ECON 320 Project - Model to Estimate the Price of Houses/Apartments in New York

Tariq Attarwala

03/05/2019

Research Proposal

My project focuses on creating a model that predicts the Price (Dependant Variable) of Houses/Properties in New York based on 4 Independent variables: (1) Living Area and (2) Number of Bathrooms (3) Number of Bedrooms and (4) Age of the House.

Expected Findings:

I expect there to be a strong positive correltaion between the Living Area and the House price, given the fact that bigger houses are more costly.

I also expect the number of Bedrooms and Bathrooms to have a positive correlation with the price.

Inversely, I expect there to be a negative correlation between the Age of the house and the Price, based on the understanding that older houses may be in worse conditions, and hence may be valued at a lower price.

This will help me understand if there is a causal relationship between the independent variables(Living Area, Number of Bathrooms, Number of Bedrooms and Age), and the dependent variable(Price).

Data Set

I have acquired my data from DASL - The Data and Story Library. Link: https://dasl.datadescription.com/

I decided to use the Dataset for Housing Prices in New York because it contains several important variables such as Living area, Bedrooms, Bathrooms, Fireplaces, Lot Size, Age etc.

It would be also be useful to use this dataset since the information is available on 1057 houses, which is a big enough data set to prevent any biases which may occur.

The link to the dataset: https://dasl.datadescription.com/datafile/housing-prices/?_sf_s=house&_sfm_cases=4+59943

Libraries Loaded:

1. library(tidyverse):The tidyverse package is designed to make it easy to install and load core packages.
2. library(ggplot2):The ggplot2 package is used for data visualiation.
3. library(dbplyr):The dbplyr package is used for data manipulation.
4. library(stargazer):The stargazer package is useful because of the large number of models it supports, its ease of use, and its beautiful aesthetics when making tables.
5. library(cars): The cars package is used to run regression analysis.
6. library(readxl): The readxl package is used to read and import the data from excel into R.
7. library(knitr): The knitr package is used to create the html documents for reporting purposes.

Summary of the Data Set

```
library(readxl)
Housing = read_excel("~/Desktop/Housing.xlsx", col_types = c("numeric",
    "numeric", "numeric", "numeric", "numeric", "numeric",
    "numeric", "numeric"))

HousingData = Housing %>% select(Price, Living.Area,
    Bathrooms, Bedrooms, Age)
```

In the chunck of code above, I have created a variable known as “Housing” to store all the data. Furthermore, I have specified that all the variables should have a “numeric” value. After that, I have created a variable “HousingData” in order to choose the Price, Living Area, Bathrooms, Bedrooms and Age parameters, using the select function. I am going to use these variables for my analysis.

```
kable(summary(HousingData), digits = 2, caption = "Summary of the Housing Data in New York")
```

Summary of the Housing Data in New York

Price	Living.Area	Bathrooms	Bedrooms	Age
Min. : 16858	Min. : 672	Min. :1.000	Min. :1.000	Min. : 0.00
1st Qu.:112400	1st Qu.:1342	1st Qu.:1.500	1st Qu.:3.000	1st Qu.: 6.00
Median :152404	Median :1675	Median :2.000	Median :3.000	Median : 18.00
Mean :167902	Mean :1819	Mean :1.929	Mean :3.184	Mean : 28.09
3rd Qu.:206512	3rd Qu.:2223	3rd Qu.:2.500	3rd Qu.:4.000	3rd Qu.: 34.00
Max. :599701	Max. :5228	Max. :4.500	Max. :5.000	Max. :247.00

```
stargazer(as.data.frame(HousingData), type = "html",
    flip = TRUE, summary.stat = c("n", "mean", "sd",
        "median", "min", "max"), title = "Housing Statistics for New York")
```

Housing Statistics for New York

Statistic	Price	Living.Area	Bathrooms	Bedrooms	Age
N	1,057	1,057	1,057	1,057	1,057
Mean	167,901.900	1,819.498	1.929	3.184	28.090
St. Dev.	77,158.350	662.941	0.650	0.740	34.928
Median	152,404	1,675	2	3	18
Min	16,858	672	1	1	0
Max	599,701	5,228	4	5	247

I have also shown the summary of the Housing Dataset using kable and stargazer. I have used this to display the important statistics that

are required for the analysis of my dataset.

Description of the Data Set

```
str(HousingData)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    1057 obs. of  5 variables:
##  $ Price      : num  142212 134865 118007 138297 129470 ...
##  $ Living.Area: num  1982 1676 1694 1800 2088 ...
##  $ Bathrooms  : num  1 1.5 2 1 1 2 1.5 1 1 1.5 ...
##  $ Bedrooms   : num  3 3 3 2 3 3 2 2 2 3 ...
##  $ Age        : num  133 14 15 49 29 10 12 87 101 14 ...
```

The dataset contains 1057 observations of 4 variables:

Price - Price of the House

Important note - *I have decided to use the logarithmic function for Price to make the highly skewed distribution less skewed. This is valuable, for both making patterns in data more interpretable, and for helping to meet my assumptions of inferential statistics.*

Bedrooms - Number of Bedrooms in the house

Bathrooms - Number of Bathrooms in the house

Age - Age of Home in Years

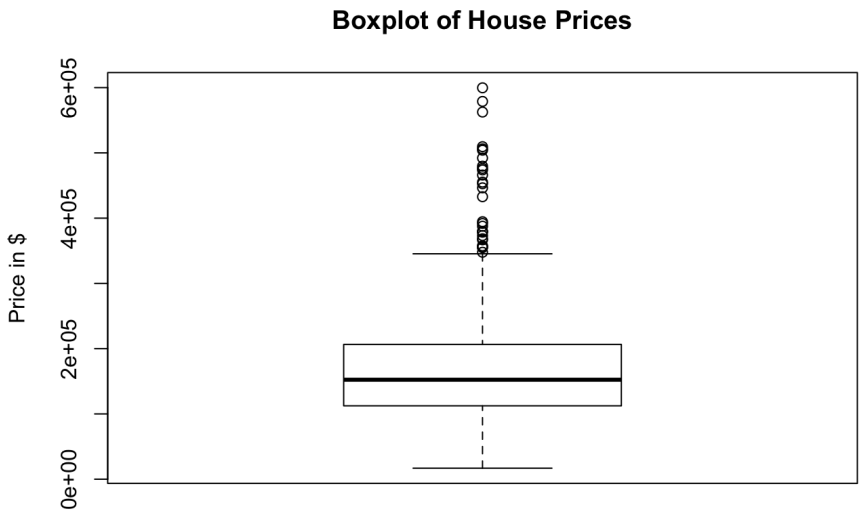
Living Area - Size of Living area by Square Feet

In this project, the dependent variable will be the Price and the independent variables will be Bedrooms, Bathrooms, Age and Living Area.

I want to investigate if the Price can be explained by the independent variables listed above.

Boxplot of House Prices in Dataset

```
boxplot(HousingData$Price, main = "Boxplot of House Prices",
    xlab = "", ylab = "Price in $")
```



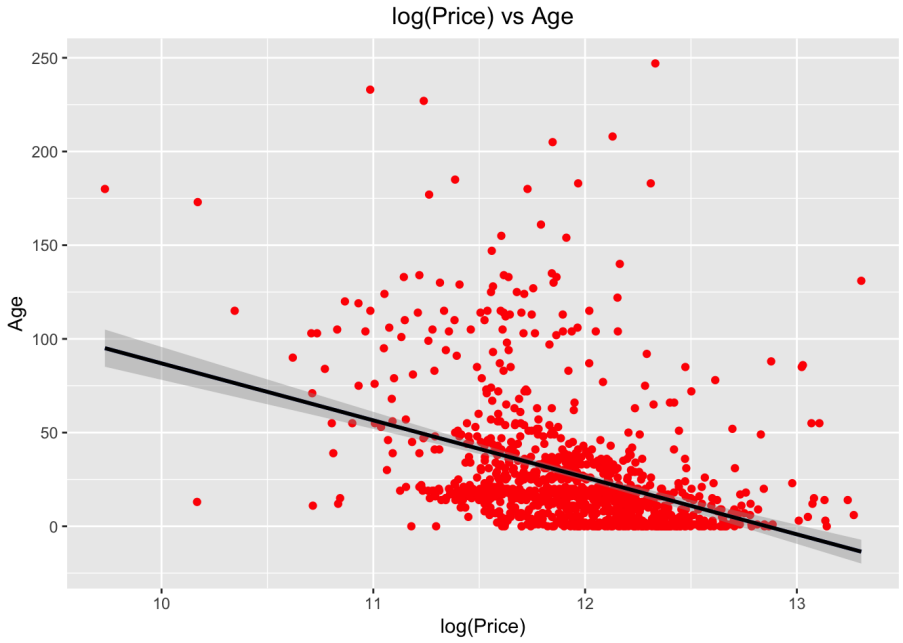
The graph above shows us the Price range of the houses in the Dataset.

Scatter Plots for Each Independent Variable

```
ggplot(HousingData, aes(log(Price), Living.Area)) +
    geom_point(color = "red") + geom_smooth(method = "lm",
    se = FALSE) + ggtitle("log(Price) vs Living Area") +
    xlab("log(Price)") + ylab("Living Area") + theme(plot.title = element_text(hjust = 0.5)) +
    stat_smooth(method = "lm", col = "black")
```

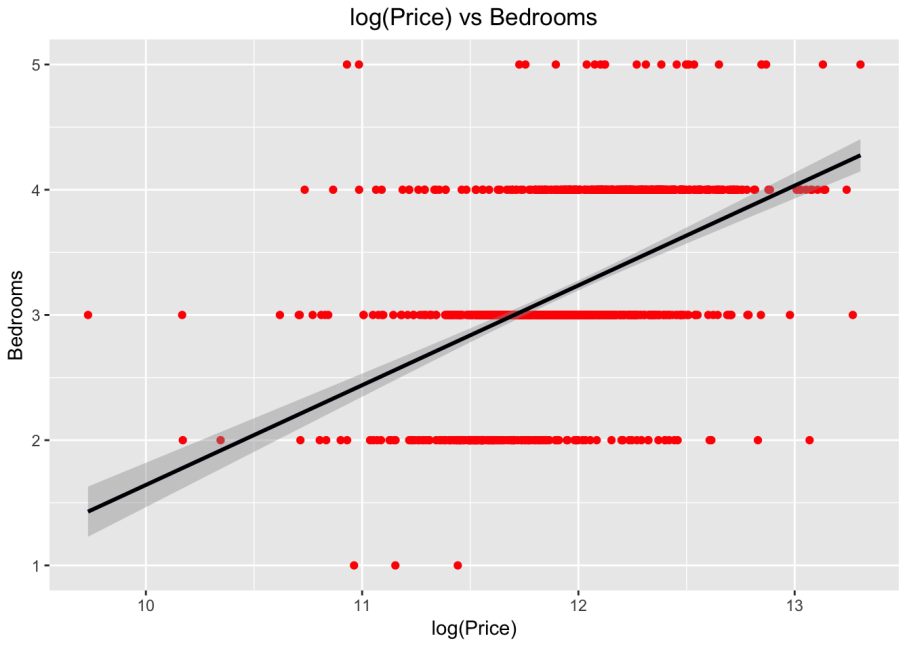
The graph above shows us that there is a **Strong and Positive** correlation between the Price and size of Living Area.

```
ggplot(HousingData, aes(log(Price), Age)) + geom_point(color = "red") +
    geom_smooth(method = "lm", se = FALSE) + ggtitle("log(Price) vs Age") +
    xlab("log(Price)") + ylab("Age") + theme(plot.title = element_text(hjust = 0.5)) +
    stat_smooth(method = "lm", col = "black")
```



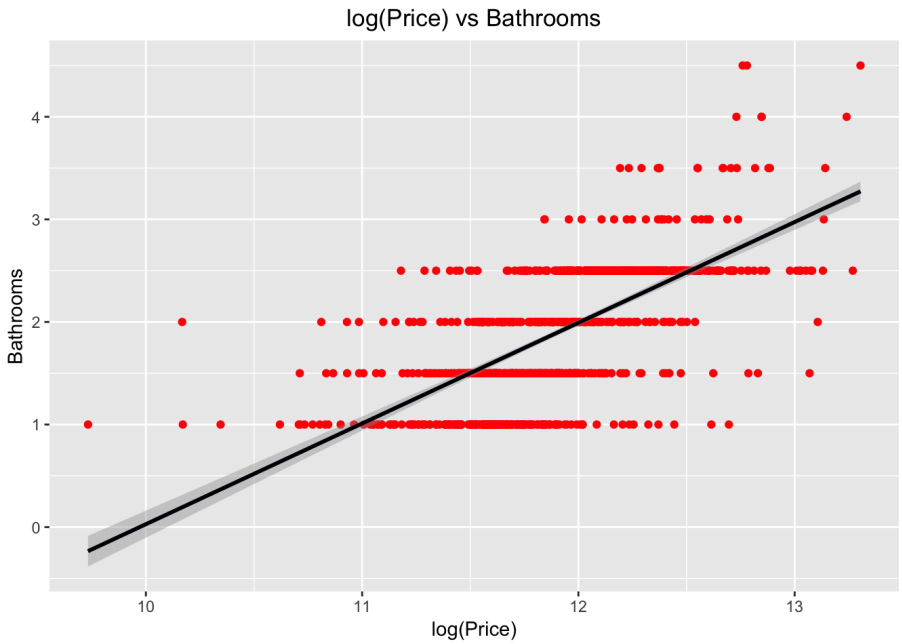
The graph above shows us that there is a **Negative** correlation between the Price and Age of the House.

```
ggplot(HousingData, aes(log(Price), Bedrooms)) + geom_point(color = "red") +
    geom_smooth(method = "lm", se = FALSE) + ggtitle("log(Price) vs Bedrooms") +
    xlab("log(Price)") + ylab("Bedrooms") + theme(plot.title = element_text(hjust = 0.5)) +
    stat_smooth(method = "lm", col = "black")
```



The graph above shows us that there is a **Positive** correlation between the Price and Number of Bedrooms in the House.

```
ggplot(HousingData, aes(log(Price), Bathrooms)) + geom_point(color = "red") +
    geom_smooth(method = "lm", se = FALSE) + ggtitle("log(Price) vs Bathrooms") +
    xlab("log(Price)") + ylab("Bathrooms") + theme(plot.title = element_text(hjust = 0.5)) +
    stat_smooth(method = "lm", col = "black")
```



The graph above shows us that there is a **Positive** correlation between the Price and Number of Bathrooms in the House.

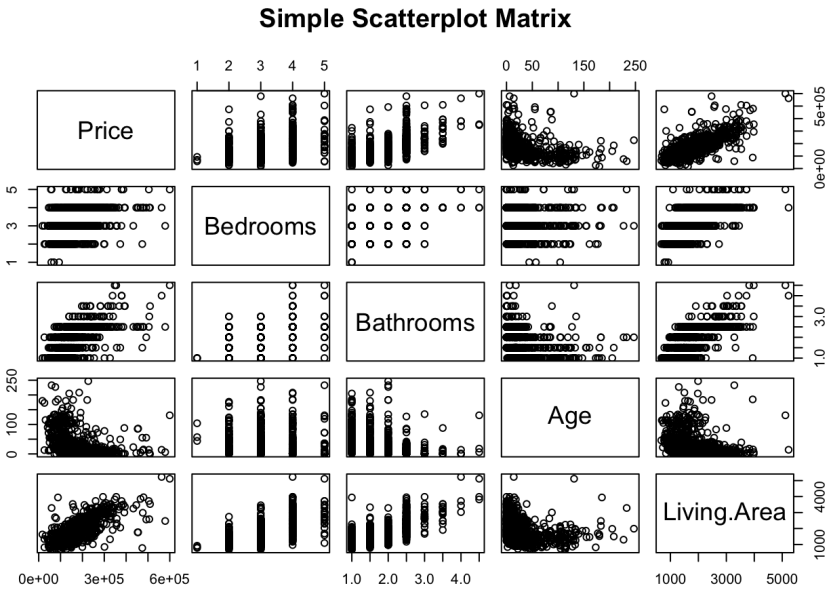
Correlation and Visualisation

I will now create a correlation matrix for the 5 variables in order to find the magnitude of correlation identified in the graphs above.

```
(cormat <- round(cor(HousingData[, c("Price", "Bedrooms",
    "Bathrooms", "Age", "Living.Area")]), 2))
```

```
##             Price Bedrooms Bathrooms   Age Living.Area
## Price        1.00     0.46      0.65 -0.30        0.76
## Bedrooms     0.46     1.00      0.51 -0.06        0.66
## Bathrooms    0.65     0.51      1.00 -0.42        0.73
## Age         -0.30    -0.06     -0.42  1.00       -0.25
## Living.Area  0.76     0.66      0.73 -0.25        1.00
```

```
pairs(~Price + Bedrooms + Bathrooms + Age + Living.Area,
    data = HousingData, main = "Simple Scatterplot Matrix")
```



Regression Analysis - T Test

We will now calculate the T Statistic for our independent variables, in order to see which variables to add to our model.

On the basis of the Correlation Matrix created above, we will include variables one at a time in the order of those that I believe will explain the Price variable the most.

1. Regression of Price with Living Area

```
Reg1 <- lm(log(Price) ~ Living.Area, data = HousingData)
(summary(Reg1)$adj.r.squared)
```

```
## [1] 0.531233
```

We can see over here that Living Area is a significant variable for this model. It explains **53.12%** of the variation in Price of the House, indicated by the Adjusted R-Squared value.

2. Regression of Price with Living Area and Number of Bathrooms

```
Reg2 <- lm(log(Price) ~ Living.Area + Bathrooms, data = HousingData)
(summary(Reg2)$adj.r.squared)
```

```
## [1] 0.5688318
```

By adding another variable to the model, we can see over here that Number of Bathrooms is also a significant variable for this model. Along with Living Area, it explains **56.88%** of the variation in Price of the House, indicated by the Adjusted R-Squared value. Therefore, it should

be included in our model.

3. Regression of Price with Living Area, Number of Bathrooms, and Age of the House

```
Reg3 <- lm(log(Price) ~ Living.Area + Bathrooms + Age,
    data = HousingData)
(summary(Reg3)$adj.r.squared)
```

```
## [1] 0.590508
```

By adding another variable to the model, we can see over here that the Age of the House is also a significant variable for this model. Along with Living Area and Number of Bathrooms, it explains **59.05%** of the variation in Price of the House, indicated by the Adjusted R-Squared value. Therefore, it should be included in our model.

3. Regression of Price with Living Area, Number of Bathrooms, Age of the House and Number of Bedrooms

```
Reg4 = lm(log(Price) ~ Living.Area + Bathrooms + Age +
    Bedrooms, data = HousingData)
(summary(Reg4)$adj.r.squared)
```

```
## [1] 0.5901416
```

By adding another variable to the model, we can see over here that the Number of Bedrooms is **not** a significant variable for this model. Along with Living Area, Number of Bathrooms and Age of the House, it still only explains **59.01%** of the variation in Price of the House, indicated by the Adjusted R-Squared value. Therefore, it should **not** be included in our model.

The Multiple Regressions run above have been put into a table format for easy interpretation:

```
stargazer(Reg2, Reg3, Reg4, type = "text")
```

```
##
## ===============================================================================================
##                                          Dependent variable:
##                     -----------------------------------------------------------------------
##                                              log(Price)
##                               (1)                 (2)                 (3)
## -----------------------------------------------------------------------------------------------
## Living.Area                0.0003***           0.0004***           0.0004***
##                            (0.00002)           (0.00002)           (0.00002)
##
## Bathrooms                  0.195***            0.136***            0.136***
##                             (0.020)             (0.021)             (0.021)
##
## Age                                            -0.002***           -0.002***
##                                                (0.0003)            (0.0003)
##
## Bedrooms                                                            0.004
##                                                                     (0.016)
##
## Constant                  10.931***           11.075***           11.069***
##                            (0.029)             (0.034)             (0.043)
##
## -----------------------------------------------------------------------------------------------
## Observations                1,057               1,057               1,057
## R2                          0.570               0.592               0.592
## Adjusted R2                 0.569               0.591               0.590
## Residual Std. Error     0.290 (df = 1054)   0.283 (df = 1053)   0.283 (df = 1052)
## F Statistic         697.580*** (df = 2; 1054) 508.602*** (df = 3; 1053) 381.125*** (df = 4; 1052)
## ===============================================================================================
## Note:                                                           *p<0.1; **p<0.05; ***p<0.01
```

Inference

We will now calculate the values for the variables at a Confidence Interval from 0.5% to 99.5%

```
confint(lm(log(Price) ~ Living.Area + Bathrooms + Age +
    Bedrooms, data = HousingData), level = 0.99)
```

```
##                      0.5 %        99.5 %
## (Intercept) 10.9591260493 11.1785284939
## Living.Area  0.0003004589  0.0004149569
## Bathrooms    0.0810620239  0.1907347096
## Age         -0.0028157731 -0.0013691929
## Bedrooms    -0.0372802401  0.0449922476
```

We see over here that the Confidence Interval of the Number of Bedrooms ranges from -0.0372802401 to 0.0449922476, and hence crosses the 0 value. Therefore, we can safely remove the variable from our model.

It can be speculated that the number of rooms in the house is not a significant variable for predicting the Price of Houses in our model because, houses of the same size with more number of rooms will not necessarily cause the Price of the House to increase.

Results

The best model we have estimated for predicting the Housing Prices in Reg3:

$$HousingPrice = \beta_0 + \beta_1 * LivingArea + \beta_2 * Bathrooms + \beta_3 * Age + \mu$$

The multiple linear regression model we have created for predicting the Housing prices in New York is written below:

```
(Reg3$coefficients)
```

```
##   (Intercept)   Living.Area     Bathrooms           Age
## 11.0749738929  0.0003603178  0.1364493290 -0.0020810356
```

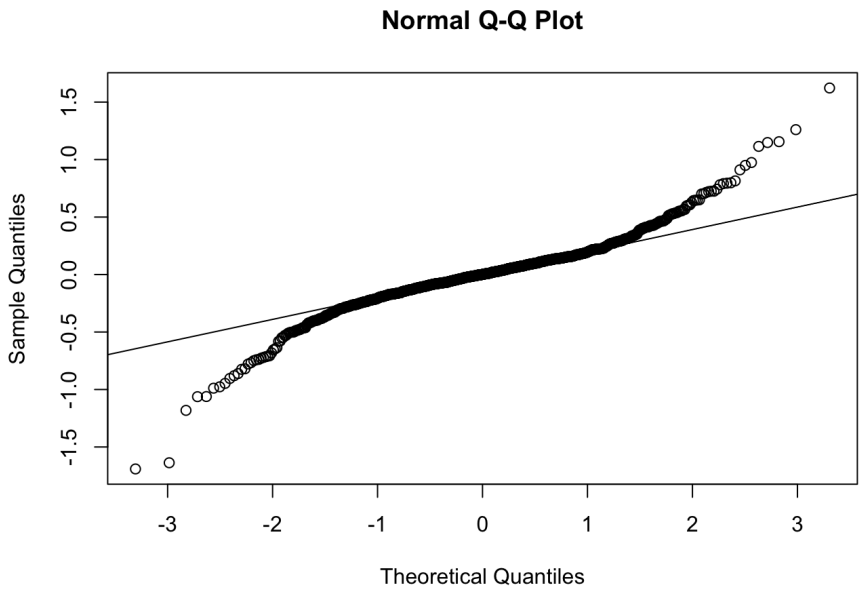
$$HousingPrice = 11.0749738929 + 0.0003603178 * LivingArea + 0.1364493290 * Bathrooms - 0.0020810356 * Age + \mu$$

```
summary((lm(log(Price) ~ Living.Area + Bathrooms +
    Age, data = HousingData)))
```

```
##
## Call:
## lm(formula = log(Price) ~ Living.Area + Bathrooms + Age, data = HousingData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.69121 -0.13070  0.00097  0.13314  1.62259
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.107e+01  3.407e-02 325.106  < 2e-16 ***
## Living.Area  3.603e-04  1.938e-05  18.596  < 2e-16 ***
## Bathrooms    1.364e-01  2.112e-02   6.461 1.58e-10 ***
## Age         -2.081e-03  2.761e-04  -7.536 1.04e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2826 on 1053 degrees of freedom
## Multiple R-squared:  0.5917, Adjusted R-squared:  0.5905
## F-statistic: 508.6 on 3 and 1053 DF,  p-value: < 2.2e-16
```

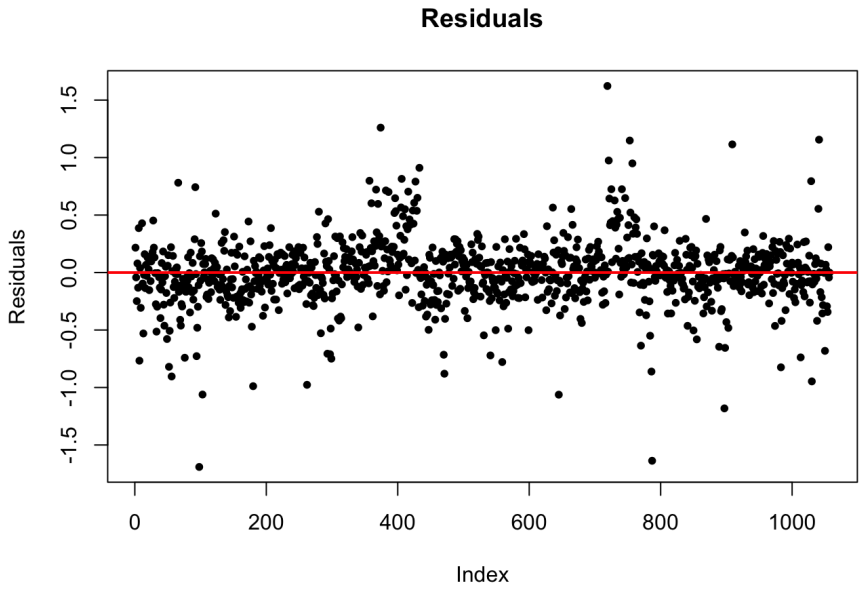
Residuals Plot

```
qqnorm(residuals(Reg3))
qqline(residuals(Reg3))
```



As we can see from the graph above, although the upper and lower ends don't fit the line, the majority of the points fit well, and hence the residuals are normally distributed.

```
Residuals <- Reg3$residuals
plot(Residuals, main = "Residuals", pch = 20)
abline(0, 0, lwd = 2, col = "Red")
```



The purpose of residuals is to show the difference between the actual values, and the values fitted my the model we have used. Since the 0, this indicates that the model is a good fit.

TESLA Stock Price Prediction

Tariq Attarwala

10/22/2018

Research Proposal:

RESEARCH PROPOSAL:

Through this project, I aim to predict the Market price of TSLA after 8 years, i.e. the Market Price of TSLA on 2036-10-17. By using the Market Price for TSLA from 2007-01-01 to 2018-10-17, I will create a time series, and calculate the mean and standard deviation of the market price data. After creating the time series, I will implement Technical Analysis tools such as Bollinger Bands and Moving Averages, which are included in the quantmod package. Also, I will use the price data to determine the mean log return of the stock, and create a Probability Distribution Function in order to calculate the probability of return in each quantile. Furthermore, I will finally use the mean log return to extraplote prices for the next 2 years.

DATA SET:

For this project, I aim to analyse the Price and Volume data for TESLA stock. I will be using the data directly from the Yahoo Finance page to make my calculations. I will use the quantmod package to directly input data from Yahoo Finance into R. The data is collected from 1/1/17 to 17/10/18.

`getSymbols("TSLA", from = "2007-01-01", to = "2018-10-17")` is used to load the data into R using quantmod.

Libraries:

1. library(tidyverse) - The tidyverse package is designed to make it easy to install and load core packages.
2. library(quantmod) - This package is loaded in order to provide a framework for quantitative financial modeling.
3. library(xts) - It is a constructor fucntion used to create a time series object
4. library(ggplot2) - This package is loaded in order to create graphs and charts

Summary and Structure of Data set:

TSLA.Open is the Opening Market Price of TESLA in the day. TSLA.Volume is the number of stocks that are traded in the day.

```
summary(TSLA$TSLA.Open)
```

```
##       Index              TSLA.Open     
##  Min.   :2010-06-29   Min.   : 16.14  
##  1st Qu.:2012-07-24   1st Qu.: 32.46  
##  Median :2014-08-22   Median :200.66  
##  Mean   :2014-08-22   Mean   :168.59  
##  3rd Qu.:2016-09-19   3rd Qu.:252.06  
##  Max.   :2018-10-16   Max.   :386.69
```

```
summary(TSLA$TSLA.Volume)
```

```
##       Index              TSLA.Volume      
##  Min.   :2010-06-29   Min.   :  118500  
##  1st Qu.:2012-07-24   1st Qu.: 1504850  
##  Median :2014-08-22   Median : 4009400  
##  Mean   :2014-08-22   Mean   : 4912742  
##  3rd Qu.:2016-09-19   3rd Qu.: 6564750  
##  Max.   :2018-10-16   Max.   :37163900
```

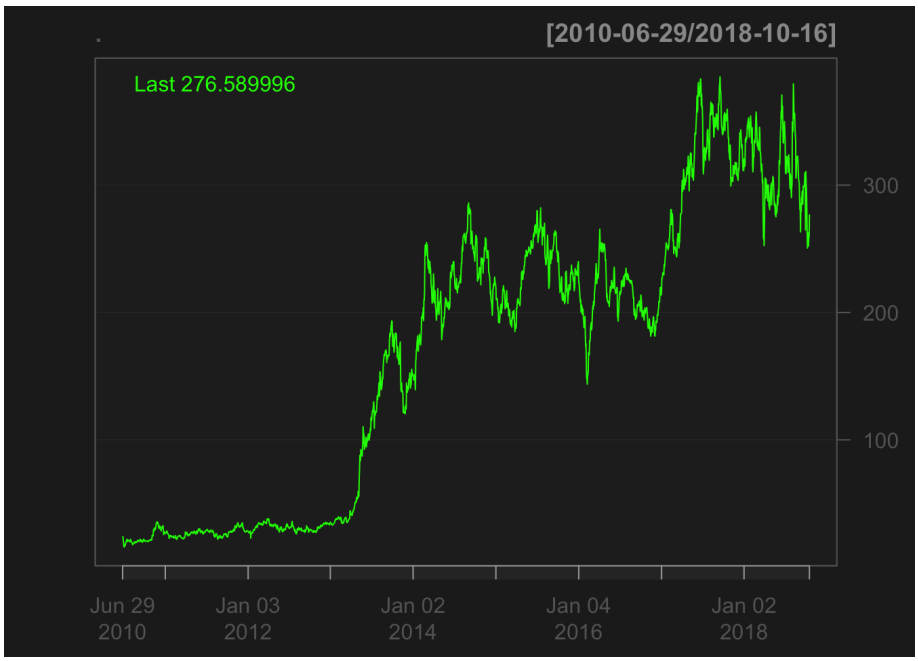
```
str(TSLA)
```

```
## An 'xts' object on 2010-06-29/2018-10-16 containing:
##   Data: num [1:2091, 1:6] 19 25.8 25 23 20 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : NULL
##   ..$ : chr [1:6] "TSLA.Open" "TSLA.High" "TSLA.Low" "TSLA.Close" ...
##   Indexed by objects of class: [Date] TZ: UTC
##   xts Attributes:
## List of 2
##  $ src    : chr "yahoo"
##  $ updated: POSIXct[1:1], format: "2018-11-27 15:04:01"
```

Graphs - Part 1

The graph below shows us the Market Price of TSLA stock price from 1/1/17 to 17/10/18. As we can see, TSLA has seen a steady increase in price, and it's trend can be categorised as "Bullish".

```
TSLA %>% Ad() %>% chartSeries()
```



The graph below shows us the trend in Market Price of TSLA, along with other Technical Indicators such as Bollinger Bands, Volume Graphs and Moving Averages.

```
TSLA %>% chartSeries(TA = "addBBands();
              addBBands(draw=\"p\");
              addVo();
              addMACD()",
    subset = "2016", theme = "white")
```

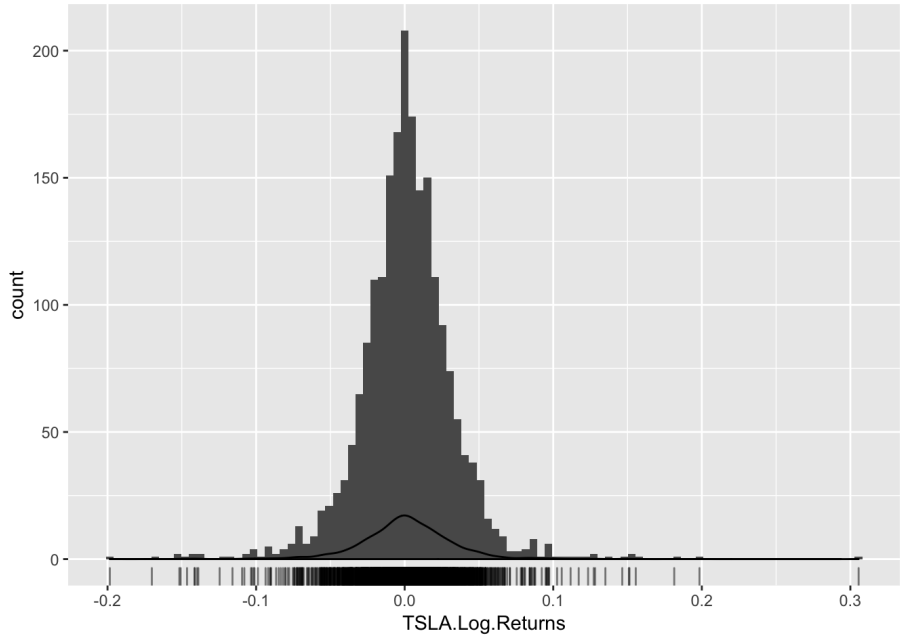
We will now use the TSLA Market price data from 1/1/17 to 17/10/18, in order to find out the Percentage return of the stock in it's natural logarithm.

```
TSLA_log_returns <- TSLA$TSLA.Open %>% dailyReturn(type = "log")
names(TSLA_log_returns) <- "TSLA.Log.Returns"
```

Graphs - Part 2

I have used the ggplot function in order to create a histogram, with 100 bins. From the graph below, we can tell that the Percentage log return of TSLA stock is crowded around 0%, however further analysis needs to be done to find out the potential return.

```
TSLA_log_returns %>% ggplot(aes(x = TSLA.Log.Returns)) +
    geom_histogram(bins = 100) + geom_density() + geom_rug(alpha = 0.5)
```



Over here, I decided to create a quantile in order to represent the log return in another format, which would be more conducive to infeering the information.

From the Quantiles below we find the probability of the Log return at Each Quantile,

```
probs <- c(0.25, 0.5, 0.75)
dist_log_returns <- TSLA_log_returns %>% quantile(probs = probs,
    na.rm = TRUE)
dist_log_returns
```

```
##          25%          50%          75%
## -0.015205373  0.001032562  0.017706482
```

Analysis

Using the Log Market price data from above, we can calculate the mean and standard devation of return.

```
mean_log_returns <- mean(TSLA_log_returns, na.rm = TRUE)
sd_log_returns <- sd(TSLA_log_returns, na.rm = TRUE)
```

Predicting the stock prices for 1000 days:

```
# Price on 27th November, 2018 - Predicting for
# 27th November, 2036

price <- 345.34

set.seed(50)
for (i in 2:length(TSLA$TSLA.Open)) {
    price[i] <- price[i - 1] * exp(rnorm(1, mean_log_returns,
        sd_log_returns))
}
random_data <- cbind(price, 1:length(TSLA$TSLA.Open))
colnames(random_data) <- c("Price", "Day")
random_data <- as.data.frame(random_data)
random_data %>% ggplot(aes(Day, Price)) + geom_line() +
    labs(title = "Tesla (TSLA) price simulation for 8 years") +
    theme_bw()
```

Inference and Limitations of Model

As we can see from the graph above, my model predcits that the Market Price of TESLA will follow a “Bullish” trend and will have a value in between $3750 and $5000. This expectional growth rate of 1500% seems too good to be true. However, if we look at the trend in Prices of

TESLA since it’s Initial Public Offering, it has increased from $17 to $345.34. This is an increase in percentage of nearly 2000%.

Hence, my model predicts that TSLA is currently undervalued based on it’s expected growth rate in the next 4 years, and it therefore confirms the hypothesis that investing in TSLA is a good decision.

The limitations of my model are that:

1. As you can see, I have set seed to 50 while simulating the prices.
2. This model only takes into account the quantitive technical factors that would affect the stock price. This is not always true since qualitative factors can greatly affect the performance of the company, which may not adhere to the mean and standard deviation in it’s price historically.