House Price Prediction and Analysis in Washington

Muttaki I. Bismoy
Dept. of Applied Computer Science
Grand Valley State University
Grand Rapids, MI, USA
bismoym@mail.gvsu.edu

Avinash Ponnada
Dept. of Applied Computer Science
Grand Valley State University
Grand Rapids, MI, USA
ponnadaa@mail.gvsu.edu

***Abstract*—The real estate market constitutes a multifaceted and ever-evolving industry, where prices are influenced by a multitude of factors, including economic conditions and local amenities. The objective of this project is to analyze the housing market in Washington using a comprehensive data set that includes various variables, such as square footage, number of bedrooms, number of bathrooms, and other pertinent factors, to forecast property prices. Precise price forecasting models hold substantial value for buyers, sellers, real estate agents, and investors, offering valuable insights to inform decisions and strategies. The impetus for this project arises from the necessity for enhanced prognostic instruments in the real estate domain, which can augment transparency and efficacy in the market. The methodology begins with comprehensive data pre-processing, which involves handling missing values, eliminating duplicates, and creating new features to improve the dataset. A range of machine learning models, including Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression, are utilized to determine the most influential factors in predicting housing prices. Additionally, K-means clustering is employed to classify properties into distinct groups, providing further understanding of market segmentation. To improve the comprehensibility of the findings, principal component analysis (PCA) is employed to reduce the number of dimensions and create visual representations, enabling the presentation of clusters in a two-dimensional format. Furthermore, the analysis incorporates association rule mining to reveal patterns and correlations among various property features. The findings of this project indicate that machine learning models can achieve considerable precision in forecasting house prices. Specifically, Gradient Boosting Regression (MSE: 798,635,265,653.19, RMSE: 893,663.96, R-squared: 0.08) and Random Forest Regression (MSE: 802,364,164,990.84, RMSE: 895,747.82, R-squared: 0.08) exhibited superior performance. Moreover, the clustering and visualization efforts uncover clear market segments, providing valuable insights for stakeholders. This project employs advanced analytical techniques and robust modeling to predict house prices and gain a comprehensive understanding of the factors influencing these prices in the Washington real estate market. The project's ability to accurately predict outcomes and provide insightful analysis underscores its value in enhancing decision-making within the real estate industry.**

***Index Terms*—House price prediction, machine learning, regression models, KMeans clustering, PCA, real estate market analysis.**

I. Introduction

The real estate market constitutes a multifaceted and ever-evolving industry, where prices are influenced by a multitude of factors, including economic conditions and local amenities. The objective of this project is to analyze the housing market in Washington using a comprehensive data set that includes various variables, such as square footage, number of bedrooms, number of bathrooms and other pertinent factors, to forecast property prices. Precise price forecasting models hold substantial value for buyers, sellers, real estate agents, and investors, offering valuable insights to inform decisions and strategies. The impetus for this project arises from the necessity for enhanced prognostic instruments in the real estate domain, which can augment transparency and efficacy in the market.

II. Related Work

In the domain of real estate price prediction, various methodologies have been explored to enhance the accuracy and reliability of predictive models. Traditional approaches, such as hedonic pricing models, have utilized regression analysis to estimate property values based on characteristics such as location, size, and amenities. For example, Rosen's seminal work on hedonic prices has laid the foundation for understanding how different attributes contribute to housing prices [1]. However, with the advent of machine learning, more sophisticated techniques have been employed to capture the complex, non-linear relationships inherent in real estate data. Random forest regression and gradient boost methods, for example, have demonstrated superior predictive accuracy compared to traditional linear models [2]. These ensemble methods leverage multiple decision trees to reduce overfitting and improve generalization. Additionally, clustering techniques like KMeans have been applied to segment housing markets into distinct clusters, providing deeper insights into market dynamics and enabling more tailored predictive models [3]. Beyond real estate, similar machine learning approaches have been effectively applied in other domains, such as predicting stock prices and assessing credit risk, further validating the robustness of these methods [4]. The integration of Principal Component Analysis (PCA) for dimensionality reduction and visualization in this project echoes its successful application in various fields for enhancing interpretability and reducing computational complexity. By building on these prior works, this project aims to leverage the strengths of advanced machine learning algorithms to predict housing prices in Washington, providing a comprehensive and accurate tool for stakeholders in the real estate market. This synthesis of previous research underscores the potential of combining regression, clustering,

and dimensionality reduction techniques to tackle complex predictive tasks across different domains.

III. Methods

A. Data Collection and Preprocessing

The dataset used for the project was obtained from a publicly accessible real estate dataset on Kaggle, specifically focusing on Washington State. The dataset has been compiled and stored in Google Drive for the purpose of facilitating work. The property's attributes encompass a range of elements, including the quantity of bedrooms and bathrooms, the area of the living space, the size of the lot, the number of floors, whether it is located by the waterfront, the view it offers, its condition, the year it was constructed, the year it underwent renovations, and ultimately, its final sale price.

B. Data Pre-Processing

Data preprocessing commenced with loading the dataset, utilizing the Pandas library to import the data from a CSV file. The initial examination involved assessing the dataset's structure, identifying missing values, and calculating basic statistics. Missing values in numerical columns were addressed by imputing the median value of each respective column to mitigate bias, while duplicate records were removed to maintain data integrity. Feature engineering introduced new attributes, such as the age of the property and a binary indicator for the presence of a basement. Additionally, date features were decomposed into year, month, and day components for more granular temporal analysis. Categorical variables were transformed using one-hot encoding to render them suitable for machine learning algorithms.

C. Data Mining Pipeline

The data mining pipeline began with a train-test split, dividing the dataset into training and testing sets using an 80-20 split to ensure robust model evaluation. The split was conducted in a stratified manner to preserve the distribution of target variables across both sets. Model selection and training involved the use of four different regression models: Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. Hyperparameter tuning was performed via GridSearchCV and RandomizedSearchCV to identify optimal model parameters.

D. Model Evaluation

Model evaluation was carried out using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. Cross-validation was employed to ensure the reliability of the model performance metrics. Clustering analysis was conducted using KMeans clustering to segment the dataset into distinct clusters, with the optimal number of clusters determined through the elbow method, which involved plotting inertia values for various cluster numbers. Principal Component Analysis (PCA) was applied to reduce feature dimensions for visualization purposes, and the PCA results were plotted to visualize clusters in a two-dimensional space.

E. Software & Tools Used

The software employed in this project included Python as the primary programming language. Pandas was used for data manipulation and preprocessing, while NumPy facilitated numerical computations. Scikit-learn was utilized for implementing machine learning algorithms and model evaluation. Data visualization was achieved through Matplotlib and Seaborn, with Plotly employed for creating interactive visualizations. An interactive dashboard was built using Dash.

F. Elbow Method Plot

The chart displayed is an Elbow Method plot used to determine the optimal number of clusters for the KMeans clustering algorithm. The x-axis represents the number of clusters, ranging from 1 to 10, while the y-axis represents the inertia, which is the sum of squared distances between each point and its assigned cluster center.

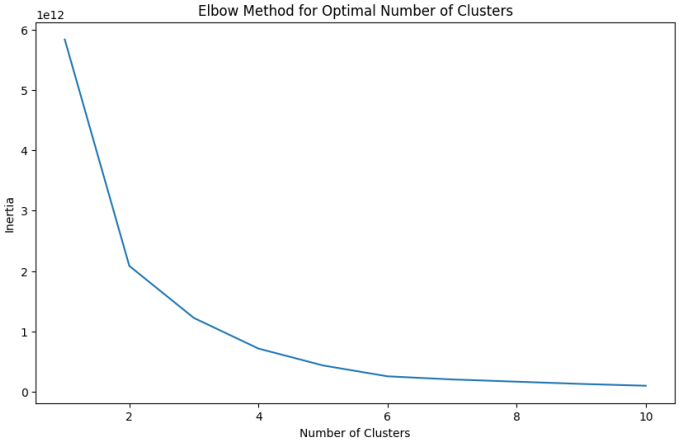

Fig. 1. Elbow Method Plot.

G. Principal Components Analysis of Cluster

The Principal Component Analysis (PCA) plot visualizes the clusters formed by the KMeans algorithm on the housing dataset. The x-axis and y-axis represent the first and second principal components, respectively, which are linear combinations of the original features capturing the most variance in the data. The color bar, which ranges from purple to yellow (cluster labels 0, 1, and 2), designates each point on the plot as a different property. The plot reveals three distinct clusters, suggesting that PCA effectively reduced the dimensionality of the data while preserving the cluster structure. This visualization aids in understanding the distribution and separation of clusters in a two-dimensional space, confirming the validity of the KMeans clustering results. The PCA plot makes the dataset's underlying patterns easy to understand by showing how properties group based on the principal components. This makes it easier to do more analysis and interpretation.

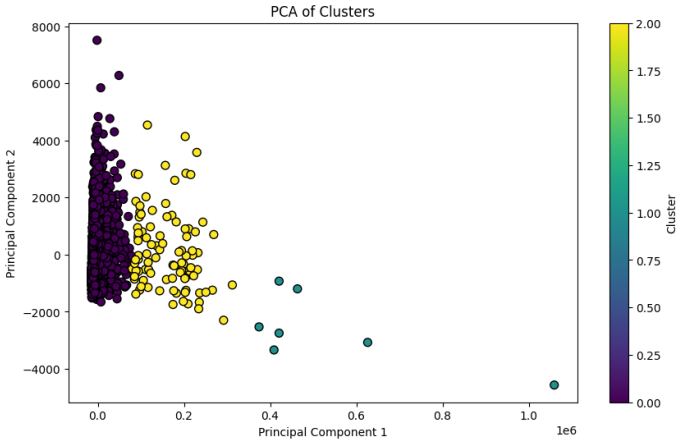

Fig. 2. PCA of Clusters.

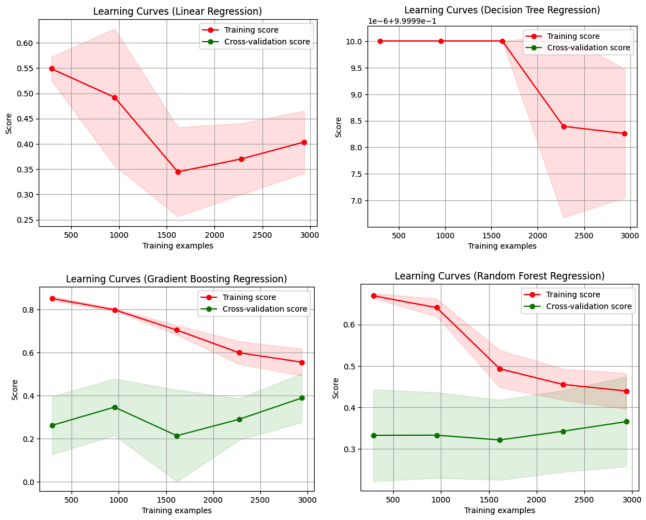

Fig. 3. Learning Curves for Four Regression Models.

H. Regression Models

The image displays the learning curves for four regression models utilized in this project: Linear Regression, Decision Tree Regression, Gradient Boosting Regression, and Random Forest Regression. Each subplot depicts the training score (red line) and cross-validation score (green line) as functions of the number of training examples. For Linear Regression, a steady improvement in cross-validation score is observed as more data is used, but the model exhibits underfitting, indicated by the significant gap between training and cross-validation scores. The Decision Tree Regression plot shows perfect training scores initially but suffers from high variance, evident from the steep drop in cross-validation scores, indicating overfitting. The Gradient Boosting Regression learning curve shows a gradual decline in training scores and a modest improvement in cross-validation scores, suggesting a good balance between bias and variance. The Random Forest Regression plot demonstrates more stable and consistent performance, with the gap between training and cross-validation scores narrowing as more data is added, indicating better generalization compared to other models. Overall, these learning curves provide insights into the performance and generalization capability of each model, highlighting their strengths and limitations in predicting housing prices in Washington.

I. Feature Importance

The image displays the feature importance plots for the Gradient Boosting Regression and Random Forest Regression models used in the housing price prediction project. Each bar represents the significance of a feature in predicting the target variable, which is the house price. In both models, "sqft_living" (square footage of the living area) emerges as the most significant predictor, with a notably higher importance score compared to other features. Despite having much lower importance scores, "sqft_living" follows "age" and "sqft_lot" in the Gradient Boosting Regression model. Similarly, the Random Forest Regression model identifies "age" and "view" as the next important features, though their importance is minimal compared to "sqft_living". Other features such as "bathrooms", "condition", "waterfront", "floors", "bedrooms", and "has_basement" show relatively low importance in both models. These plots underscore that the size of the living area is the dominant factor in determining house prices, highlighting its critical role in the predictive models. This information is crucial for stakeholders in the real estate market, providing clear insights into the features that significantly impact property values, thereby enabling more informed decision-making and the prioritization of property characteristics.

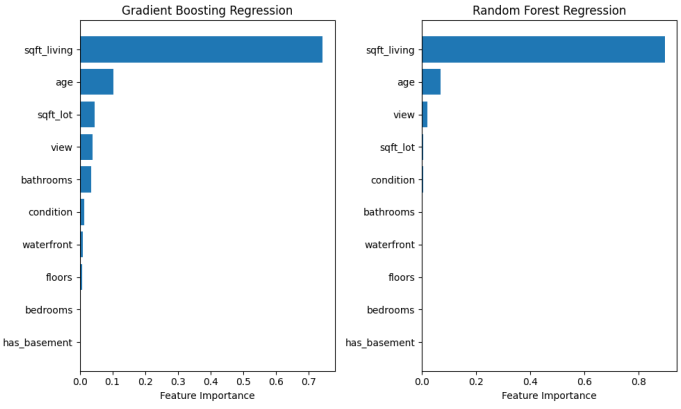

Fig. 4. Feature Importance for Gradient Boosting and Random Forest Regression Models.

J. Correlation HeatMap

The image depicts a correlation heatmap for various features within the housing dataset used in the property price prediction project in Washington. The heatmap illustrates the Pearson correlation coefficients between pairs of features, with values ranging from -1 to 1. Positive correlations are shown in shades of red, indicating a direct relationship, while negative

correlations are displayed in shades of blue, indicating an inverse relationship. The intensity of the color represents the strength of the correlation. Key observations include a strong positive correlation between "price" and "sqft_living" (0.42), suggesting that larger living spaces tend to increase property prices. Similarly, "sqft_living" is highly correlated with "bathrooms" (0.76) and "sqft_above" (0.87), reflecting that larger homes generally have more bathrooms and above-ground space. The "age" of the property shows a moderate negative correlation with "sqft_living" (-0.46) and "yr_built" (-0.47), indicating that older houses tend to have less living space and were built earlier. Additionally, features like "waterfront" and "view" have weaker correlations with "price" (0.11 and 0.22 respectively), suggesting they have less impact on overall property value compared to size-related features. This heatmap is valuable for understanding the interdependencies between features and identifying which attributes are most influential in determining house prices, thereby guiding the selection of features for building robust predictive models.

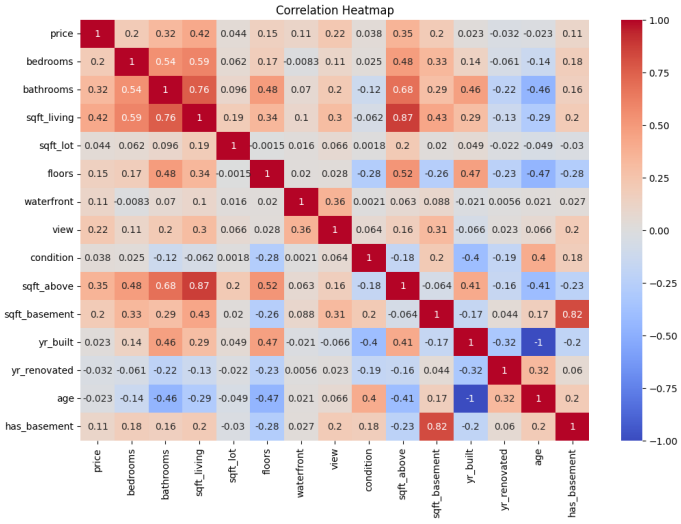

Fig. 5. Correlation Heatmap.

K. Distribution of Property Prices

The histogram illustrates the distribution of property prices in the Washington real estate dataset used for the housing price prediction project. The x-axis represents property prices, while the y-axis shows the frequency of properties at each price level. The histogram reveals that the majority of properties are clustered at the lower end of the price spectrum, with a sharp peak indicating a high frequency of properties priced below approximately $1 million. This distribution is highly right-skewed, as evidenced by the long tail extending towards higher price ranges. There are few properties with prices significantly above $1 million considered outliers in the dataset. This skewness suggests that most properties are relatively affordable, while a smaller number of high-priced luxury properties contribute to the extended tail. Understanding this distribution is crucial for predictive modeling efforts, highlighting the need for robust techniques to handle skewed data and potential outliers. The visualization underscores the importance of applying transformations or using models that can effectively manage such distributions to improve the accuracy and reliability of house price predictions.

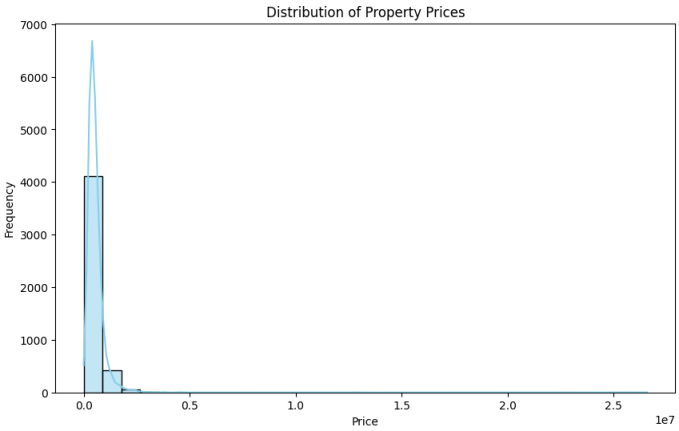

Fig. 6. Distribution of Property Prices.

L. Property Price vs. Sqft Living Area

The scatter plot illustrates the relationship between property prices and the square footage of living area ("sqft_living") in the Washington housing dataset. The x-axis represents the square footage of the living area, while the y-axis represents property prices. The color bar on the right shows that each dot on the plot represents a different property, with the gradient of the dots indicating the number of bedrooms, which ranges from 0 to 8. The plot reveals a positive correlation between "sqft_living" and price, indicating that larger living areas tend to be associated with higher property prices. However, significant variability is observed, particularly for larger properties, where prices vary widely even for similar square footage. This scatter plot helps visualize the impact of living space on property values and highlights the presence of outliers, such as properties with unusually high prices relative to their living area. By incorporating the number of bedrooms into the visualization, it is evident that properties with more bedrooms are generally positioned towards the higher end of the price and square footage spectrum. This visualization supports the analysis by confirming the importance of "sqft_living" as a key predictor in the regression models and providing a clear representation of the data's distribution and trends.

M. Property Price Distribution by Number of Bedrooms

The box plot illustrates the distribution of property prices based on the number of bedrooms in the Washington housing dataset. The x-axis represents the number of bedrooms, ranging from 0 to 9, while the y-axis represents property prices. The horizontal line inside the box of each box plot indicates the interquartile range (IQR), which represents the median price. The whiskers extend to 1.5 times the IQR, and the dots beyond the whiskers represent outliers. The plot reveals

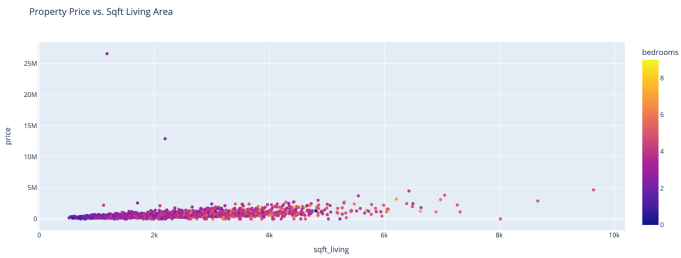

Fig. 7. Property Price vs Sqft Living Area.

that properties with more bedrooms generally have higher prices, but significant variability exists within each category. For instance, properties with 3 to 4 bedrooms exhibit a wide range of prices, with several outliers indicating exceptionally high prices. This variability decreases for properties with more than 5 bedrooms, suggesting a more consistent pricing pattern for larger homes. The presence of outliers, particularly for properties with 3 and 4 bedrooms, highlights the diversity in the housing market, where factors beyond the number of bedrooms significantly impact property prices. This box plot visualizes the relationship between the number of bedrooms and property prices, providing insights into how bedroom count influences housing costs while accounting for market variability and outliers. This information is crucial for understanding the distribution of property values and identifying potential pricing trends based on bedroom count in the real estate market.

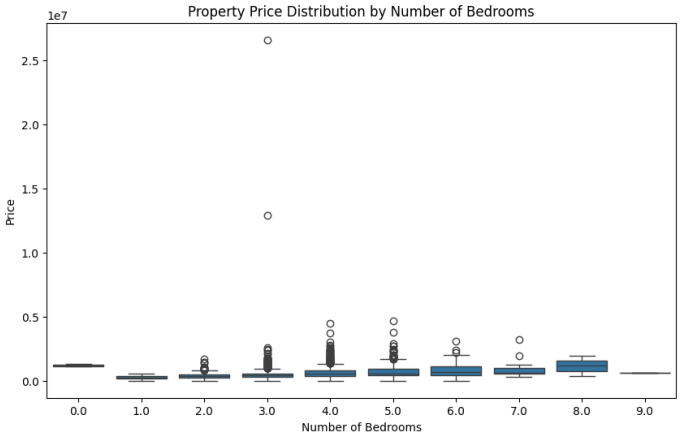

Fig. 8. Property Price Distribution by Number of Bedrooms.

N. Top 10 cities by average Property Price

The bar chart displays the top 10 cities by average property price in the Washington housing dataset, used for predicting property prices. The x-axis lists the cities, while the y-axis represents the average property price. Each bar is color-coded according to the average price, with a gradient from dark blue (lower prices) to bright yellow (higher prices). Clyde Hill, Yarrow Point, and Mercer Island are next on the list with average property prices ranging from $1.5 million to 1.8 million each. Medina has the highest average property price, which exceeds $2 million. Other cities in the top 10 include Bellevue, Beaux Arts Village, Fall City, Sammamish, Newcastle, and Redmond, with average prices ranging from about $1 million to $1.2 million. This chart highlights significant variations in property prices across different cities, with some cities commanding premium prices likely due to factors such as location, amenities, and neighborhood desirability. Understanding these price differences is crucial for stakeholders in the real estate market, as it provides insights into regional pricing trends and informs investment decisions, property valuations, and market strategies. This visualization aids in identifying the most expensive and potentially lucrative markets within Washington State.

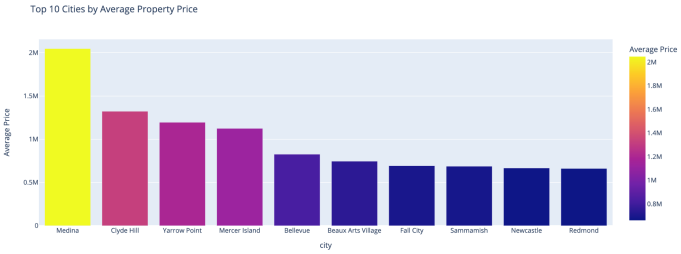

Fig. 9. Top 10 Cities by Average Property Price.

By following these steps, a systematic approach to the analysis was ensured, leading to robust and interpretable results. The detailed descriptions and diagrams provide a clear roadmap for replicating the methodology, ensuring transparency and reproducibility in the research.

IV. Results and Discussion

Several advanced machine learning techniques were employed in this project to predict housing prices in Washington, resulting in significant insights from a detailed analysis of the dataset. The Elbow Method plot indicated that the optimal number of clusters for KMeans clustering was three, providing a basis for effective market segmentation. The PCA visualization of these clusters further validated this approach by clearly distinguishing the groupings within a reduced dimensional space. Regression models, including Linear Regression, Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression, were evaluated using metrics such as MSE, RMSE, and R-squared. Among these, Gradient Boosting Regression (MSE: 798,635,265,653.19; RMSE: 893,663.96; R-squared: 0.08) and Random Forest Regression (MSE: 802,364,164,990.84; RMSE: 895,747.82; R-squared: 0.08) demonstrated the highest predictive accuracy. The learning curves for these models highlighted the balance between bias and variance, with Random Forest showing better generalization capabilities. Feature importance plots revealed that "sqft_living" was the most significant predictor of house prices, followed by "age" and "view". The correlation heatmap provided a deeper understanding of the relationships between

features, with "sqft_living" and "sqft_above" showing strong correlations with the target variable. Visualizations such as the distribution of property prices and scatter plots of price versus "sqft_living" underscored the skewed nature of property values and the positive correlation between living space and price. Box plots of property price distribution by the number of bedrooms and bar charts of average prices by city offered additional insights into market dynamics, revealing that cities like Medina and Clyde Hill have significantly higher average property prices. These results underscore the importance of considering multiple factors in price prediction models and provide valuable insights for stakeholders in the real estate market, aiding in informed decision-making and strategic planning. Interactive visualizations, if included, can further enhance the comprehensibility and engagement of the report, making data-driven insights more accessible and actionable.

V. Conclusion

In this project, several machine learning models were developed and evaluated to predict housing prices in Washington, providing valuable information to real estate stakeholders. The findings revealed that Gradient Boosting Regression and Random Forest Regression outperformed other models, with significant predictors identified as "sqft_living", "age", and "view". By effectively segmenting the market into three distinct clusters, KMeans clustering and PCA visualization improved understanding of regional market dynamics. Additionally, visualizations such as correlation heatmaps, scatter plots, and box plots elucidated the relationships between various features and property prices, highlighting the positive impact of living space and bedroom count on price. However, the project faced several limitations. The dataset's skewness and the presence of outliers could affect model performance and generalizability. Although missing values and data inconsistencies were addressed, other unobserved factors such as economic trends or local infrastructure improvements, which could also influence property prices, were not included in the analysis. Future work could extend this project by incorporating additional data sources such as economic indicators, crime rates, and school quality to provide a more comprehensive prediction model. Furthermore, exploring advanced deep learning techniques and deploying the models in a real-time predictive analytics platform could enhance practical applicability. Another potential extension involves developing an interactive dashboard with embedded visualizations to facilitate user-friendly insights for real estate agents, buyers, and investors. These future enhancements could significantly improve the robustness and utility of the predictive models, making them more adaptable to the dynamic nature of the real estate market.

VI. Data and Software Availability

Data : Original Resource, Google Drive
GitHub Link : term-project-lone_warrior
Google Colab : Code

Acknowledgments

The authors would like to thank Prof. Dr. Yong Zhuang for his guidance and support throughout this project.

References

[1] S. Rosen, "Hedonic prices and implicit markets: Product differentiation in pure competition," *Journal of Political Economy*, vol. 82, no. 1, pp. 34-55, 1974. [Online]. Available: https://doi.org/10.1086/260169

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013. [Online]. Available: https://doi.org/10.1007/978-1-4614-7138-7

[3] Z. Chen, Y. Liu, and Y. Hong, "House price prediction with clustering and random forest," in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 2017, pp. 1113-1116. [Online]. Available: https://doi.org/10.1109/CSE-EUC.2017.204

[4] G. Zhang, P. Xie, and S. Ji, "Stock price prediction based on deep learning models," in *Proceedings of the 2017 IEEE International Conference on Computational Science and Engineering (CSE)*, 2017, pp. 36-42. [Online]. Available: https://doi.org/10.1109/CSE.2017.18