HOUSE SALE IN NORTHWESTERN COUNTY

INTRODUCTION

- Real estate can be defined as land, natural resources and any permanent structures attached to the land including the rights to use and occupy them
- Real estate may be classified into different categories which include ; Residential Real Estate which deals with properties used for living purpose
- The value of the real estate can be influenced by number of factors which include location, size, condition, infrastructure surrounding the real estate, economic conditions, population and government policies

PROBLEM STATEMENT

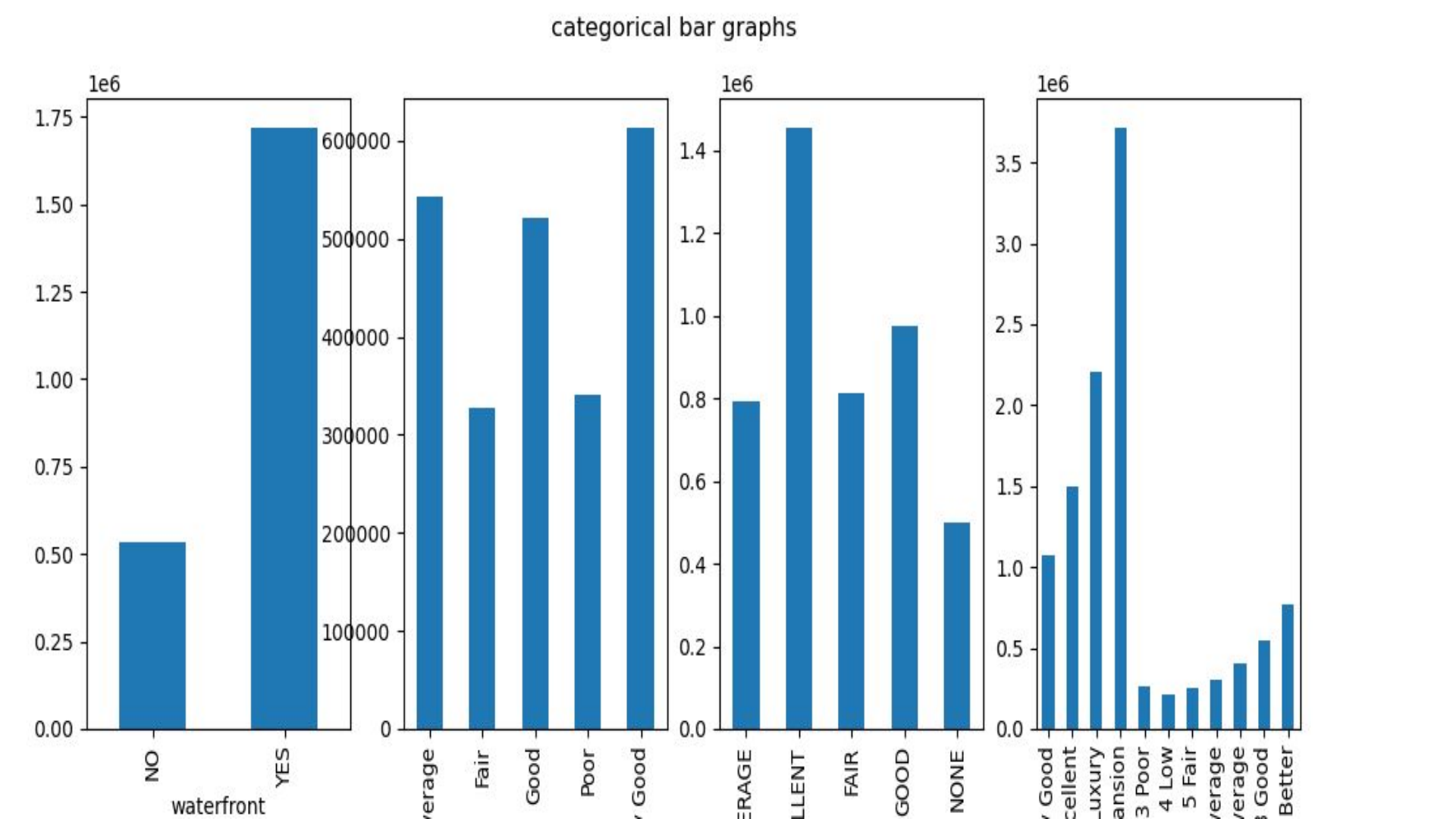
- Real estate has a number of stakeholders.These stake holders include ; buyers, sellers, brokers, developers, financial institutions and the government.
- All these stakeholders are interested in the value of the property and the factors most affecting the value
- They need to easily and with good accuracy, predict the value of the properties they are interested in.This will help the stakeholders arrive on agreement more easily and quickly

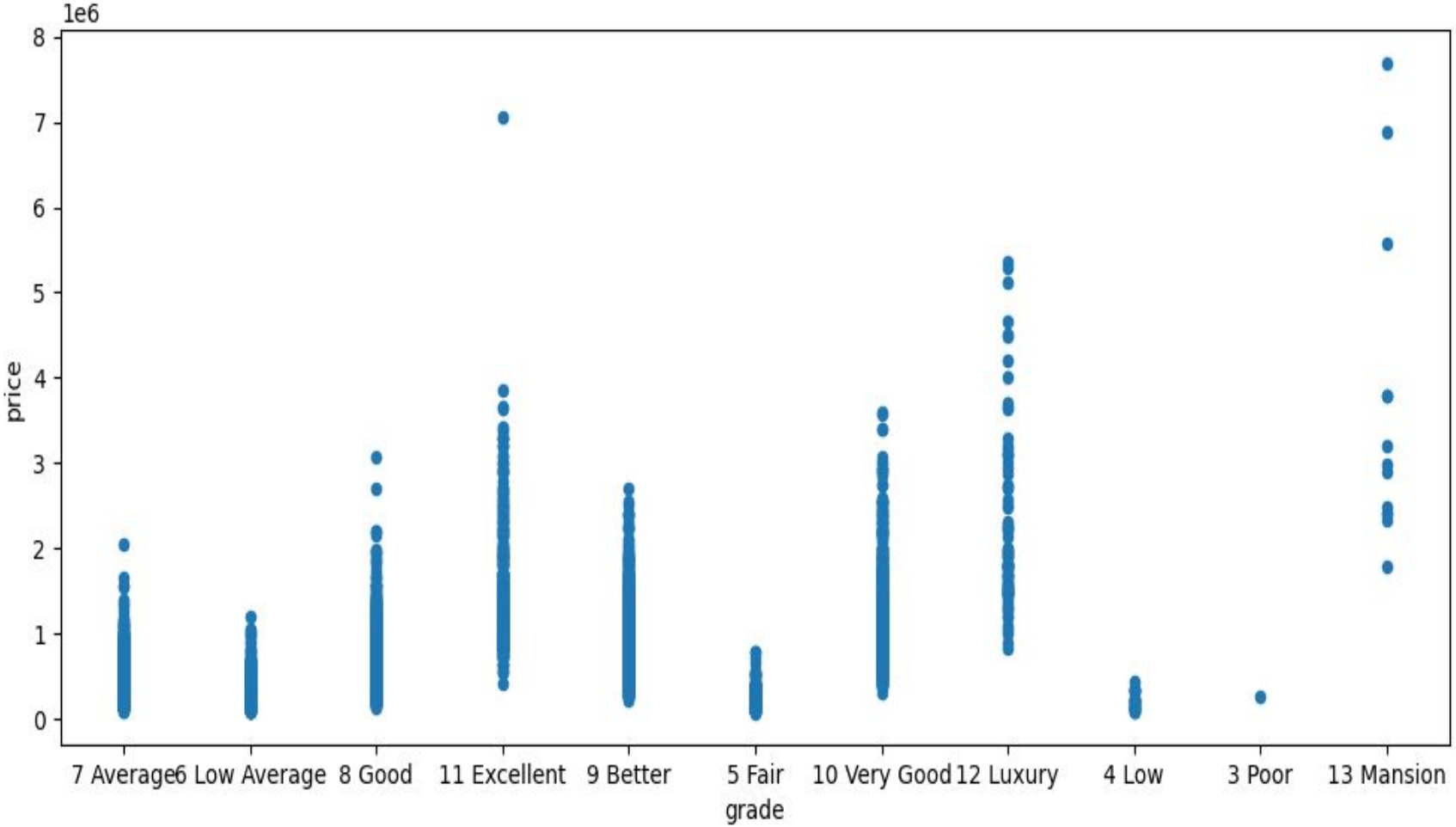
DATA UNDERSTANDING

- This project uses the King County House Sales dataset and contains data with 21597 rows and 20 columns
- This data contain information on houses including ; dates, prices,number of rooms and bathrooms,living room , lot and basement areas, waterfronts,the views from the house, house grade, areas of the 15 neighbouring houses.
- All this information from this data set has an impact on the overall price of the houses

DATA PREPARATION

- replacing missing data was done on view and waterfront columns.They were replaced with none in each case.
- checking for numeric column most correlated to price and correlation between them was done to find a best predictor and avoid multicollinearity .
- Transforming Categorical Variables .the data had 4 important categorical variables.view,waterfront,condition and grade.
- Checking their impact using graphs;

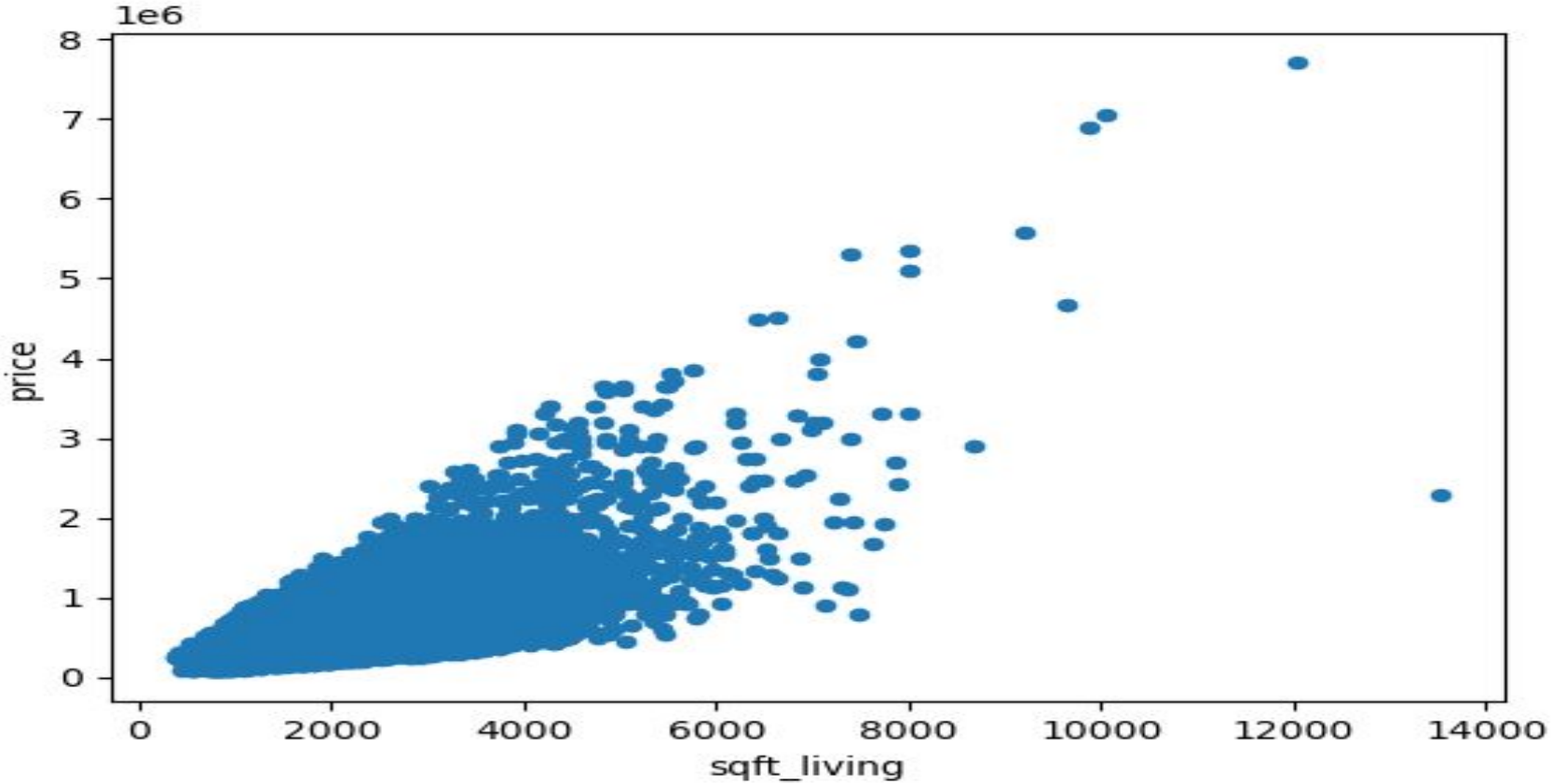

categorical bar graphs
1e6
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
NO
YES
waterfront
600000
500000
400000
300000
200000
100000
0
Fair
Good
Poor
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
FAIR
GOOD
NONE
1e6
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Poor
Low
Fair
Good
Better


Scatter Plot grade vs price
1e6
8
7
6
5
4
3
2
1
0
price
7 Average
6 Low Average
8 Good
11 Excellent
9 Better
5 Fair
10 Very Good
12 Luxury
4 Low
3 Poor
13 Mansion
grade

MODELING

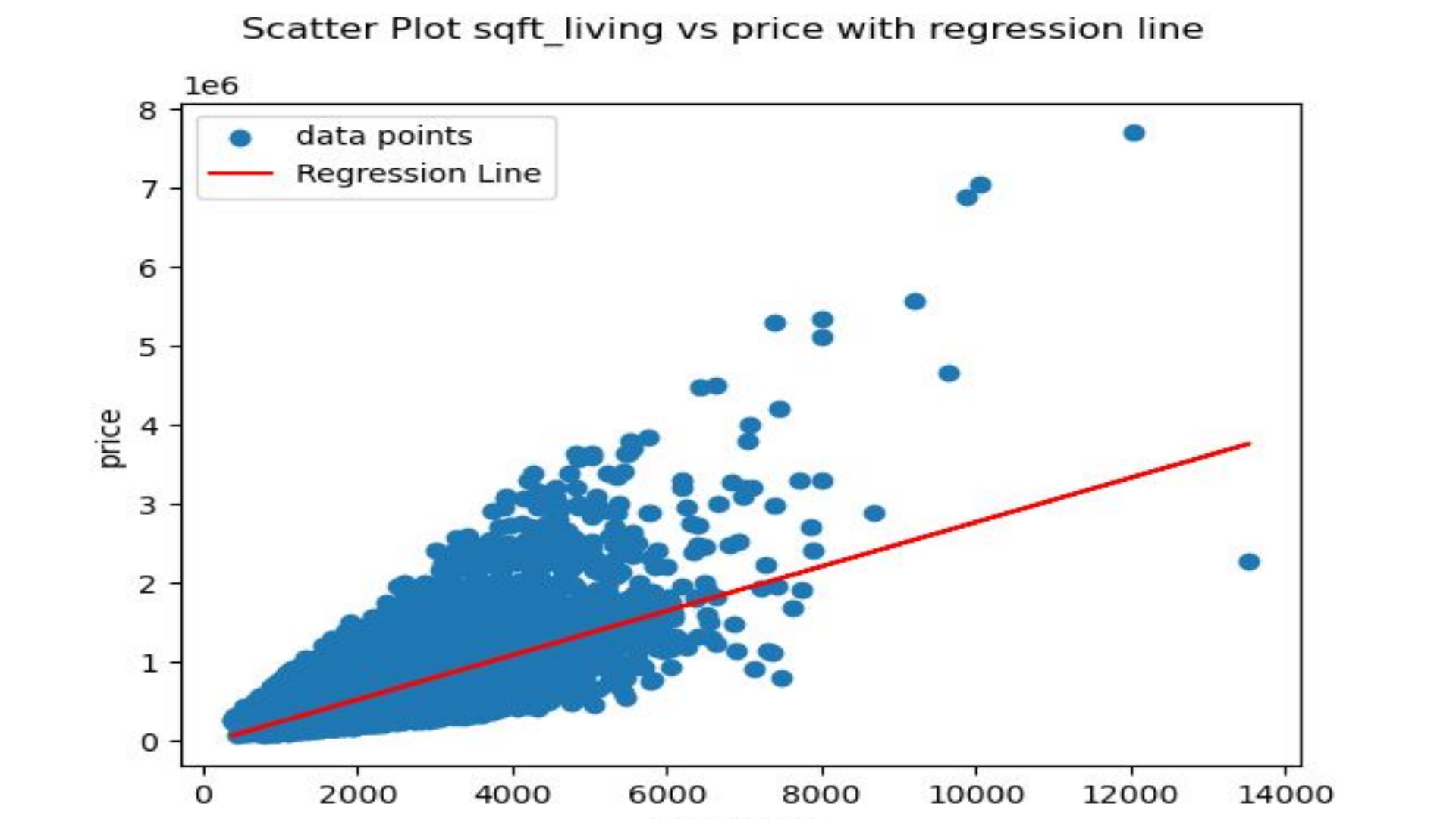
- We will start modeling with a simple linear regression model between the area of the housing living room(sqft_living) against price.
- This is because sqft_living column has the highest correlation to price compared to other numeric variables.
- First we will have a scatter plot preview of sqft_living vs price to get a picture of the correlation between the 2.

scatter plot showing distribution of sqft_living against price



First simple linear regression model

- **price=281(sqft_living)-43989**
- It has R_squared of about 0.49. This means 49% of the variation in price is explained .itl has an overall p_value of 0 which is below the typical alpha of 0.05.This mean that the model is statistically significant.
- The model has slope of 281 which mean an increase of sqft_living by 1 increase the price by 281 units.
- the y intercept is -43989 which is means when the living area is zero the price is -43989.That does not make sense but we will transform our more advanced models for an interpretable y intercept.


Scatter Plot sqft_living vs price with regression line
1e6
data points
Regression Line
price
0
1
2
3
4
5
6
7
8
0
2000
4000
6000
8000
10000
12000
14000

Improving the model

- In our final model we use sqft_living squared in place of sqft as it seems to improve the r_squared therefore improving our variation explanation while improving the p values.
- Introduction of categorical variables in our model help improve the r_squared further explaining the variation in price.
- Transforming the x axis values to be centred around the mean of sqft_living squared improves our explanation of the y intercept.

Final model Regression Results

- **R_squared** and **F_pvalue** Our final regression model has R_squared of about 0.653. This means about 65% of the variation in price is explained by our model. Our model has an overall p_value of 0 which is below the typical alpha of 0.05.This means that our model is statistically significant.
- **Y_intercept** Our y_intercept is 418800 ,that is to say the price of an average size cost is about 418800 dollars on price.This is because our data is centered around the mean of sqft_living

Regression results continued

- **coefficient of sqft_living** In the final model sqft_living is squared therefore 0.0256 is coefficient of sqft_living squared therefore gradient translates to 2(0.0256)sqft_living.which can be simply interpreted as a increase of 1% in Sqft_living will translate in about 0.0512% increase in price
- **view** there is an increase of 164800 for view fair compared to view none,an increase of 96690 for view average compared to none,an increase of 155800 for view good as compared to none and an increase of 293100 for excellent view compared to none
- **waterfront** There is an increase of 533600 in price for houses with waterfront compared to those that do not have a waterfront.

Regression results continued

- **condition** we can't say anything about about condition poor and fair as they are not statistically viable.However an increase of 51490 is observed for good condition compared to average condition while an increase of 134600 is observed for houses with very good conditions compared to average
- **Grade** decrease of 122300,121300,67130 in prices is observed for houses of grade 4-low,5-fair and 6-low average respectively compared to houses of grade_7 Average. increase of 86460 ,218900, 386600, 591300, 910900 ,1764000 is seen for house of grade 8 Good,9 Better,10 Very Good, 11 Excellent, 12 Luxury, 13 Mansion respectively compared to houses of grade_7 Average.

Conclusions

- Statistically significant - the final model has a has an overall p_value of 0 which is below the typical alpha of 0.05.This means that our model is statistically significant.
- Coefficients - an increase in area (living room space which is correlated to other areas) show a improvement of 0.0512% in price for 1% increase in area.grade,condition,view and waterfront seem to greatly improve the prices of the houses.
- limitation - our model provided only about 65% explanation of variation which might be small in some practical cases.

Recommendations

- All the factors considered in this model should be given much considerations when dealing with house as they have proved to be statistically significant.
- House Grade seems to have the largest impact among categorical data and should be given priority compared to the other categorical data.
- The model should be used in less variation sensitive cases such as price estimation but not price setting as the r_squared is not that high.