A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Renata Oliveira Souza |
| *Student Number* | 2025064 |
| *Module Title* | HDIP Machine Learning - CA1 – 2 Semester |
| *Assessment Title* | HDIP in Science in Data Analytics for Business CA1- 2º Semester |
| *Assessment Due Date* | 31/10/2025 |
| *Date of Submission* | 31/10/2025 |

**Declaration**

By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

**Integrated CA1 – 2º Semester**

**Programme Title:** HDIP in Science in Data Analytics for Business

**Student Name:** Renata Oliveira Souza

**ID Number:** 2025064

**CCT College Dublin** – February 2025

**Modules:** HDIP Machine Learning - CA1

**Professor:** **Dr. Muhammad Iqbal**

**Submission Date:** 31 October 2025

**Part I - Introduction, Motivation, Problem Statement, Objectives, and Dataset Justification**

**Introduction**

The housing market is a key and complex component of every economy since it affects household wealth, investment and government policies. Real estate professionals, investors and urban planners benefit from an understanding of the characteristics of a property, its price, and the relationship of the property and its price over time. With housing market datasets becoming more available, machine learning is an effective way of uncovering the patterns and making datadriven decisions.

This project examines King County (Seattle, USA) residential property data to determine the most important price drivers and how they change over time. This study employs the techniques of unsupervised learning (clustering) and time series forecasting (ARIMA) to investigate, segment, and predict housing markets to provide insights relevant to the Housing and Zoning sector.

**Motivation**

Motivation stems from the growing need for data-driven solutions to be incorporated into real estate management. Traditional methods of statistics can struggle to capture the more complex, nonlinear and interrelated associations between location, quality, and price which machine learning is able to.

To target specific audiences with appropriate marketing strategies which account for changes in bottlenecks and customer behaviors for housing services (e.g. leveraging customer profiles for demand-survey and marketing to lower customer acquisition costs). Time series forecasting will benefit investors, policymakers, and developers by predicting future market behaviors based on past price trends and market cycles.

**Problem Statement**

There are economic, spatial, and physical dimensions to all property values. Ascertaining valuable dimensions and tracking their value over a period will always remain difficult.

The primary research question is:

In what way might business and policy decisions in the property market be improved using clustering and time series techniques to analyze and predict property market trends and characterize value metrics?

This question is directed towards real estate businesses, housing authorities, and financial institutions, and private investors with interest in the price, potential value and investment in the property market and zoning to be prioritized.

**Project Objectives**

The objectives are:

To examine and preprocess the datasets related to housing and ascertain the level of data quality alongside the relevant quantitative and qualitative variables.

To execute clustering techniques (K-Means and Hierarchical/Agglomerative Clustering) on the data and spatially classify the economic and structural dimensions of the property.

To assess the quality of segmentation and the optimal number of clusters using the Silhouette Score and the Davies–Bouldin Index as clustering metrics.

To analyze average monthly house prices over time to uncover trends and seasonal patterns, and to create an ARIMA model for forecasting.

To effectively communicate results, integrating and contextualizing the analysis to the dynamics of the housing market and the implications for the business.

By integrating spatial segmentation with time projection, the ARIMA model provides a complete perspective on the market’s inter-temporal dynamics.

Dataset Description and Rationale (King County - Seattle)

The dataset for this project comprises house sales records within Seattle and its surrounding areas from 2014 to 2015, as King County was the focus. It contains a complete temporal component to support time series analysis as well as extensive structural, qualitative, and geographical components for each housing unit.

The primary components include:

* price – the final selling price (target variable)
* bedrooms, bathrooms – the number of bedrooms and bathrooms
* sqft\_living, sqft\_lot – sizes of the living area and the lot
* floors – the number of stories
* waterfront, view, condition, grade – qualitative descriptors
* sqft\_above, sqft\_basement – structural allocation
* lat, long, zipcode – geographical location
* date – date of sale (used to extract year/month

Reasons for selection:

This relates to the Housing and Zoning domain mentioned within the CA1 guidelines.

There is a mix of the different data types (numerical, ordinal, and categorical) which supports the various activities of clustering and correlation analysis.

There is a date field so a monthly average price series can be constructed for ARIMA modelling.

This dataset is widely used within machine learning benchmarks which aids in reliability and reproducibility.

Feature engineering examples include house\_age =year of sale − year built (age at sale) and price\_per\_sqft = price / sqft\_living (value per square foot).

This dataset provides a solid context in which to apply machine learning in the housing analytics field in the various clustering of the different types of a property to forecasting the price trends which meets the learning objectives of the module in its entirety.

**Part II - Clustering Algorithms and Comparison**

2.1. Clustering

Clustering is a type of unsupervised learning which involves the grouping of data points with similar characteristics, without using predefined labels. It is widely used in market segmentation, customer profiling, and property classification.

In this instance, the clustering is used to separate a group of houses that share similar physical and economic characteristics (size, quality, and price). This classification provides value with regards to pricing, zoning, and overall real estate analytics.

2.2. Selected Variables (King County)

For clustering, only relevant numerical features were chosen and normalized to enable fair comparisons across features of different scales.

Variable Description

price Final sale price

bedrooms Number of bedrooms

bathrooms Number of bathrooms

sqft\_living Living area (square feet)

sqft\_lot Lot area (square feet)

floors Number of floors

waterfront Waterfront property (0/1)

view Quality of view (0–4)

condition Overall condition (1–5)

grade Construction and design quality

sqft\_above Area above ground

sqft\_basement Basement area

lat, long Geographical coordinates

house\_age Age of the property at sale

price\_per\_sqft Price per square foot

2.3. Data Preparation and Standardization

Cleaning: Removed records with missing critical values (price, sqft\_living, lat/long, etc.).

Feature Engineering: Created house\_age and price\_per\_sqft.

Outlier Treatment: Trimmed price\_per\_sqft between the 1st and 99th percentiles for robustness.

Standardization: Applied z-score normalization to ensure all variables contribute equally to distance calculations.

2.4. Determining the Optimal Number of Clusters and Algorithm Comparison

A range of K values (2–10) was tested using K-Means, with performance evaluated via Silhouette Score (higher = better cohesion) and Davies–Bouldin Index (lower = better separation).

The best K was selected by maximizing Silhouette and minimizing DBI. A Hierarchical (Agglomerative) model with the same K was then compared.

Usually the results can be interpreted as follows:

High-value clusters are bigger than average homes and newer as well, get higher grades, and are located in more desirable areas.

Mid-range clusters are average-sized homes that have a balanced mix of attributes.

Low-value clusters are located in the suburbs, and consist of older and smaller homes.

This allows for focused marketing, pricing, and investment to be more effective.

**Part III - Time Series Analysis and ARIMA Forecasting**

3.1. Constructing the Time Series

A continuous time series was created based on the sale date variable and calculated monthly average prices. Considering Seattle’s real estate cycle, the data showed a consistent upward trend with minor, short-term fluctuations.

3.2. ARIMA Methodology

An ARIMA(p, d, q) model was implemented to forecast prices based on historical trends. Parameters were tuned using grid search minimizing the Akaike Information Criterion (AIC), which balances fit quality and model simplicity.

The data was split into a training set and a 12-month test set for validation. Model performance was assessed with Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).  
After evaluation, the model was refitted on the full dataset and used to generate a 12-month forecast.

Model rationale:

* AR (Autoregressive): captures dependencies among past observations.
* I (Integrated): ensures stationarity by differencing.
* MA (Moving Average): models error correlations.

For future improvement, the model could be extended to SARIMA (seasonal), ARIMAX (exogenous variables such as interest rates), or advanced models such as Prophet or LSTM networks for nonlinear trends.

3.3. Results and Discussion

The ARIMA model achieved satisfactory accuracy, with error levels consistent with similar housing studies. The 12-month forecast indicated moderate, continuous price growth, reflecting strong demand and limited supply in King County.

However, macroeconomic variables such as interest rates, inflation, and construction costs can affect long-term accuracy. Periodic model updates and inclusion of external regressors are recommended to improve robustness.

**Part IV - Conclusions**

The combined application of clustering and ARIMA enhanced the understanding of King County’s housing market by:

* Segmenting the market into coherent clusters defined by price, size, and location;
* Forecasting average monthly prices, providing forward-looking insights for planning and investment.

Clustering supported market segmentation and strategic decision-making, while ARIMA provided short-term predictive value. Limitations include the lack of macroeconomic and spatial dependencies in the current models.

Future research could explore SARIMA/SARIMAX models, include economic indicators, or apply geospatial clustering at the ZIP code level.  
Overall, this study demonstrates the value of combining machine learning techniques for data-driven decision-makingin the housing and real estate sector.

References

Hyndman, R.J. & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice.* 3rd ed. O Texts.  
Kaufman, L. & 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>.  
Harlfoxem (n.d.). *House Sales in King County, USA.* Kaggle. Available at: <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction> (Accessed: today).