INTRODUCTION

The King County House dataset that was used contains information about various house features. This presentation aims towards explaining how I built a multiple linear regression using Python to predict house prices. Multiple linear regression was used to describe the relationship that exists between various variables and house prices. The results will help home owners interested in selling their homes by informing them on important factors to consider hence improve sale prices.

The column names used in the dataset are explained below;

- Price -sale price(prediction target)
- Bedrooms -number of bedrooms
- Bathrooms -number of bathrooms
- Sqft_living -square footage of living space in the home.
- Grade - overall grade of the house related to the construction and design of the house.
- Sqft_above -square footage of the house apart from the basement
- Yr_renovated -year house was renovated
- Waterfront -whether the house is on a waterfront
- Sqft_living15 -the square footage of the interior housing living space for the nearest 15 neighbors.

1.1 DATA PREPARATION

The data was loaded into a Pandas data frame and the shapes columns and data types were checked. The dataset contained more than 21613 entries and 21 columns with majority having numeric data. There were also some columns which were spotted with missing values since their entries did not sum to 21613.

1.1.1 Data cleaning

The value counts for each of the columns with missing values; waterfront, view and year renovated were checked. For year renovated since the variables are of numeric type, the missing

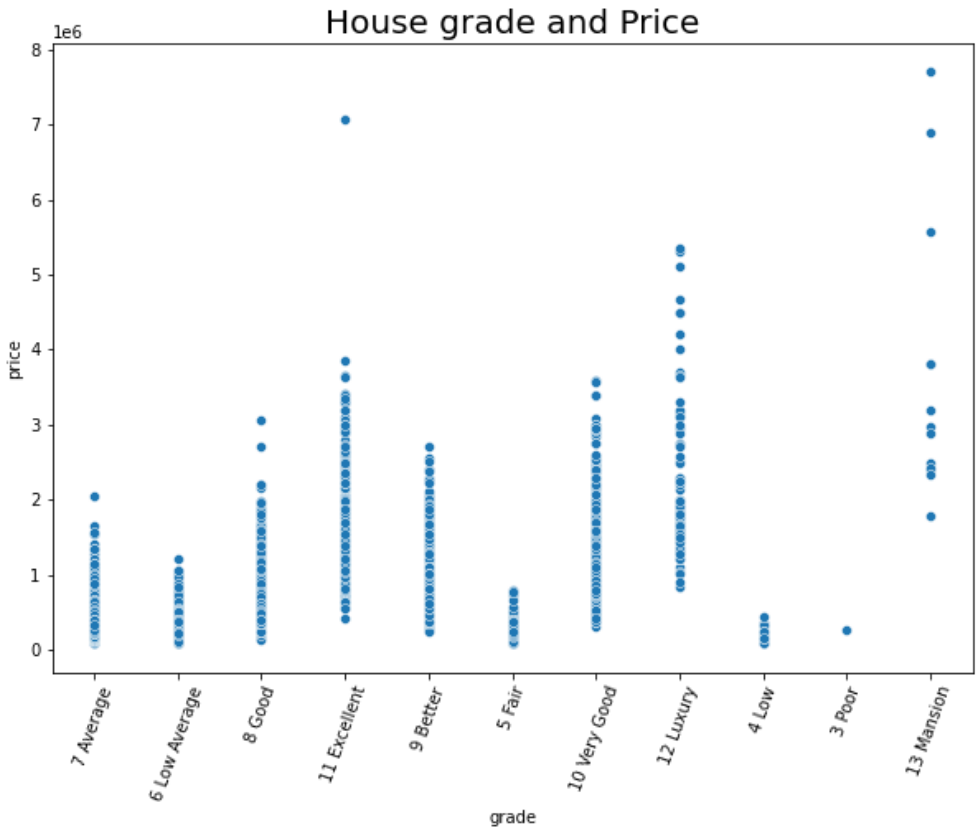
values were filled with their median. On the other hand, waterfronts, view which contain non numeric variables the missing values were filled with none to indicate lack of data. The value counts were later checked to ensure the missing values were eliminated from the columns.

1.1.2 Dealing with outliers

A function was defined to check for outliers in the bedroom and bathroom columns so that houses that lack both are eliminated. To eliminate these houses which are considered as outliers, another function was defined to drop the rows containing these values.

1.1.3 One Hot Encoding

To clearly explain the relationship between house grades and prices, the values for grade column were one hot encoded. The values were encoded in a 3 to 13 scale where 3 represents poor and 13 represents mansion. The scatter plot below shows that mansion tends to have a higher price which may be due to the materials used and design implemented during construction. Houses graded as poor tend to have a low price which may be attributed to lack of design and less expensive materials used in construction.



A code was used to create dummy variables for the grade column and the first value was dropped to avoid the dummy variable trap. The new variables were joined to the data frame.

2.1 CORRELATIONS AND MULTICOLLINEARITY

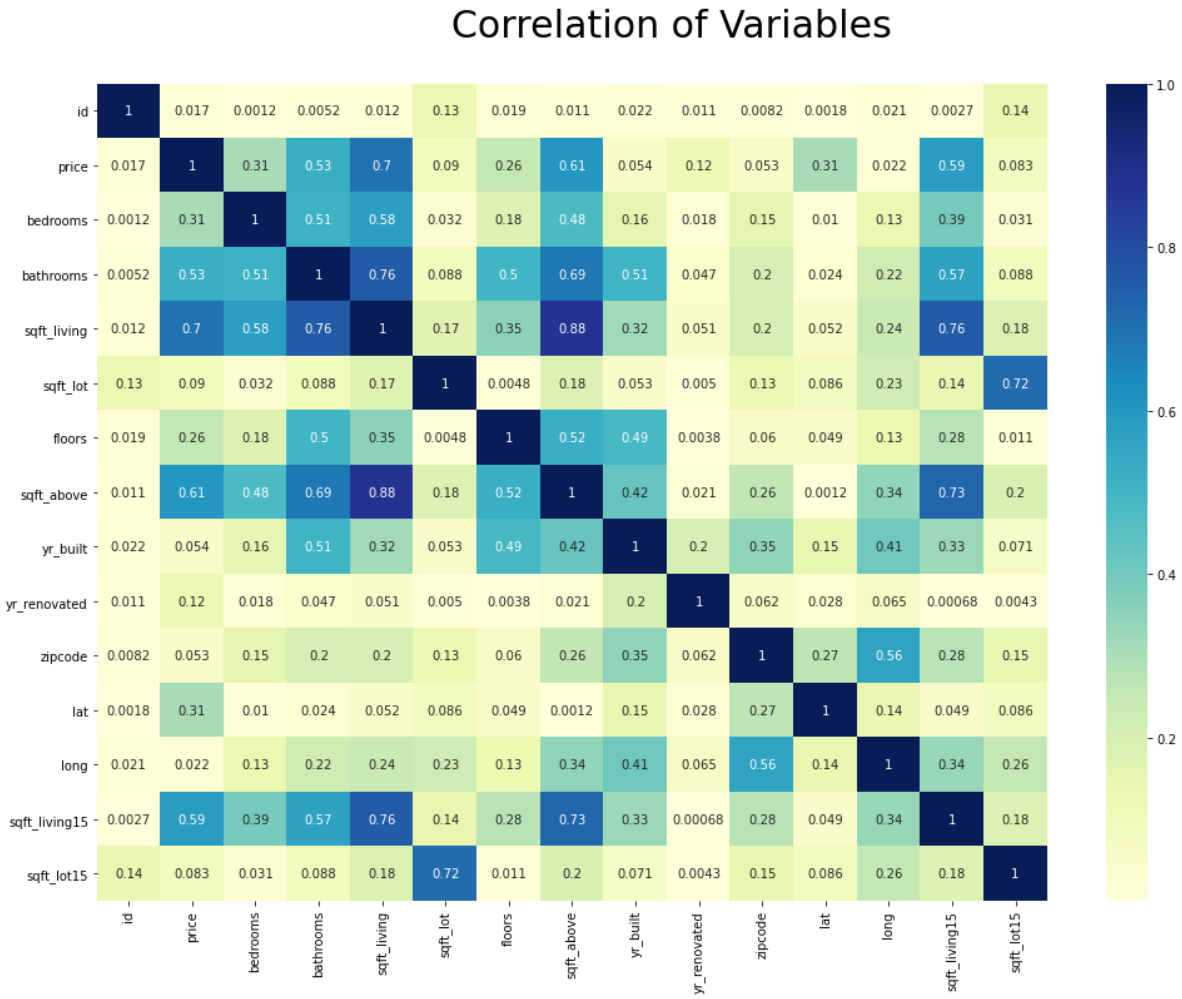
There are assumptions that must be checked before creating a multiple regression model;

- A linear relationship should exist between the response variable and predictor variable.
- Multicollinearity should be avoided between features hence the data should be independent.
- Variability should be equal across values of independent variables hence homoscedasticity.
- Residuals should follow a normal distribution.

2.2.1 Multicollinearity

A heatmap of correlations between each of the variables was created using seaborn. Additionally, the absolute value of correlations was calculated to determine the strength of the linear relationship.

According to the heatmap below, many of the variables related to size of homes like number of rooms and square foot are highly correlated with each other. However, we preferably want to include variables that have a high correlation with the y variable.



Variables whose correlation with one another exceeds 0.75 will not be included in our model since it implies multicollinearity.

Multicollinearity above 0.75	
Variables	Correlations
sqft_living, sqft_above	0.876448
bathrooms, sqft_living	0.755758

The table above displays the sets of variables that are highly correlated with each other hence violating the multicollinearity assumption. The most appropriate approach is to drop one variable from each pair since they cannot be included in the same model.

2.2.2 Correlations with price

The table below displays the correlation of each of the variables with price.

Correlations with price	
Correlation	**Variables**
0.701917	sqft_living
0.605368	sqft_above
0.585241	sqft_living15
0.525906	bathrooms
0.308787	bedrooms

Since sqft_living has a high correlation with price, it is likely to be used in the multiple regression model. Sqft_above and bathrooms will most likely be eliminated from our model due to their multicollinearity with sqft_living despite their correlation with price.

3.1 SIMPLE LINEAR REGRESSION

A simple linear regression has one explanatory/independent variable and one response/dependent variable. The assumption of linearity, normality and homoscedasticity require a regression model since they are mainly concerned with residuals.

Before coming up with the final regression model, simple linear regression analysis was conducted for each of the variables that remained after elimination due to multicollinearity. The variables include; sqft_living, bedrooms and sqft_living15. While conducting the analysis, the variables failed to fully satisfy normality and homoscedasticity. Log transformation was therefore conducted to update the variables to their natural log.

3.1.1 Test for linearity

Linearity was tested for each of the variables using the linear rainbow test from stats models. The null hypothesis states that the relationship between the variables is considered linear while the alternative hypothesis states that the relationship is not considered linear.

The p values obtained were compared to the standard alpha value of 0.05. A low p value indicates that the model is not linear hence the null hypothesis is rejected contrary to what normal p values indicate. On the other hand a high p value indicates that the relationship is linear hence we fail to reject the null hypothesis.

The test returns a test statistic based on the F test and a p value of the test.

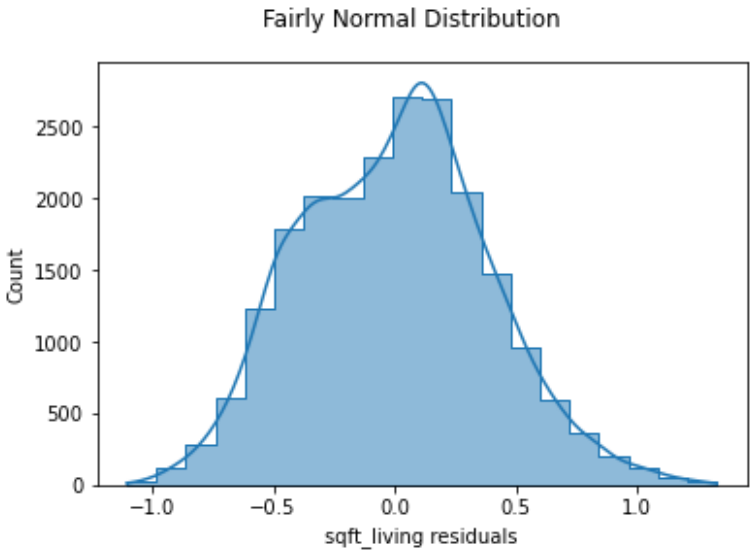
Linear Rainbow Test Results		
Sqft_living	Bedrooms	Sqft_living15
0.81189	0.88228	0.92861

As displayed in the above table all the simple linear regression models displayed a p value higher than 0.05 hence a linear relationship.

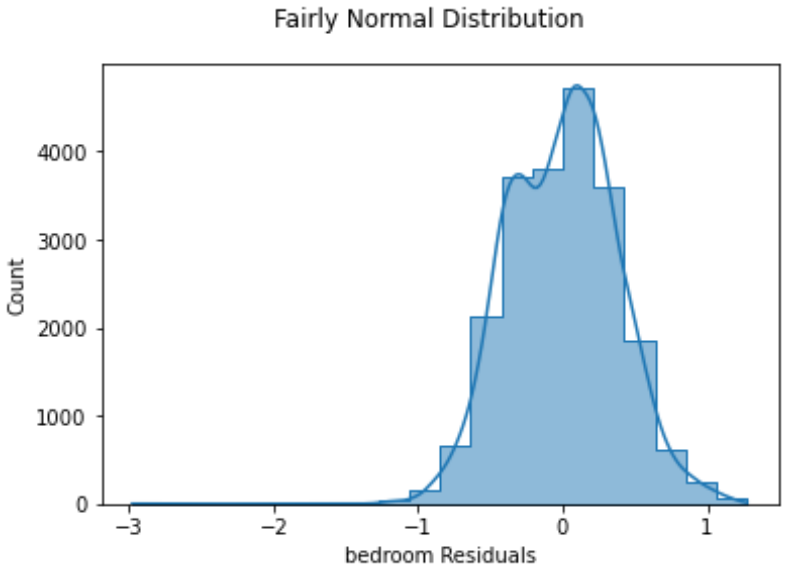
3.1.2 Test for normality

The simple linear regression models were tested for normality by plotting a histogram. All the models displayed a fairly normal distribution as shown in the histograms below.

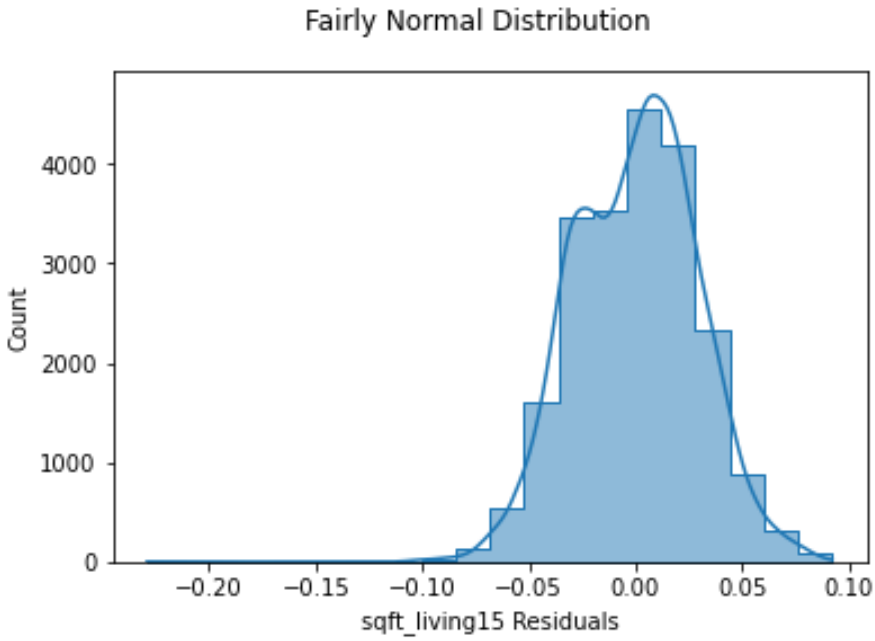
Histogram for square footage of living space in the home



Histogram for number of bedrooms



Histogram for the square footage of interior housing living space for nearest 15 neighbors.



3.1.3 Test for Homoscedasticity

Homoscedasticity was tested using goldfeld quandt test. The null hypothesis states that equality of variance exists between the dependent and independent variable while the alternative hypothesis states that inequality of variance exists between the dependent and independent variable.

The p values obtained were compared to the standard alpha value of 0.05. A low p value indicates inequality in variance hence heteroscedasticity and rejection of the null hypothesis. A high p value on the other hand, indicates equality in variance hence homoscedasticity and failure to reject the null hypothesis.

Goldfeld test results		
Sqft_living	Sqft_living15	bedrooms
0.4987	0.3213	0.9996

The above table shows that the p values obtained were all higher than 0.05 hence homoscedasticity of all the simple regression models.

4.1 MULTIPLE LINEAR REGRESSION

A multiple linear regression model has one dependent variable and two or more independent variables. The three variables that had a high correlation with price without violating the assumption of multicollinearity were included in the model as independent variables.

$$Price = \beta_0 + \beta_1(sqft\_living) + \beta_2(bedrooms) + \beta_3(sqft\_living15) + \varepsilon$$

Where;

Price = sale price

Sqft_living = square footage of living space in the home.

Bedrooms = number of bedrooms

Sqft_living15= square footage of interior housing living space for the nearest 15 neighbors.

ε = Error term

Model Summary

Variables	**Coefficient**	**P value**
Price	0.7079	0.000
Sqft_living	0.0221	0.000
bedrooms	-0.0076	0.000
Sqft_living15	0.0102	0.000
R squared	0.488	
Adjusted R squared	0.488	
Prob.(F-statistic)	0.00	

The adjusted R squared which is used to establish how predictive the model is had a value of 48.8%.This shows that sqft_living, bedrooms and sqft_living15 contributed to 48.8% variations

in sale price. The remaining 51.2% is left unexplained showing that sale price is also influenced by other variables apart from those used in our model. This shows the need for further studies to establish these factors.

A probability F-statistic value of 0.000 was obtained. This confirms the fitness of the regression model in implying the relationship in existence between sale price and predictor variables since, the value obtained was less than $\alpha = 0.05$.

4.1.1 Hypothesis Testing

Null hypothesis: no relationship exists between the chosen explanatory variables and response variable.

Alternative hypothesis: a relationship exists between the chosen explanatory variables and response variable.

The model coefficients all display a p value 0.00 that is below 0.05 hence the null hypothesis is rejected. This evidently shows that a relationship exists between the explanatory variables and response variable. The model is therefore statistically significant.

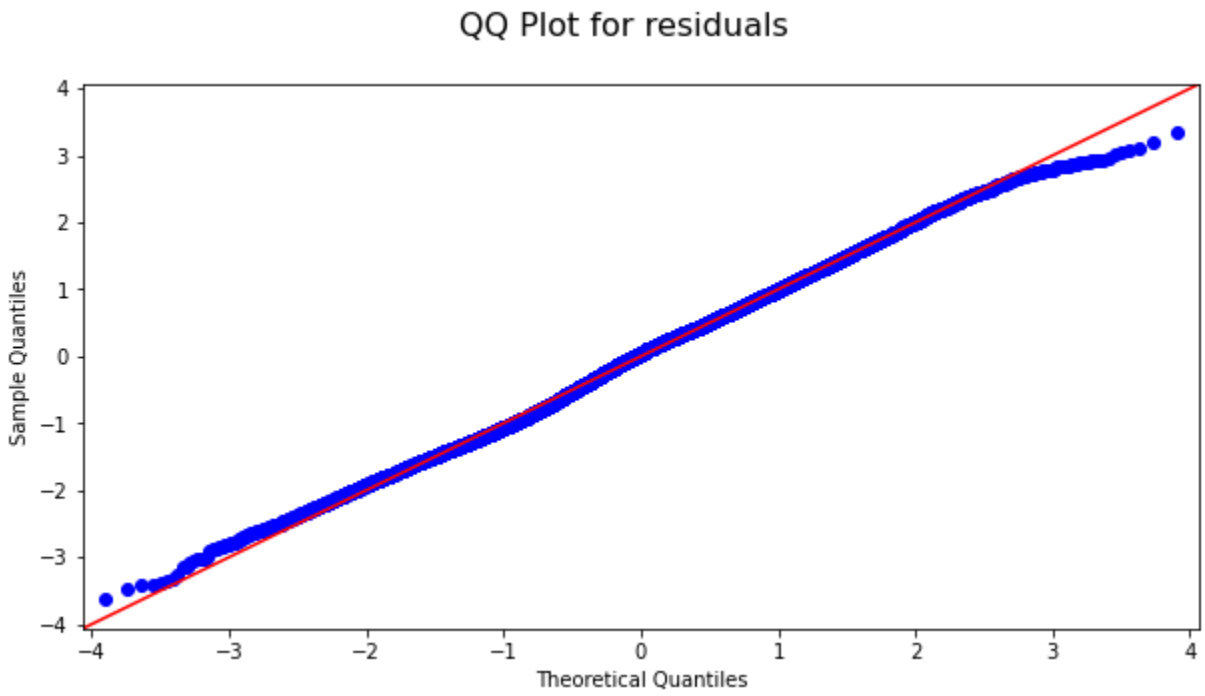
The established multiple linear regression for the study was:

$$Y= 0.7079 + 0.0221\text{sqft\_living} + 0.0102\text{sqft\_living15} - 0.0076\text{bedroom} + \varepsilon$$

Since the model has log transformed predictors and a log transformed target, the coefficient values are interpreted as percentages. The regression equation shows that if there is no change in sqft_living, bedrooms and sqft_living15, sale price would be 70.79%.
The results show that, when all other factors are held constant, 1% increase in square footage of living space in the home is associated with a 2% increase in sale price. The results also indicate that, 1% increase in number of bedrooms holding all factors constant results in a 0.76% decrease in sale price. The results further show that, all other factors held constant, 1% increase in square footage of interior living space for the nearest 15 neighbors is associated with a 1% increase in sale price.

4.1.2 Normality test

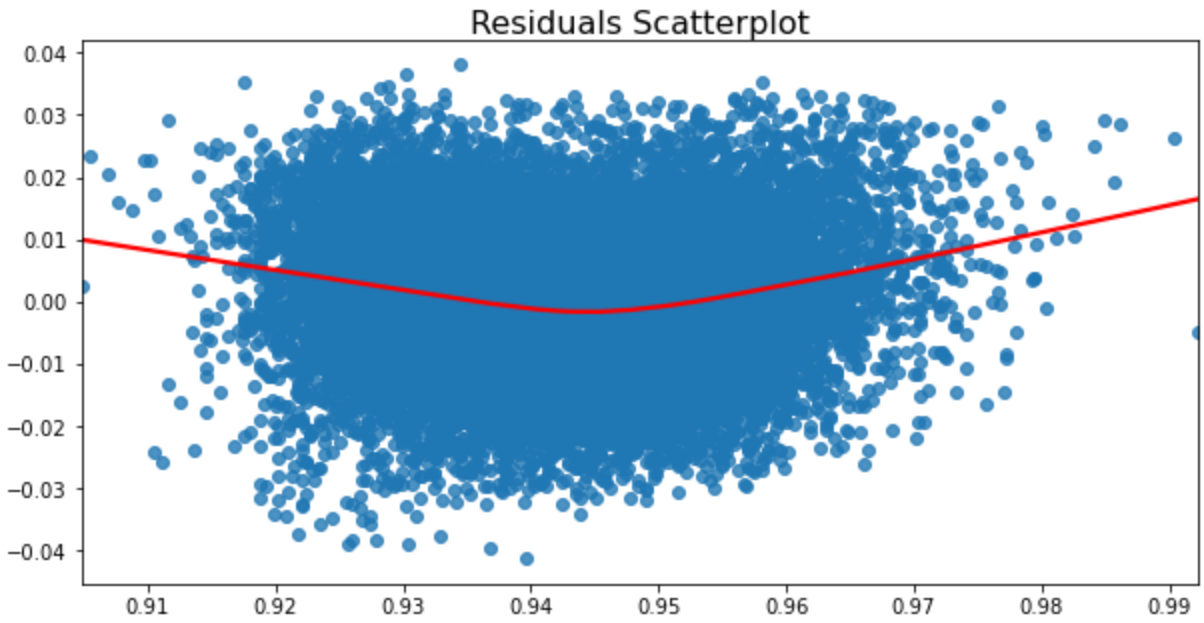
The homoscedasticity assumption was checked for each predictor variable when diagnosing the simple linear regression models. To check the normality of the model residuals, a QQ-plot was created to confirm that the residual fall on a straight line.



Since majority of the data points fall along a straight line according to the QQ plot, the normality of assumption is satisfied.

4.1.3 Homoscedasticity

The test for homoscedasticity was done using a scatter plot. The scatter plot was visualized with fitted values on the x axis and model residuals on the y axis. The homoscedasticity assumption holds when the shape of the points is roughly symmetrical across a line.



The points display a roughly symmetrical blob-like shape which is consistent across the x axis. The model therefore satisfies the assumption of homoscedasticity.

4.1.4 Model Evaluation

The model is evaluated using Adjusted R squared and Root Mean Squared Error (RMSE). The lower a RMSE value the better the model.

Adjusted R squared

Adjusted r squared value	0.48752
Non adjusted r squared	0.48759

Adjusted R squared increases only when increase in variance is explained more than what we would expect to see due to chance. The non-adjusted R squared and adjusted R squared both display similar results.

Root Mean Squared Error

The value displayed 0.01 indicates that the model is off by 0.01 price in a given prediction. The average prediction error rate is 1%.

5.1 CONCLUSION

Square footage of living space in the home, number of bedrooms and square footage of interior living space for nearest 15 neighbors effectively predict house prices in King County.

5.1.1 RECOMMENDATIONS

In order for homeowners to sell their homes at a higher price they should expand the square footage of living space. The square footage of neighbors' living space is a positive predictor of price, but homeowners have less control over this factor. However, they can increase sale price by encouraging neighbors to expand the square footage of their living space. Moreover, they should consider reducing the number of bedrooms since the analysis suggests that additional bedrooms reduce the sale price.

5.1.2 LIMITATIONS

The model has some limitations:

The variables were log transformed to satisfy regression assumptions. Therefore, any data to be used with the model would have to be subjected to the same preprocessing. Regional differences in housing prices limit applicability of the model to data from other countries. Moreover removal of outliers from the model may make it less appropriate in predicting large values.