THE 2.2'S

DEVELOPERS

- Rosemary Mburu
- Paul Kamau
- Fiona Njuguna
- Brian Nderu
- Ian Tulienge
- Paul Musau

MZALENDO LTD

GROUP 2.2 PHASE 2 PROJECT

REAL ESTATE TRENDS IN KING COUNTY, WASHINGTON: ANALYSIS OF KC HOUSE DATA

<<<<

AGENDA

>>>>

- Overview
- Business Problem
- Data Understanding
- Data Analysis
- Modelling
- Results
- Conclusions
- Recommendations
- Limitations

DATA OVERVIEW

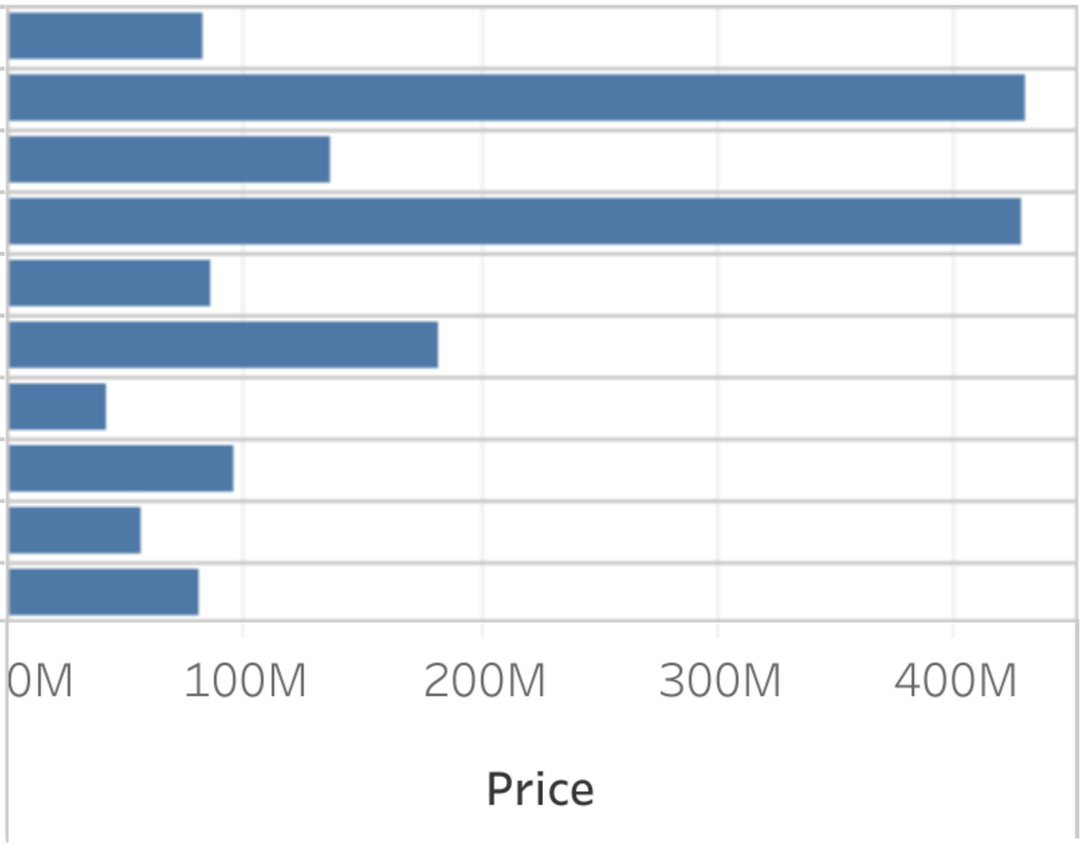
- We had a {21597 rows, 21 columns} number of properties in the dataset, the time period is {May 2014-May 2015}, the type of properties were {luxury, mansion, average, good, very good} just to mention but a few and location was{King County-Washington}.

KEY FINDINGS

- According to our findings we were able to interpret that in the month of May, there was a record of high number of sales.
- This is influenced by the summer break.
- Average houses are reflecting high prices thus why are in high demand.

Geographical region, zipmap and price geo_bar

Zipcode	City name
98003	Federal Way
98004	Bellevue
98005	Bellevue
98006	Bellevue
98007	Bellevue
98008	Bellevue
98010	Black Diamond
98011	Bothell
98014	Carnation
98019	Duvall



Geographical region

+.+.+.

METHODOLOGY

+.+.+.

DATA PRE_PROCESSING

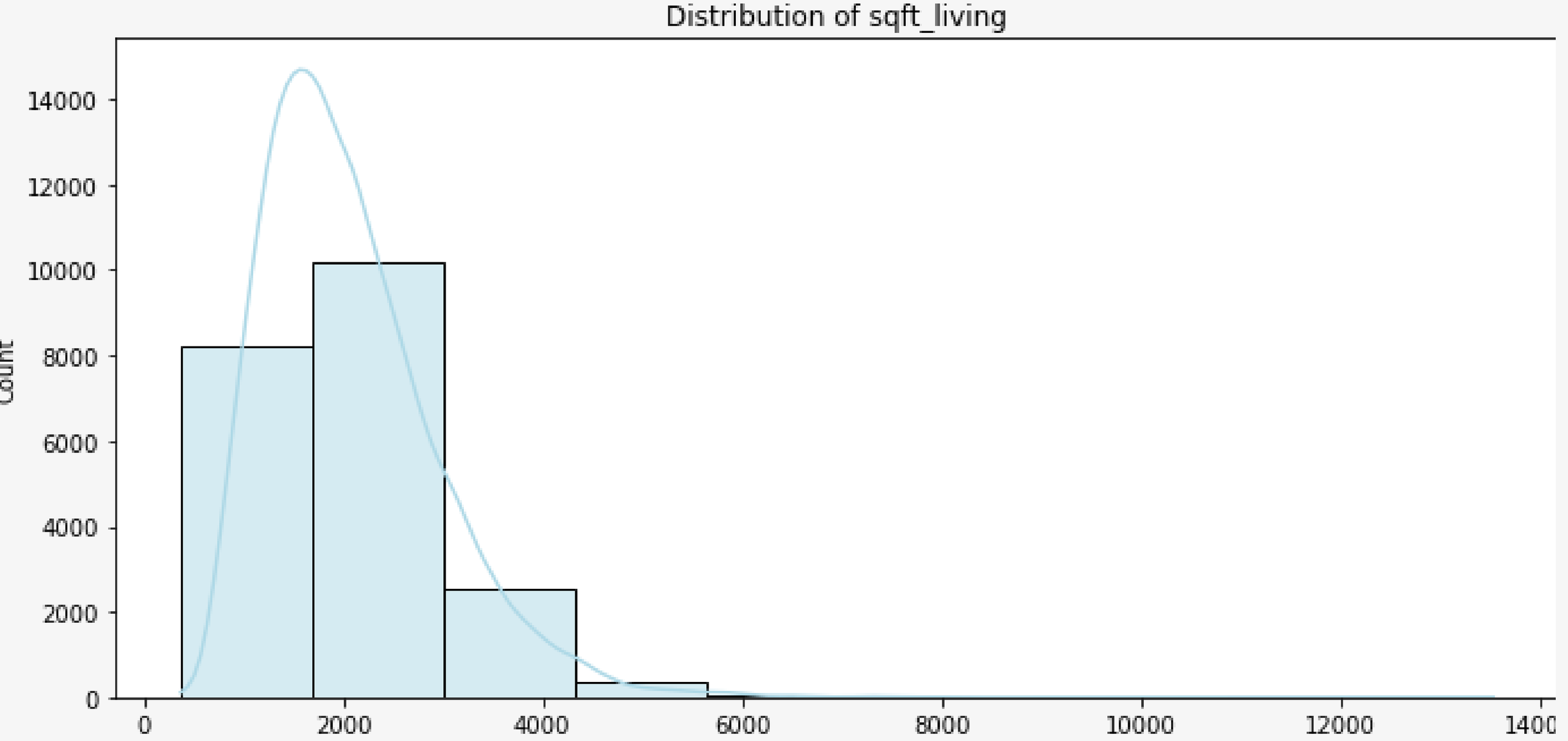
- Duplicates- There were duplicates in the dataset.
- Outliers- Price was the most affected, thus filtered out the outliers for the price variable less than 99TH percentile {only the top 0.5% are included}.{<0.5TH percentile only bottom 0.5% are inluded}
- Dealing with missing values. We decided to fill with the mode. Water-front and view are categorical variables so we checked the value counts and filled with the most occurring feature.
- Dropping Columns- We dropped the yr_renoveted column with 18% of missing values since it was not useful to our model

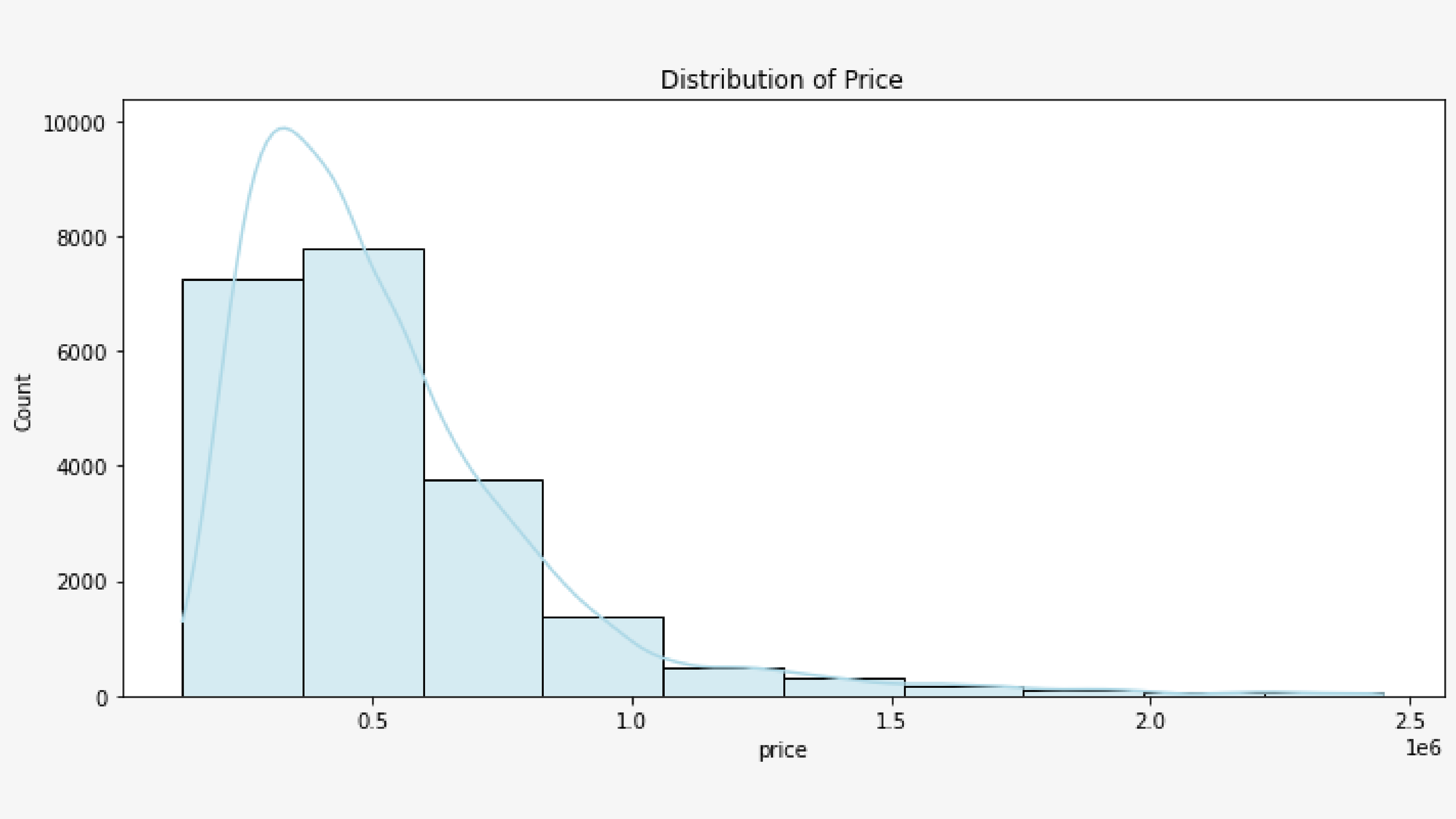
DATA ANALYSIS

- Univariate Analysis
- Bivariate Analysis
- Multi-variate Analysis

UNIVARIATE ANALYSIS

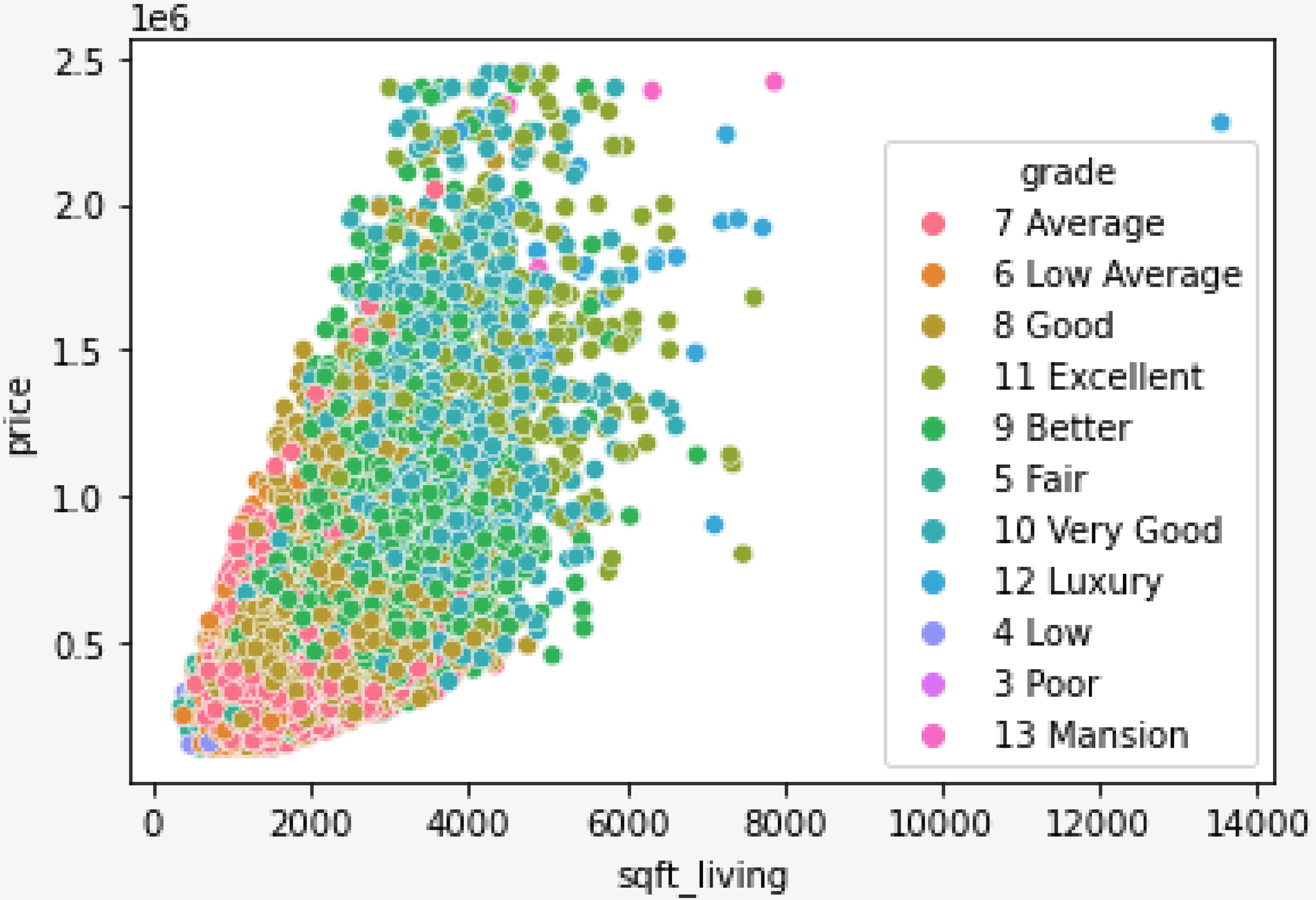
- We explored the dataset to identify the categorical and numeric columns.
- We also plotted visualizations to display the counts of the values in the columns to get a better understanding of how they are distributed.
- We also inspected correlation of the columns to a specific column selected for the project analysis which didn't have much of a high correlation.
- Slide 11, 12 are categorical representing the condition and grade columns.
- Slide 13,14 are numeric representing the Distribution of value counts in price and sqft_living columns.

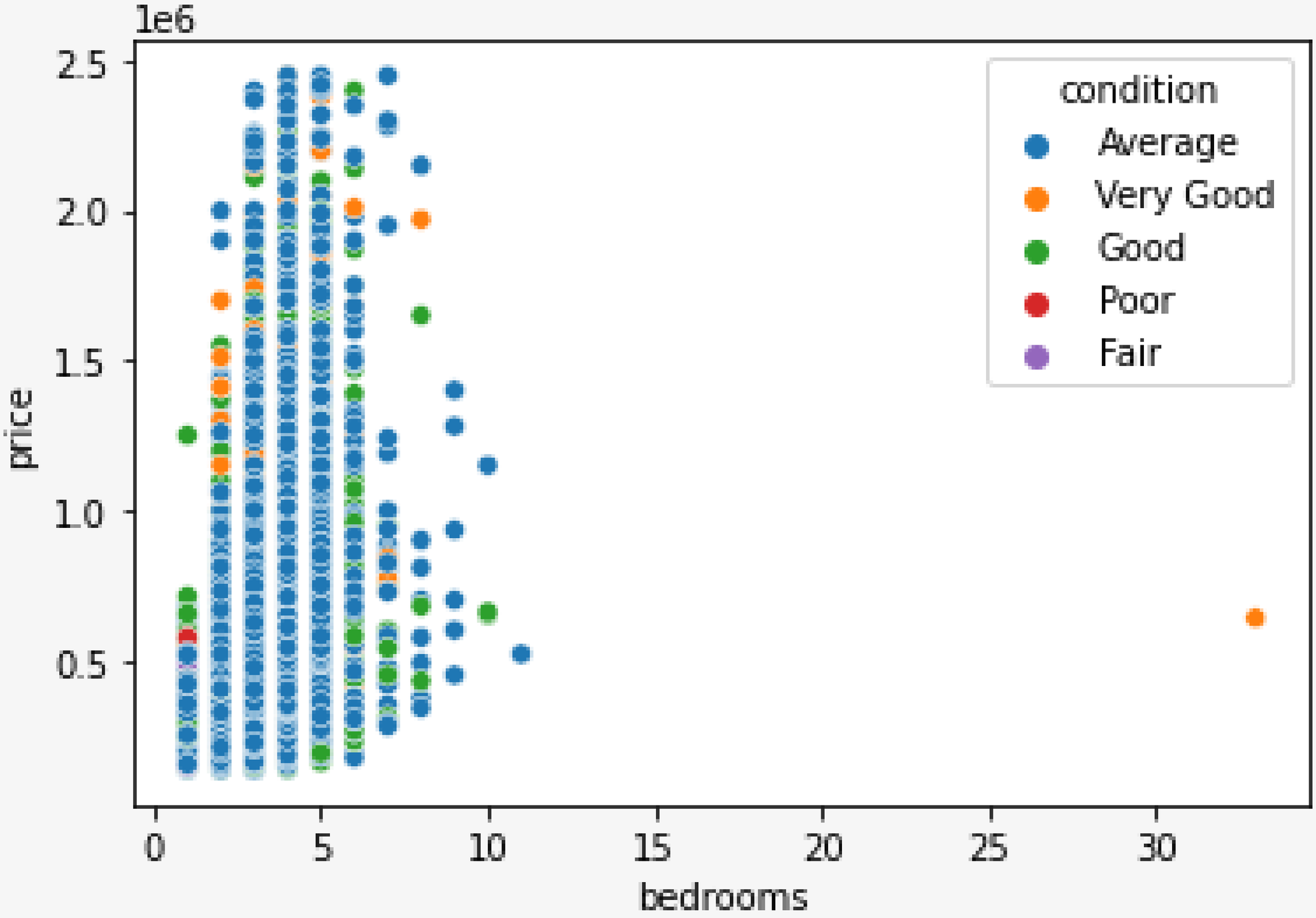

Distribution of sqft_living
Count
14000
12000
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000
12000
1400


Distribution of Price
10000
8000
6000
4000
2000
0
Count
0.5
1.0
1.5
2.0
2.5
price
1e6

BIVARIATE ANALYSIS

- We looked at specific correlations between two variables each to see how one affects the other.
- Price had high correlation to the sqft_living and bedrooms columns thus were plotted.
- We also applied bivariate analysis on the simple linear regression model, using price and sqft_living which had the highest correlation to the target variable which was price.
- Slide 16- scatterplot for continuous variable sqft_living and price with the grade as hue, 17- scatterplot for discrete variable bedrooms and price with hue as grade.


1e6
2.5
2.0
1.5
1.0
0.5
price
0
2000
4000
6000
8000
10000
12000
14000
sqft_living
grade
7 Average
6 Low Average
8 Good
11 Excellent
9 Better
5 Fair
10 Very Good
12 Luxury
4 Low
3 Poor
13 Mansion


1e6
2.5
2.0
1.5
1.0
0.5
price
0
5
10
15
20
25
30
bedrooms
condition
Average
Very Good
Good
Poor
Fair

MULTI-VARIATE ANALYSIS

- We then build multiple regression models to improve the simple regression model.
- We prepared the categorical variables for this by using various steps to assist in building suitable models.
- We did multiple linear regression, one hot encoding and log transformation.

MODELLING

- The metric of success we used are the R-squared adjusted and the MSE because they provide a more interpretable measure of the goodness of fit of the model.

- Further,R-squared also takes into account the number of variables included in the model, which can be helpful in comparing models with different numbers of variables.

RESULTS

- The base linear model had a adjusted R squared of 47%
- The second model was a multiple regression model which was one hot coded with condition column and had an adjusted rsquared of 53%.
- The third model was a log transformed model of the second model which had an adjusted rsquared of 50%. This model was lower compared to the second model and we opted an approach of one hot coding another categorical variable column called grade.
- The fourth model was a multiple one hot coded model of grade column and had an adjusted r squared of 63% which achieved our target of above 60%.
- All our models were statistically significant as they had a p value less than the significant value.
- The Rsquared of the model improved from the baseline linear model to multiple regression model {one hot encoded} but dropped on the third log transformed model but improved to the forth to 63%.

RESULTS

- We choose R-squared over RMSE because it provides a more interpretable measure of the goodness of fit of the model. Further, R-squared also takes into account the number of variables included in the model, which can be helpful in comparing models with different numbers of variables.
- We calculated the Mean Squared Error {MSE} and found that the trained model had an MSE of 0.124 thus indicating that it neither overfits nor underfits the training data.

CONCLUSIONS

- We have performed extensive analysis on the housing dataset, exploring its features, identifying patterns and trends, and building multiple linear regression models to predict housing prices.
- We found that the most significant predictors of housing prices were the square footage of living space, the number of bedrooms, and the number of bathrooms.
- The multiple regression model is better than the simple regression model since it has a higher Rsquared value than the simple linear regression model.
- From this we gathered the categorical variables have a positive effect on the price predictor variable.
- The agency should consider the sqft_living variable when selling since it has a high linear relationship to the prices predictor.

RECOMMENDATIONS

- Based on our findings, we recommend that prospective homebuyers focus on properties with larger living spaces and more bedrooms and bathrooms, as these tend to have higher values.
- Additionally, we recommend that real estate agents use our model to help set appropriate listing prices for homes, taking into account the features of the property.

LIMITATIONS

- Limited variables: The dataset used in this analysis only included a limited number of variables. There could be other variables that could also impact housing prices that were not included in the dataset.
- Data quality: The dataset used in this analysis is assumed to be accurate, complete, and representative of the population. However, there may be errors or biases in the data that were not detected.
- Model assumptions: The regression models used in this analysis make certain assumptions, such as linearity, independence, and normality. Violations of these assumptions could impact the accuracy and reliability of the results.
- Limited geographic scope: The dataset used in this analysis only covers housing prices in King County, Washington. The results may not be generalizable to other locations.
- Timeframe: The dataset used in this analysis only covers housing prices from May 2014 to May 2015. Housing prices may have changed since that time, and the results may not be applicable to the current housing market.

NEXT STEPS

- Feature engineering: try creating new variables based on domain knowledge or statistical analysis to see if they improve the model performance.
- Collect more data: gather additional data on housing characteristics or socioeconomic factors that could be included in the model to improve its accuracy.
- Deploy the model: if the model proves to be accurate and useful, it could be deployed in a web application or integrated into a larger data pipeline to provide real-time predictions for potential home buyers or sellers.
- Log transformation can be performed on variables that have huge skewness and kurtosis and see if it improves the model.

THANK YOU!