEGORY	CLASSIFICATION	WEIGHT
Main roads	Positive	0.10
Airports	Positive	0.08
Schools	Positive	0.12
Universities	Positive	0.09
Healthcare	Positive	0.14
Supermarkets	Positive	0.12
Shopping centres	Positive	0.08
Restaurants	Positive	0.08
Parks	Positive	0.08
Beaches	Positive	0.08
Mountains/peaks	Positive	0.03
Industrial land	Negative	0.14
Military land	Negative	0.10
Waste facilities	Negative	0.12
Power plants	Negative	0.14

Table 2: Amenities' and Disamenities' weights

Each normalised criterion was then assigned a weight (user-defined) reflecting its importance in influencing residential property value, positively or negatively. As shown in Table 2, the higher the weight, the stronger the influence on the final score; for example, industrial areas and power plants were assigned the highest negative weight (0.14), reflecting the noise, air pollution, and elevated accident risk associated with proximity to such land uses.

Sensitivity analysis was then performed using Monte Carlo with Spearman Rank Correlation. Monte Carlo generates thousands of scenarios of alternative weighting (Harrison, 2010). The Spearman Rank Correlation compares the outputs ranking from each simulated scenario against the original baseline ranking (Eden et al., 2022). The results show that the baseline ranking is robust overall, maintaining strong central alignment with a mean Spearman correlation of 0.827 (median 0.853). However, it shows left-tail sensitivity: in about 33% of tests, changing priorities caused the list to scramble the original ranking. Consequently, while general positions are stable under minor variations, larger priority shifts could reorder top or borderline rankings.

Finally, the MCE score was calculated by subtracting the negative accessibility sub-score from the positive one. Districts with better access to amenities and lower exposure to disamenities receive a higher total MCE score.

The MCE score's redundancy with the variable city was tested using eta squared against a 50% threshold. Eta squared is an effect-size measure that shows how much of the total variation in the dependent variables is associated with a categorical factor (Kim, 2016). The result of 0.56 showed that the city already explains the variation in the MCE score, so the score does not provide additional information.

3.4 Descriptive statistics

The descriptive statistical analysis summarises the main features, providing a better understanding of the dataset and its feature distributions.

Numerical Variables

To gain deeper insight into the behaviour of the numerical variables, we calculated main descriptive statistics, followed by inspection of the variables' histograms and boxplots.

As shown in Table 3, the variable under study, price, indicates a median property price of €360,000 and a considerably higher mean of approximately €614,217, with a high coefficient of variation (145.15%).

In terms of space, the average internal area of the house is 158.9 sqm, with the median even lower at 107 sqm, while the average land size is 638 sqm, with the median significantly lower at 380 sqm. Both variables suggest significant variance, reflecting that the dataset contains diverse residential properties.

VARIABLE	MEAN	MEDIAN	VARIANCE	STD. DEVIATION	COEFFICIENT OF VARIATION	MIN	MAX
price	614,216.93	360,000	7.9482E+11	891,526.33	**145.15**	33,000	16,845,000
bedrooms	2.53	2	1.26	1.12	44.38	0	12
bathrooms	2.25	2	1.27	1.13	49.95	1	12
property_size_sqm	158.9	107	79915.05	282.69	**177.91**	30	16895
land_size_sqm	638.4	380	1446955.75	1202.89	**188.42**	38	21740
veranda_covered	23.71	20	292.68	17.11	72.16	2	200
veranda_uncovered	36.5	21.3	1633.51	40.42	110.74	2	350
construction_year	2021.12	2025	129.29	11.37	0.56	1880	2029

Table 3: Descriptive statistics of the most important numerical variables

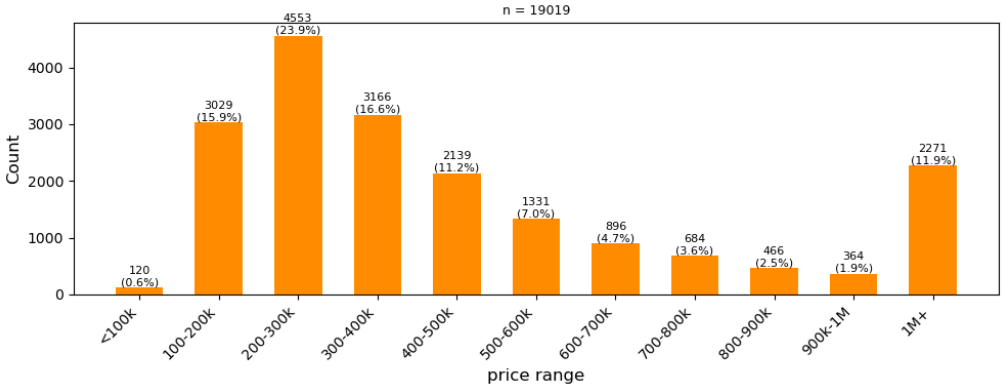


Figure 4: Bar chart of price in ranges

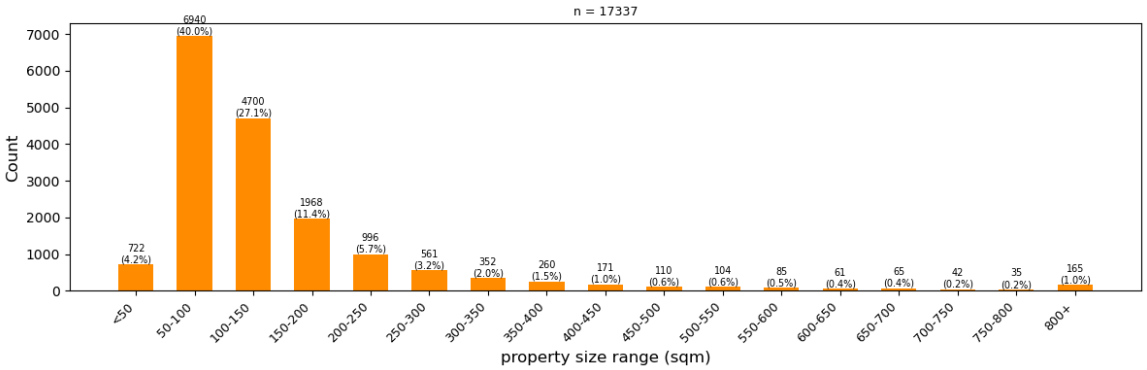


Figure 5: Bar chart of property_price_sqm in ranges

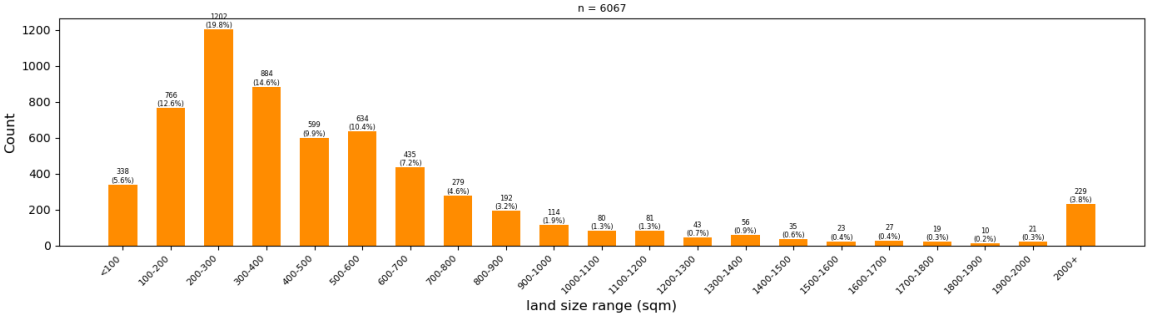
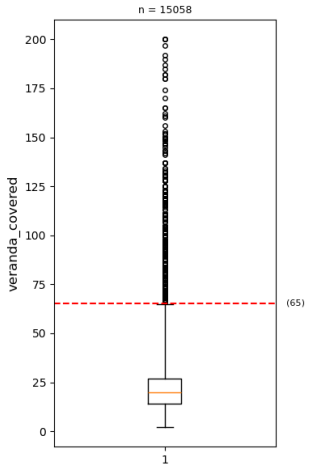


Figure 6: Bar chart of land_size_sqm in ranges

After removing the extreme observations in veranda_uncovered, the maximum values are now capped at realistic values of 200 sqm and 350 sqm, respectively.

Due to the high variability of the price, the variable was better visualised by categorising the values into ranges. In Figure 4, a right-skewed distribution is observed, with most values concentrated on the left side of the plot in the lower and



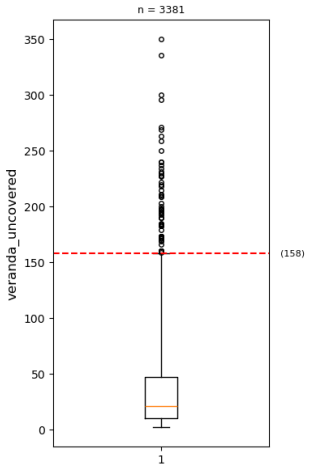


Figure 7: Boxplots of veranda_covered *Figure 8: Boxplot of veranda_uncovered*

middle price brackets. Notably, the 100-400k range accounts for 56.4% of the dataset, representing the primary consumer market.

Similarly, for the space variables property_size_sqm and land_size_sqm, as shown in Figures 5 and 6, their values were categorised into ranges to better visualise their distributions.

The internal living space of most of the properties is between <50 sqm and 100 sqm, accounting for 44.2% of the dataset. When combined with the 101-150 m^2 bracket (27.1%), these categories comprise more than 2/3 (71.3.1%) of the entire dataset. Beyond 150 sqm, the distribution decays rapidly.

Unlike the sharp spikes seen in the property_size_sqm plot, land sizes are distributed more evenly across the categories. The buckets starting from 101 sqm to 600 sqm collectively account for 67.3% of the properties' land data, reflecting standard suburban parcel sizes. After 601 sqm, the tail gradually slopes downwards.

Figure 7 and 8 present the boxplots for the covered and uncovered veranda variables. Both variables exhibit right-skewed distributions, with most observations concentrated at lower values. As shown in Figure 8, 75% of the observations for veranda_covered are below 27 sqm. Veranda_uncovered has 75% of its values below 47 sqm, which is 20 sqm higher than the corresponding value for veranda_covered.

Categorical Variables

Continuing with categorical variables, bar charts were generated to inspect variables' behaviours and distributions

As illustrated in Figure 9, most properties are located in Limassol (26.7%). Larnaca and Nicosia are similarly represented, each accounting for approximately 19% of the properties. Paphos and Famagusta have the lowest representation, with 10.6% and 6.2% of the observations, respectively.

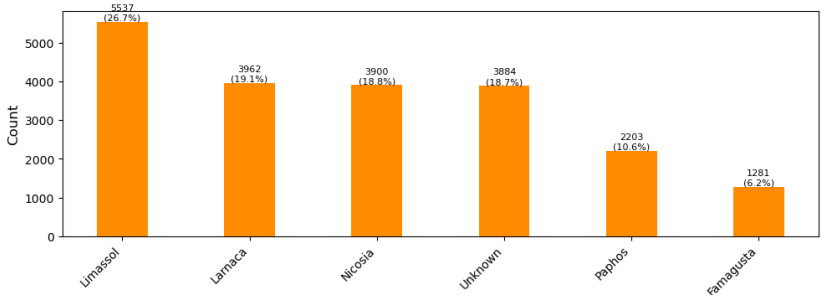


Figure 9: Bar chart of city

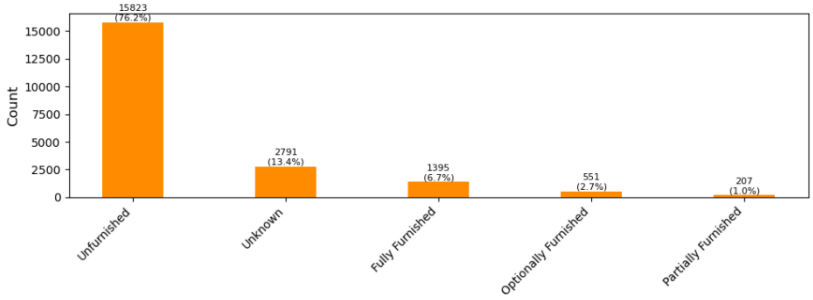


Figure 10: Bar chart of furnished

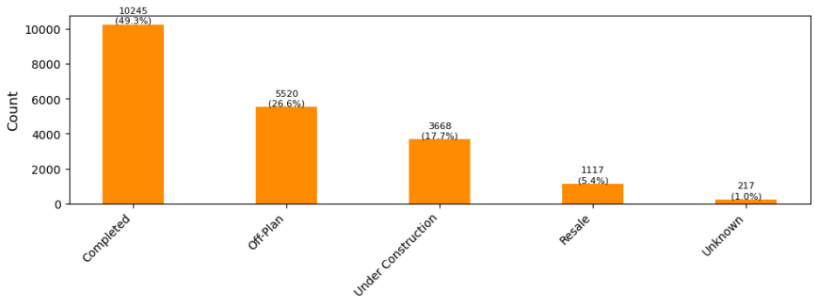


Figure 11: Bar chart of property status

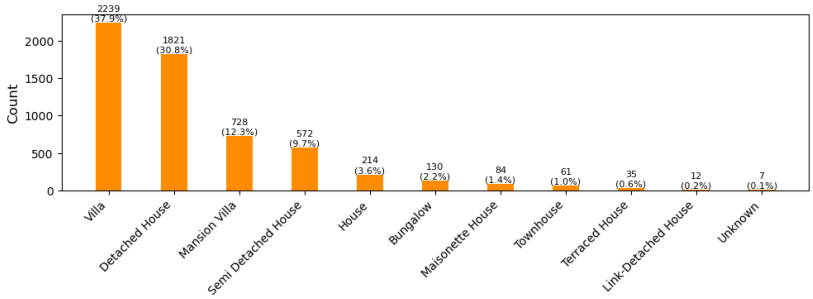


Figure 12: Bar Chart of house_type

Figure 10 demonstrates the bar chart of the variable furnished. Most properties in the dataset are unfurnished, accounting for 76.2% of all observations.

The dataset contains mostly completed properties, representing 49.3% of the dataset as shown in Figure 11. 26.6% of the properties are off-plan, and 17.7% are under construction. Only 5.4% are resale properties.

Figure 12 visualises the frequency distribution of the categories of the variable house_type. The bar chart indicates that villas and detached houses dominate the housing market in the dataset, together comprising more than 2/3 of all houses.

Binary variables

The binary variables were examined to assess the distribution of observations across the two possible categories. For each variable, the percentage frequency of each characteristic was calculated and visualised using bar charts.

Figure 13 shows bar charts for pool, double_glazing, disabled_access, and veranda. As shown, 33.3% of properties have a pool, whether private or shared with other houses in a complex, and 39% have double-glazed windows, an important feature that improves the house's energy efficiency.

Figure 13 also presents the distribution of properties with disabled access. As demonstrated, only 4.1% of properties include disabled-access features. According to Eurostat (2024), 22% of the population aged 16 years and over in Cyprus has a disability, based on self-reported long-term activity limitations. A clear gap exists between housing supply and accessibility needs.

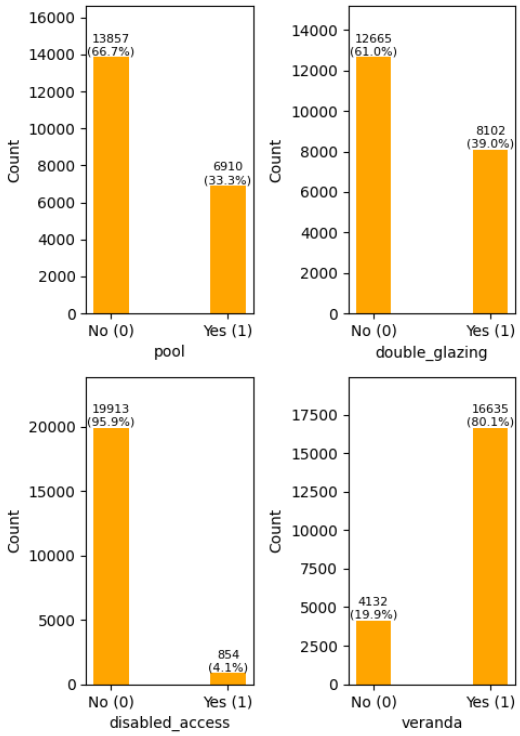


Figure 13: Bar charts of binary variables

The results in Figure 13 indicate that 80.1% of the properties include either a covered or uncovered veranda. This high prevalence reflects the importance of outdoor living spaces in Cyprus, where the Mediterranean climate allows residents to enjoy them for much of the year.

3.5 Relationship Identification

Following the descriptive analysis of the individual variables, the next step is to analyse and identify the relationship between them. Exploring these relationships helps identify potential associations and patterns among property characteristics, providing a better understanding of how different features co-exist within residential properties. This analysis also offers valuable insights into the dataset's structure and can reveal dependencies between variables that may be useful for subsequent modelling.

Numerical Variables

To analyse the relationships between continuous numerical variables, Pearson's Correlation Coefficient was used. The test measures the strength and direction of a linear relationship between two continuous numerical variables (Sedgwick, 2012). The coefficient reflects how closely the data points align with a straight line (Sedgwick, 2012).

Even though the variables do not follow normal distributions, for sample sizes greater than 30, Pearson's correlation (PC) is generally considered reliable even if the normality test is violated due to the central limit theorem (Greene, 2018).

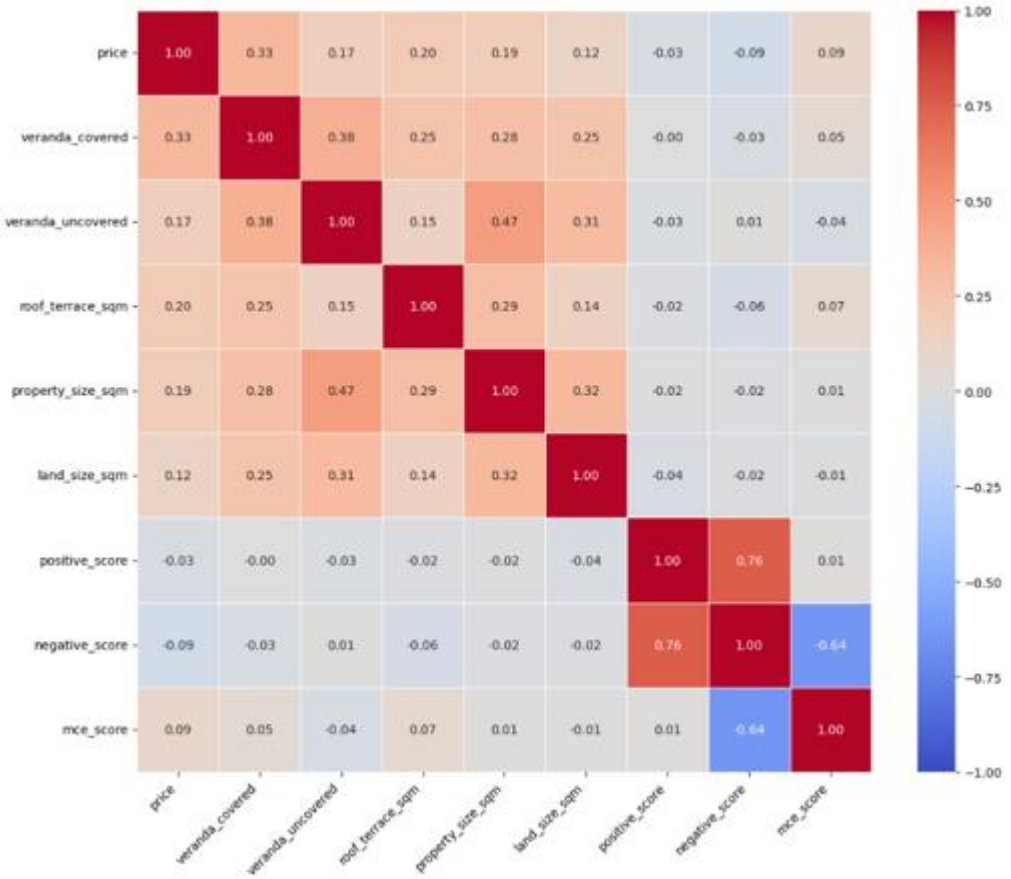


Figure 14: PC Heatmap

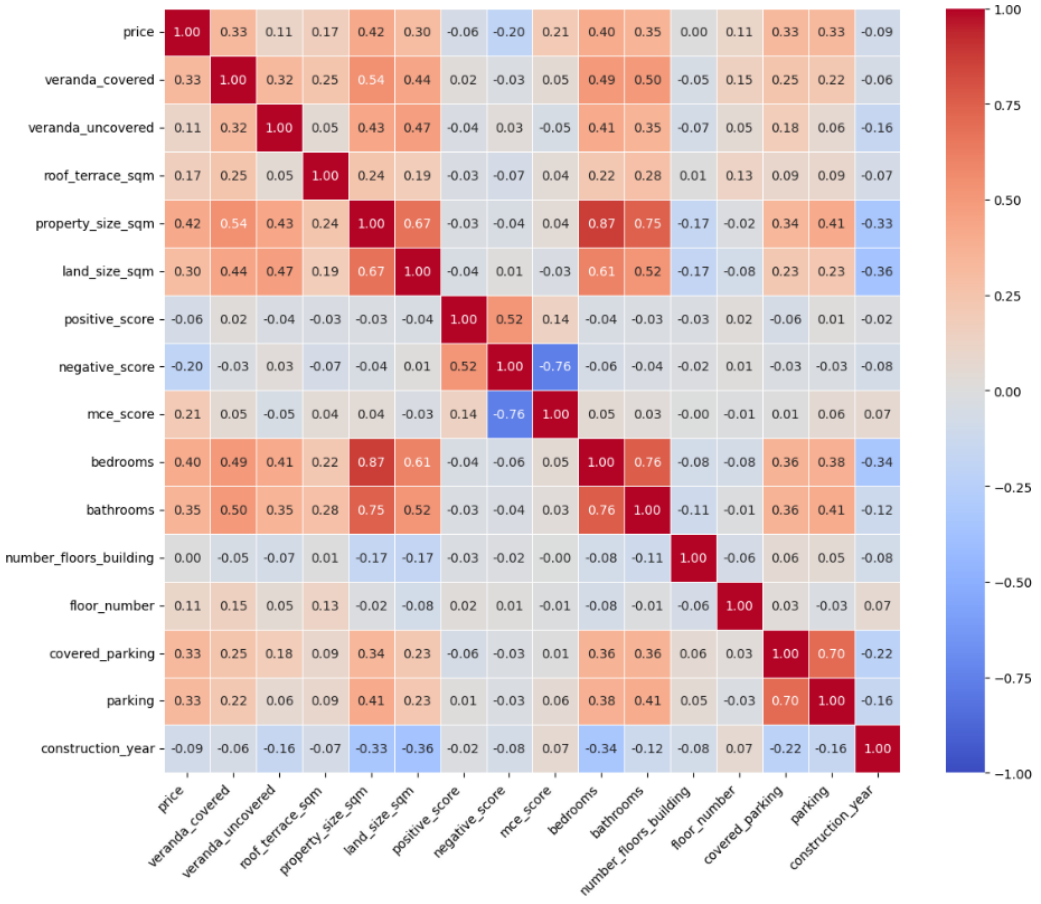


Figure 15: Spearman's Rank Correlation

Figure 14 presents the PC results. The weak to moderate associations indicate that the numerical variables are not linearly related. The only 2 strong relationships observed are between mce_score and negative_score (r = -0.64), indicating that higher negative_scores are associated with lower overall mce_scores, and between positive_score and negative_score (r = 0.76), suggesting that areas with a high concentration of amenities also tend to be located near disamenities. This finding is reasonable, as highly urbanised areas often contain a mix of desirable facilities such as schools, shops, and healthcare services, alongside disamenities such as busy roads and industrial areas.

Spearman's rank correlation was additionally employed to examine relationships among numerical variables (both continuous and discrete). Spearman's rank correlation is a nonparametric measure that assesses the strength and direction of a monotonic association between two variables using their ranked values. Unlike PC, it

does not assume normally distributed data or linear relationships and is less sensitive to outliers.

Figure 15 presents the result of Spearman's rank correlation. Compared with Pearson's correlation, Spearman's rank correlation revealed several relationships. Moderate positive associations were identified between price and property_size_sqm (r=0.42), bedrooms (r=0.40), and bathrooms (r=0.35). Weak-to-moderate positive relationships were also observed with veranda_uncovered (r=0.33), covered_parking (r=0.33), and parking (r=0.33). These findings suggest that larger properties with more bedrooms, bathrooms, and parking spaces generally command higher prices. The weak to moderate relationships suggest these variables increase property value, but not significantly.

Several strong relationships ($r \geq 0.5$) were also identified among the explanatory variables. The strongest positive association was found between property_size_sqm and bedrooms (r = 0.87), indicating that larger properties generally contain more bedrooms. A similarly strong positive relationship was observed between property_size_sqm and bathrooms (r = 0.75). Land_size_sqm was positively associated with property_size_sqm (r=0.67), bedrooms (r=0.61), and bathrooms (r=0.52). In addition, property_size_sqm and veranda_covered showed a moderate-to-strong positive relationship (r=0.54), suggesting that larger homes are more likely to have a larger uncovered veranda area. The strongest negative relationship identified by Spearman's correlation was between mce_score and negative_score (r = -0.76), consistent with Pearson's correlation.

Categorical Variables

Associations between categorical variables were also examined to identify additional relationships.

VARIABLE 1	VARIABLE 2	CRAMÉR'S V (V)
property_type	apartment_type	0.99
city	district	0.99
website	energy_efficiency_grade	0.98
district	floor_type	0.75
website	house_type	0.71
property_type	district	0.60
website	property_status	0.57
website	district	0.51
city	floor_type	0.49
district	parking_type	0.49
property_type	floor_type	0.47
district	house_type	0.41

Table 4: Cramer's V associations

To begin with, the Chi-square test of independence was applied to all pairs of the 12 categorical variables remaining in the dataset after the cleaning process: property_type, city, furnished, website, apartment_type, district, kitchen_type, energy_efficiency_grade, floor_type, house_type, parking_type, property_status. In total, 66 possible pairs were tested, of which 53 were considered significant with p-values<0.05.

For the pairs considered significant, Cramer's V was then calculated to determine the strength of association. Its values range from 0 to 1, where 0 indicates no association and 1 indicates a perfect association (Geert, 2017).

Table 4 presents associations between categorical pairs with V > 0.4. The table reveals near-perfect associations for two structurally expected pairs: property_type and apartment_type, city and district. The associations were expected, as most observations correspond to apartments and each district belongs to a single city.

Beyond these structural pairs, the strongest association was between website and energy_efficiency_grade (V = 0.98), indicating substantial dataset bias. Website also

showed a strong link with house_type (V = 0.71) and moderate links with property_status (V = 0.57) and district (V = 0.51).

District was likewise moderately to strongly associated with property_type (V = 0.60), as expected, since apartments are more common in urban areas. A less expected finding was the strong association between district and floor_type (V = 0.75), which may point to regional variation in construction practices, architectural conventions, or buyer preferences.

X	Y	TAU (Y|X)	TAU (X|Y)	MAX TAU
property_type	apartment_type	0.7996	0.9968	0.9968
city	district	0.1518	0.9914	0.9914
website	energy_efficiency_grade	0.6676	0.9684	0.9684
website	house_type	0.4977	0.6912	0.6912
district	floor_type	0.6250	0.1176	0.6250
website	property_status	0.3002	0.5802	0.5802

Table 5: Goodman–Kruskal Tau Test

VARIABLE 1	VALUE 1	VARIABLE 2	VALUE 2	COUNT	% OF TOTAL	ADJ. RESIDUAL	MISSING % (V1)	MISSING % (V2)
1	website	Website B	property_status	Completed	574	2.79%	−107.74	0.00%
2	website	Website B	energy_efficiency_grade	G	9034	95.95%	+95.42	0.00%
3	website	Website A	property_status	Completed	7832	38.11%	+80.97	0.00%
4	website	Website C	energy_efficiency_grade	A	258	2.74%	+79.26	0.00%
5	website	Website B	property_status	Under Construction	3668	17.85%	+77.07	0.00%
6	website	Website A	house_type	Villa	2239	37.96%	+62.53	0.00%
7	website	Website A	apartment_type	Apartment	4147	19.97%	−43.09	0.00%
8	website	Website B	apartment_type	Apartment	6643	31.99%	+41.20	0.00%
9	website	Website B	kitchen_type	Fitted Kitchen	7869	95.98%	+37.64	0.00%
10	city	Nicosia	property_status	Under Construction	1621	9.72%	+35.03	18.70 %

Table 6: 10 most significant adjusted residuals and contingency tables

Goodman-Kruskal tau was also calculated. According to Minitab (2026) the test measures the percentage improvement in predictability of the dependent variable given the value of another variable.

Table 5 presents the most significant results from the Goodman–Kruskal Tau test. To begin with, knowing apartment_type predicts property_type almost perfectly (Tau = 0.9968). Energy_efficiency_grade and website (Tau = 0.9684) indicate that, given energy_efficiency_grade, we can predict the website. The remaining pairs (website-house_type, district-floor_type, website_property_status) show more moderate, though still asymmetric associations.

Table 6 summarises the 10 most significant adjusted residuals and contingency tables found in the analysis. Website accounts for 9 out of 10 strongest associations, reinforcing the earlier finding that much of the observed structure reflects differences among data collection sources rather than underlying market characteristics.

Website B is strongly under-represented among "Completed" listings and over-represented among "G" energy efficiency. In contrast, Website A is strongly linked with "Completed" listings and Villas, but it is under-represented among apartments. Website C is represented mostly by A-efficiency houses. Lastly, Nicosia mostly sells "Under Construction" properties.

Binary Variables

For binary variables, we used a similar process to that for categorical variables to identify significant relationships. 48 binary variables were tested in all possible pairs (1128) using the Chi-square test of independence to determine whether the variables were associated or independent. Of all pairs, 862 pairs indicated a p-value < 0.05 and were considered significant.

For the significant pairs, Cramer's V was calculated to assess the strength of the relationship between the variables. The most significant pairs with an association greater than 0.50 are presented in Table 20 in the Appendix.

The variables internet_access and telephone_access showed the most significant pair (V=0.9261). The strong association between these variables may be due to their sharing the same contract/wiring; therefore, they seem to provide the same information.

Beyond that, the variables double_glazing, garden, pressurized_water_system, storage_space, internet_access, telephone_access, elevator, and en-suite are strongly or moderately associated with one another. Taken together, these variables describe a specific property type rather than 8 independently informative variables.

Chapter 4

Clustering and Within-Cluster Prediction

Chapter 4 presents the clustering and per-cluster prediction analysis of the Cypriot residential property market. The relationships identified in Chapter 3, together with the associations between price and the remaining explanatory variables examined in this chapter using Kruskal-Wallis and Mann-Whitney U tests and Random Forest, interpreted through SHapley Additive exPlanations (SHAP) values, inform the variable selection process, narrowing the candidate set down to those variables that best and most independently separate properties into distinct, meaningful clusters. Two clustering methods, K-means and Ward's method, are applied using the same variables and number of clusters to compare models. The chapter then presents the clustering results, describing the resulting property segments and their characteristics. Finally, it applies Decision Tree (DT), RF, Extreme Gradient Boosting (XGBoost) and Gradient Boosting Machine (GBM) models, individually and in combination, to predict property prices separately within each segment and examine which property characteristics most influence the price.

4.1. Clustering and Variable Selection

Dinh et al. (2021) defines cluster analysis as a core technique in data analysis that groups data objects into clusters based on their similarity. Objects within the same cluster are more similar to each other, while objects belonging to different clusters are dissimilar, according to a defined similarity or distance measure.

Applying clustering to the dataset identifies market segments within the Cypriot residential real estate market that share similar characteristics. Clustering reveals how certain property characteristics tend to co-occur, supporting comparisons of price behaviour across the resulting segments.

As clustering algorithms group observations based on a distance or dissimilarity metric computed across selected variables, the choice of which variables to include is critical to the validity of the resulting clusters. Including correlated variables effectively assigns greater implicit weight to the underlying dimension they both represent, distorting the distance calculation and causing bias. For this reason, the variable selection was guided by the relationship findings established in Chapter 3 (Section 3.5).

The variable price was intentionally excluded, as the objective of the clustering was to identify property segments based on their characteristics rather than their market value. However, the variables were required to show a meaningful individual relationship with price, low missing values, and, critically, not significant correlation or association with another variable.

To identify which property characteristics have a significant relationship with the target variable, relationships and the effect of the property characteristics on real estate prices were explored.

For the categorical variables with the fewest missing values, we performed the Kruskal-Wallis test, a nonparametric test used to determine whether three or more independent groups differ significantly. The test does not require normally distributed data (Ramachandran & Tsokos, 2015).

VARIABLE	H-STATISTIC	P-VALUE
property_type	4139.162000	0.00e+00
city	3577.285000	0.00e+00
house_type	906.445000	2.53e-189
property_status	513.989000	4.43e-111
furnished	132.232000	1.79e-28
energy_efficiency_grade	83.265000	2.97e-15

Table 7: Kruskal-Wallis Results: Price vs Categorical Variables

All six variables are highly significant under the Kruskal-Wallis test ($p < 0.05$ for each), confirming that price distributions differ across categories, as shown in Table 7. Property_type has the highest H value (H = 4139.2), followed by city (H = 3577.3), indicating that the two variables differ more in how their prices rank relative to each other. In contrast, the energy_efficiency_grade and furnished groups differ the least (H = 83.3 and H = 132.3, respectively).

Continuing with the binary variables, a Mann-Whitney U test was performed, a nonparametric test used to compare two independent groups (in this case, the presence and absence of the feature) (Mann & Whitney, 1947). Of the 48 binary variables tested, only 2 had p-values> 0.05: tennis_court and kitchen_appliances. The 2 variables did not allow rejection of the null hypothesis; therefore, the test did not detect any difference.

To assess the impact on price when all variables are included together, a RF with SHAP was used. RF is a machine learning model that builds decision trees. Each tree is trained on a slightly different random subset of the data and features, and their predictions are then averaged (Breiman, 2001). SHAP is a technique that breaks down each individual prediction into how much each feature contributed, positively or negatively, to push that specific prediction away from the average (Lundberg & Lee, 2017). The hyperparameters used were: number of estimators: 200, maximum depth: 15, minimum sample per leaf: 3, number of jobs: -1, and random state: 42.The model does not allow missing values; therefore, missing values were imputed with the median for numerical variables and with the category "Unknown" for categorical variables. The median was chosen as it is robust to outliers and skewed distributions.

The model explains about 41% of the price variance (R=0.4088). The Mean Absolute Percentage Error (MAPE) is 47.59%, meaning that the model's predictions are off by an average of about 47.6% of the actual price. The error is relatively large, indicating that the model misses much of what determines the price.

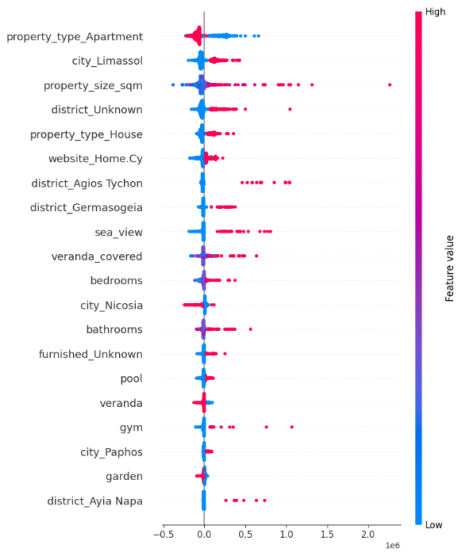


Figure 16: SHAP Summary Plot, RF Model

Figure 16 shows Property_type_Apartment as the most influential predictor, and its pattern indicates that being an apartment consistently pushed the price down. As expected, City_Limassol shows a positive trend, indicating a strong premium location in Cyprus's main coastal business city. On the other hand, the city_Nicosia demonstrated that properties in the capital are priced below the model's average. The property_size_sqm variable shows the widest spread of any feature, with a mild negative effect on price for small properties and a positive effect of over €+2.000.000 for extremely large properties. Districts near Limassol, Agios Tychon, and Germasogeia, along with Ayia Napa, each show red dots clustering at strongly positive SHAP values, reflecting well-known upscale or tourist-premium locations. Standard luxury features behave as expected: sea_view, pool, and garden all show red dots (feature present) skewing positive and blue (absent) sitting at zero or slightly negative, meaning the presence of the feature increases the price. In contrast, the absence of the feature does not affect the price, or slightly negatively affects it. Finally, bedrooms and bathrooms similarly indicated that more units drive up the price.

The resulting set of ten variables was therefore selected: 3 numerical (property_size_sqm, construction_year, veranda_covered), 4 categorical (property_type, city, house_type, and furnished), and 3 binary (pool, elevator, and air_condition).

4.2. K-means Clustering

K-means clustering is a partition-based clustering algorithm that divides a dataset into a predefined number of clusters by iteratively assigning observations to clusters and updating cluster centroids (MacQueen, 1967) . The goal is to partition the data so the sum of squared distances between each observation and its cluster centroid is minimised. This ensures that data points within the same cluster are as similar as possible, while points in different clusters are as distinct as possible. (Ahmed et al., 2020).

Using the variables mentioned in the prior section, a K-means clustering model was trained. To handle missing values, numerical variables were imputed with the median, and for categorical variables, a new dummy variable, "Unknown," was created. Numerical variables were standard-scaled, and categorical variables were one-hot encoded. The number of clusters was chosen using Gower Distance with silhouette score and the elbow method. Gower distance measures similarity between observations described by different variable types (Liu et al., 2024). The silhouette coefficient measures how well each observation fits its own cluster compared with other clusters, while the elbow method identifies the point where increasing the number of clusters produces only a small reduction in within-cluster variation (Sen & Cohen, 2025).

VARIABLE	STANDARD DEVIATION OF CLUSTER MEANS (SCALED DATA)
property_size_sqm	3360.493235
veranda_covered	27.065305
construction_year	13.916093

Table 8: Standard Deviation of cluster mean on scaled data

FEATURE	TYPE	CHI² SCORE	P-VALUE
house_type	Categorical	84560.959996	0.000000e+00
property_type	Categorical	3566.457112	0.000000e+00
elevator	Binary	2111.816563	0.000000e+00
air_condition	Binary	982.281122	2.468616e-211
pool	Binary	489.676141	1.145054e-104
furnished	Categorical	423.278464	2.593539e-90
city	Categorical	17.061697	1.880317e-03

Table 9: Chi-square test for clustering

Since the Gower-silhouette did not yield high values (0.233), an elimination process was used based on statistical tests performed for each variable type to determine whether the variables' values vary across clusters. For numerical variables, the mean value of each variable and then the standard deviation of those per-cluster means were computed within each cluster. A high score means that the variable takes noticeably different average values across the clusters.

As shown in Table 8, property_size_sqm has a much higher value than the other variables. For that reason, the other variables were eliminated.

For categorical variables and binary variables, a chi-square test was performed to measure how strongly each categorical variable's distribution is associated with cluster membership.

As shown in Table 9, the cluster membership is driven overwhelmingly by property structure rather than features and location. The house_type variable is the most important for clustering; clusters are essentially organised around the type of house the observation is. Property_type and elevator follow, indicating they are both meaningful for separating clusters. Pool, furnished and city, on the other hand, showed weak Chi-square values and were eliminated.

Continuing the elimination process, the best final clustering was obtained with only 2 variables: property_size_sqm and property_type.

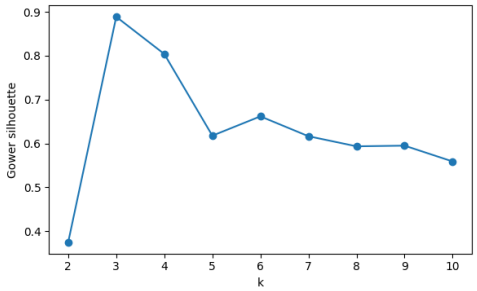


Figure 17: Gower-silhouette

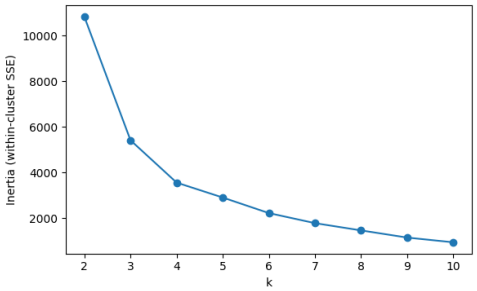


Figure 18: Elbow Method

To determine the optimal number of clusters (k), plots of the Gower-silhouette and Elbow method were inspected (Figures 17 and 18). Both methods suggested 3 clusters: the Gower-silhouette is maximised at k = 3 (0.889), while the Elbow method showed a noticeable change in the rate of reduction in within-cluster variation at k = 3.

4.3. Ward's Method

Using the best variables identified by K-means, Ward's Method was also developed to compare the model's performance. Ward's Method is a hierarchical clustering technique that merges clusters based on a minimum variance criterion. It is a hierarchical agglomerative clustering algorithm, first proposed by Ward in 1963 (Murtagh & Legendre, 2011). Ward's method decides which clusters to merge at each step by minimising the increase in total within-cluster variance (sum of squares) (Murtagh & Legendre, 2011).

To compare the 2 clustering processes, the same variables as in K-means were used, as they are the most informative for this process. Figure 19 shows the dendrogram of the partitioning process using Ward's Method. Visual inspection of the dendrogram, together with Table 10, supports k = 3 or 4 table . K = 3 was selected as it is preferable for comparing the 2 clustering methods used.

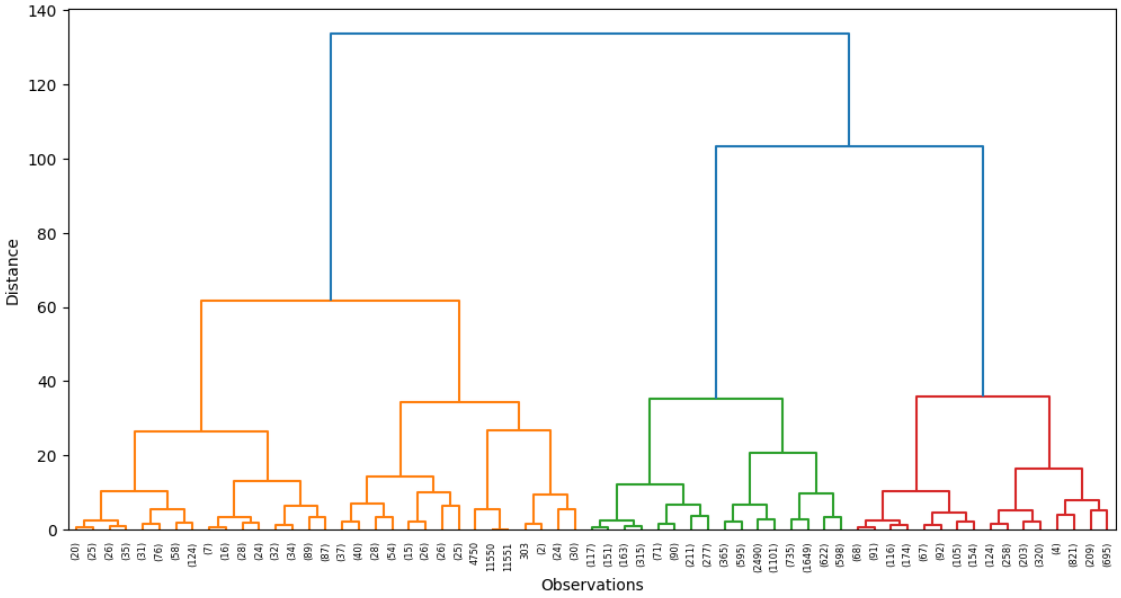


Figure 19: Dendogram of Ward's Method

K	GOWER - SILHOUETTE
2	0.351
3	0.845
4	0.846
5	0.763
6	0.551
7	0.552
8	0.552
9	0.614
10	0.568

Table 10: Gower Silhouette for Ward's Methods

4.4. Clustering Results

To assess the robustness of the clustering solution, the results from K-means need to be compared with those from Hierarchical Ward clustering to identify the best clustering method. The comparison considers two dimensions: validation metrics and substantive composition of clusters.

To begin with the validation metrics, the two clustering methods provided nearly identical results, with K-means slightly outperforming: K-means yielded a 0.889 Gower– Silhoute and Ward's method 0.845.

Continuing with the cluster composition, the two clustering methods produced broadly similar distributions. The main point of divergence is Cluster 3, the Luxury/Large segment. As shown in Table 11, K-means defines a tighter, luxury cluster of 673 properties, whereas Ward's method includes an additional 350 borderline properties (182 from the Mid-Size cluster and 168 from Standard Urban Apartments). Overall, the two methods agree with 97.5% of the data.

K-means is known for its computational efficiency; it scales well to large datasets, produces tighter and more exclusive cluster boundaries, and can be rerun with different k-values without recomputing from scratch. For these reasons, K-means was selected as the best clustering method for studying the real estate residential market. K-means also slightly outperformed Ward's method in Gower-silhouette scores, indicating better cluster separation. Figure 20 illustrates the K-means data partition, and Table 12 summarises the cluster characteristics.

CLUSTER NUMBER	CLUSTER LABEL	K-MEANS	WARD'S METHOD
1	Standard Urban Apartments	n = 9,722 (69.09%)	n = 9,550 (67.86%)
2	Mid-Size Properties	n = 3,679 (26.14%)	n = 3,501 (24.88%)
3	Luxury/Large Properties	n = 673 (4.78%)	n = 1,023 (7.27%)

Table 11: Cluster composition per method

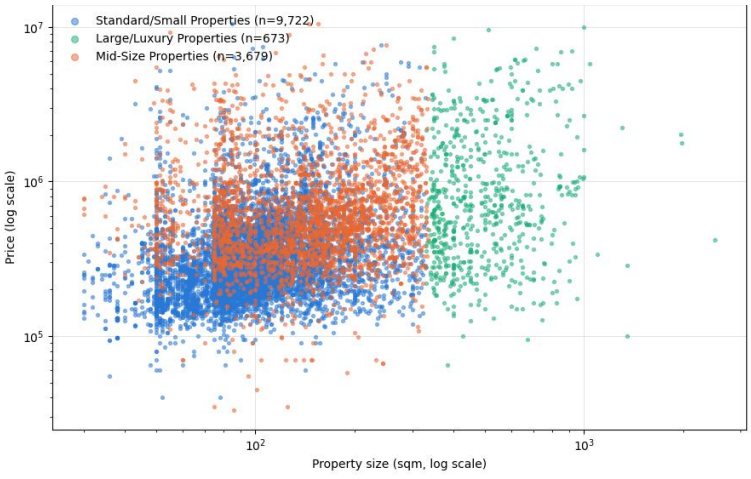


Figure 20: K-Means Clusters (k=3): Property Size vs Price

CLUSTER	LABEL	N	AVG SIZE	AVG PRICE	DEFINING FEATURE
1	Standard Urban Apartments	9,722	107.3 sqm	€421,563	Apartment-dominant, affordable
2	Mid-Size Properties	3,679	153.1 sqm	€904,784	House-dominant, mainstream
3	Luxury/Large Properties	673	522.3 sqm	€1,243,040	House-dominant, high-amenity, premium

Table 12: Summary of the Clusters' characteristics using K-Means

Cluster 1 can be titled "Standard Urban Apartments" as it contains 100% apartments, of which 89.3% are typical apartments and 6.9% are penthouses. It represents 69.09% of the properties. The average property size is 107.3 sqm, the average price is €421,563, and the typical range (IQR) is €210,000 to €440,000, with a median of €290,000. Cluster 1 properties are mainly located in Limassol (35.8%), Larnaca (26.5%) and Nicosia (23.8%), the three most populated cities of Cyprus. Most apartments are completed (44.8%); however, there are also apartments off-plan (31.6%) and under construction (20.7%). Nearly all the apartments are unfurnished, with only 6.6% being furnished – fully or partially. 85.1% of the apartments have a veranda of approximately 21.6 sqm (median), 38.9% have an elevator in the building, 28.7% have a common or private pool, 48.7% have a garden, 44.7% have double glazing and 19.2% have luxury fittings. Cluster 1 can be mainly characterised as mass-market new-build apartments, mainly in urban centres.

Cluster 2 represents Mid-Size Family Houses. This cluster contains 100% houses: 46.6% are villas, 22.9% are detached, and 9.1% are mansion/villas. The average house size is 153.1 sqm, and the average price is €904,784, with a typical range of €385,000 to €971,500 and a median of €575,000. The houses are located mostly in Paphos, Limassol, and Nicosia (28.6%, 22%, and 18.1%, respectively). The houses are mostly complete (64.1%). 65.8% of the houses have a veranda slightly larger than the apartments, averaging 24.35 sqm; 36.6% have a pool (private or shared), 28.4% have a garden, 23.8% have double glazing and 13.5% have luxury fittings. The average number of bedrooms is 3, and the average number of bathrooms is 2-3. The average construction year is 2019. The cluster can be characterised as mainstream residential housing across touristic and suburban areas, typical for families.

Luxury/Large Properties are represented in Cluster 3. This cluster contains both houses and apartments, with apartments representing only 30.9% of the cluster population. 69.1% are houses, of which 49.5% are detached houses and 31.8% are mansion/villas. The average size is 522.3 sqm, and the average price is €1,243,040, with a median of €697,050 and a notably wide typical price range of €369,750 to €1,500,000. The properties are located mostly in Limassol (31.9%); however, 25.8% are in Paphos and 21.3% in Nicosia. 52.1% are completed, and 14.9% are resales. The houses are large and/or luxurious: 70% have a pool, 47.1% a garden, 36.6% double-glazed windows, 30.8% luxury fittings, 8.3% a gym, and 9.4% a sea view. The most common configuration is 4 bedrooms, 4 bathrooms and a construction year of 2025.

4.5. Per-Cluster Price Prediction

Rather than fitting a single price model across the entire Cyprus real estate dataset, independent prediction models were produced for each K-means cluster. This approach assumes that different groups of properties are influenced by different price determinants. This section aims to identify what drives price in each segment by predicting the price of 20% of the data (test set) in each cluster.

To predict property prices within each cluster, price was the target variable, and predictors were selected based on their established associations with property prices. The predictors are: 7 numerical (property_size_sqm, parking, construction_year, floor_number, veranda_covered, bedrooms, bathrooms), 7 categorical (property_type, city, district, property_status, house_type, furnished, apartment_type) and 6 binary (veranda, garden, double_glazing, pool, elevator, luxury_fittings). Most binary variables were extremely sparse; therefore, they were not included.

The target variable was transformed to log(price) as the price distribution is right-skewed, and all its missing values were dropped. Missing values in numeric variables were imputed with the median, and categorical variables were replaced with

"Unknown". Numerical values were standard-scaled, and categorical values were one-hot encoded.

The dataset was split into 80% training and 20% test. 5-fold Cross-Validation was used: the training data was split into 5 equal parts; 4 were used to train, and 1 was held out for validation; the process was repeated for all 5 combinations and averaged (Kohavi, 1995).

The chosen models to predict per cluster were used individually and in combination (2 at a time) to achieve the most suitable model per cluster. The description of each model is as follows:

- DT: The model splits the data into branches by asking a sequence of yes/no questions, ending in a predicted price per branch (Breiman et al., 1984).
- RF: The model trains many decision trees, each on a random subset of rows/features, and averages their predictions (Breiman, 2001).
- XGBoost: The model builds many trees, but sequentially: each new tree corrects the errors of the ones before it ("gradient boosting"), rather than averaging independent trees (Hakkal & Lahcen, 2024).
- GBM: The model builds decision trees, with each new tree learning to correct the prediction errors of the previous ones (Friedman, 2001).

MODEL	HYPERPARAMETERS SEARCHED
DT	max_depth: [3, 5, 8, 12, None] · min_samples_leaf: [1, 3, 5, 10] · min_samples_split: [2, 5, 10]
RF	n_estimators: [100, 200, 300] · max_depth: [5, 10, 15, None] · min_samples_leaf: [1, 3, 5] · max_features: ["sqrt", "log2", None]
XGBOOST	n_estimators: [100, 200, 300] · max_depth: [3, 5, 7] · learning_rate: [0.01, 0.05, 0.1] · subsample: [0.7, 0.85, 1.0] · colsample_bytree: [0.7, 0.85, 1.0]
GBM	n_estimators: [100, 200, 300] · max_depth: [3, 5, 7] · learning_rate: [0.01, 0.05, 0.1] · subsample: [0.7, 0.85, 1.0] · max_features: ["sqrt", "log2", None]

Table 13: Models' Hyperparameters

The models were tuned via RandomizedSearchCV, with the hyperparameter search ranges presented in Table 13, to optimise model performance and avoid overfitting. Two variants were compared per cluster: full data and outlier-streamed data (rows with prices above the cluster's IQR upper fence were eliminated). Outlier removal was performed to assess whether outliers affect model performance. As a result, 8.35% of cluster 1, 10.99% of cluster 2, and 9.18% of cluster 3 were eliminated.

CLUSTER	VARIANT	BEST MODEL	R^2	MAE (€)	MAPE
Standard Urban Apartments	Full data	RF+XGBoost	**0.359**	**123,107**	**23.98%**
Standard Urban Apartments	Outlier-trimmed	XGBoost	0.112	136,984	22.29%
Mid-Size Properties	Full data	RF+GBM	**0.346**	**415,510**	**44.46%**
Mid-Size Properties	Outlier-trimmed	XGBOOST	0.005	458,336	40.10%
Luxury/Large Properties	Full data	GBM	**0.659**	**454,225**	**37.80%**
Luxury/Large Properties	Outlier-trimmed	DT+GBM	0.284	682,942	45.27%

Table 14: Accuracy matrix per cluster prediction

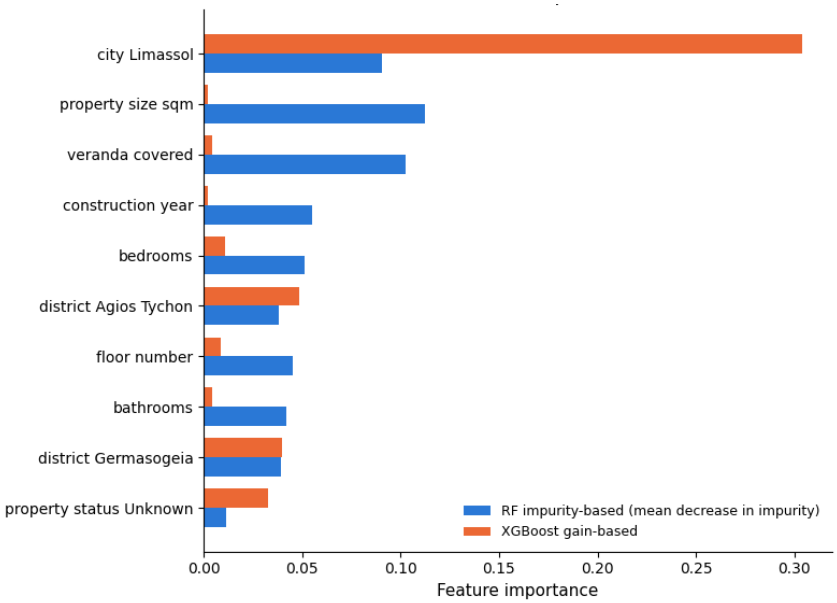


Figure 21: 10 most important features Standard Urban Apartments

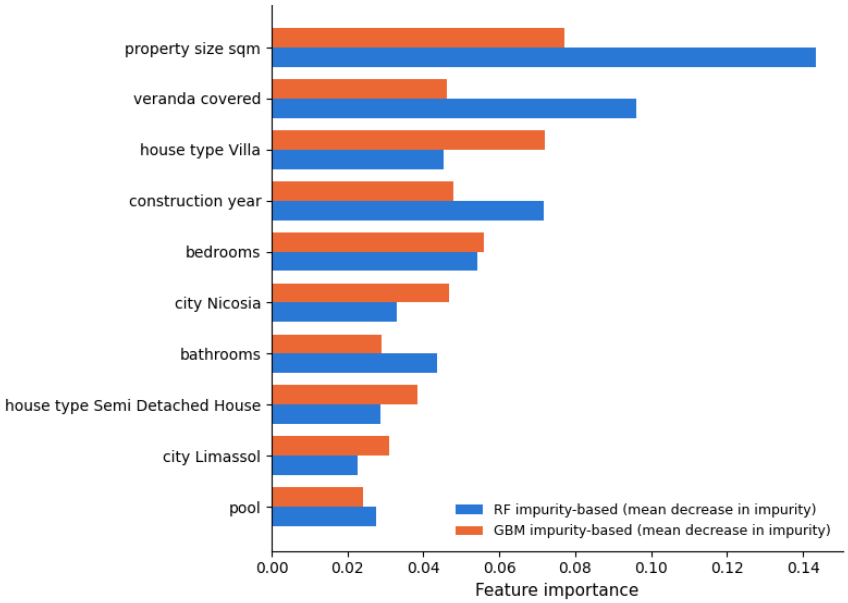


Figure 22: 10 most important features Mid-Size Properties

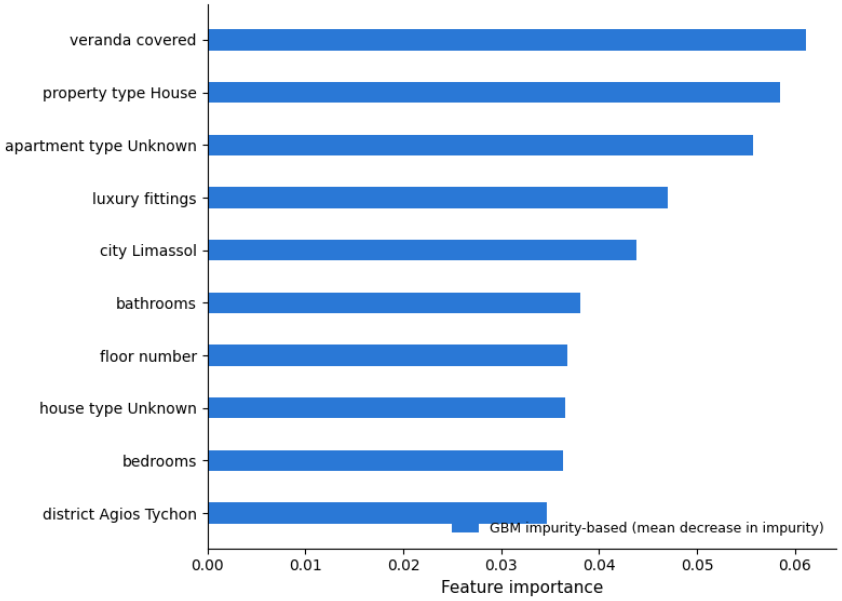


Figure 23: 10 most important features Luxury/Large Properties

To evaluate model performance, the coefficient of determination (R^2), MAE, and MAPE were calculated. R^2 measures how much price variation the model explains (James, 2021). MAE (mean absolute error) is the average size of the model's prediction error in raw € terms, and MAPE is in percentage terms (James,

2021). The outlier-removal process consistently worsens R^2 in every cluster; therefore, removing the outliers does not produce better results.

As summarised in Table 14, the best-performing model for Standard Urban Apartments is the RF and XGBoost prediction, explaining 35.9% of price variation with a mean error of 23.98%. For Mid-Size Properties, the RF and GBM combination was the best model, explaining 34.6% of price variation with a mean error of 44.46%. Luxury/Large Properties have the strongest fit: GBM explains 65.9% of price variation, with a mean error of 37.8%.

Figures 21-23 illustrate which features carry the most weight in each cluster's prediction. Because the best models for Standard Urban Apartments and Mid-Size Properties are an average of 2 different models, both models' drivers are shown separately.

As shown in Figure 21, for Standard Urban Apartments, the two models show different drivers of price: XGBoost shows that Limassol city affects price the most, whereas RF states that price is driven by a combination of features (property size, covered veranda, and Limassol), with the size being the most important driver. Taken together, location and property size are consistently the two leading factors.

For Mid-size Properties, property size is clearly the dominant feature for both models, followed by covered veranda, which is the second most important feature, as shown in Figure 22. Villa is the third most influential driver, followed by construction year, bedrooms, and city (Nicosia, Limassol), which also drive price, but to a lesser extent.

For Luxury/Large Properties, a combination of features drives the property price, as their weights are similar, as shown in Figure 23. Covered veranda is at the top, followed closely by the property being a house, luxury fittings, and being located in Limassol. Bathrooms, floor number, and bedrooms are also important; however, they affect the price to a lesser extent.

Because the clustering was based on property_size_sqm and property_type, their importance in within-cluster prediction appears lower than it actually is. Nevertheless, these two variables remain the most effective for partitioning the Cyprus residential real estate market.

Taken together, the per-cluster prediction results show that veranda covered is consistently important to all three segments. The results also confirm that each property segment is driven by a different combination of factors: Standard Urban Apartments' prices are driven by location and size; Mid-size Properties are led by size and covered veranda, with house type villa as a secondary factor; and Luxury/Large Properties spread importance across veranda size, the property type, and luxury-specific amenities.

Chapter 5

External Factors and Forecasting of the HPI

Chapter 5 investigates the Cyprus HPI at a macro level, examining the extent to which macroeconomic, labour market, and housing supply factors show significant relationships with the HPI over time. The chapter begins with an introduction to the external factors used in the analysis, then addresses the required time-series preprocessing, including transforming monthly series to quarterly, imputing internal missing values, and assessing and differentiating non-stationarity across the series.

The relationships between the HPI and the candidate predictors were then identified using Granger causality tests and Cross-Correlation Function (CCF) analysis. Finally, two forecasting approaches were compared: an ARDL conditional model that predicts the HPI from external factors whose feature paths are forecast (using Holt-Winters, Random Walk with Drift, and ARIMA), and univariate models (the same three methods) that predict the HPI from its own past values alone.

5.1. External Factors

Using the HPI, this study also aims to examine the external factors that affect residential property prices in Cyprus. Specifically, the analysis investigates whether, and to what extent, a set of macroeconomic, labour market and housing supply indicators exhibit a significant relationship with HPI, using time-series econometric techniques.

As shown in Table 21 in the Appendix, 5 variables spanning two frequencies (quarterly and monthly) were compiled from the Cyprus Statistical Service. These variables were grouped into 4 conceptual categories: Housing Market Indicators,

Macroeconomic Indicators, Labour Market and Income Indicators, and Housing Supply & Construction Indicators. Demographic factors were not included because they were available only at an annual frequency.

5.2. Time series Preprocessing

Time series preprocessing is the systematic cleaning, transforming and structuring of sequential data. Preprocessing is essential as raw historical data contain missing values, non-stationarity and seasonal noise that might violate core econometric assumptions and compromise model accuracy (Tunnicliffe Wilson, 2016)

As an initial step in the preprocessing pipeline, a local averaging technique was applied to fill missing points in the time series. Using a backwards 3-period rolling mean, 17 internal gaps were imputed, replacing each missing value with the average of the three most recent available observations (Hyndman & Athanasopoulos, 2018). As per Nielsen (2019) this method is preferred because it relies exclusively on past observations, avoiding future bias. Additionally, a short 3-period window tracks each series' local level and short-run trend, and iterative imputation ensures that within blocks of consecutive missing values, each newly imputed value could itself serve as an input for the next, ensuring a smooth transition.

Figure 24 illustrates the time series plots of the external factors. The HPI reached its lowest point around 2015 before entering a sustained upward trend that rose sharply beyond 150 by 2025, while average earnings percentage change mostly fluctuates between -1% and 3%, suffering a brief drop below -2% during the COVID-19 period before rebounding. GDP shows heavy quarterly swings between -2 and 8%, with significant downturns in 2006, 2020 and 2024. The labour force expanded from 280,000 in 1998 to over 500,000 by 2025. Finally, the number of permits declined after peaking around 2007 to a low near 2015, then recovered steadily before extreme fluctuations from late 2024 to 2025.

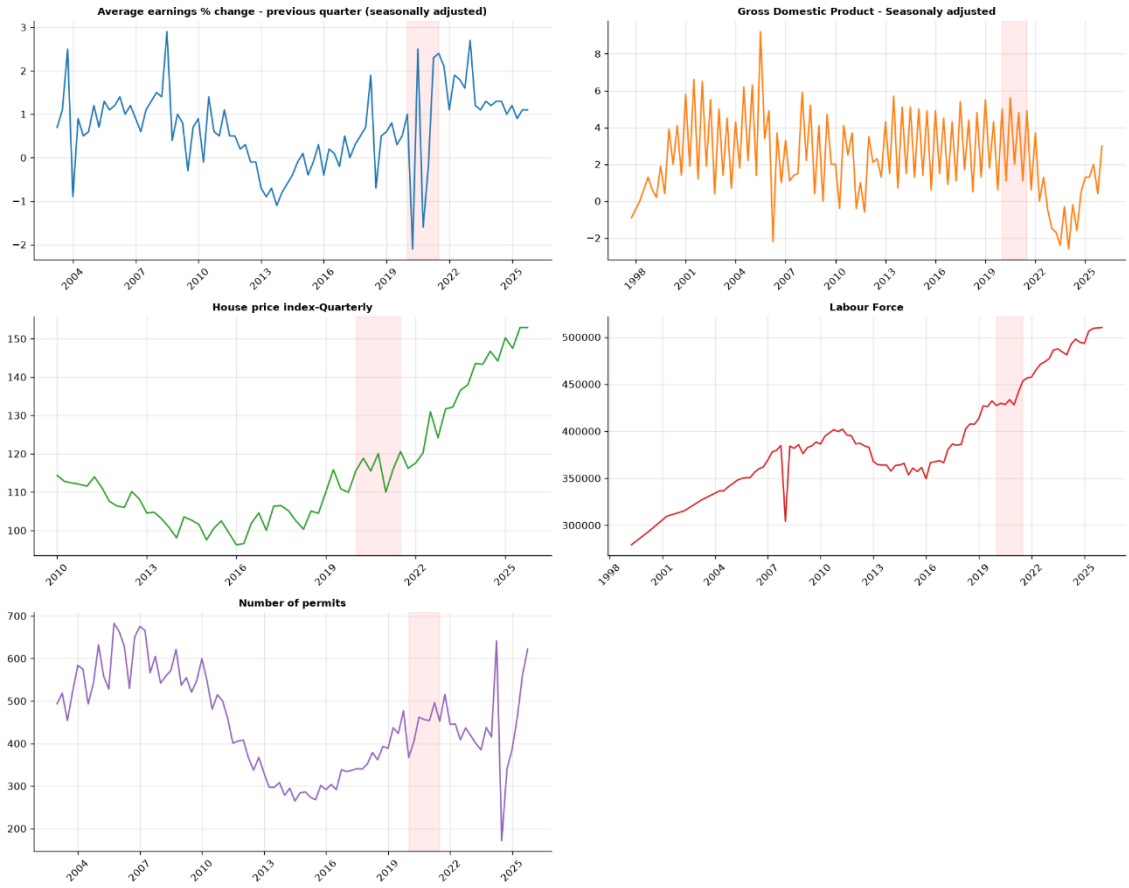


Figure 24: Average eargings % change, GDP - seasonaly adjusted, HPI, Labour Force, Number of permits Time-series plots

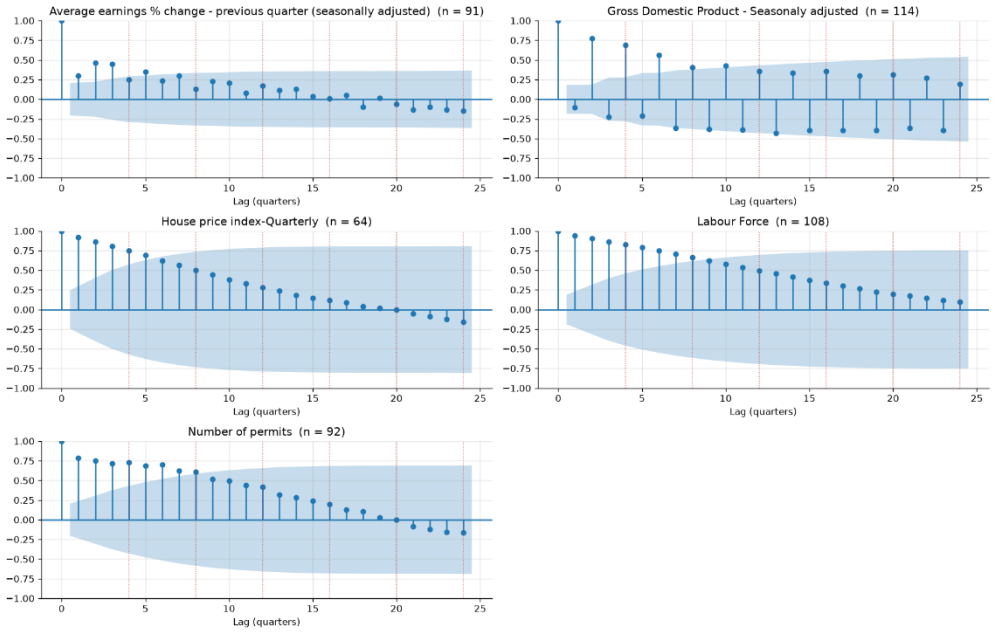


Figure 25: ACF Plots

The autocorrelation function (ACF) measures how strongly a series correlates with its own past values at each lag k, and its shape is diagnostic: a slow, persistently positive decay points to a trend or unit root, repeated peaks indicate seasonality, and a quick drop towards zero points to stationarity (Hyndman & Athanasopoulos, 2018). Any bar extending beyond the shaded $\pm 1.96\sqrt{n}$ band (the range expected under pure random noise) is statistically significant at the 5% level (Bartlett, 1946).

Figure 25 shows the ACF plots of the time series. The HPI, Labour Force, and Number of Permits exhibit non-stationarity, with a slow, near-linear decay that stays positive for roughly 8-12 lags. No seasonality is visible in the panels, reflecting trend dominance. Average earnings percentage change decays into the confidence band after only 2-3 lags. GDP is seasonally adjusted; however, it shows a decay pattern of negative values at odd lags and positive values at even lags.

Lag plots are a graphical tool used in time-series analysis to examine the relationship between observations of a series and their past values. A lag plot represents the current observation (Y_t) against the earlier observation, (Y_{t-k}), where k denotes the lag (Hyndman & Athanasopoulos, 2018). Lag plots were employed as a graphical diagnostic to investigate potential seasonality and temporal dependence in the quarterly time series of HPI, Labour Force, and Number of Permits using 8 lags (Hyndman & Athanasopoulos, 2018).

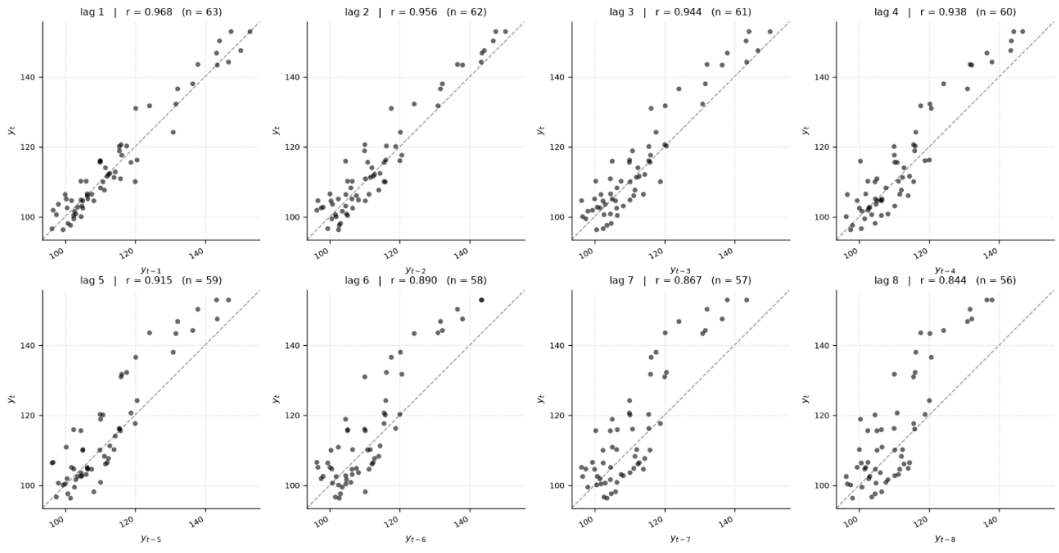


Figure 26: Lag Plots of HPI

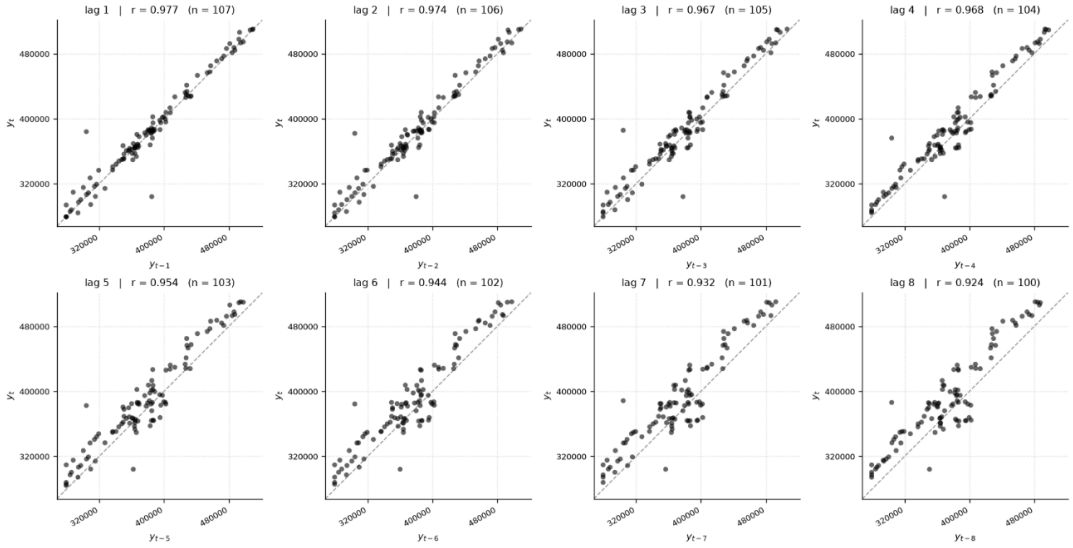


Figure 27: Lag Plots of Labour Force

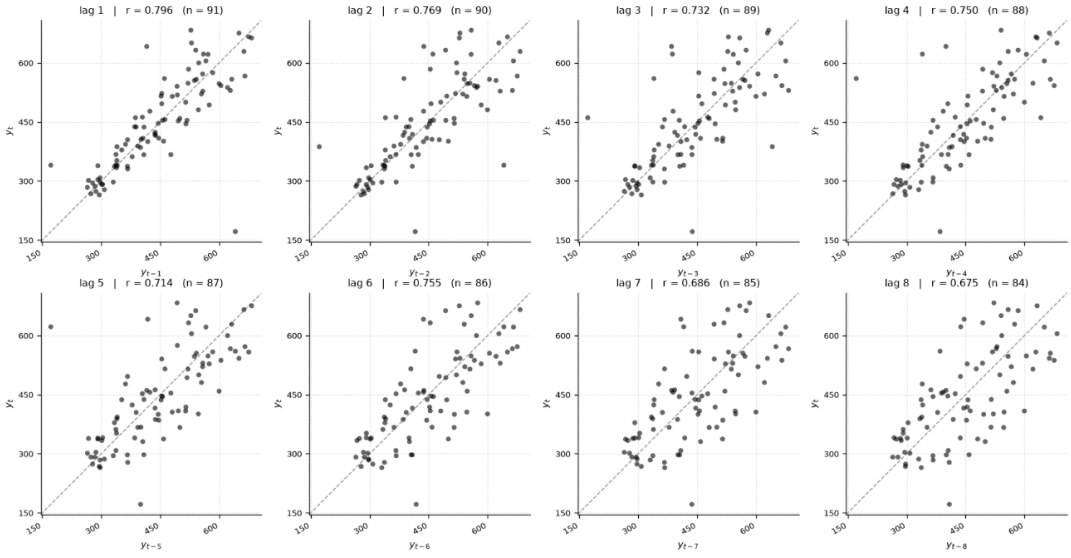


Figure 28: Lag Plots of Number of permits

The lag plots in Figures 26-28 show each time series plotted against its own values one to eight quarters earlier. For the HPI and the Labour Force, the points lie in a narrow band along the 45-degree reference line (r = 0.844 and r = 0.997 at lag 1, and the band widens slowly as the lag grows, with the correlation still at 0.844 and 0.924, respectively, after two full years. Both series change minimally from quarter to quarter and remain closely tied to their own past values, showing a strong trend but not seasonality. As shown in Figure 28, the Number of permits is more dispersed and less

persistent than the other two series, with the correlation decreasing from (r = 0.796) at lag 1 to (r= 0.675) at lag 8. One unusually low observation, corresponding to the COVID-19 pandemic shutdown, is clearly visible across the lag plots. For all three series, no stronger relationships are observed at lags 4 and 8, providing no visual evidence of seasonality.

Each series was then tested for stationarity, since several of the econometric models require stationary inputs. Stationarity was assessed using the Augmented Dickey–Fuller test (ADF test) at the 5% significance level, where the null hypothesis states that the time series is non-stationary (Said & Dickey, 1984). Consistent with the ACF patterns, the House Price Index, Labour Force, Number of permits and Average earnings percentage change were non-stationary as expected. These four time series were therefore first differenced (Nelson & Plosser, 1982).

As shown in Figure 29, after first differencing, the autocorrelations of all four series collapse into the 95% confidence band within one or two lags. The ACF plots show no slow decay and no near-unity spike, confirming that the trend has been removed and the differenced series are now stationary. The only exception is Labour Force, whose autocorrelation at annual lag 4 (≈0.27) lies just outside the marginal band.

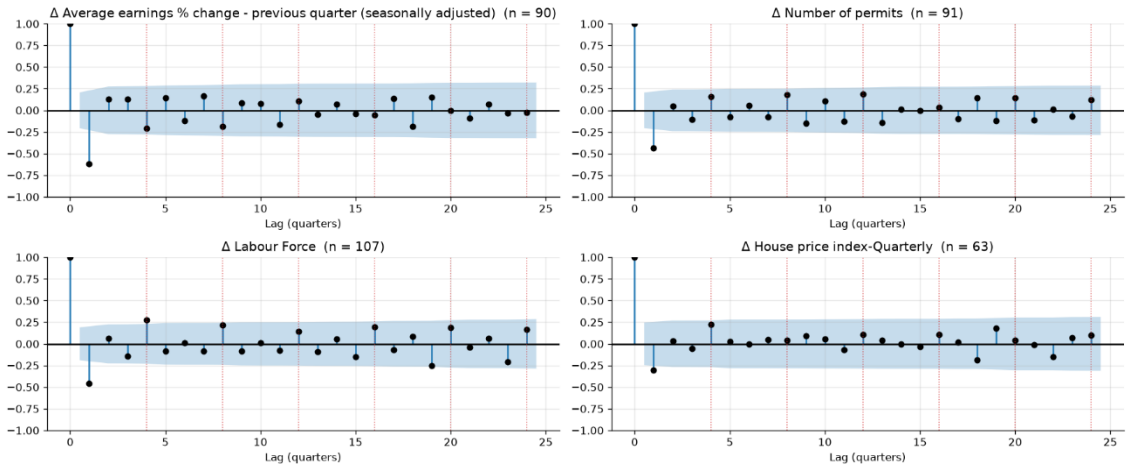


Figure 29: ACF on the first difference

5.3. Granger Causality and Cross-Correlation Analysis

According to Granger (1969), Granger Causality is a statistical concept that assesses whether the past values of one time series contain information that helps predict another. Formally, a variable X "Granger-causes" the HPI if adding historical values of X significantly improves the predictions of HPI compared to using past HPI data alone. This improvement is evaluated using an F-test of the joint significance of X's lagged terms.

The test requires stationary inputs, so we used stationary versions of each series. The HPI was tested directly against the time series, using lag lengths of 1 to 4 quarters (approximately 1 year). All tests were conducted at the 5% significance level.

The results revealed that two predictors Granger-caused the HPI at the 5% significance level: Average earnings percentage change and Labour Force. The non-significance of building permits remains somewhat unexpected.

As shown in Table 15, Average earnings percentage change is significant at every lag, indicating that recent wage growth movements carry predictive information for the HPI over horizons from one quarter up to a year. Labour force is significant at lags 3 and 4, so workforce changes Granger-cause HPI with a delay of about a year.

A CCF analysis measures the correlation between two stationary time series at varying lags, allowing the identification of the specific number of periods by which one series leads or lags the other (Hyndman & Athanasopoulos, 2018). The CCF was computed on the stationary version of each variable, with correlation coefficients calculated across symmetric windows of ±8 quarters (approximately 2 years) for each predictor. A positive lag indicates that the predictor leads HPI, while a negative lag indicates that HPI leads the predictor (Hyndman & Athanasopoulos, 2018). To assess statistical significance, each correlation was benchmarked against a 95% confidence interval of $\pm 1.96/\sqrt{n}$, where n is the number of aligned observations; correlations exceeding this threshold at a given lag were considered statistically significant.

Figure 30 shows the CCF results. Most bars fall inside the ±0.25 band. GDP is significant only at negative lags (-2 and -4 quarters), meaning HPI leads GDP movements. Average earnings percentage change is significant mainly at lag +1 (r ≈ 0.40) and +2 (r ≈ -0.44); however, the sign of r is inconsistent, indicating no single stable directional effect. The number of permits is significant at ±7 quarters (both positive), indicating a weak two-way association. Finally, labour force shows a significant positive correlation at lags -8, 0 and 4; therefore, no single stable directional effect is shown.

PREDICTOR	P (LAG 1)	P (LAG 2)	P (LAG 3)	P (LAG 4)
Average earnings percentage change (SA)	**0.003**	**0.006**	**0.003**	**0.0003**
Labour Force	0.076	0.067	**0.039**	**0.026**
GDP (SA)	0.596	0.164	0.198	0.612
Number of permits	0.199	0.394	0.307	0.538

Table 15: Granger causality predictors – HPI

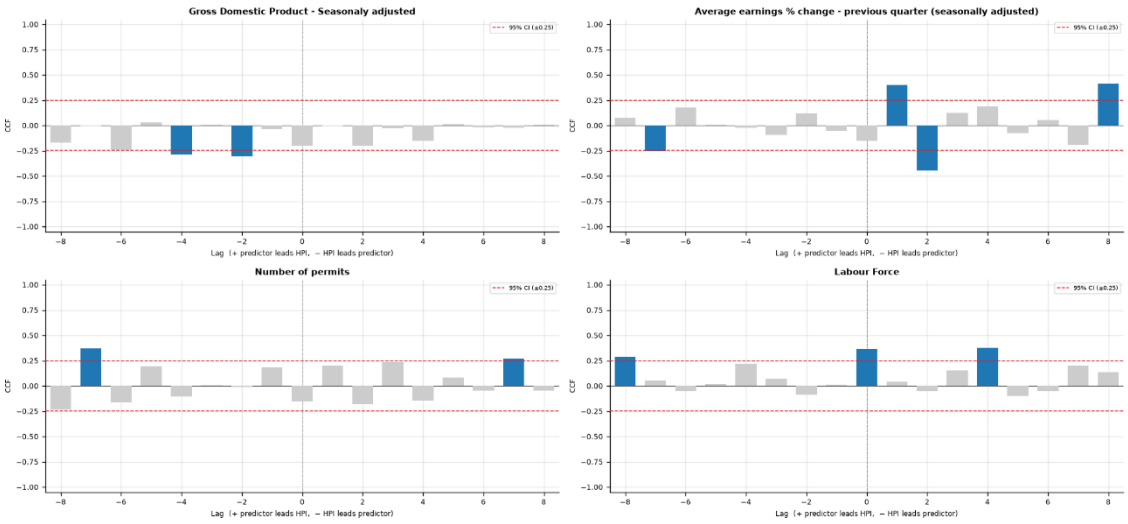


Figure 30: Cross Correlation with HPI

5.4. Forecasting the HPI

The final part of the analysis forecasts the HPI six quarters ahead, from 2026Q1 to 2027Q2, to assess whether the HPI is better predicted using external indicators or its own past values. Two approaches are therefore compared: an ARDL model that

forecasts HPI using GDP, Labour Force, and Average earnings % change, and univariate models that predict HPI based on its own past values. To evaluate predictive performance a test set of the final 6 quarters is used, matching the 6-quarter forecasting window.

The ARDL models link HPI to the other external drivers by comparing lags of HPI with the current and lagged values of the explanatory variables, capturing both short-run and long-run dynamics and relationships (Pesaran et al., 2001). According to Pesaran et al. (2001), the ARDL is valid when the variables are a mixture of I(0) and I(1) (stationary and first-difference stationary), and it performs well in small samples, which suits the constraint of only 64 quarterly HPI observations (2010-2025).

The model serves two purposes in this study: testing, through the bound test, whether a stable long-run relationship exists between the HPI and its drivers, and forecasting HPI conditional on the future path of those drivers using the variables in their original level. An ARDL model is:

$$HPI_t = c + \sum_{i=1}^{p} \varphi_i HPI_{t-1} + \sum_{j=1}^{k} \sum_{l=0}^{qj} \beta_{j,l} X_{j,t-l} + \varepsilon_t$$

where $X_{j,t}$ is the j-th regressor, c_0 an unrestricted constant with no trend term (the Pesaran–Shin–Smith "Case III" specification), and ε_t a white-noise error.

The ARDL bounds test was conducted at the 5% significance level using asymptotic critical values (Pesaran et al., 2001):

$$F_{I(1)} = 4.35 < F_{stat} = 5.21$$

Since the calculated F-statistic exceeds the upper critical bound of 4.35, the null hypothesis of no cointegration is rejected. The ARDL bounds test provides evidence of a long-run relationship among the variables.

Because the ARDL is conditional on its regressors, forecasting HPI six quarters ahead requires forecasts of GDP, Labour Force and Average earnings percentage change over the horizon. Each was projected independently using ARIMA (order selected by Akaike Information Criterion), Holt-Winters, or Random Walk with Drift, and each set of predictor paths was substituted into the fitted ARDL. The Akaike Information Criterion (AIC) is a model-selection score that rewards goodness of fit while penalising the number of estimated parameters k, so that the model with the lowest AIC offers the best expected out-of-sample predictive accuracy (Akaike, 1974).

A combination of the models was also used to select the best model for each driver. As per Hyndman and Athanasopoulos (2018), Holt-Winters captures trend and seasonal patterns through exponentially weighted smoothing; a Random Walk with Drift assumes future values follow the series' historical average growth rate plus random noise, and the ARIMA model leverages past values (autoregression) and past forecast errors (moving averages) to predict future prices.

The accuracy of the forecasting was assessed using the Mean Absolute Error (MAE) and MAPE. MAE is the average size of the forecast errors in the series' own units, while the MAPE expresses that average error as a percentage of the actual values (Hyndman & Koehler, 2006).

As per Table 16, the best model used only Holt-Winters, with a MAPE of 2.93%. The best model followed the equation below:

$$\begin{aligned} HPI_t = {} & -13.404 + 0.410HPI_{t-1} + 0.177HPI_{t-2} - 0.122HPI_{t-3} + 0.296HPI_{t-4} \\ & - 0.236EARN_t + 0.122EARN_{t-1} - 1.907EARN_{t-2} + 0.579EARN_{t-3} \\ & + 1.576EARN_{t-4} - 0.093GDP_t + 0.0001LF_t + \varepsilon_t \end{aligned}$$

Where HPI = House Price Index, $EARN$ = Average earnings % change, GDP = GDP (seasonally-adjusted growth rate), LF = Labour Force.

The HPI dominates the model, with coefficients that net to +0.76. The average earnings percentage change shows real explanatory power, moving the HPI down after two quarters and back up after a year. The cumulative effect, though, is essentially zero with no clear direction. GDP adds a small negative effect equal to 0.093. The Labour force, on the other hand, affects the HPI significantly (+1 per 10,000 additional person in the labour force).

The ARDL model is only valid if the model's errors are uncorrelated, exhibiting white noise (Pesaran et al., 2001). As shown in Figure 31, only lag 8 is approaching the confidence band; however, since there is no systematic decay or seasonal spikes, the residuals are effectively white noise.

PREDICTOR FORECASTS	MAE	MAPE
Holt-Winters	4.41	**2.93%**
Best mix (ARIMA for GDP & EARN, RW+drift for LF)9	5.30	3.53%
Random Walk + drift	5.69	3.79%
ARIMA	6.83	4.54%

Table 16: ARDL forecast accuracy

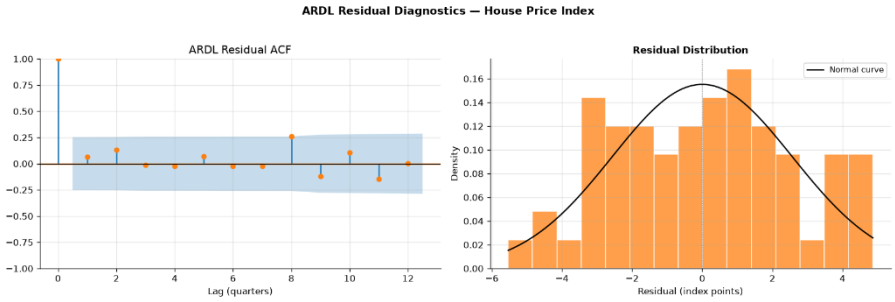


Figure 31: ARDL residual diagnostics

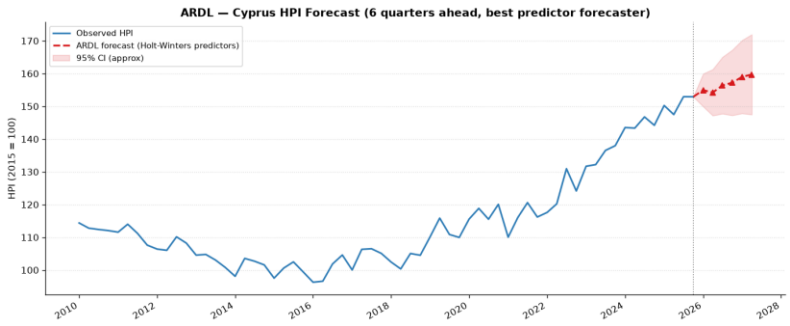


Figure 32: Forecast of the HPI - ARDL model

Figure 32 shows the HPI forecast using the ARDL model. HPI will continue increasing from 153 in 2025Q4 to almost 162 by 2027Q2 (an increase of around 6%). The 95% confidence interval widens steadily with the forecast horizon, reflecting growing uncertainty further out.

Forecasting using only HPI was also conducted to verify whether past values alone can produce better accuracy. The univariate models tested on the first-difference time series used were ARIMA (order selected by AIC), Holt-Winters, and Random Walk with Drift.

Table 17 summarises the accuracy results for the three models. All the models predicted the HPI with MAPE < 3%. However, ARIMA had the lowest error (1.37%).

The ACF shows all lags (1–12) falling within the shaded 95% confidence band, indicating no significant autocorrelation remains in the ARIMA(1,1,2) residuals, as shown in Figure 33. The residual distribution is roughly bell-shaped and centred near zero but slightly right-skewed, suggesting the model occasionally underestimates sharp positive price movements.

MODEL	MAE	MAPE
ARIMA(1,1,2)	2.02	**1.37%**
Holt-Winters	3.44	2.31%
Random Walk + Drift	4.02	2.66%

Table 17: Accuracy Results of HPI Forecast

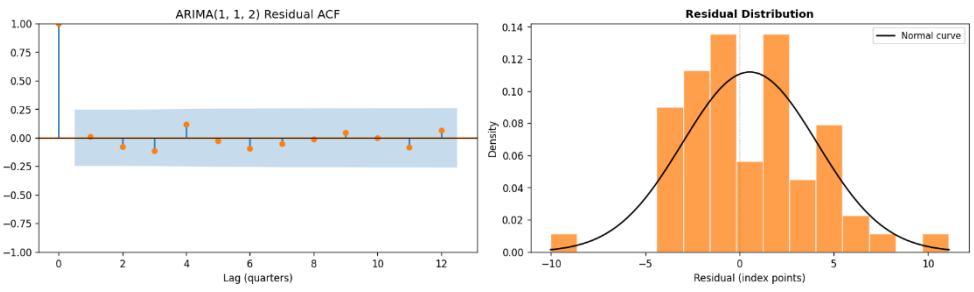


Figure 33: ACF residual plot and Residual distribution – ARIMA

The ARIMA(1,1,2) forecast extends the observed HPI's upward trend smoothly, projecting a continued rise from roughly 153 to 162 over the six-quarter horizon (2026Q1–2027Q2) as demonstrated in Figure 34. The 95% band widens steadily with the horizon, reflecting growing uncertainty further out.

Both the ARDL model and ARIMA(1,1,2) projected a consistent six-quarter rise in the HPI from roughly 153 to 162; however, the ARIMA(1,1,2) model slightly outperformed the ARDL, achieving a lower MAPE (1.37% vs 2.93%) and MAE (2.02 vs 4.41)4. This close performance is because the ARDL model is heavily dominated by historical HPI lags rather than external drivers. Moreover, as the ARDL model relies on secondary forecasts, it introduces additional external noise without offering much predictive power. Consequently, while the ARDL provides useful structural insight into external relationships, the ARIMA(1,1,2) model delivers a simpler and more precise forecasting tool.

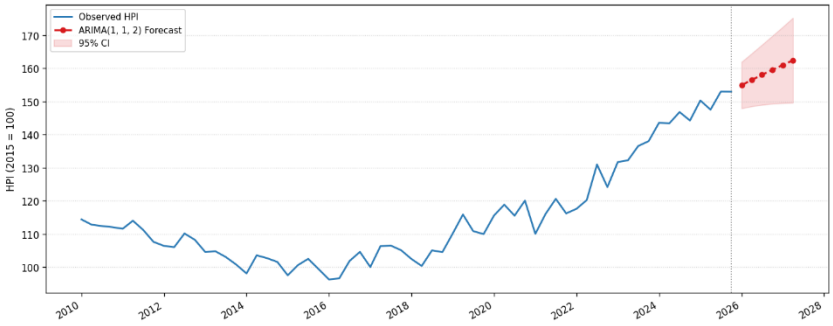


Figure 34: ARIMA (1,1,2) forecast of HPI

Chapter 6

This chapter discusses the study's conclusions and limitations, as well as directions for future research.

6.1. Main conclusions

The dataset for this study was web-scraped using Selenium from three different real estate websites in Cyprus: Website A (48.4%), Website B (43.1%), and Website C (8.6%). In total, 21,647 listings were web-scraped, and 88 variables were defined after applying text-mining techniques. Of the 88 variables, 16 were numerical, 14 were categorical, and 54 were binary. During preprocessing, 2 variables were dropped because more than 90% of their values were missing.

To capture the influence of location on price, a composite location desirability score (MCE score) was derived from 15 distance-decayed, weighted amenity and disamenity categories. The score was sourced from OSM and provided a district-level proxy for locational quality, computed from 237 geocoded individual points across Cyprus. However, the score was highly collinear with the city variable (eta-squared = 0.56) and therefore provided no additional information.

Descriptive statistics show that price is heavily right-skewed, with a median of €360,000 and a mean of €614,217. Approximately 56.4% of the properties in the dataset fall in the price range €100,000-€400,000, whereas 11.9% exceed €1m. The relationship identification process revealed that property size, bedrooms, bathrooms, and land size are strongly correlated, indicating overlapping "size-related" information.

For the categorical and binary variables, nearly every variable is significantly associated with price, as indicated by the Kruskal-Wallis and Mann-Whitney tests.

Only 2 binary variables were not statistically associated: tennis_court and kitchen_appliances. Finally, the combined RF and SHAP model indicated that property type and location are the top drivers of price. The model could only explain 41% of the price variance. These relationships, together with the associations examined in Chapter 3, informed the variable selection of the clustering process.

Continuing with clustering, after the reduction process, only 2 variables were selected for clustering (property_size_sqm, property_type), as their combination best separated the clusters. K-means slightly outperformed Ward's method with a Gower-silhouette score of 0.889. The two methods agreed on 97% of the data; however, K-means was selected for its computational efficiency and slightly better results.

Using K-means, 3 distinct segments were defined: Standard Urban Apartments (69,14%), Mid-Size Houses (26.14%) and Luxury/Large Properties (4.78%). Per-cluster prediction was then performed to identify the main price drivers for each cluster. The RF-XGBoost combination was the best model for Standard Urban Apartments, explaining 35.9% of the price variation; the RF-GBM combination was the best model for Mid-Size Houses, explaining 34.6% of the price variation, and GBM alone was the best model for Luxury/Large Properties, explaining 65.9%.

From the per-cluster prediction process, it was identified that price in each segment is driven by different characteristics: Apartment prices are primarily driven by location under XGBoost, with Limassol affecting prices the most, while RF indicated property size as the top characteristic; Mid-size houses' prices are driven mainly by property size and veranda area, with house type Villa as a secondary factor; Luxury/Large Properties' prices are driven by a broader mix of features such as veranda size, property type and luxury-specific amenities (double-glazed windows and luxury fittings). The results confirm that a single undifferentiated pricing model would be inappropriate for the Cypriot residential real estate market.

The effect of macro indicators on the HPI was also examined using external variables, collected from the Cyprus Statistical Service. The Granger causality and CCF tests revealed that Average Monthly Earnings % change has stronger predictive power for the HPI than GDP, Labour Force or housing supply proxies (building

permits). Specifically, wage growth showed immediate and short-run predictive power (Granger causality HPI at lags 1 through 4). In contrast. GDP Labour Force and building permits failed to show strong direct predictive lead effects. The ARDL bounds test provided evidence of a true long-run cointegrating relationship at the 5% significance level, with an F-statistic of 5.21 exceeding the upper bound.

For forecasting, ARIMA (1,1,2) achieved the highest predictive precision among all models tested, yielding the lowest MAPE of 1.37%. Within the multivariate framework, the ARDL model using Holt-Winters for exogenous driver predictions achieved a competitive MAPE of 2.93%. Because the ARDL model's predictions were heavily dominated by internal historical lags of the HPI itself (with autoregressive coefficients equal to +0.76), and because adding secondary driver inputs increased forecast variance, the ARIMA(1,1,2) was considered a better model for forecasting the HPI. The results match the results of the study of Yılmaz and Kestel (2020), where ARIMA indicated higher accuracy. The results suggest that HPI can be forecast more accurately using its own historical values alone than by incorporating external predictors.

In summary, these results suggest that real estate companies should avoid a uniform pricing approach across all residential real estate markets and instead use different pricing models and marketing strategies for each segment. Policymakers must navigate the complex, lagged impact of income growth, which depresses housing prices in the short term before pushing them back up after about a year, yielding a net cumulative effect near zero. GDP indicates a modest negative effect, while labour force changes affect the HPI significantly. Furthermore, the ARIMA-based forecasting indicates that HPI can be monitored using past data alone. Lastly, the results clarify which property characteristics most affect prices within each market segment, supporting more informed buying and selling decisions for the public.

6.2. Limitations and Future Work

Despite the results presented in this chapter, the study has several limitations that future work could address. Starting with data collection, although the number of

observations collected was high, the web scraping process was restricted to three Cypriot real estate websites, which showed strong associations between variables in the dataset, such as energy_efficiency_grade and house_type, suggesting that the websites advertise specific residential segments. In future work, the scraping process should be expanded to additional platforms to reduce the source-driven bias.

As per the Amenity-Based Location Desirability Index, 26.1% of the properties could not be geocoded to an official district, limiting the coverage of the location desirability index. Furthermore, the MCE index captures only physical accessibility due to a lack of open data, and it is based on *a priori* defined weights. Future work could improve the geocoding process by sourcing exact property addresses, enabling neighbourhood-level location. This could enable better neighbourhood-level research. To define weights, a survey could capture the public's perceptions of what matters most, reducing the score's subjectivity.

Continuing with the RF and SHAP model, it explained only 41% of the price variation, indicating that property characteristics alone leave much price variation unexplained. Future work could improve the model's explanatory power by incorporating variables not captured in this study. Testing alternative or more flexible model architectures might improve accuracy.

Additionally, predictive accuracy across the clusters remained modest, particularly for Standard Urban Apartments and Mid-Size Houses. This suggests that future work could explore additional predictors, examine price movement dynamics more closely, or apply more advanced modelling approaches.

Future work should collect more extensive datasets, check additional external indicators, and use a longer time horizon. In addition, annual variables were excluded from forecasting due to frequency mismatch; future work should apply mixed-frequency techniques to retain such predictors. Finally, no scenario or stress testing was conducted to assess model robustness under market shocks or policy changes, which future work should address.

Bibliography

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics*, *9*(8), 1295.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Analytics, M. s. (2024, 2026). *Cyprus House Price Index*. Economy.com. Retrieved 17/01/2026 from https://www.economy.com/cyprus/house-price-index

Anselin, L. (1988). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, *20*(1), 1–17. https://doi.org/https://doi.org/10.1111/j.1538-4632.1988.tb00159.x

Bale, A., Ghorpade, N., Kamalesh, S., S, R., R, R., & S, R. (2022). *Web Scraping Approaches and their Performance on Modern Websites*. https://doi.org/10.1109/ICESC54411.2022.9885689

Bartlett, M. S. (1946). On the Theoretical Specification and Sampling Properties of Autocorrelated Time-Series. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 27–41. https://doi.org/10.2307/2983611

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees* (1 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315139470

Case, B., Pollakowski, H. O., & Wachter, S. M. (1991). On choosing among house price index methodologies. *Real estate economics*, *19*(3), 286–307.

Case, K. E., & Shiller, R. J. (1989). The efficiency of the market for single-family homes. *The American Economic Review*, *79*(1), 125–137.

Dinh, D.-T., Huynh, V.-N., & Sriboonchitta, S. (2021). Clustering mixed numerical and categorical data with missing values. *Information Sciences*, *571*, 418–442. https://doi.org/https://doi.org/10.1016/j.ins.2021.04.076

Dubin, R. A. (1998). Spatial Autocorrelation: A Primer. *Journal of Housing Economics*, *7*(4), 304–327. https://doi.org/https://doi.org/10.1006/jhec.1998.0236

Eden, S. K., Li, C., & Shepherd, B. E. (2022). Nonparametric estimation of Spearman's rank correlation with bivariate survival data. *Biometrics*, *78*(2), 421–434. https://doi.org/10.1111/biom.13453

Estate, D. R. (2024). *Which countries' citizens most often buy property in Cyprus?* Retrieved 30/01/2026 from https://dom.com.cy/en/live/digest/which-countries-do-citizens-of-most-often-buy-property-in-cyprus/

Estates, L. (2025). *What to know before buying real estate in Cyprus: Definitions, types, regions, market, and investment options*. https://www.leptosestates.com/blog/cyprus/real-estate-in-cyprus/

Eurostat. (2024). *Disability statistics*. European Commission https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Disability_statistics

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232, 1144.

Geert, V. d. B. R. (2017). *Cramér's V – What and Why?* Retrieved 16/01/2026 from https://www.spss-tutorials.com/cramers-v-what-and-why/

Goodman, A. C., & Thibodeau, T. G. (1998). Housing Market Segmentation. *Journal of Housing Economics*, *7*(2), 121–143. https://doi.org/https://doi.org/10.1006/jhec.1998.0229

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424–438. https://doi.org/10.2307/1912791
Greene, W. H. (2018). *Econometric analysis*. Pearson.
Hakkal, S., & Lahcen, A. A. (2024). XGBoost To Enhance Learner Performance Prediction. *Computers and Education: Artificial Intelligence*, *7*, 100254. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100254
Harrison, R. L. (2010). Introduction To Monte Carlo Simulation. *AIP Conf Proc*, *1204*, 17–21. https://doi.org/10.1063/1.3295638
Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd edition ed.). OTexts, Melbourne, Australia.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/https://doi.org/10.1016/j.ijforecast.2006.03.001
Investments, C. (2025). *Cyprus Golden Visa*. Cyprus Investments. https://www.cyprusinvestments.com/articles/cyprus-golden-visa
James, G. W., Daniela; Hastie, Trevor; Tibshirani, Robert. (2021). *An Introduction to Statistical Learning: with Applications in R* (2 ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1
Kathimerini. (2025). *Cyprus real estate market draws record numbers of foreign buyers in 2025* KNEWS – Kathimerini Cyprus. Retrieved 06/01/2026 from https://knews.kathimerini.com.cy/en/business/cyprus-real-estate-market-draws-record-numbers-of-foreign-buyers-in-2025
Kim, H. Y. (2016). Statistical notes for clinical researchers: Sample size calculation 3. Comparison of several means using one-way ANOVA. *Restor Dent Endod*, *41*(3), 231–234. https://doi.org/10.5395/rde.2016.41.3.231
Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection* Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2, Montreal, Quebec, Canada.
Küp, E. T., Sözdinler, M., Işık, A. H., Doksanbir, Y., & Akpınar, G. (2025). Enhancing Real Estate Listings Through Image Classification and Enhancement: A Comparative Study. *Engineering Proceedings*, *92*(1), 80.
Lahoz, A. E. p. (2007). *An analysis of how geographicalfactors affect real estate prices* (Publication Number 07-12) [Student thesis, DiVA. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-199839
Liu, P., Yuan, H., Ning, Y., Chakraborty, B., Liu, N., & Peres, M. A. (2024). A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02427-8
LiVE, D. (2025). *Cyprus Real Estate Market Continues to Break Records*. DOM LiVE. Retrieved 06/01/2026 from https://dom.com.cy/en/live/digest/cyprus-real-estate-market-continues-to-break-records/
Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA.
MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Berkeley, CA.
Malczewski, J. (2004). GIS-based land-use suitability analysis: a critical overview. *Progress in Planning*, *62*(1), 3–65. https://doi.org/https://doi.org/10.1016/j.progress.2003.09.002
Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In T. G. O'Sullivan, K. (Ed.), *Housing Economics and Public Policy* (pp. 67–89). Blackwell Science.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, *18*(1), 50–60, 11.

Minitab, L. (2026). *What are the Goodman-Kruskal statistics?* Minitab, LLC. Retrieved 15/07/2026 from https://support.minitab.com/en-us/minitab/help-and-how-to/statistics/tables/supporting-topics/other-statistics-and-tests/what-are-the-goodman-kruskal-statistics/

Mohan, S., Hutson, A., MacDonald, I., & Lin, C. C. (2019). Impact of macroeconomic indicators on housing prices. *International Journal of Housing Markets and Analysis*, *12*(6), 1055–1071. https://doi.org/https://doi.org/10.1108/IJHMA-09-2018-0070

Murtagh, F., & Legendre, P. (2011). Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm. *arXiv preprint arXiv:1111.6285*.

Nelson, C. R., & Plosser, C. R. (1982). Trends and random walks in macroeconmic time series: Some evidence and implications. *Journal of Monetary Economics*, *10*(2), 139–162. https://doi.org/https://doi.org/10.1016/0304-3932(82)90012-5

Nelson, J. (2024). *The Importance of Pricing a Property Correctly When Selling*. Forbes. https://www.forbes.com/sites/jamesnelson/2024/09/18/the-importance-of-pricing-a-property-correctly-when-selling/

Nielsen, A. (2019). *Practical Time Series Analysis: Prediction with Statistics and Machine Learning* (1st ed.). O'Reilly Media, Inc.

Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds Testing Approaches to the Analysis of Level Relationships. *Journal of Applied Econometrics*, *16*(3), 289–326.

Ramachandran, K. M., & Tsokos, C. P. (2015). Chapter 12 - Nonparametric Tests. In *Mathematical Statistics with Applications in R (Second Edition)* (pp. 589–637). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-417113-8.00012-6

Realty, F. (2025). *Cost of Buying a House in Cyprus (2025): Prices, Fees & Taxes Explained*. FOX Realty (foxrealty.com.cy). Retrieved July 7, 2025 from https://foxrealty.com.cy/blog/cost-of-buying-a-house-in-cyprus-2025-prices-fees-taxes-explained/?gad_source=1&gad_campaignid=22104759836&gbraid=0AAAApDUGP0S78SXvOqpuTRGXnhgy2cX_&gclid=Cj0KCQiAkPzLBhD4ARIsAGfah8inO02Q9TlwNPBHdOh0Ux43eHFUUMaK0opL_KVdOJUj9VgVVZhiABQaAvd3EALw_wcB

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, *82*(1), 34–55. https://doi.org/10.1086/260169

Saar-Tsechansky, M., & Provost, F. (2007). Handling missing values when applying classification models.

Said, E. S., & Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, *71*(3), 599–607. https://doi.org/10.2307/2336570

Santos, J. M. A. (2018). *Real Estate Market Data Scraping and Analysis for Financial Investments* Universidade do Porto].

Sedgwick, P. (2012). Pearson's correlation coefficient. *BMJ*, *345*, e4483. https://doi.org/10.1136/bmj.e4483

Sen, S., & Cohen, A. S. (2025). On the Use of Elbow Plot Method for Class Enumeration in Factor Mixture Models. *Appl Psychol Meas*, *49*(7), 379–395. https://doi.org/10.1177/01466216251344288

Sharma, K., & Borkar, G. (2024). Comparative Analysis of Dynamic Web Scraping Strategies: Evaluating Techniques for Enhanced Data Acquisition. In (pp. 241–252). https://doi.org/10.56155/978-81-955020-7-3-22

Soylemez, İ., Nalici, M. E., & Ünlü, R. (2026). Fluctuations in the European Housing Market: Forecasting the House Price Index Change with Time-Series Models. *Politeknik Dergisi*, 1–1.

Stundziene, A., Pilinkienė, V., & Grybauskas, A. (2021). The impact of external factors on housing prices in Lithuania. *International Journal of Housing Markets and Analysis*, *15*(4), 833–851. https://doi.org/10.1108/IJHMA-04-2021-0048

Tunnicliffe Wilson, G. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *Journal of Time Series Analysis*, *37*, 709–711. https://doi.org/10.1111/jtsa.12194

UK, H. C. o. C. i. t. (2026). *General information about Cyprus*. High Commission of Cyprus in the UK. Retrieved 06/01/2026 from https://cyprusinuk.com/general-information/

Wiersma, S., Just, T., & Heinrich, M. (2021). Segmenting German housing markets using principal component and cluster analyses. *International Journal of Housing Markets and Analysis*, *15*(3), 548–578. https://doi.org/10.1108/IJHMA-01-2021-0006

Yılmaz, B., & Kestel, A. s. S. (2020). FORECASTING HOUSE PRICES IN TURKEY: GLM, VAR AND TIME SERIES APPROACHES. *Journal of Business Economics and Finance*, *9*(4), 274–291. https://doi.org/10.17261/Pressacademia.2020.1310

Appendix

Reference	Title of the study	data	Method	Main findings	Limitations
(Santos, 2018)	Real Estate Market Data Scraping and Analysis for Financial Investments	Time stamps, location, prices, photos and typologies	Python web-scraping libraries (Beautiful Soup, Selenium and Screen-Scraper), K-means, Random Forest, ANN	RF outperformed ANN	IP blocking issues; small dataset; limited location detail; simple ANN architecture; limited preprocessing
(Mohan et al., 2019)	Impact of macroeconomic indicators on housing prices	House Price Index, crude oil price, 30-year mortgage interest rate, consumer price index, Dow Jones Industrial Average, and unemployment rate.	Vector Autoregression (VAR)	Changes in HPI explained up to 92.7% of its own variance one month ahead, and still about 74.5% at 12 months.	Focused on Amherst, NY; only five macro indicators; structural shifts (2007–08) not considered
(Stundziene et al., 2021)	The impact of external factors on housing	Macroeconomic, labour market,	Correlation, regression analysis, causality	Real estate cycles are closely related to economic cycles. 3 most	Considered only external factors; limited to Lithuania; time-

	prices in Lithuania	business, financial, real estate indicators, consumer and business expectations	tests, and ARDL models	significant indicators: quarterly changes in GDP, consumer spending and household cash and deposits	period specific; not generalisable
(Lahoz, 2007)	An analysis of how geographic alfactors affect real estate prices	Accessibility variables, physical environment, social environment, neighbourhood quality and tax assessment.	Geographic Information System and MCE	Most influential factors: Income and crime rate, Foreigners in the neighbourhood, proximity to water, green, and the airport.	Assumed linearity without testing; ambiguous interpretation of variables; incomplete pollution and green data; municipal boundary issues; tax value distortions; inconsistent measurement periods
(Wiersma et al., 2021)	Segmenting German housing markets using principal component and cluster analyses	Market size and dynamics, price level and dynamics, socioeconomic indicators and dynamics and demographics	Principal Component Analysis, Ward's method and K-means	The German real estate market can be segmented into 4 groups.	Germany-only study; recommends adding migration and economic indicators; limited generalisability.
(Yılmaz & Kestel, 2020)	Forecasting House Prices In Turkey: GLM, VAR, and Time Series	House Price Index, macroeconomic and financial indicators	Generalised Linear Models, VAR model and Univariate Time series	GLM and VAR models can identify key housing drivers, and ARIMA and Exponential smoothing	Dependent on the accuracy of explanatory variables, a wide CI for long horizons, sensitive to

	Approaches		models, including AutoRegressive Moving Average, ARIMA and exponential smoothing models.	performed the highest accuracy forecasts	structural market changes
(Soylemez et al., 2026)	Fluctuations in the European Housing Market: Forecasting the HPI Change with Time-Series Models	HPI of 27 European Countries	Autoregressive Integrated Moving Average, Holt Linear Trend, Additive Damped Trend and Exponential Smoothing	The results highlight the importance of country-specific forecasting approaches.	No macroeconomic indicators; only classical models; country-specific results; limited policy insight; not generalisable

Table 18: Summary of the studies' limitations

VARIABLE	CATEGORY	DESCRIPTION	SOURCE(S)
page_url	Identifier	Unique URL of the property listing page	All
title	Text	Title of the listing as per the website	All
description	Text	Description of the listing	All
features	Text	Additional features of the listing	All
price	Numerical (continuous)	Property price in EUR	All
bedrooms	Numerical (discrete)	Number of bedrooms	All
bathrooms	Numerical (discrete)	Number of bathrooms	Website B
number_floors_building	Numerical (discrete)	Total number of floors in the building	Website B
uncovered_parking	Numerical (discrete)	Number of uncovered parking spaces	All
floor_number	Numerical (discrete)	Floor on which the unit is located	Website B, Website C
covered_parking	Numerical (discrete)	Number of covered parking spaces	Website B, Website C
parking	Numerical (discrete)	Total number of parking spaces (covered and uncovered)	All
construction_year	Numerical (discrete)	Year the property was built	All

renovation_year	Numerical (discrete)	Year of last renovation	All
veranda_covered	Numerical (continuous)	Covered veranda area in m²	All
veranda_uncovered	Numerical (continuous)	Uncovered veranda area in m²	Website A, Website C
roof_terrace_sqm	Numerical (continuous)	Roof terrace/roof garden area in m²	Website A
monthly_communal_fees	Numerical (continuous)	Monthly communal/maintenance fees in EUR	Website B
property_size_sqm	Numerical (continuous)	Internal/covered property area in m²	All
garden_area	Numerical (continuous)	Garden area in m²	Website B
land_size_sqm	Numerical (continuous)	Total land/plot size in m²	Website B
energy_efficiency_grade	Categorical (ordinal)	Energy efficiency rating from A (highest) to G (lowest)	Website B
district	Categorical (high cardinality)	Sub-district or neighbourhood	All
kitchen_type	Categorical (binary)	Kitchen layout: Fitted Kitchen or Open Plan	All
electricity_type	Categorical (binary)	Type of electricity supply: Single Phase or Three Phase	All
construction_type	Categorical (binary)	Primary building material: Concrete or Metal	All
property_type	Categorical	General property type: house or apartment	All
city	Categorical	City where the property is located (Nicosia, Limassol, Larnaca, Paphos, Famagusta)	All
furnished	Categorical	Furnishing status: Unfurnished, Fully Furnished, Partially Furnished, Optionally Furnished	Website A, Website C
website	Categorical	Source website the listing was scraped from (3 platforms)	All
apartment_type	Categorical	Apartment subtype: Standard, Penthouse, Maisonette, Ground Floor, etc.	All
floor_type	Categorical	Flooring material: Parquet, Marble, Laminate, Tiles, Wooden Floors, etc.	Website B
house_type	Categorical	House subtype: Villa, House, Bungalow, Semi-Detached, Link-Detached, etc.	Website B, Website C
parking_type	Categorical	Type of parking available: Covered, Uncovered, Garage, or Underground	All
property_status	Categorical	Listing/construction status: Completed, Resale, Off-Plan, Under Construction	All
alarm	Binary	Property has an alarm system (1 = yes, 0 = no)	All
cctv	Binary	Property has CCTV (1 = yes, 0 = no)	Website C

elevator	Binary	Building has an elevator (1 = yes, 0 = no)	All
pool	Binary	Property has a shared or private swimming pool (1 = yes, 0 = no)	Website B, Website C
sea_view	Binary	Property has a sea view (1 = yes, 0 = no)	Website B
fireplace	Binary	Property has a fireplace (1 = yes, 0 = no)	Website B, Website C
balcony	Binary	Property has a balcony (1 = yes, 0 = no)	Website B
gym	Binary	Complex includes a gym (1 = yes, 0 = no)	Website B
home_office	Binary	Property has a designated home office (1 = yes, 0 = no)	Website B
en-suite	Binary	Property has an en-suite bathroom (1 = yes, 0 = no)	Website B, Website C
sauna	Binary	Property has a sauna (1 = yes, 0 = no)	Website B
secure_door	Binary	Property has a security/armoured door (1 = yes, 0 = no)	Website B, Website C
pergola	Binary	Property has a pergola (1 = yes, 0 = no)	Website B, Website C
disabled_access	Binary	Property has disabled access features (1 = yes, 0 = no)	Website B
maids_room	Binary	Property has a maid's room (1 = yes, 0 = no)	Website B, Website C
double_glazing	Binary	Windows have double glazing (1 = yes, 0 = no)	Website B
automatic_light_control	Binary	Property has automatic/smart light control (1 = yes, 0 = no)	Website B
electrical_garage_gate	Binary	Property has an electrical garage gate (1 = yes, 0 = no)	Website B
basement	Binary	Property has a basement (1 = yes, 0 = no)	Website B, Website C
walk-in_wardrobe	Binary	Property has a walk-in wardrobe (1 = yes, 0 = no)	Website B
luxury_fittings	Binary	Property has luxury fittings/finishes (1 = yes, 0 = no)	Website B
gated_complex	Binary	Property is within a gated complex (1 = yes, 0 = no)	Website B
internet_access	Binary	Property has wired internet access (1 = yes, 0 = no)	Website B
pressurized_water_system	Binary	Property has a pressurized water system (1 = yes, 0 = no)	Website B, Website C
provision_for_split_units	Binary	Provision for split A/C units is installed (1 = yes, 0 = no)	Website B
smart_home	Binary	Property has smart home automation features (1 = yes, 0 = no)	Website B
pets_allowed	Binary	Pets are permitted in the property (1 = yes, 0 = no)	Website B
door_phone	Binary	Property has a door phone (1 = yes, 0 = no)	Website B

city_view	Binary	Property has a city view (1 = yes, 0 = no)	Website B
telephone_access	Binary	Property has telephone line access (1 = yes, 0 = no)	Website B
satellite_tv	Binary	Property has satellite TV provision (1 = yes, 0 = no)	Website B, Website C
storage_space	Binary	Property includes a storage space (1 = yes, 0 = no)	Website B
electrical_shutters	Binary	Property has electrical/motorised shutters (1 = yes, 0 = no)	Website B
garden	Binary	Property has a garden (1 = yes, 0 = no)	Website B
mountain_view	Binary	Property has a mountain view (1 = yes, 0 = no)	Website B, Website C
loudspeakers	Binary	Property has built-in loudspeaker system (1 = yes, 0 = no)	Website B
air_condition	Binary	Property has air conditioning (1 = yes, 0 = no)	Website B, Website C
solar_water_heater	Binary	Property has a solar water heater (1 = yes, 0 = no)	Website B, Website C
laundry_room	Binary	Property has a dedicated laundry room (1 = yes, 0 = no)	Website B
jacuzzi	Binary	Property has a jacuzzi (1 = yes, 0 = no)	Website C
spa	Binary	Property has a spa (1 = yes, 0 = no)	Website C
tennis_court	Binary	Property has a tennis court (1 = yes, 0 = no)	Website C
kitchen_appliances	Binary	Property includes kitchen appliances (1 = yes, 0 = no)	Website C
entrance_gate	Binary	Property has an entrance gate (1 = yes, 0 = no)	Website C
barbeque	Binary	Property has a barbeque/outdoor cooking area (1 = yes, 0 = no)	Website B, Website C
veranda	Binary	Property has a veranda (1 = yes, 0 = no)	Website B
door_camera	Binary	Property has a door/entry camera (1 = yes, 0 = no)	All
title_deed	Binary	Title deed is available for the property (1 = yes, 0 = no)	Website B
provision_for_storage_heating	Binary	Property has infrastructure provision for storage heating (1 = yes, 0 = no)	Website B
ceiling_fans	Binary	Property has ceiling fans installed (1 = yes, 0 = no)	Website B
cul-de-sac	Binary	Property is situated on a cul-de-sac / dead-end road (1 = yes, 0 = no)	Website B, Website C
projectors	Binary	Property has built-in projectors (1 = yes, 0 = no)	Website C
solar_photovoltaic_panels	Binary	Property has solar photovoltaic (PV) panels for electricity generation (1 = yes, 0 = no)	Website C
has_heating	Binary	Property has a heating system (1 = yes, 0 = no)	Website C

Table 19: Dataset Description

VARIABLE 1	VARIABLE 2	CRAMÉR'S V
internet_access	telephone_access	0.9261
double_glazing	garden	0.8412
double_glazing	pressurized_water_system	0.8024
balcony	air_condition	0.7941
pressurized_water_system	garden	0.7352
double_glazing	storage_space	0.6792
internet_access	garden	0.6514
telephone_access	garden	0.6443
pressurized_water_system	storage_space	0.6419
double_glazing	internet_access	0.6382
elevator	double_glazing	0.6382
double_glazing	telephone_access	0.6307
internet_access	pressurized_water_system	0.6298
pressurized_water_system	telephone_access	0.6231
elevator	pressurized_water_system	0.6179
storage_space	garden	0.6110
luxury_fittings	door_phone	0.6095
elevator	storage_space	0.6048
en-suite	double_glazing	0.5994
elevator	garden	0.5958
luxury_fittings	solar_water_heater	0.5589
luxury_fittings	garden	0.5511
double_glazing	luxury_fittings	0.5485
en-suite	pressurized_water_system	0.5470
en-suite	storage_space	0.5433
en-suite	garden	0.5388
elevator	internet_access	0.5208
luxury_fittings	provision_for_split_units	0.5151
gym	spa	0.5036
door_phone	solar_water_heater	0.5034

Table 20: Cramer's V for Binary Variables

CATEGORY	VARIABLE	FREQUENCY	TIME FRAME	DESCRIPTION
Housing Market Indicators (target)	House Price Index	Quarterly	2010 Q1 – 2025 Q4 (n=64)	Dependent variable. Residential property price index for Cyprus, base 2015 = 100.
Macroeconomic Indicators	GDP (SA)	Quarterly	1997 Q4 – 2026 Q1 (n=112)	Seasonally adjusted GDP
Labour Market & Income Indicators	Average earnings percentage change (prev. quarter, SA)	Quarterly	2003 Q2 – 2025 Q4 (n=91)	Quarter-on-quarter wage growth rate

Labour Market & Income Indicators	Labour Force	Quarterly	1999 Q2 – 2026 Q1 (n=93)	Total number of people employed or seeking employment
Housing Supply & Construction Indicators	Number of Building Permits	Monthly	Jan 2003 – Jan 2026 (n=277)	Count of building permits issued

Table 21: External Indicators