DSO499_Lab5

Jose Canela

10/9/2019

Multiple Linear Regression, ggplot2, and Real Estate

We begin by uploading the NYC condo evaluations data for fiscal year 2011-2012 and looking at the variables that were evaluated.

```
library(ggplot2)
housing = read.csv("housing.csv", stringsAsFactors=F)
names(housing)
```

```
##  [1] "Neighborhood"            "Building.Classification"
##  [3] "Total.Units"             "Year.Built"
##  [5] "Gross.SqFt"              "Estimated.Gross.Income"
##  [7] "Gross.Income.per.SqFt"   "Estimated.Expense"
##  [9] "Expense.per.SqFt"        "Net.Operating.Income"
## [11] "Full.Market.Value"       "Market.Value.per.SqFt"
## [13] "Boro"
```

Next, we rename the column names of the dataset to make analysis easier.

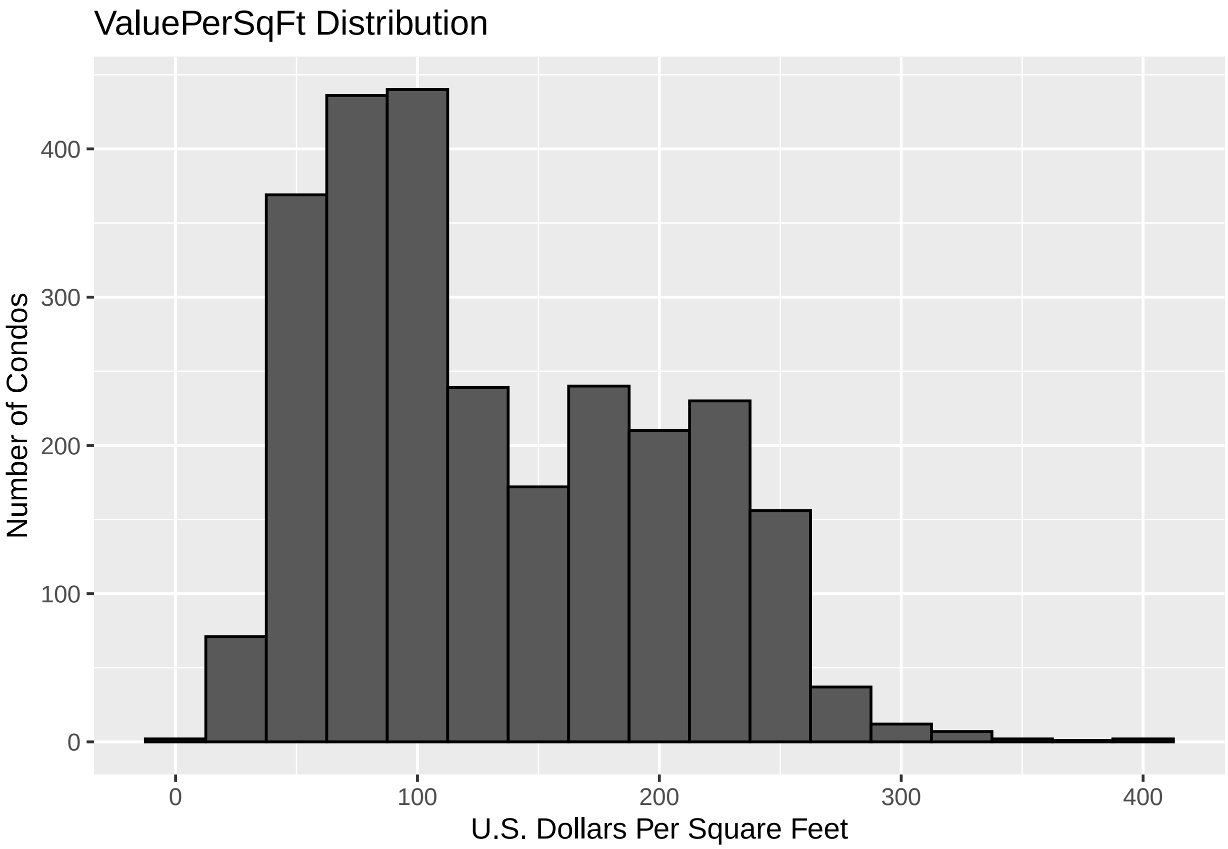
```
names(housing)=c("Neighborhood","Class","Units","YearBuilt","SqFt","Income","IncomePerSqFt","Expense","

names(housing)
```

```
##  [1] "Neighborhood"   "Class"          "Units"          "YearBuilt"
##  [5] "SqFt"           "Income"         "IncomePerSqFt"  "Expense"
##  [9] "ExpensePerSqFt" "NetIncome"      "Value"          "ValuePerSqFt"
## [13] "Boro"
```

Observing The Distribution Of ValuePerSqFt

1. Create a histogram of ValuePerSqFt using ggplot. Use the labs layer to properly label x and y axes. Use the binwidth option to select an appropriate binwidth. Comment on your findings.

```
ggplot(housing, aes(x = ValuePerSqFt)) +
  geom_histogram(binwidth = 25, col = "black") +
  ylab("Number of Condos") +
  xlab("U.S. Dollars Per Square Feet") +
  labs(title = "ValuePerSqFt Distribution")
```



The distribution of the condo value per square feet is somewhat right skewed. Hence, from the histogram we can see that the median number of U.S. dollars per square feet is approximately 112.50 U.S. dollars per square foot and the mean number of U.S. dollars per square feet that is probably approximately 130.50 U.S. dollars per square foot. In fact, we can verify whether these values are close to the actual median and means:

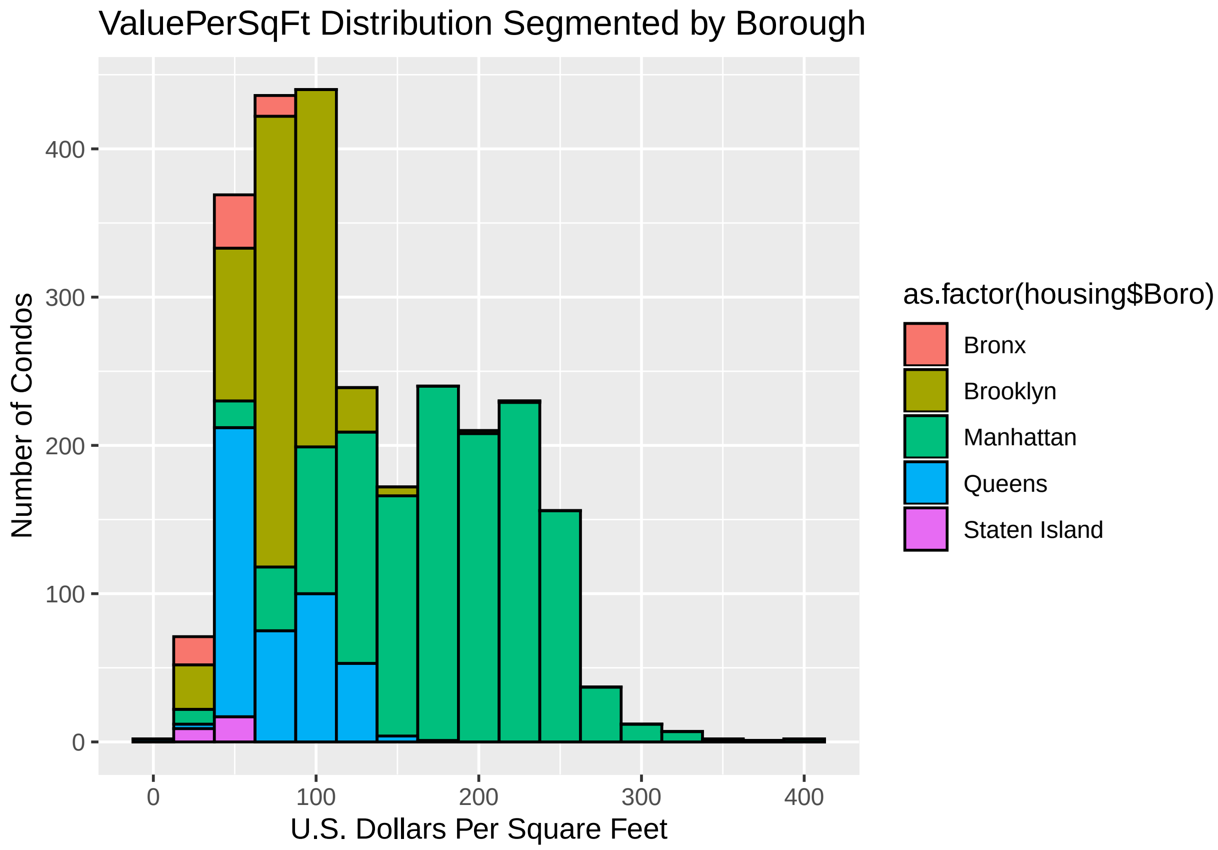
```
## [1] "Median Value per Square Foot: $112.22"
```

```
## [1] "Mean Value per Square Foot: $131.187204874334"
```

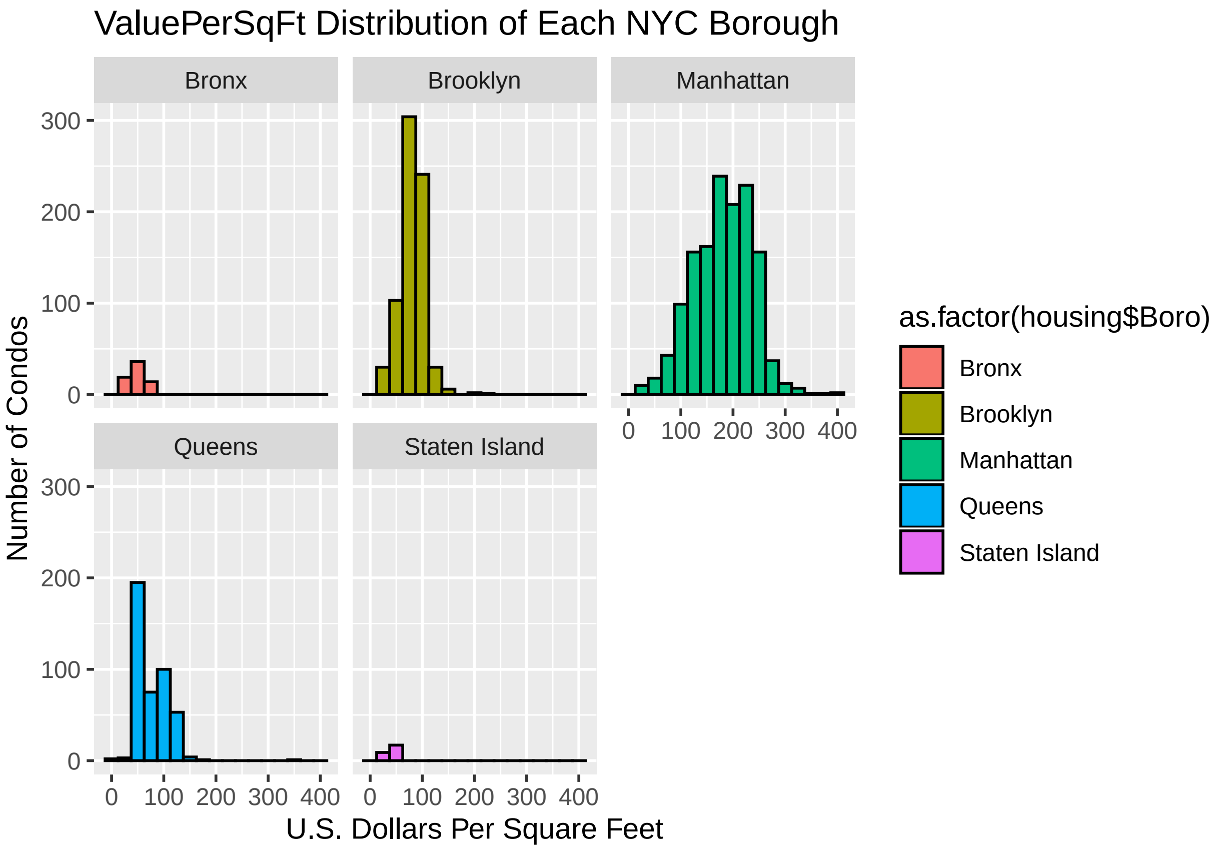
Thus, we can see that the observed values were very close!

Histograms That Take Location Into Account

2. First, map color to Boro using fill option. Then create the multiple histograms faceting on Boro and keeping the color coding on the location. Comment on your findings.



When we segment the overall value per square feet distribution by borough, we can see that the vast majority of the condos with high value per square footage are located in Manhattan. In this case, most of them have a value per square foot above the median and mean. However, if you live outside of Manhattan, the value per square footage of our condo will most likely be below the median and mean. If one were to rank the boroughs by value per square footage most of their condos have, the ranking in descending order would be: Manhattan, Brooklyn, Queens, Staten Island, Bronx.

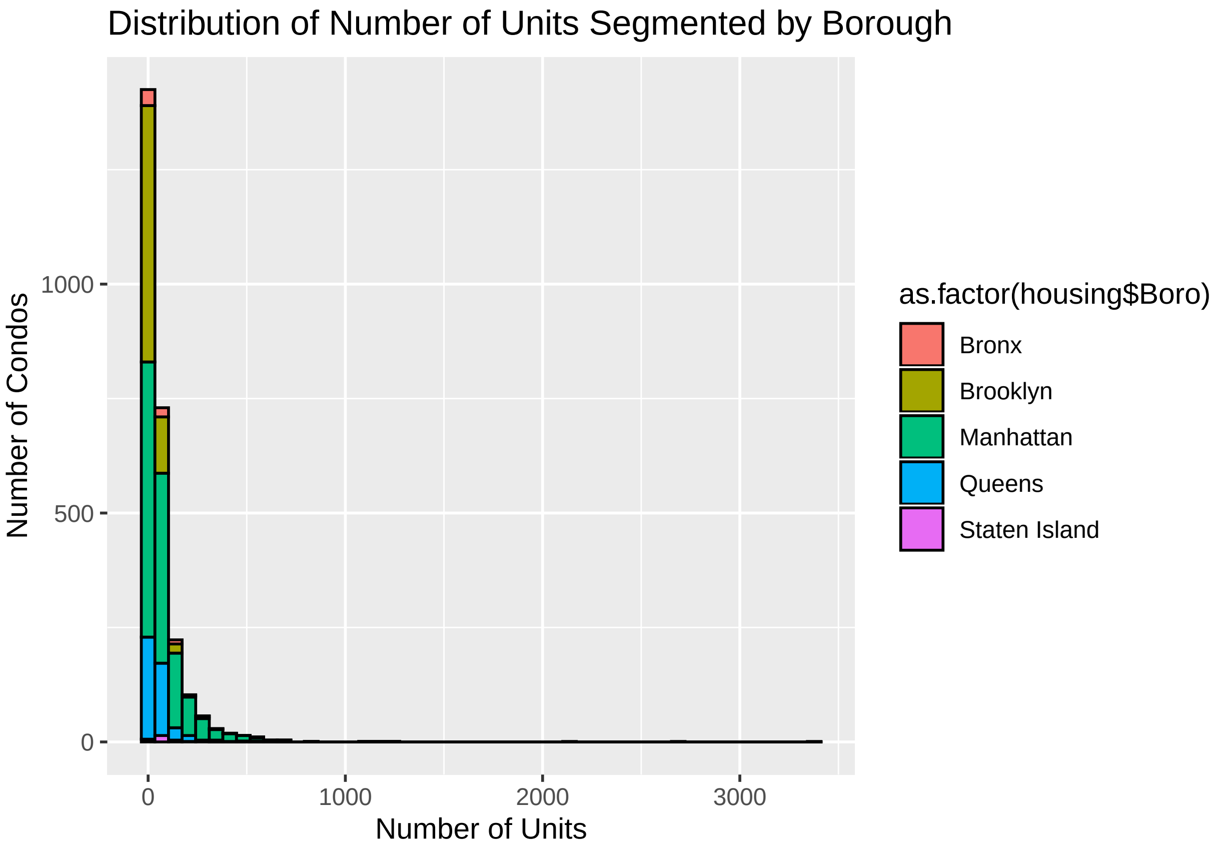


When we observe the value per square feet distribution of each individual borough, we notice that the distributions do not match the distribution of value per square foot when one combines all of the boroughs in one's distribution analysis. Overall, we can predict that Manhattan will hold most of the variability. In addition, we can see that the reason why the valuePerSqFt distribution was right-skewed was due to the most of the condos in Manhattan having a balue per square foot above the mean and median of value per square footage for the entire dataset. The distribtuions of all the non-Manhattan boroughs are basically right-skewed whereas the Manhattan value per square footage distribtuion is slightly left-skewed. In addition, when looking at the individual distributions for each of the non-Monhattan boroughs, we can see that almost all of the condos in each histogram are less than the mean and median of the value per square foot of the entire dataset.

Observing The Distributions Of Square Footage And The Number Of Units

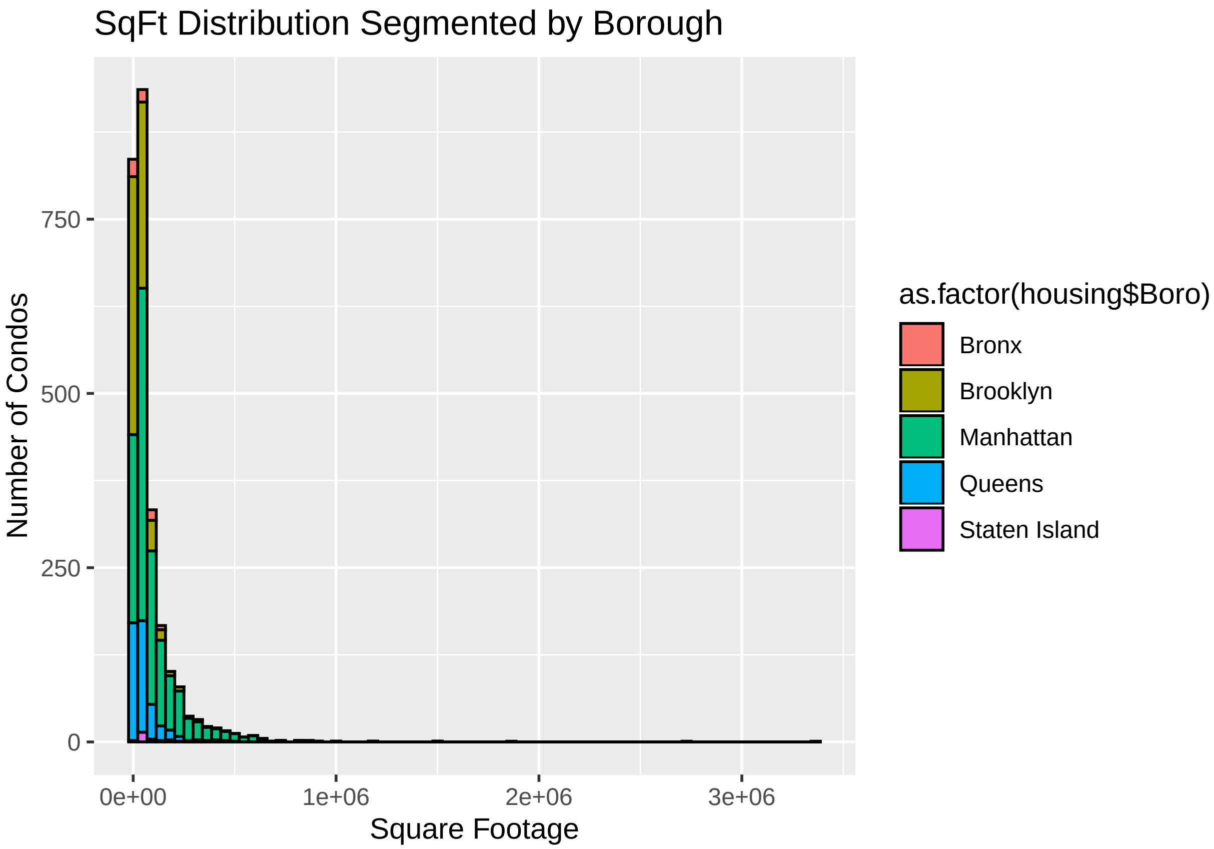
3. Next create histograms for square footage and the number of units. Comment on the distributions.

```
ggplot(housing, aes(x = Units)) +
geom_histogram(bins = 50, aes(fill = as.factor(housing$Boro)), col = "black") +
ylab("Number of Condos") +
xlab("Number of Units") +
labs(title = "Distribution of Number of Units Segmented by Borough")
```



The distribution of the number of units is heavily right-skewed. The median number of units in New York City condos was 30. The mean number of units in New York City condos was 70.1839299. When we color segment the distribution by borough, we see that condos with larger number of units are generally located in Manhattan. In addition, it looks like the condos with the lowest number of units are generally located in Manhattan as well.

```
ggplot(housing, aes(x = SqFt)) +
geom_histogram(bins = 75, aes(fill = as.factor(housing$Boro)), col = "black") +
ylab("Number of Condos") +
xlab("Square Footage") +
labs(title = "SqFt Distribution Segmented by Borough")
```

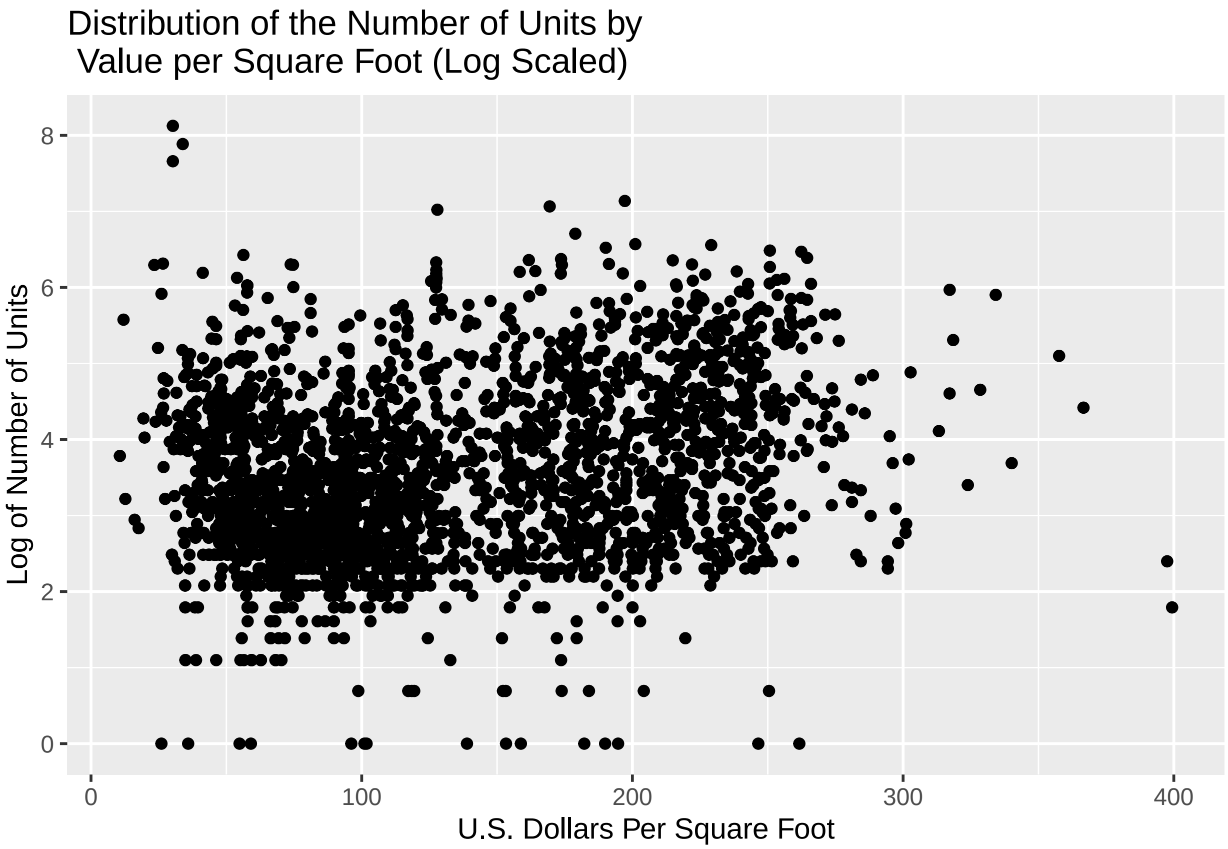


The distribution of condo square footage is heavily right-skewed. The median square footage of New York City condos was 30. The mean square footage of New York City condos was 70.1839299. When we color segment the distribution by borough, we see that condos with larger square footage are generally located in Manhattan. Condos with lower square footage seem to generally be located in either Brooklyn or Manhattan.

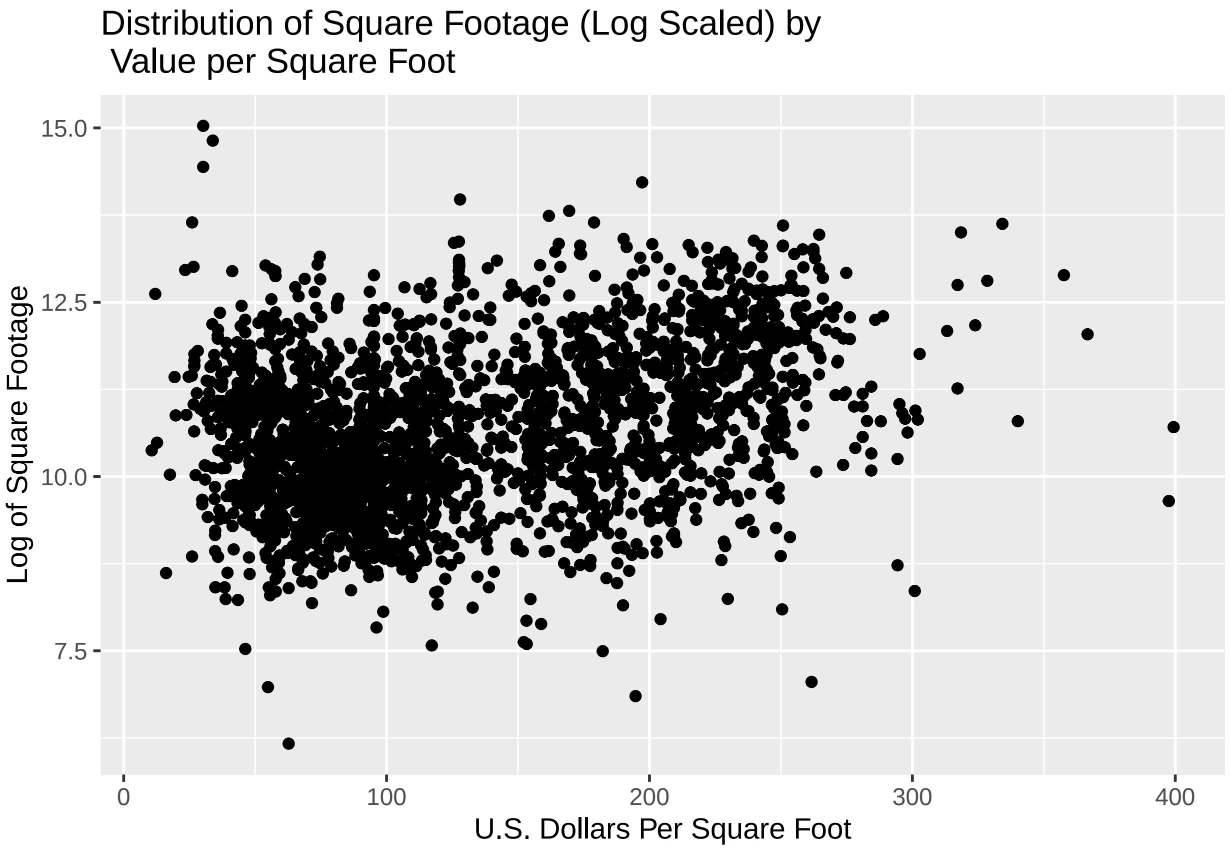
Observing The Respective Distributions Of The Number Of Units And Square Footage BY Value Per Square Foot

4. Plot a scatterplots of the value per square foot versus log of number of units and log of square footage.

```
ggplot(housing, aes(x = ValuePerSqFt, y = log(Units))) +
geom_point() +
xlab("U.S. Dollars Per Square Foot") +
ylab("Log of Number of Units") +
labs(title = "Distribution of the Number of Units by\n Value per Square Foot (Log Scaled)")
```



```
ggplot(housing, aes(x = ValuePerSqFt, y = log(SqFt))) +
geom_point() +
xlab("U.S. Dollars Per Square Foot") +
ylab("Log of Square Footage") +
labs(title = "Distribution of Square Footage (Log Scaled) by\n Value per Square Foot")
```



Filtering Through The Dataset Due To The Large Number of Units For Some Building

5. You should notice that there are quite a few buildings with an incredible number of units. How many buildings have at least 1000 units?

```
nrow(subset(housing, Units>=1000))
```

```
## [1] 6
```

There are 6 condos that have at least 1000 units. Hence, we will remove these condos from our dataset and focus on predictive modeling.

```
housing_lt1000 = subset(housing, Units < 1000)
```

Multiple Linear Regression

6. Build a multiple linear regression model to predict the property value per sq ft. We already saw that accounting for different boroughs will be an important and the various scatterplots indicated that Units and SqFt will be important as well. First, Fit the model using the formula interface in the *lm* function and use I(log(Units)) and I(log(SqFt)) sintax for the log transformed inputs. Next, use the *summary* command to analyze the model.

```
?lm
linear = lm(ValuePerSqFt ~ I(log(Units)) + I(log(SqFt)) + Boro, housing_lt1000)
summary(linear)
```

```
##
```

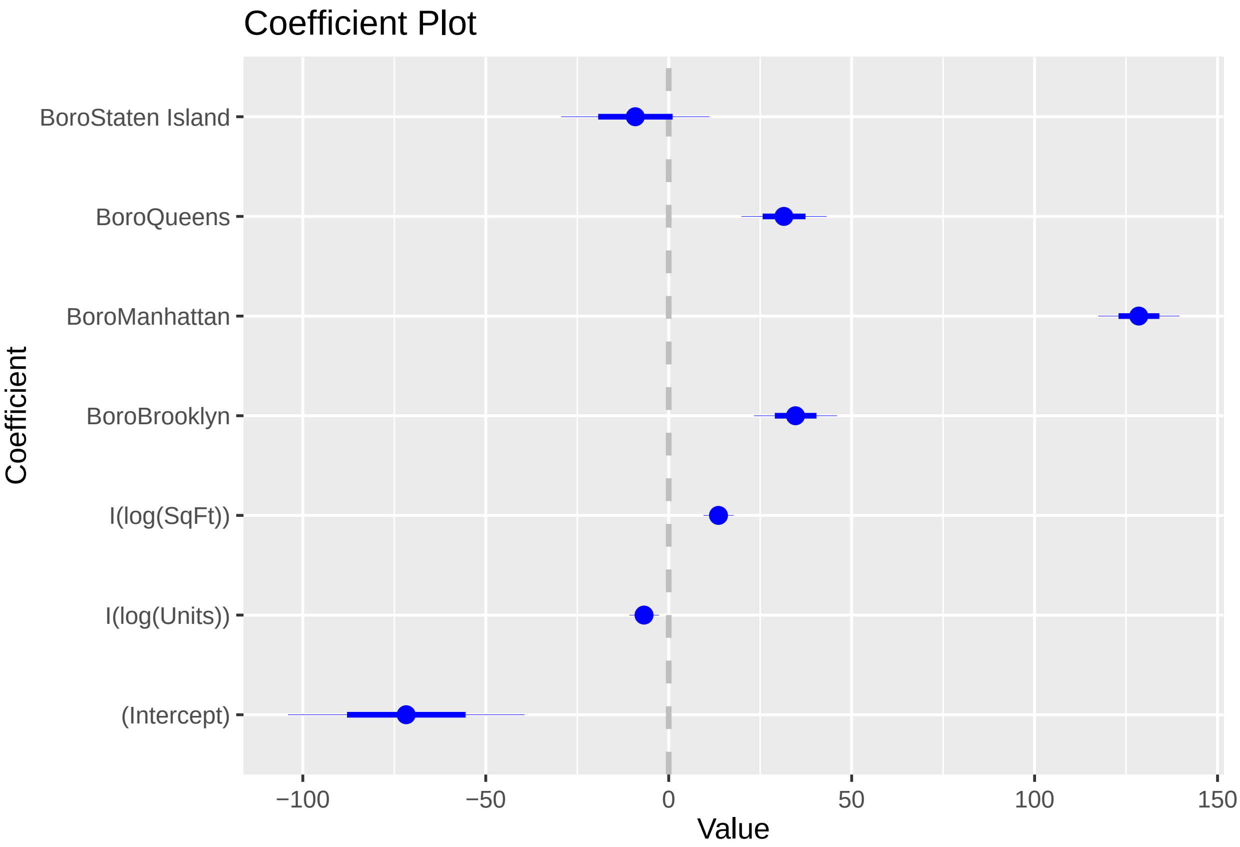
```
## Call:
## lm(formula = ValuePerSqFt ~ I(log(Units)) + I(log(SqFt)) + Boro,
##     data = housing_lt1000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.972  -24.215    1.291   28.285  258.432
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -71.749     16.140  -4.446 9.13e-06 ***
## I(log(Units))       -6.745      1.981  -3.405 0.000671 ***
## I(log(SqFt))        13.605      2.015   6.752 1.79e-11 ***
## BoroBrooklyn        34.652      5.625   6.160 8.40e-10 ***
## BoroManhattan      128.471      5.504  23.343  < 2e-16 ***
## BoroQueens          31.484      5.773   5.454 5.39e-08 ***
## BoroStaten Island   -9.133     10.103  -0.904 0.366061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.58 on 2613 degrees of freedom
## Multiple R-squared:  0.5965, Adjusted R-squared:  0.5955
## F-statistic: 643.7 on 6 and 2613 DF,  p-value: < 2.2e-16
```

The response variable (i.e. dependent variable) in this multiple linear regression model is value per square foot ((in U.S. dollars) of a condo and the predictors (i.e. independent variables) are the log of the number of units in a condo, the log of a condo's square footage, and the borough the condo is located in. After summarizing/analyzing the model, we can see that a condo being located in Staten island is statistically insignifcant when predcting a condo's value per quare foot. In general, the log of the number of units of a condo, the log of a condo's square footage, a condo being located in Manhattan, a condo being located in Brookyln, and a condo being located in Queens all had large statistical significant. A condo being located in Manhattan had the largest statistical signifcance when it came to predicting a condo's value per square footage with a p-value of less than p-value: < 2.2e-16.

R-squared represents the proportion of the variance for a dependent variable that is explained by the independent variable(s) in a regression model. Adjusted R-squared adjusts for the number of terms in a model. Hence, if you keep on adding statistically insignificant variables to a model, adjusted R-squared will decrease. Our adjusted R-squared was 0.5955. This is not bad given that we only used 6 independent variables. Our residual standard error was $43.58.

Visualizing Model Information

7. Use the *coefplot(your.model)* function to visualise the coefficients. Interpret your results.



From the coefficient plot, we can state that according to the results of the model:

- Keeping the other independent variables constant, every increase of one for the log of the number of units, the value per square foot of a condo decreases by $6.745 per square foot.
- Keeping the other independent variables constant, every increase of one for the log of the square footage of a condo, the value per square foot of the condo increases by $13.605 per square foot.
- Keeping the other independent variables constant, if a condo is located in Staten Island, the value per square foot of the condo decreases by $9.133 per square foot. However, it is important to note that this independent variable was not statistically significant when it came to predicting a condo's value per square foot.
- Keeping the other independent variables constant, if a condo is located in Brooklyn, the value per square foot of the condo increases by $34.652 per square foot.
- Keeping the other independent variables constant, if a condo is located in Queens, the value per square foot of the condo increases by $31.484 per square foot.
- Keeping the other independent variables constant, if a condo is located in Manhattan, the value per square foot of the condo increases by $128.471 per square foot.
- The intercept exists due to bias in the model. In this case, the interpretability of the intercept is not clear at all. In general, it would be that if a condo doesn't exist at all (due to the fact that the log of square footage is equal to zero, and every other independent variable in the model is equal to zero), the value per square foot of a condo is -$71.749 per square foot. There is no meaningful way to interpet this. Instead, it seems that the intercept coeffcient is just used to take care of the bias of the linear regression model.

Predicting ValuePerSqFt Using A New Data Set

To show that regression can be used to predict values, we will first upload new condo evaluations data that is available at http://www.jaredlander.com/data/housingNew.csv and create a new dataset called *housingNew*.

```
housingNew=read.table("http://www.jaredlander.com/data/housingNew.csv",sep=",",header=T,stringsAsFactor
```

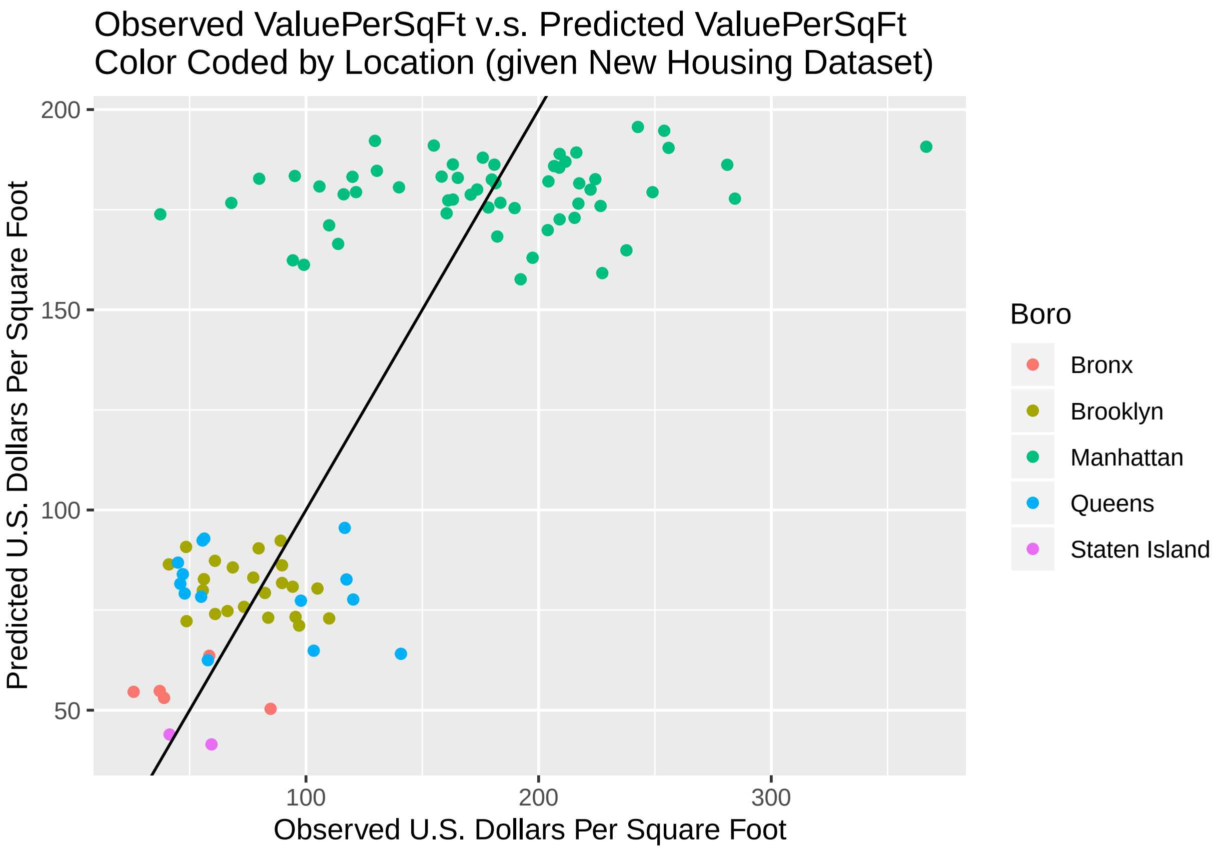
After uploading the new data, we then make a prediction with the new data with 95% confidence bounds:

```
housePredict=predict(linear,newdata=housingNew, sefit=T,interval="prediction", level=0.95)
```

Visualizing Observed Values v.s. Predicted Values

8. Use ggplot function to visualize observed values vs. predicted. Color code by the location, add a y=x line using the *geom_abline* function. Comment on your findings.

```
# head(housePredict)
ggplot(housingNew, aes(x = housingNew$ValuePerSqFt, y = housePredict[,1])) +
geom_point(aes(col = Boro)) +
geom_abline() +
xlab("Observed U.S. Dollars Per Square Foot") +
ylab("Predicted U.S. Dollars Per Square Foot")+
labs(title = "Observed ValuePerSqFt v.s. Predicted ValuePerSqFt\nColor Coded by Location (given New Hou
```



From the plot we can see that our model seems to predict the value per square footage for condos in the non-Manhattan boroughs pretty well. However, the model has trouble predicting the value per square footage for condos in Manhattan. There is a lot of variation still not being captured for the value per square footage of Manhattan condos. Thus, it looks like the there isn't a constant variance across boroughs when predicting the market value per square footage of condos. In other words, there seems to be some heterskedasticity present. Hence, an alternative option would be to create two models: one for predicting the value per square footage for condos in Manhattan and another model dedicated to the condos in the other four boroughs in NYC.