Project: Multiple linear regression modeling to analyze house sales in King County, Washington.

Presenter: Ahmed Ali

Business understanding

Stakeholder : Customers who are looking into the King County for affordable houses.

Business Problem: Predicting the price of houses based on multiple features

Data Understanding

After Loading the Dataset into a pandas DataFrame, we found out that the dataset had 20 columns and 21597 rows. Each row presenting a house with unique id.

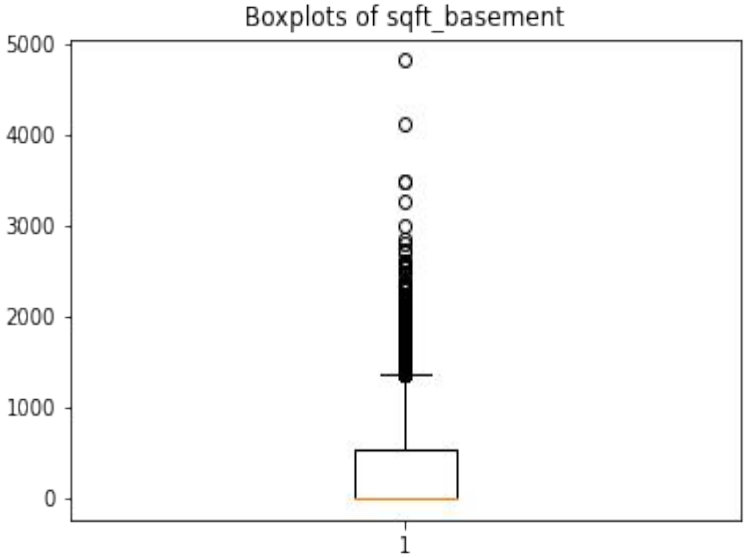
The Data Contains Multiple Features of Houses that are in King County. Some of them include: Dates houses were sold and renovated. The prices,bedrooms,bathrooms, their co-ordinates(latitudes and longitudes) and more.

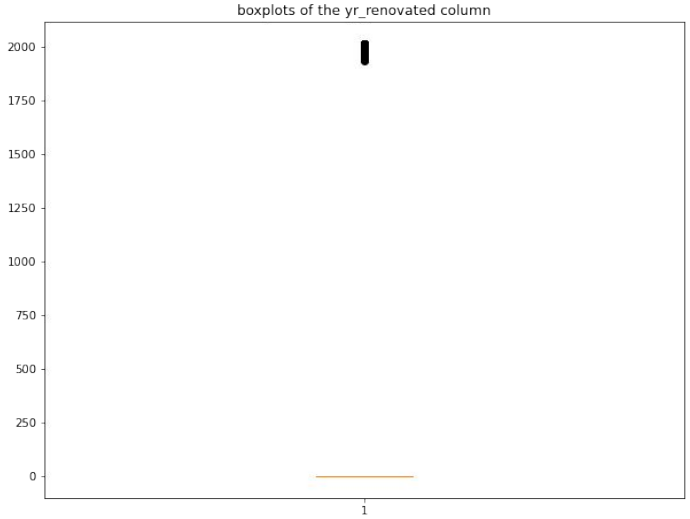
After further exploration we found out that, Almost all the data columns were numerical, either **float** or **integer**. With five of them containing **object** dtypes including date columns.

The waterfront column contained 2376 null values, view column contained 63 null values and the yr_renovated column contained 3842 null values .

We handled the missing data appropriately and found out that some columns contained outliers which would have given lesser precise models and why we decided to drop them.

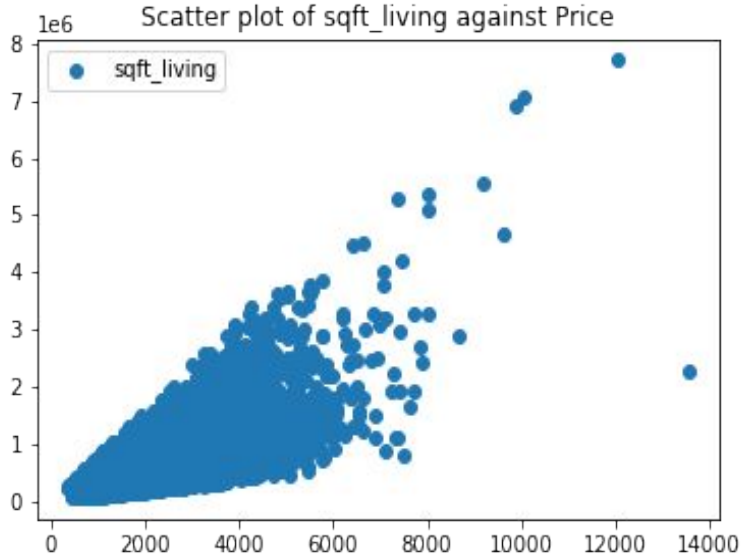
Visualization of the yr_renovated and sqft_basement column

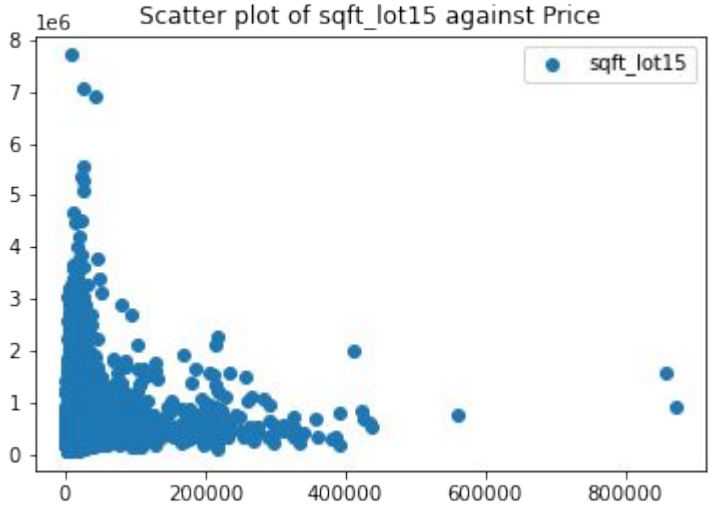




To Understand our data further, we had to plot scatter plots of the data columns against the Target Variable(price) to see which followed Linearity.

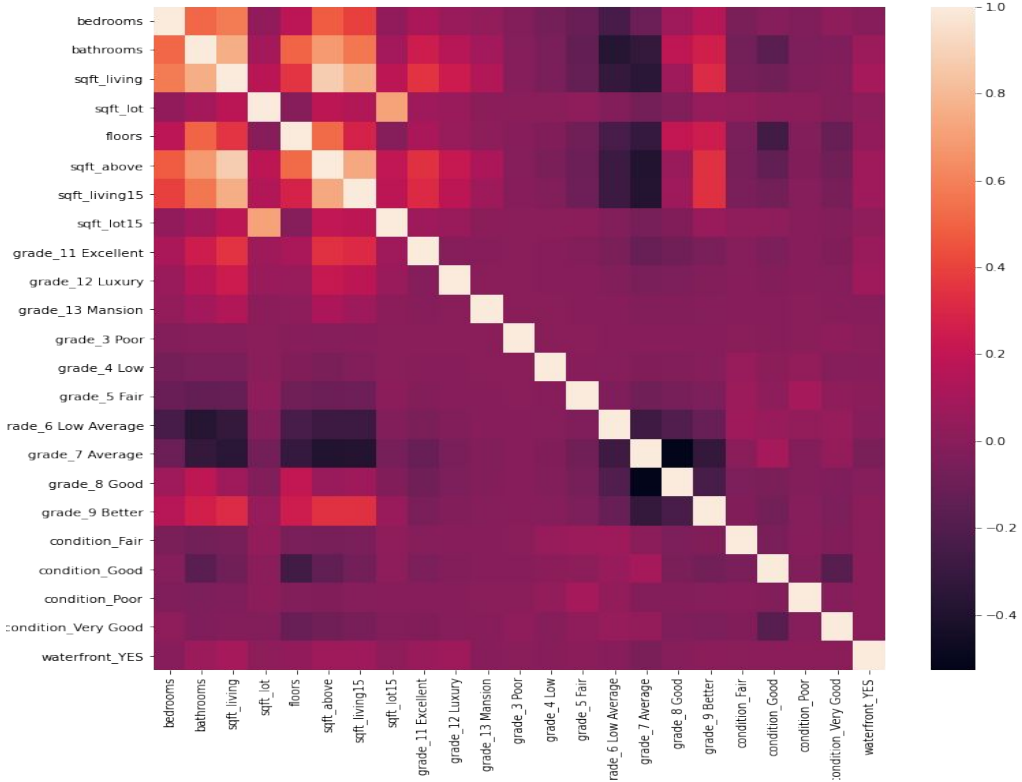
Here are Some:


Scatter plot of sqft_living against Price
1e6
sqft_living
0
1
2
3
4
5
6
7
8
0
2000
4000
6000
8000
10000
12000
14000


Scatter plot of sqft_lot15 against Price
1e6
sqft_lot15
0
1
2
3
4
5
6
7
8
0
200000
400000
600000
800000

We also checked for multicollinearity and dropped the appropriate columns that were highly correlated against each other.

Here is the heat map of the correlation data.


bedrooms
bathrooms
sqft_living
sqft_lot
floors
sqft_above
sqft_living15
sqft_lot15
grade_11 Excellent
grade_12 Luxury
grade_13 Mansion
grade_3 Poor
grade_4 Low
grade_5 Fair
rade_6 Low Average
grade_7 Average
grade_8 Good
grade_9 Better
condition_Fair
condition_Good
condition_Poor
:ondition_Very Good
waterfront_YES
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

We created dummy dataframe to feed to our regression model by one hot encoding the categorical data.

After preparing the data thoroughly, we proceeded to model The Multi Linear Regression.

Regression Modelling

The first model gave us base understanding of the regression model then we proceeded to drop the columns that were not statistically significant that is they had pvalues of above 0.05.

The second model was much better.

Interpretation of the Model.

R-squared:	***0.626***
Adj. R-squared:	***0.626***
F-statistic:	***2010.***
Prob (F-statistic):	***0.00***

Compared to our previous model of F-statistic of 1723, the F-statistic of this model was high showing statistical significance

R-Squared: Indicates that 63% of the Target variable can be explained by our model.

High F-statistic of 2010 and pvalue below 0.05 indicates that our model is statistically significant in predicting the target variable.

Const: Intercept is 5.633e+05. When all the other predictors are 0.The house price is $563300.

For each additional bedroom, the house price decreases by approximately $23,350.

For each additional bathroom, the house price increases by approximately $1735.

For each additional of one square footage of living space, the price is set to increase by $169.

For each additional floor in the house the price is set to decrease by about $9027.

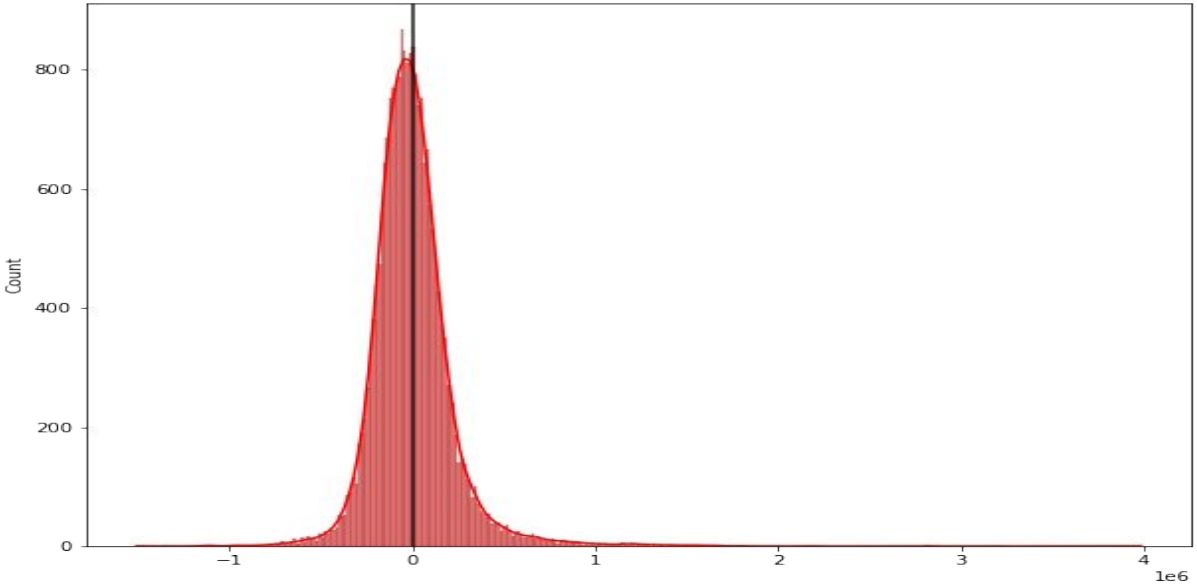
For Houses with grade_11 (Excellent), its price is set to increase by $276300 compared to other grades.

Houses with good condition are set to increase by about $57800.

Houses with very good condition are set to increase by about $140800.

Houses with waterfront are set to increase by about $683600 than houses without waterfront.

Normality of errors (Residuals)



The errors follow a normal distribution.Justified by the bell curve and the mean situated around 0 with the errors 1 standard deviation away from the mean. Meaning Approximately 68% of the data falls within one standard deviation of the mean,which is proves the assumption in our case.

Homoscedasticity

After conducting a Breusch-Pagan test here were the results:

Lagrange multiplier statistic: 4595.046421005922

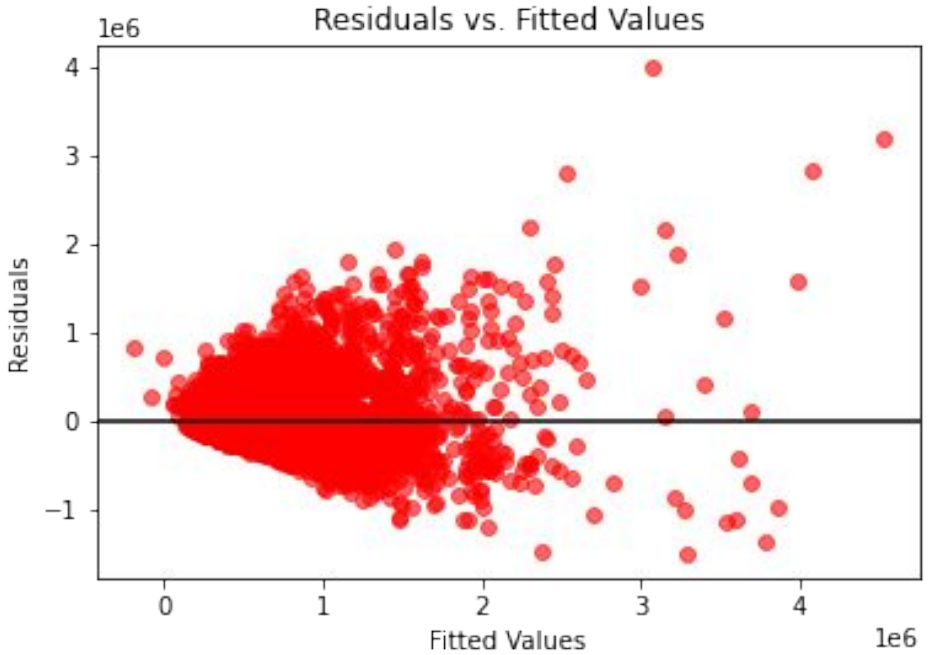
Lagrange multiplier p-value: 0.0

Breusch-Pagan test statistic: 323.9886235288215

Breusch-Pagan test p-value: 0.0

The p-values below our statistically significant alpha of 0.05 indicates we reject the null hypothesis for **Breusch-Pagan Test** and conclude that there is heteroscedasticity in our errors.

Visualization of the errors (Checking for Homoscedasticity)



Most of the residuals are situated around 0. Our model is biased and our coefficients will be less precise in actually predicting our target Variable(price)

Recommendations.

From our model, i recommend to customers who are looking for affordable house to look into houses with less bedrooms.

Each additional bathroom had an increase of $1735 per house so look for lesser bathroom houses.

Each additional square footage of houses had an increase of $169 , that looks affordable.

Houses with waterfront are set to increase by about $683600 than houses without waterfront, that looked a lot so i recommend houses without waterfronts like lakes, bays and rivers.