**Data Collection**

**Description of data source –**

The dataset includes statistics on home prices in King County, Washington. This project helps to address the issue of potential purchasers in the housing market in this county, which is predominantly urban and suburban. The sheer number of factors that influence price could easily cause confusion. This research will provide deeper understanding of the elements that influence prices, enabling potential buyers, realtors, and builders to make data-driven decisions.

The dataset includes statistics on property sale prices in King County, Washington. The collection includes historical data in more than 21,000 rows and 21 columns, including the price, size, location, year of construction or remodeling, number of bedrooms and bathrooms, and King County-based rating given to the residence (1-poor,13-excellent).

The built-up area has several extra characteristics that affect a house's price, pricing strategy, and marketing. The main objective is to predict home sales in King County, Washington.

**Data Set Variables –**

Dataset consists of 21 Variables and 21613 Observations.

|  |  |
| --- | --- |
| **Variables** | **Brief description** |
| id | Unique Identifier |
| date | Sold date of House |
| price | Price of the House |
| bedrooms | Number of Bedrooms |
| bathrooms | Number of Bathrooms |
| sqft\_living | Area of the House (in sqft) |
| sqft\_lot | Area of the lot (in sqft) |
| floors | Levels in the House |
| waterfront | House with waterfront |
| view | Number of House views |
| condition | Condition of the House |
| grade | The overall grade is given based on King County |
| sqft\_above | square footage of the House apart from the basement |
| sqft\_basement | square footage of the basement |
| yr\_built | Built Year |
| yr\_renovated | Year when House was renovated |
| zip code | zip code |
| lat | Latitude coordinate |
| long | Longitude coordinate |
| sqft\_living15 | square footage of interior HVAC Area of nearest 15 neighbors |
| sqft\_lot15 | square footage of the land of the nearest 15 neighbors |

**Data Description –**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Summary of data | Minimum | Quadrant 1 | Median | Mean | Quadrant 3 | Maximum |
| id | 1000000.00 | 2123000000.00 | 3905000000.00 | 4580000000.00 | 7309000000.00 | 9900000000.00 |
| price | 78000 | 322000 | 450000 | 540297 | 645000 | 7700000 |
| bathroom | 0.5 | 1.75 | 2.25 | 2.116 | 2.5 | 8 |
| floor | 1 | 1 | 1.5 | 1.494 | 2 | 3.5 |
| lat | 47.16 | 47.47 | 47.56 | 47.57 | 47.68 | 47.78 |
| long | -122.5 | -122.3 | -122.2 | -122.2 | -122.1 | -121.3 |
| date | 1 | 3 | 3 | 3.373 | 4 | 33 |
| grade | 3 | 7 | 7 | 7.658 | 8 | 13 |
| sqft\_above | 370 | 1190 | 1560 | 1789 | 2210 | 9410 |
| sqft\_basement | 0 | 0 | 0 | 84.46 | 0 | 2015 |
| yr\_built | 1900 | 1951 | 1975 | 1971 | 1997 | 2015 |
| yr\_renovated | 0 | 0 | 0 | 84.46 | 0 | 2015 |
| sqft\_living15 | 399 | 1490 | 1840 | 1987 | 2360 | 6210 |
| sqft\_lot15 | 651 | 5100 | 7620 | 12758 | 10083 | 871200 |
| sqft-lot | 520 | 5040 | 7618 | 15099 | 10685 | 1651359 |
| sqft-living | 370 | 1430 | 1910 | 2080 | 2550 | 13540 |
| waterfront | 0 | 0 | 0 | 0.007547 | 0 | 1 |
| view | 0 | 0 | 0 | 0.2343 | 0 | 4 |
| condition | 1 | 3 | 3 | 3.41 | 4 | 5 |
| bedroom | 1 | 3 | 3 | 3.373 | 4 | 33 |
| zipcode | 98001 | 98033 | 98065 | 98078 | 98118 | 98199 |

Table

Description automatically generated

**Table

Description automatically generated**

Code review:

setwd("/Users/Team-2/Project")

project.data <- read.csv("kc\_house\_data.csv")

library(dplyr)

is.na(project.data)

summary(project.data)

variable.names(project.data)

str(project.data)

 A close-up of a document

Description automatically generated with medium confidence

Text

Description automatically generated

**Data Preprocessing, Visualization and Exploration**

**Data Preprocessing & Visualizing –**

The data source does not contain any undesirable or missing values. The outliers are a representation of the extreme numbers in the data source that will affect the forecast. We can replace them with the median value to avoid them and stop outliers from impacting the data source.

For evaluating the price dependent variable, all visualizations were finished. The data in this study, which is drawn from a variety of urban and suburban areas, has several outliers or exceptions. The source of the data is not entirely explained because of the disparity in housing prices between various communities.

We used the scatterplot method to visualize the correlations between different variable subsets of a dataset.

These plots of visualization are just a few examples. Each feature is compared to the cost of the house, including square footage, floors, views, and the number of bedrooms and bathrooms.

This plot shows the price of the houses with the attribute number of bedrooms and its evident that that the price of the house is higher, and most sales have been taken place for the 4–5-bedroom houses.

Chart, histogram, scatter chart

Description automatically generated

This plot shows the price of the houses with the attribute number of bathrooms and its clear from the plot that the price of the house is higher for the 8-bathroom house, but the sales have been more prominent for the 2.5-4.5-bathroom houses.

Chart, scatter chart

Description automatically generated

This plot shows the price of the houses in comparison with the attribute number of floors. It’s clear from the plot that the price of the houses has been higher for the 2.5 floor houses, but the sales are prominent for the 2-floor houses.

Chart, box and whisker chart

Description automatically generated

This plot depicts the prices of the houses in comparison with the square footage of the houses and its clear that the prices have been higher for the houses that has square footage in the range 10000-12000.

Chart, scatter chart

Description automatically generated

This plot portrays us the prices of the houses with the number of views in the house. Sales have been prominent for all the number of views, but the prices have gone up for the houses that have 2-3 views in the houses.

Chart, box and whisker chart

Description automatically generated

**Data Exploration –**

Correlation Matrix -

The correlation matrix between the variables in this dataset was discovered using a heatmap. The correlation matrix and relationship between the variables are shown in the heatmaps below. We performed a correlation matrix for a different set of variables.

Chart

Description automatically generated

Figure – 1 - Correlation for "price, yr\_rebuilt, yr\_renovated" variables.

Shape, square

Description automatically generated

Figure – 2 - Correlation for "Condition, grade, floors" -Variables.

**Cluster Analysis**

In the first stage, we created a new data set using variables from our data source and ran a Cluster analysis on it. We identified clusters 3,4,5,6 using the K-means method.

To determine how many clusters exist in our data source and to compare the numbers, we decided to perform a cluster analysis. When performing clustering, we observed that several other criteria have substantial connections between the variables price, year rebuilt, year renovated, condition, grade, and floors. We chose k=4 as a better number of clusters than k=5 based on the results.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

We used the elbow approach and the silhouette method to run a k-means clustering analysis on the number of clusters. We included other factors like year built, year rennovated, grade, and condition because we were unable to forecast the number of clusters using the "price" variable alone.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

**Hierarchical Method –**

We even performed a hierarchical method of clustering for our data set to know the hierarchical relationship between the clusters in the dendrogram.

Fig - Hierarchical single method clustering

Chart, histogram

Description automatically generated

Fig - Hierarchical complete method clustering

Chart

Description automatically generated

Fig - Hierarchical Average method of clustering

Chart

Description automatically generated with medium confidence

**Association Rule Mining –**

Parallel Coordination Rule –

To spot recurring patterns and examine a variety of performance indicators, we applied a Parallel coordination rule. Additionally, the arrow and color display which rules have a high level of support and confidence from the data source as well as the elements that have a significant impact on price.

Chart

Description automatically generated

Apriori algorithm –

We used an Apriori technique to determine the frequent set of rules for variables in the data source, with a support of 0.03 and confidence of 0.9. To make it simple for buyers and real estate agents to estimate costs and make judgments, we even executed the ECLAT algorithm to compare, define the rule, and see the support and confidence of the influencing elements on price.

Chart, scatter chart

Description automatically generated

Table

Description automatically generated

**Predictive Analysis**

**Logistic Regression –**

To forecast the categorization and the dwelling price as well as the factors affecting the price, we used logistic regression. We produced test data, and the yr renovated variable was recorded there as NA. We installed the necessary libraries and ran the model using glm. We accomplished 10 cross-fold validation too.

|  |  |
| --- | --- |
| Accuracy | 61% |
| Precision | 0.96 |
| Recall | 0.99 |
| F-score | 0.97 |
| AUC | 0.859 |

Chart

Description automatically generated

**Decision Tree Model –**

With train and test data, we ran a decision tree model to identify the factors affecting the pricing.

|  |  |
| --- | --- |
| Accuracy | 36% |
| Precision | 38% |
| Recall | 3.08 |
| F-score | 68% |
| AUC | 0.926 |

Chart

Description automatically generated

**SVM model –**

We performed an SVM model for classification with test and train data sets.

|  |  |
| --- | --- |
| Accuracy | 1 |
| Precision | 0.00016 |
| Recall | 0.00015 |
| F-score | 0.00016 |
| AUC | 0.958 |

**Chart

Description automatically generated**

**Random Forest Model –**

We performed a random forest model for classification, which helps to distinguish the positive class values from the negative class values.

|  |  |
| --- | --- |
| Accuracy | 1 |
| Precision | 0.00016 |
| Recall | 1 |
| F-score | 0.00032 |
| AUC | 0.795 |

Chart

Description automatically generated

**Conclusion –**

We first preprocessed the data to get rid of all the missing values. We discovered that the data was clean after the preprocessing step was complete, and we then carried out data visualization and exploration by creating various scatter plots to understand the relationship between variables, carrying out silhouette and cluster analysis, and using a hierarchical method to determine the number of clusters and their relationship with variables, as well as association rule mining to understand the set of rules influencing the price variable.

All four models were executed by our project (Logistic Regression model, Decision Tree Model, SVM Model, and Random Forest Model). Due to its superior precision metrics of 96% when compared to other models, we concluded that Logistic Regression is the best model. We discovered that the variables like grade, condition, yr built, and yr rennovated have the most influence on the price variable.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Precision** | **Recall** | **F-score** | **AUC** | **Accuracy** |
| Logistic Regression | 0.9611696 | 0.9986932 | 0.9795722 | 0.859 | 0.6114 |
| Decision Tree | 0.3835348 | 3.081365 | 0.6821615 | 0.926 | 0.3683 |
| SVM model | 0.000163345 | 0.000156863 | 0.000160038 | 0.958 | 1 |
| Random Forest | 0.000163345 | 1 | 0.000326637 | 0.795 | 1 |