# Machine Learning for Business – Housing Market Analysis (King County Dataset)

Student: Renata Oliveira Souza  
Module: Machine Learning for Business  
Date: 2025-10-27

## 1. Introduction and Objectives

The real estate market is a complex environment influenced by economic, social, and geographic factors. Accurate analysis of property data is essential for pricing strategies, investment planning, and identifying emerging market trends. This report examines the King County housing market (Seattle area) using the ‘House Sales in King County’ dataset. The dataset provides detailed information about housing prices, physical characteristics, and geographic location, including both numerical and categorical variables, which make it ideal for applying machine learning techniques in business analytics.

The aim of this study is twofold: first, to identify hidden patterns among properties through unsupervised learning using clustering algorithms; and second, to forecast average monthly house prices through time series modelling with the ARIMA approach. These objectives align with the Machine Learning for Business module outcomes, which focus on the ability to apply clustering and predictive models to real-world business data. The rationale for selecting this dataset lies in its relevance to a high-value industry where small predictive advantages can translate into significant financial gains. By combining clustering and ARIMA forecasting, this analysis provides a holistic view of both spatial and temporal housing market dynamics.

## 2. Clustering Analysis

Clustering, an unsupervised learning technique, groups data points based on similarity without predefined labels. In this study, KMeans and Agglomerative Clustering algorithms were applied to identify segments of similar houses. The feature engineering process created new variables such as ‘house\_age’ (year of sale minus year built) and ‘price\_per\_sqft’, which normalises property value by living space, enabling comparison across differently sized houses. To ensure model stability, extreme outliers were trimmed using the 1st and 99th percentiles. Numerical features were standardised using z-score normalisation to remove bias caused by scale differences between variables like price and square footage.

The clustering evaluation was based on two internal validation metrics: the Silhouette Score and the Davies–Bouldin Index (DBI). The Silhouette Score measures the degree of separation between clusters, where values closer to 1 indicate well-formed clusters. Conversely, the DBI evaluates average similarity between clusters, where lower values suggest better separation. KMeans clustering was tested for different numbers of clusters (K from 2 to 10). The optimal K was selected based on the highest Silhouette Score and lowest DBI, representing the balance between compactness and separation.

KMeans, a centroid-based algorithm, was preferred for its computational efficiency and ability to handle large datasets. Agglomerative Clustering, a hierarchical approach, was also explored to validate results through an alternative linkage-based method. After determining the optimal K, both models were compared. Results showed that KMeans produced slightly superior Silhouette scores and clearer boundaries between clusters. Each cluster represented a distinct housing segment, with the first cluster including smaller, lower-priced houses; the second containing mid-range properties; and the third dominated by luxury homes with higher grades, larger square footage, and waterfront locations.

This segmentation offers valuable business insights. Developers and real estate agencies can target marketing campaigns more efficiently by tailoring them to specific clusters. For example, luxury segments may prioritise architectural design and location views, whereas affordable clusters might focus on financing options. The analysis demonstrates that unsupervised learning methods such as clustering can transform raw property data into actionable strategic intelligence.

## 3. Time Series Analysis and ARIMA Forecasting

While clustering provided a structural understanding of property groups, time series analysis enabled temporal forecasting of price evolution. The dataset’s ‘date’ attribute allowed the creation of a monthly average price series, revealing overall market trends. The series exhibited a persistent upward trajectory with moderate volatility, consistent with Seattle’s sustained economic growth and housing demand over recent years.

The ARIMA model—short for Autoregressive Integrated Moving Average—was selected due to its ability to model linear dependencies and trends in time series data. The AR (autoregressive) component captures relationships between an observation and its lagged values; the I (integrated) component ensures stationarity through differencing; and the MA (moving average) component models the error of previous forecasts. Optimal parameters (p, d, q) were determined through grid search minimising the Akaike Information Criterion (AIC). The AIC penalises overfitting by balancing model complexity and goodness of fit, ensuring the chosen model generalises well to unseen data.

The dataset was divided into a training set and a 12-month hold-out testing set. The ARIMA model trained on historical averages and produced forecasts for the testing period. Model accuracy was assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The relatively low error values indicated that the model effectively captured short-term patterns in the housing market. When retrained on the full dataset, the ARIMA model projected a moderate but continuous increase in average monthly prices for the subsequent 12 months, suggesting market stability and sustained demand.

Despite its usefulness, ARIMA has limitations. It assumes linearity and may not capture complex nonlinear relationships inherent to real estate markets. Moreover, it does not consider exogenous variables such as interest rates, inflation, or population growth, which can influence housing prices. Future work could address these gaps by extending ARIMA to SARIMA (Seasonal ARIMA) or by incorporating external regressors into an ARIMAX framework. Alternatively, machine learning methods such as Facebook Prophet or LSTM neural networks could capture nonlinear patterns and seasonality more effectively.

## 4. Conclusions

This study demonstrated the practical application of machine learning methods in understanding and forecasting housing market dynamics. Clustering successfully segmented the King County market into coherent groups with distinct price, size, and quality profiles. The segmentation results can support business strategies, from pricing and marketing to urban planning. The ARIMA model complemented this spatial analysis by providing a forward-looking perspective, predicting future price movements based on historical trends. Together, these approaches illustrate the synergy between unsupervised and predictive analytics in business contexts.

From a business intelligence perspective, the insights gained enable data-driven decision-making. Real estate agencies can use cluster definitions to profile customers and optimise sales strategies, while investors can identify undervalued segments and anticipate market fluctuations. The combination of clustering and ARIMA proves particularly valuable in markets characterised by complex interactions between local and temporal variables. Nonetheless, continuous model evaluation and incorporation of external economic indicators are essential to maintain forecast accuracy in rapidly changing conditions.

In conclusion, this project reinforces that machine learning techniques—when properly implemented—can provide a powerful toolkit for strategic analysis in the housing sector. Beyond King County, the same methodology could be applied to other cities to compare market behaviours and assess global real estate dynamics. Future improvements could include feature selection automation, ensemble forecasting, and integration with geospatial analytics to map market evolution more precisely. Overall, the findings confirm that data analytics offers substantial value in supporting business intelligence and decision-making within the property market.

## References

Hyndman, R.J. and Athanasopoulos, G. (2021) \*Forecasting: Principles and Practice.\* 3rd ed. OTexts.  
Kaufman, L. and Rousseeuw, P.J. (2009) \*Finding Groups in Data: An Introduction to Cluster Analysis.\* Wiley.  
King County Department of Assessments (n.d.) \*Property Assessment Data.\* Available at: https://kingcounty.gov/depts/assessor (Accessed: today).  
Harlfoxem (n.d.) \*House Sales in King County, USA.\* Kaggle. Available at: https://www.kaggle.com/datasets/harlfoxem/housesalesprediction (Accessed: today).