**King County Housing Market sales Prices Analysis**

King county housing sales prices have been increasing. This project is an attempt to conduct and run machine learning models using R programming. The data used here is King County data from May 2014 to May 2015 time frame. The objective of this project is to understand the factors that impact the sales price and create models to predict the price given set conditions.

Data

Data used for this study is from May 2014 to May 2015 time frame. (Source: <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>)

Data Analysis:

EDA was done on the data set. Data is verified to check for any missing values etc. Identified the data types for each variable. There were redundant variables like zip codes, latitude and longitude, all 3 indicate location. Latitude and latitude columns are removed. Filtered data set with no bathrooms. It seems to be inhospitable, so removed them.

The target sales price is converted in log scale , waterfront, view, zipcode, floors, condition and grade as factors for ease of analysis. Featured engineering was added to data- is\_basement, is\_renovated, is\_new, is\_goodview, is\_condition. See Fig 1:

*Fig 1:*

A screenshot of a computer program

Description automatically generated with medium confidence

Dataframe now looks like this: See Fig 2.

*Fig 2*

A screenshot of a computer program

Description automatically generated with low confidence

Plot 1 shows the histogram of sales price (log scale). Some of the houses have more than 220K (log value of 6 on graph) as sales price at tail end of histogram. This seemed like outlier so removed it for analysis.

|  |  |
| --- | --- |
| *Plot 1: Histogram of overall data* | *Plot 2: Filtered data histogram* |
| A picture containing diagram, screenshot, plot, pixel  Description automatically generated | A picture containing diagram, screenshot, text, pixel  Description automatically generated |
| *Plot 3: Group by zipcode data for filter data* | *Plot 4: Sqft vs Zipcode* |

EDA indicates that location (variable zipcode) and area of house (variable sqft\_living) have an impact on the sale price. Reviewing the data with area of house, it seems areas with high sales have larger houses. Either zipcode or area of house can be dependent factor. See plots 3 & 4.

EDA was continued by grouping with year and other variables. The important variables that seem to impact price are zipcode, sqft\_living, rooms, good view, year\_built,, is\_basement, grade

**Model Analysis**

The database is now converted into binary model for future analysis easiness. Data is pre-processed and removed of any near zero values. Data is partitioned for training and testing with 80:20 ratio. LASSO method and simple tree decision methods were used to identify the factors. Using tree model varimp plot, showed that zipcode may be dependent factor. So, using the factors from both methods, a linear regression model was developed. Multiple iterations was done to optimize the model.

|  |  |
| --- | --- |
| *Lasso method:*  *Plot 5: Varimp plot* | *LASSO method:*  *Plot 6: Predicted vs observation.* |
|  |  |

**Summary:**

1. The sales price of the house is influenced by these 5 variables: sqft\_living, # of rooms, # of bedrooms. Goodview, and good grade. Other factors like basement, new construction and good conditions are minor factors.
2. Year built showed some correlation with sales price during EDA. However plotting year vs area shows area of houses increased significantly since 1951. So area of house had more importance in varimp plot than year built.
3. This data is biased only on physical condition of house like area, lot size, bedroom etc. but doesn’t include neighborhood school districts ratings, or bank loan rate. There may be other additional factors that may have impacted the sales price.

**Next steps:**

This analysis is based only on May 2014- May 2015 data. Current market prices may have been influenced by other factors. Further analysis needs to be done with recent data.