**DS6372 – Project 1**

# Introduction

Price prediction is pivotal for real estate. Home sellers want to know the appropriate time to sell and how much profit they can expect from their efforts. Home buyers want to know whether they are getting a fair price, where to look for homes in their budget, and various trade-offs that accompany a purchasing decision. Real estate companies navigate both sides of real estate; hence, they too are a key stakeholder. These stakeholders utilize multiple factors related to real estate to determine the fair price for the property. These same factors can be built into a model for price prediction that assists in taking some of the guess work out of property pricing.

The purpose of this paper is to provide a predictive statistical analysis of house sales in King County, Washington, USA, using the King County housing dataset. The analysis has two main objectives with the dataset. The first objective is to build a regressing model using the dataset providing metrics off the model and to interpret the regression model. The second objective is to demonstrate an understanding of an advanced analysis workflow.

## Data Description

The data used for this analysis, described in the sections below, comes from the Kaggle website. The dataset may be found on the Kaggle website (<https://www.kaggle.com/harlfoxem/housesalesprediction>). The dataset contains housing related data for King County, Washington, USA representing homes sold from May 2014 through 2015. A detailed listing of King County Housing Sales variables and what they represent may be found on the King County Assessor website (<http://your.kingcounty.gov/assessor/eRealProperty/ResGlossaryOfTerms.html>). The total data set contains 21,613 observations with 20 features or variables. These 20 features contain information quantity and quality-based attributes of a physical property that may interest any of the key stakeholders (prospective home buyer, home seller, real estate company/agent). For example, the data provides answers to questions such as: “How many rooms in the property?”, “What is the condition of the property?”, “Is the property on the waterfront?”, “What is the location of the property?”.

The dataset contains a mix of categorical variables and numeric variables. The categorical variables indicate information a prospective stakeholder would like to understand such as a property grade, number of floors, and a condition. The numerical variables indicate information such as square footage that is above or below ground, price, square footage of the interior living space and square footage of the property lot.

## Exploratory Data Analysis (EDA)

Initial EDA determined that the variables listed below could be removed for reasons noted below:

* ID (ID) – this field is a record ID field and is not informational for analysis

Data quality checks were performed across the data. It was determined via summary statistics (Table 1) that there were no missing values within the data set that needed to be addressed.

Additional analysis was performed to evaluate QQ Plots for all variables (Figure 1) to quickly review data normality. Although not important for the data itself to be normally distributed these graphs provide an indication of data skew.

## Analysis Question 1

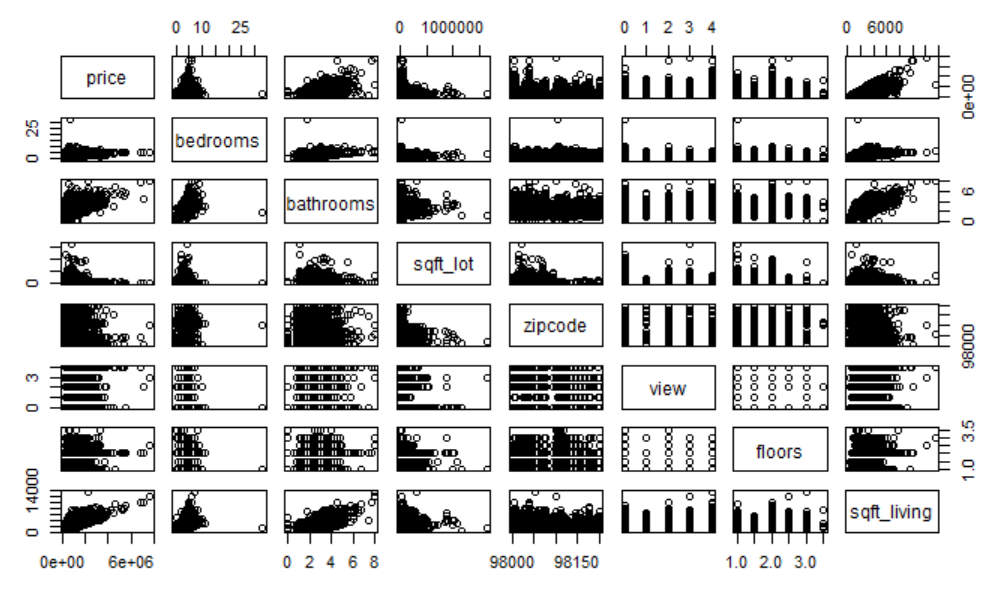
## Problem Statement

The goal of analysis question 1 is to perform a regression analysis and report on the predictive ability of the model using either a test data set, or some means through CV. This regression analysis should include an interpretation of the regression model coefficients, confidence intervals and hypothesis testing.

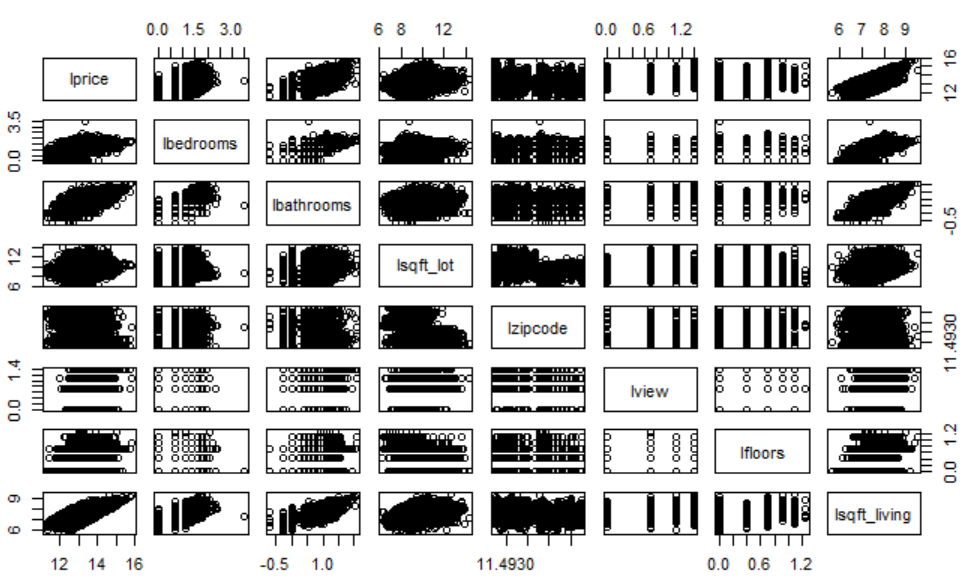
* Use alpha = 0.05, where necessary.
* Use R and provide relevant code and output.

**Step 1: Scatterplots**

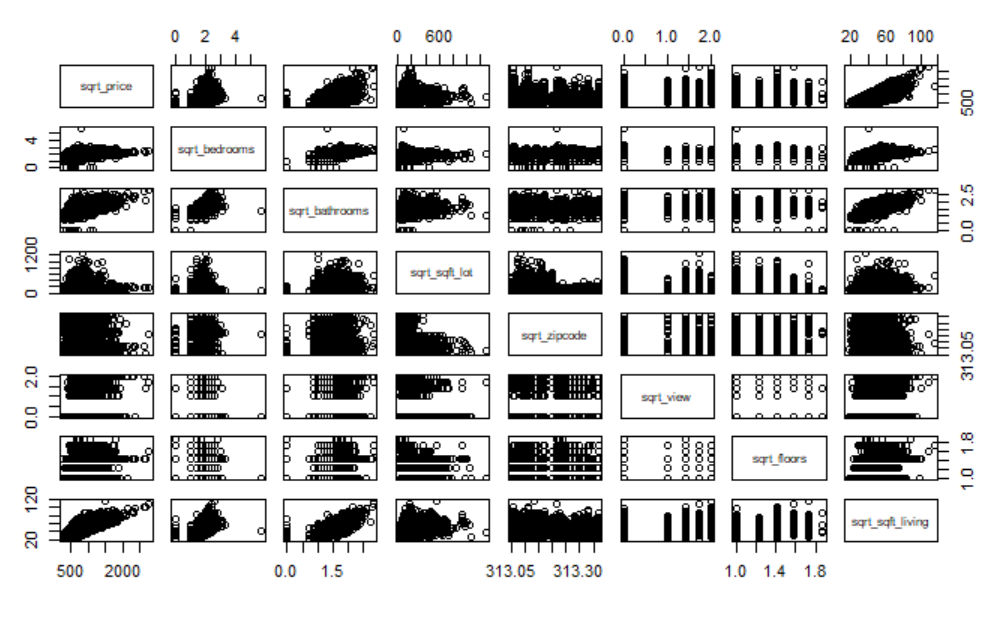
**Original data:**



**Try a transformation with log transformation with every relevant variable logged.**



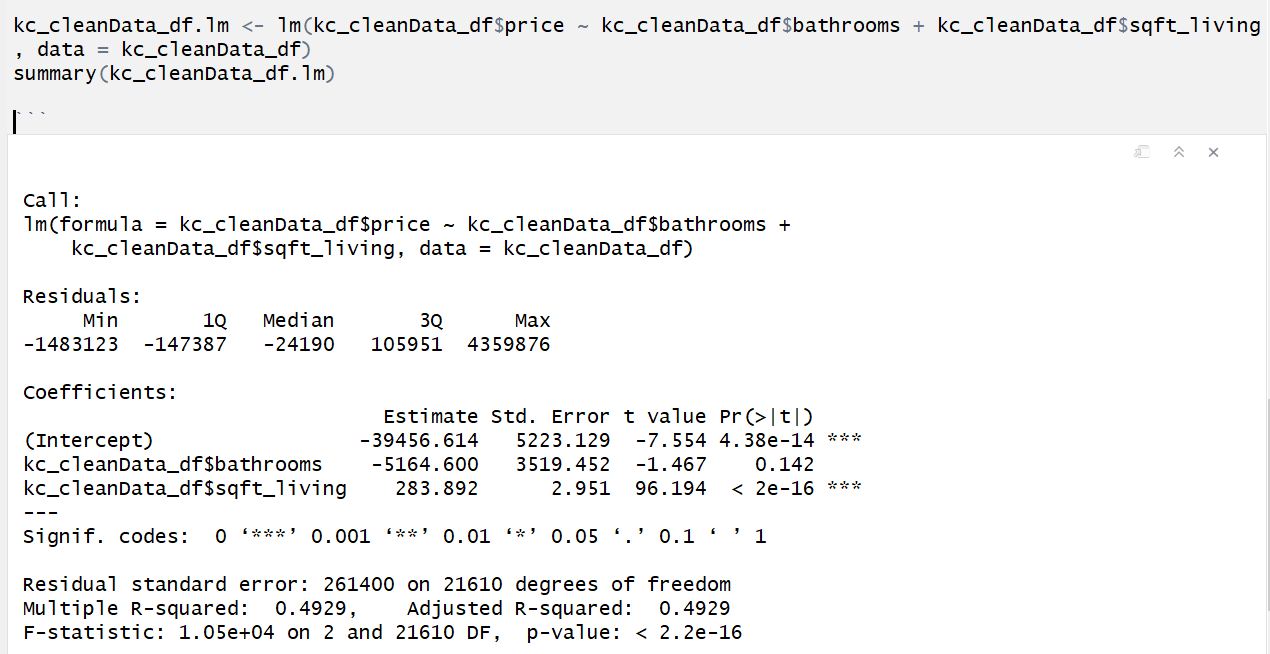
**Try a transformation with square root transformation with every relevant variable logged.**



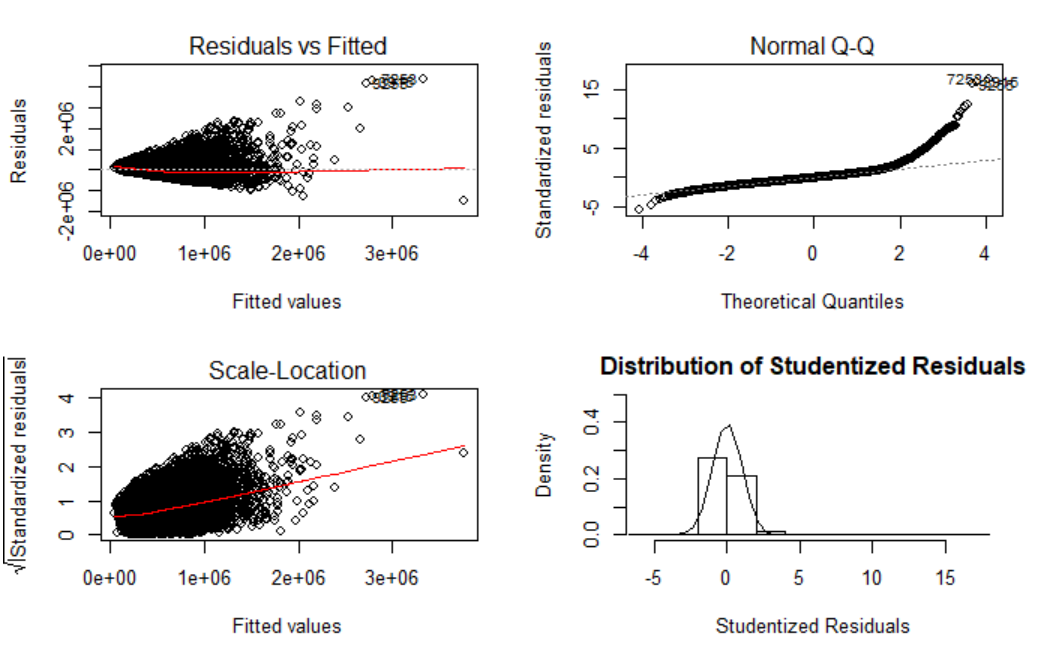
**The log transformed variables look to visually satisfy assumptions of linear trend, constant variance and conditional normality. Independence will be assumed.**

**Step 2: Build the model**

**Step 3: Fit the model**



**Step 4: Provide a residual plot, studentized residual plot, histogram of residuals, and q-q plot of residuals to provide evidence of the appropriateness of the model.**



**Residual Plot: There is a slight suspicion of non-constant variance, as we see a slight funnel shape as the predicted value increases. This is not too serious, and we will proceed with caution.**

**Studentized Residual Plot: This plot is very similar to the Residual plot although this plot identifies potential outlying observations. This may provide some evidence against the normality assumption. Further analysis may be necessary on these points.**

**Q-Q Plot of Residuals: The QQ Plot of residuals displayed below provides little evidence that the residuals are not normally distributed.**

**Histogram of Residuals: The histogram of residuals displayed does not provide strong evidence that the residuals are not normally distributed.**

## Model Selection

## Parameter Interpretation

## Conclusion

## Analysis Question 2

## Problem Statement

The goal of analysis question 2 is to perform a secondary analysis to demonstrate the ability to understand the analysis workflow and clarification of writing conclusions.

## 

## Main Analysis

## Conclusion

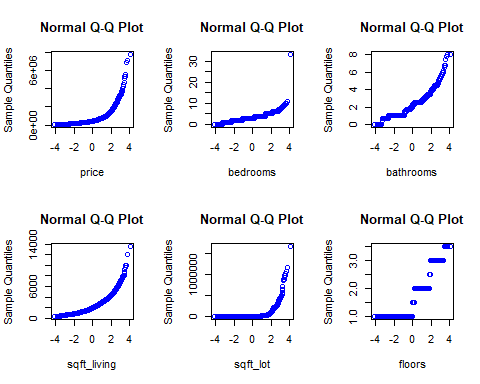
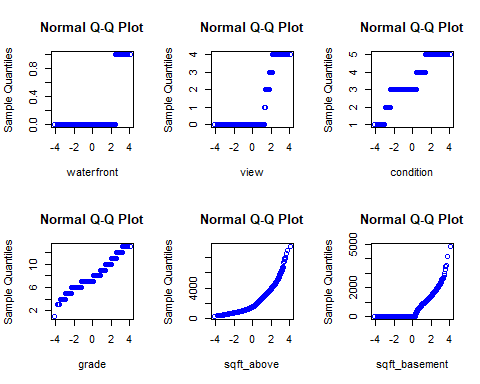
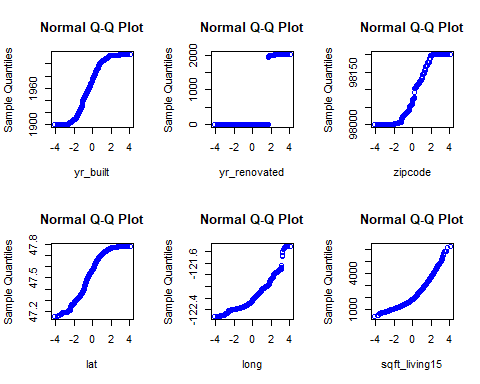
## Appendix

## List of Tables

Table 1 – EDA Summary Statistics

## id date price   
## Min. :1.000e+06 20140623T000000: 142 Min. : 75000   
## 1st Qu.:2.123e+09 20140625T000000: 131 1st Qu.: 321950   
## Median :3.905e+09 20140626T000000: 131 Median : 450000   
## Mean :4.580e+09 20140708T000000: 127 Mean : 540088   
## 3rd Qu.:7.309e+09 20150427T000000: 126 3rd Qu.: 645000   
## Max. :9.900e+09 20150325T000000: 123 Max. :7700000   
## (Other) :20833   
## bedrooms bathrooms sqft\_living sqft\_lot   
## Min. : 0.000 Min. :0.000 Min. : 290 Min. : 520   
## 1st Qu.: 3.000 1st Qu.:1.750 1st Qu.: 1427 1st Qu.: 5040   
## Median : 3.000 Median :2.250 Median : 1910 Median : 7618   
## Mean : 3.371 Mean :2.115 Mean : 2080 Mean : 15107   
## 3rd Qu.: 4.000 3rd Qu.:2.500 3rd Qu.: 2550 3rd Qu.: 10688   
## Max. :33.000 Max. :8.000 Max. :13540 Max. :1651359   
##   
## floors waterfront view condition   
## Min. :1.000 Min. :0.000000 Min. :0.0000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:0.000000 1st Qu.:0.0000 1st Qu.:3.000   
## Median :1.500 Median :0.000000 Median :0.0000 Median :3.000   
## Mean :1.494 Mean :0.007542 Mean :0.2343 Mean :3.409   
## 3rd Qu.:2.000 3rd Qu.:0.000000 3rd Qu.:0.0000 3rd Qu.:4.000   
## Max. :3.500 Max. :1.000000 Max. :4.0000 Max. :5.000   
##   
## grade sqft\_above sqft\_basement yr\_built   
## Min. : 1.000 Min. : 290 Min. : 0.0 Min. :1900   
## 1st Qu.: 7.000 1st Qu.:1190 1st Qu.: 0.0 1st Qu.:1951   
## Median : 7.000 Median :1560 Median : 0.0 Median :1975   
## Mean : 7.657 Mean :1788 Mean : 291.5 Mean :1971   
## 3rd Qu.: 8.000 3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997   
## Max. :13.000 Max. :9410 Max. :4820.0 Max. :2015   
##   
## yr\_renovated zipcode lat long   
## Min. : 0.0 Min. :98001 Min. :47.16 Min. :-122.5   
## 1st Qu.: 0.0 1st Qu.:98033 1st Qu.:47.47 1st Qu.:-122.3   
## Median : 0.0 Median :98065 Median :47.57 Median :-122.2   
## Mean : 84.4 Mean :98078 Mean :47.56 Mean :-122.2   
## 3rd Qu.: 0.0 3rd Qu.:98118 3rd Qu.:47.68 3rd Qu.:-122.1   
## Max. :2015.0 Max. :98199 Max. :47.78 Max. :-121.3   
##   
## sqft\_living15 sqft\_lot15   
## Min. : 399 Min. : 651   
## 1st Qu.:1490 1st Qu.: 5100   
## Median :1840 Median : 7620   
## Mean :1987 Mean : 12768   
## 3rd Qu.:2360 3rd Qu.: 10083   
## Max. :6210 Max. :871200   
##

## List of Figures

Figure 1 – QQ Plots of Features  

## Completed Code