Data Analysis and Price Prediction of House Sales Data

Chris Aryan [S20210020265]
ECE, IIIT Sri City
chris.a21@iiits.in

Sai Likhitha [S20210020289]
ECE, IIIT Sri City
sailikhitha.k21@iiits.in

***Abstract*— This study aims to analyze the factors influencing house prices and develop a predictive model for house prices using data analysis techniques. Our analysis revealed significant correlations between certain features and house prices, shedding light on the dynamics of the housing market. Building upon these insights, we developed a predictive model using linear regression.**

Keywords: House sales data, data analysis, price prediction, linear regression, exploratory data analysis.

I. Introduction

- The dataset contains information related to real estate properties.
- It includes various attributes that describe the properties and their characteristics.
- The dataset is focused on housing market data and related features.
- Data points cover a range of parameters that influence property prices.
- The dataset spans multiple dimensions such as location, size, condition, and amenities.
- It comprises both numerical and categorical variables.
- The dataset is intended for analysis and modeling tasks related to real estate pricing.
- Variables like 'price', 'bedrooms', 'bathrooms', and 'sqftliving' are key indicators.
- Additional attributes like 'waterfront', 'view', and 'grade' provide further insights into property quality.
- The dataset includes information on the geographical location of properties (e.g., 'zip code', 'lat', 'long').

Understanding the factors that drive fluctuations in house prices is essential for homeowners, investors, policymakers, and industry professionals alike. From economic shifts and demographic trends to locational attributes and property features, numerous factors influence the value of residential properties.

In this report, we delve into the domain of extensive data analysis and price prediction. Our study aims to dissect the various factors that shape house prices and hold the importance of predictive modeling to anticipate future price movements.

Our objectives are twofold:

1. First, to conduct a comprehensive data analysis of the determinants of house prices, uncovering underlying patterns, correlations, and trends within the data.
2. Second, to develop predictive models capable of forecasting house prices with a high degree of accuracy and reliability.

By achieving these objectives, we aim to empower stakeholders with the knowledge and tools necessary to navigate the nuances of the real estate market effectively.

II. Methodology

We have the dataset prepared. We then begin with the preprocessing steps, ensuring the integrity and quality of the dataset. We further perform an extensive exploratory data analysis, uncovering insights that shed light on the dynamics of the housing market. Next, we follow predictive

modeling, developing and evaluating models for house price prediction.

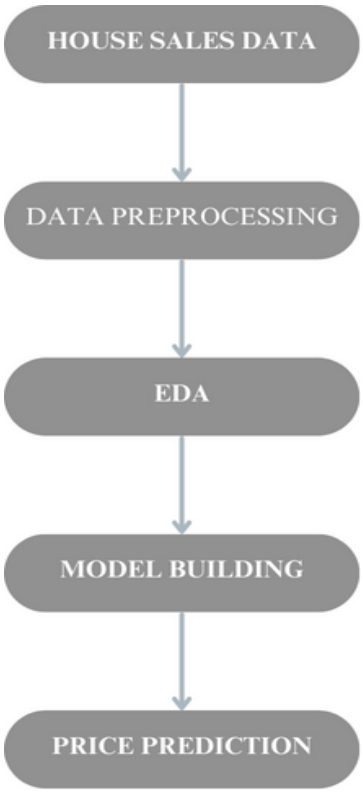

Fig. 1

2.1 Dataset Description

- The dataset consists of housing-related information.
- It contains 21613 property listings and 21 attributes.
- Numeric Features: price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, view, condition, grade, sqft_above, sqft_basement, yr_built, yr_renovated, lat, long, sqft_living15, sqft_lot15.
- Categorical Feature: waterfront (binary), zip code.

2.2 Data Preprocessing

A. Removing duplicates:
We notice that there are some duplicates of the same property. The only difference is in the valuation date. So, we prepare a dataframe which consists only of the properties with their recent valuation prices. We further create a duplicate csv, which contains the properties which have different prices at two different dates.

B. Handling Nan values:
Our dataset is independent of any Nan values.

C. Removing Outliers:
We plot boxplot for different columns and analyse the outliers. Based on the correlation of the attribute with our target variable, we proceed to remove outliers using the Interquartile Range method.

Not removing outliers completely to check the robustness of our model as we will be applying Linear Regression which itself is compatible with outliers as its assumptions, such as linearity and homoscedasticity, are not significantly affected by a few extreme data points.

Outliers in the target variable can affect the accuracy of our model significantly.

Generally, we replace the outliers in the numeric column with the mean of the data. But as we are dealing with house sales data, replacing the price outliers with the mean can affect our analysis and prediction. Therefore, we drop the outliers from our dataframe.

2.2 Exploratory Data analysis

A. From the duplicate csv we have kept only unique properties and added two different valuation columns for two different dates. Fig. 2 describes the approach.

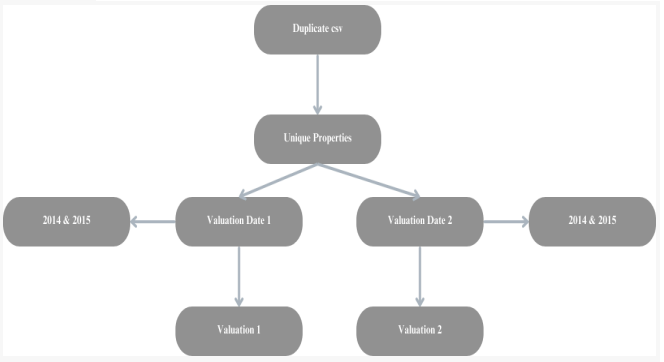

Fig. 2

Further, we have calculated the valuation change for the different valuation dates, i.e. valuation change for properties valued and revalued in 2014 and 2014 respectively, valued and revalued in 2014 and 2015 respectively and so on.

B. Fig. 3 shows the correlation between the target variable, i.e. price, and the rest of the attributes. We can clearly observe significant correlation between price and bedrooms, bathrooms, square living, view, grade, square foot above and square foot living of the neighbouring 15 properties.

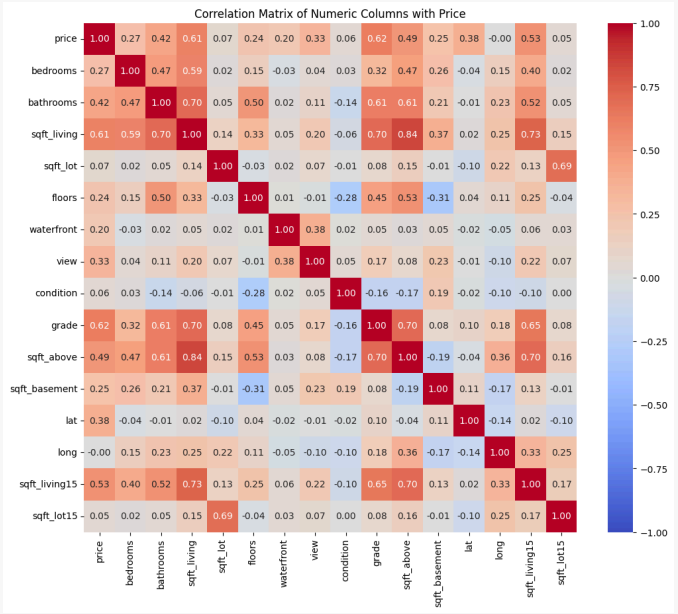

Fig. 3

To visualize the relationship between price and our significant variables, we proceed by plotting the regression line plot of these individual variables with our target variable.

Furthermore, comprehensive analysis has been conducted on the dependent variable, price, and the independent attributes.

1. Analysing the distribution of price with outliers, the graph is observed to be skewed. Removing the skewness to get normal distribution by removing outliers.
2. Dividing the properties based on the median price as seen in Fig. 4.

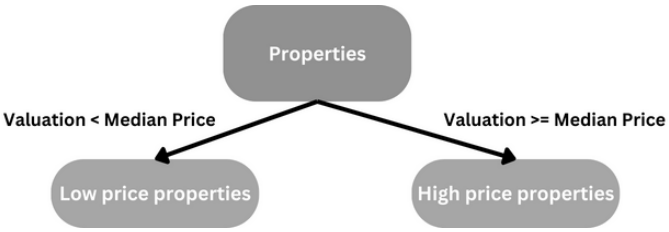

Fig. 4

3. Further analysis on the two categories has been done, like distribution of grade and view.
4. Also, we have used the folium library to visualize the distribution of high and low priced properties on map.
5. We have used scree plot to find out the optimal number of clusters for the latitudes and longitudes, which came out to be 3.
6. Following this we have applied the K-means algorithm to cluster all the properties into 3 clusters and calculated the average price for each cluster.
7. Further, we have performed analysis on zip code. Like the count of properties at each zip code and the average price of properties at each zip code.
8. In order to analyse relationships between significant attributes, we have used pairplot to plot their relation.

2.3 Model Building

Chose linear regression as the base model and decided to use hyperparameter tuning to optimize the model's performance.

A. *Data collection and preprocessing:* Cleaned the data by handling duplicate values, removing outliers and finding new attributes to remove even further outliers. The column consisting of zip codes of different places is one hot encoded and the original zipcode column is dropped.
B. *Train-test split:* The dataset is split into training and testing sets to evaluate the

performance of the model. The training set, typically comprising 80% of the data, is used to train the model, while the testing set 20% is used to assess the model's predictive performance on unseen data.

C. *Model selection and hyperparameter tuning:* Chose linear regression as the base model and used GridSearchCV to find the best hyperparameters.
D. *Model Training:* Trained the linear regression model using the training dataset.
E. Model Evaluation: Evaluated the model's performance on the testing dataset using R-squared, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and adjusted R-squared metrics.
F. *Model Interpretation:* Analyzed the model's performance metrics to assess its effectiveness in predicting house prices. Interpreted the coefficients of the linear regression model to understand the impact of each feature on house prices.

III. Results

3.1 Valuation change analysis

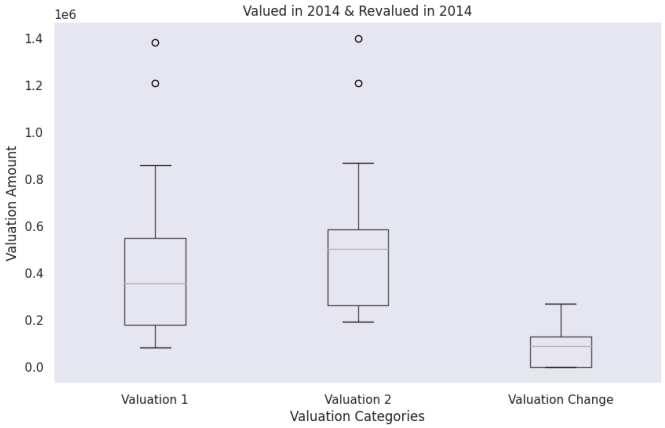

Fig. 5

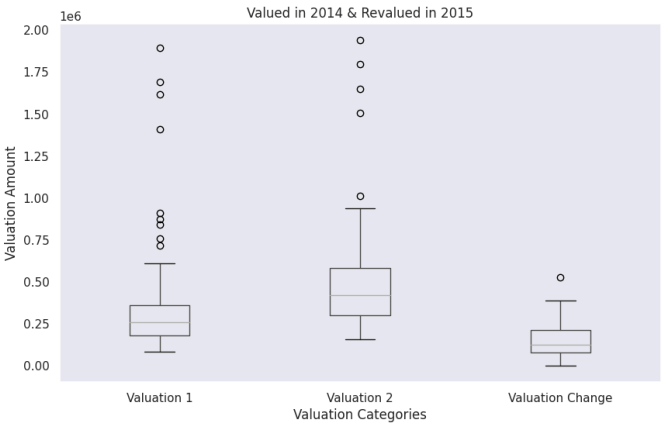

Fig. 6

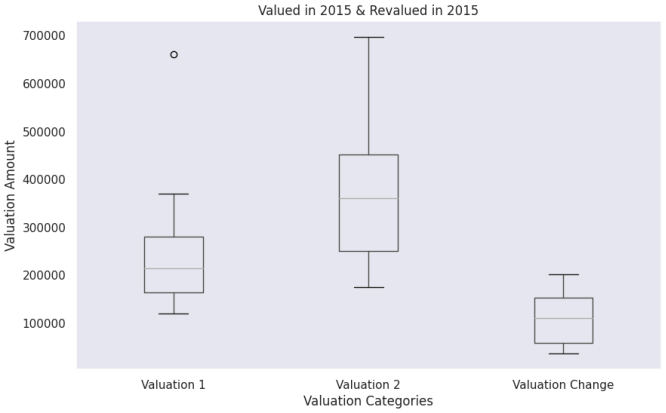

Fig. 7

From the above figures, we can see that the valuation change is minimum for properties valued and revalued in 2014 and 2015 respectively. While it is maximum for the properties valued and revalued in 2015.

Inference: We infer that the real estate market saw a relatively increased growth in the fiscal year of 2015-16.

3.2 Analysis on high and low price properties

1. As we have seen that the attribute grade had a relatively higher correlation with prices. Fig. 8 demonstrated the distribution.

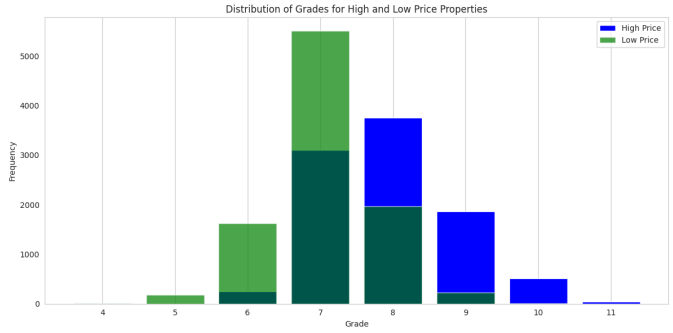

Fig. 8

From the figure above, we can clearly see that the low price properties in green have their grade distribution between 5 to 9, with maximum properties having a grade of 7.Whereas the properties with the high price tag in blue colour can be seen to have their grade distribution between 6 to 11, with maximum properties having a grade of 8.

Inference: The properties with higher price are having a better grade distribution as compared to those with lower price tag.

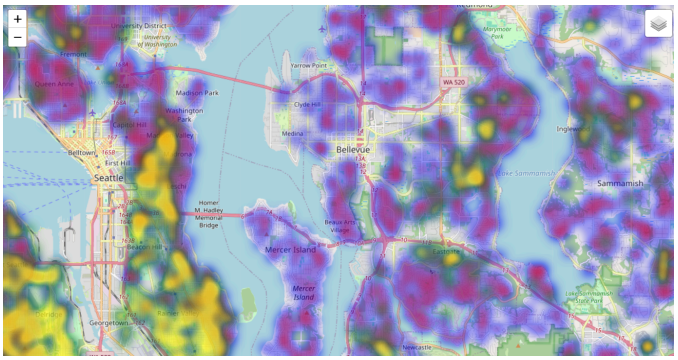

Fig. 9

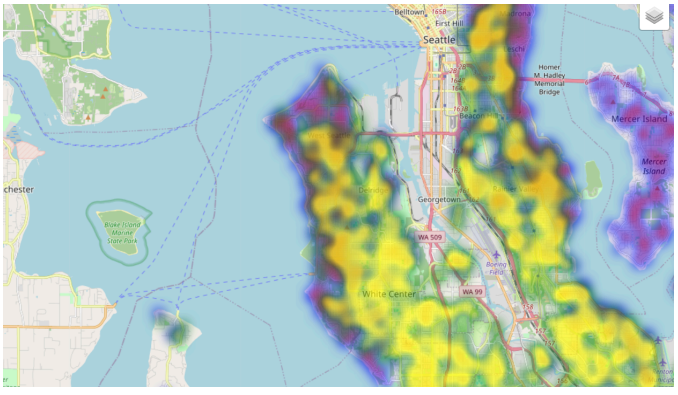

Fig. 10

2. From Fig. 9 and 10, we can see the distribution of high and low price properties on a map. The purple and blue coloured region indicates properties with high price while the green and yellow region indicates the opposite.

Inference: The properties along the coast as well as those on small islands, notice the Mercer island in Fig. 10, are having higher rates as compared to those which are landlocked.

3. Moving forward with the analysis, we have clustered the properties in accordance with the latitudes and longitudes into 3 clusters using the K-means algorithm. The optimal number of clusters was decided by plotting a scree plot. Fig. 11 shows the avg price at each cluster.

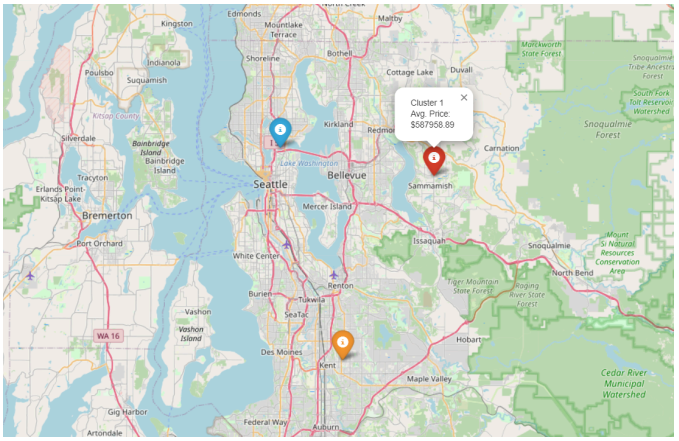

Fig. 11

Inference: Cluster 1 in red is having the highest average price, followed by cluster 2 in blue and then cluster 3 in orange. This helps the buyers to wisely choose their region of interest while investing into real estate.

3.3 Analysis on the basis of zip code:

A. Grouped the data by bedrooms and price per sqft column and calculated statistics for each group.

The scatter plot of zip code 98001:

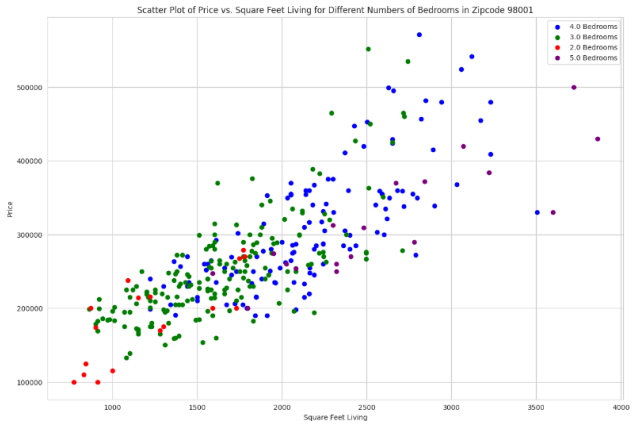

Fig. 12

The mean of every bedroom type is calculated at every zip code, and the outliers are identified by comparing the values within each group of bedroom and zip code with the mean of the previous groups.Such outliers are removed for every zip code.

The scatter plot of zip code 98001 after removing outliers:

Fig. 13

B. Furthermore, for each zip code, we have plotted the frequency distribution of properties and average rate of properties at each zip code.

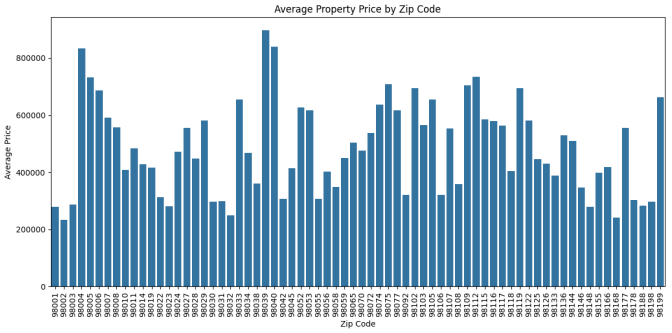

Fig. 13

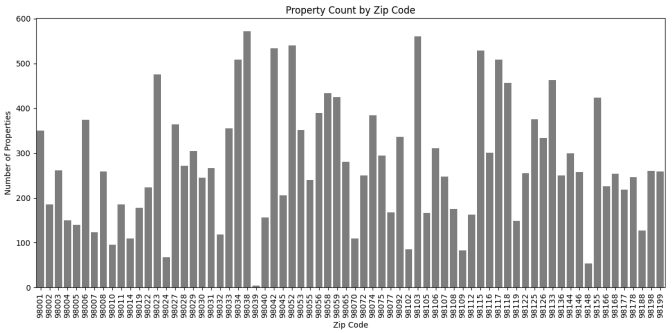

Fig. 14

From the above two figures we can see that zip code 98038 has the maximum number of properties but at the same time the average price at this zip is comparatively very low. While zip 98039 has the highest average price but the number of properties at this zip are the lowest.

Inference: This analysis helps the investor to pick out the zip code in which they want to invest according to their preference of “high standard, less populated” society or “mediocre property, with dense population”.

3.4 Square footage analysis

A. From the correlation analysis we saw significant correlation between the square footage values of the properties. Fig. 15 shows the relationship between the attributes.

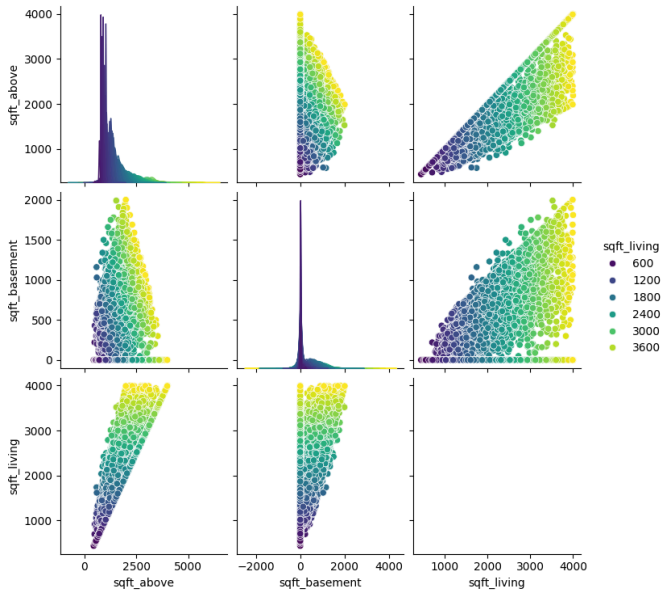

Fig. 15

We observe a linear relationship between square footage above and square footage living, while on the other hand square footage basement stays constant with the square footage living.

Inference: This makes sure that the size of the basement is not affecting my overall living size. This assures the investors that they pay for the increase in living space solely when they opt for spacious properties.

B. In some cases the living space is greater than the lot space, which generally is not possible. But there are some factors that influence this which are:
    a. The property might have been renovated.
    b. The property might have more floors.
    c. The property might be old, and due to the old architectural designs the living space is more.

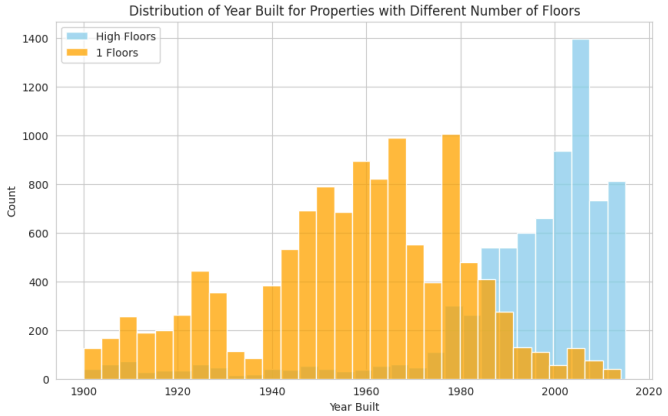

Fig. 16

The above figure clearly demonstrates the increase in the number of floors in the properties as we shift into the 21st century.

Inference: Multi Storey properties assures the investors of its recent construction. This helps the investors to diversify their investment into different kinds of properties which ultimately reduces the risk factor.

3.5 Inferences from linear regression:
After applying the Linear Regression model to the testing set, we obtained the following evaluation metrics:

Mean Absolute Error (MAE): 62266.86
Mean Squared Error (MSE): 7565546051.5
Root Mean Squared Error (RMSE): 86980
R^2 value: 0.802
Adjusted R^2 value: 0.799

Mean Absolute Error (MAE): MAE is the average of all the absolute differences between the predicted price values and the actual values and provides a straightforward way to assess how well a model's predictions match the actual values, with lower MAE indicating better performance.In this case, MAE is 62266, suggests that, on average, the model's predictions are off by approximately 62266 units.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): MSE is computed by averaging over all data points the square of the difference between the actual and predicted values

and RMSE is the square root of the MSE and represents the standard deviation of the residuals. In this case RMSE is 86980.

R^2 value: The R^2 value is a statistical measure that represents the proportion of the variance in the price value that is predictable from the independent variables in the regression model. An R^2 value closer to 1 indicates the independent variables in the model contribute to a larger proportion of the variance in the dependent variable. The R^2 value of 0.802 suggests that approximately 80.22% of the variance in price values is explained by the model.

Adjusted R^2: Adjusted R^2 considers both the number of predictors and R^2. It penalizes the addition of unnecessary predictors and is a more reliable measure of the model's goodness of fit when comparing models with different numbers of predictors. In this case, an R_a^2 value of 0.799 means that approximately 79.9% of the variance in the dependent variable is explained by the independent variables, adjusted for the number of predictors in your model.

The values of actual prices and predicted prices are as below:

Zipcode	Bedrooms	Bathrooms	Sqft_living	Sqft_lot	Floors	Waterfront	View	Condition	Grade	Sqft_above	Sqft_basement	Sqft_living15	Sqft_lot15	Actual_price	Pred_price
98178	3	1	1180	5650	1	0	0	3	7	1180	0	1340	5650	221900	206136.56
98178	3	1	1200	8000	2	1	1	3	7	1000	200	1340	5650		264930.07
98125	3	2	2570	7242	2	0	0	3	7	2170	400	1690	7639	538000	523956.06
98125	1	1	500	300	2	0	1	3	6	500	0	800	900		251784.76

This model helps buyers by providing more accurate predictions of house prices based on their specific needs and preferences.

IV. Conclusion

Market Growth in 2015-16: The observed trends suggest increased growth in the real estate market during the fiscal year of 2015-16.

Price and Grade Distribution: Properties with higher prices tend to have better grade distributions. This implies that higher-priced properties may offer superior quality and amenities, appealing to buyers seeking higher standards.

Location Influence on Rates: Properties located along the coast or on small islands command higher rates compared to landlocked properties. Location plays a significant role in property valuation and can impact investment potential.

Cluster Analysis for Investment Strategy: Clustering analysis shows that certain regions may be more desirable for investment based on price performance.

Preference-Based Investing: Investors can tailor their investment strategy based on preferences such as desiring a "high standard, less populated" society or prioritizing property affordability and population density.

Basement Size vs. Living Space: Understanding that basement size does not significantly impact overall living space assures investors that they are paying for usable living space efficiently, without overpaying for unnecessary features.

The linear regression model demonstrated a reasonably good fit to the data, with an MAE of 65335 and an R-squared value of 0.7947, indicating that approximately 79.47% of the variance in house prices is explained by the model.The analysis provides valuable insights for buyers and sellers, aiding in pricing strategies and decision-making in the real estate market.

Overall, this house price prediction contributes valuable insights into the dynamics of the real estate market, providing a foundation for informed decision-making and further research in the field.