MSDS 6372 PROJECT 1

Team Members:

Adesanya Olufemi

Anderwald Scott

Scott Marvin

Date: 09/26/2016

**Factors of Value per Square Foot for Boroughs**

**1. Introduction**

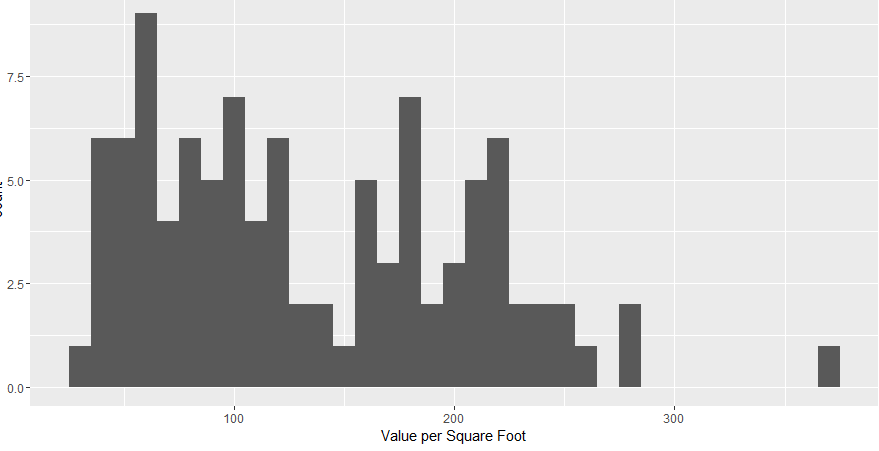
Our group decided to work on the New York City condo dataset for fiscal year 2011-2012, obtained through NYC Open Data. NYC Open Data is an initiative by New York City to make government more transparent and work better. It provides data on all manner of city services to the public. The dataset we procured include 100 observations and 13 variables. The goal of this project is to make predictions or inferences on the response using multiple predictors; how the explanatory variables affects the ValuePerSqFt variable through multiple linear regression.

Our dataset includes multiple explanatory variables like Neighborhood, Class, Unit, YearBuilt, Sqft, Income, IncomePerSqFt, Expense, ExpensePerSqFt, NetIncome, Value, ValuePerSqFt, and Boro.

**2. Exploratory Data Analysis**

For this data, the response is the ValuePerSqFt variable and the predictors are the remaining variables. However, we ignore the income and expense variables, as they are actually just estimates based on requirement that condos be compared to rentals for valuation purposes. The first step is to perform an exploratory data analysis with a histogram of ValuePerSqFt, which is shown in Figure 1.0 below

Figure 1.0 Histogram of value per square foot for NYC condos. It appears to be bimodal.



Based on the bimodal nature of our histogram in fig 1.0, we further explored by color mapping to Boro Fig 1.1 and faceting on Boro in Fig 1.2 reveal that Brooklyn and Queens make up one mode and Manhattan makes up the other, while there is a lack of data for both Bronx and Staten Island.

|  |  |
| --- | --- |
| Fig 1.1 Boro seperated by color | Fig 1.2 Boro seperated by color and facet |

**Fig 1.1 and Fig 1.2** above shows Histograms of value per square foot. These illustrate structures in the data revealing that Brooklyn and Queens make up one mode and Manhattan makes up the other, while there is not much data on the Bronx and Staten Island.

Next we looked at histograms for total square feet and the number of units. The distribution is highly right skewed.

|  |  |
| --- | --- |
|  |  |

**Fig 1.4** The distributions are highly right skewed.

Next we looked at the plot for ValuePerSqFt and SqFt, as we can see in the scatter plots in Fig 1.5 below, we have the scatterplot of value per square foot versus square footage and value versus number of units. Even after the removal of the outliers, it still seems like a log transformation was needed due to the right skewedness of the data.

Fig 1.5

|  |  |
| --- | --- |
|  |  |
|  |  |

Next, a log transformation is added to the data so a fit can be made to the potential model. As we can see in Fig 1.6 below, we plotted ValuePerSqFt against SqFt and we took the log of values. The plots indicated that taking the log of SqFt would be beneficial to the model.

|  |  |
| --- | --- |
|  |  |

**Fig 1.6 Scatterplot of ValuePerSqFt against SqFt**

**3. Regression and Coefficient Equation**

We viewed and analyzed our data in different ways from the above plots, based on our analysis, accounting for different boroughs is important and the various plots indicated that Units and SqFt will be important as well. Using a multiple regression equation. The logical extension of linear regression is multiple regression which allows for multiple predictors

Equation: Y =Xβ +ε

Y= the response vector

X = is the matrix (n rows and p n-1 predictors plus the intercept)

β = is the vector of coefficients (one for each predictor and intercept)

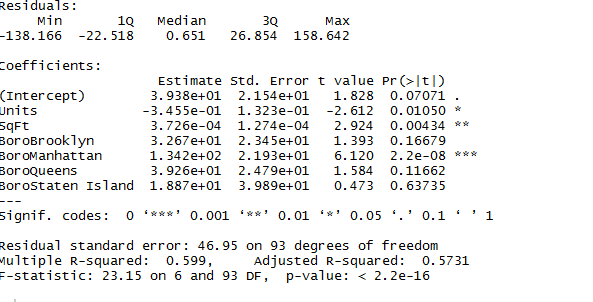
ε = is the vector of normally distributed errors

The solution for the coefficients is: β = (XtX)¯¹ XtY

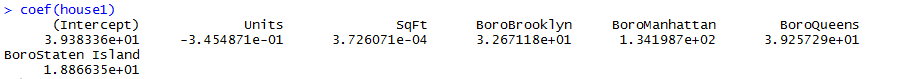
**4. Fitting The Model**

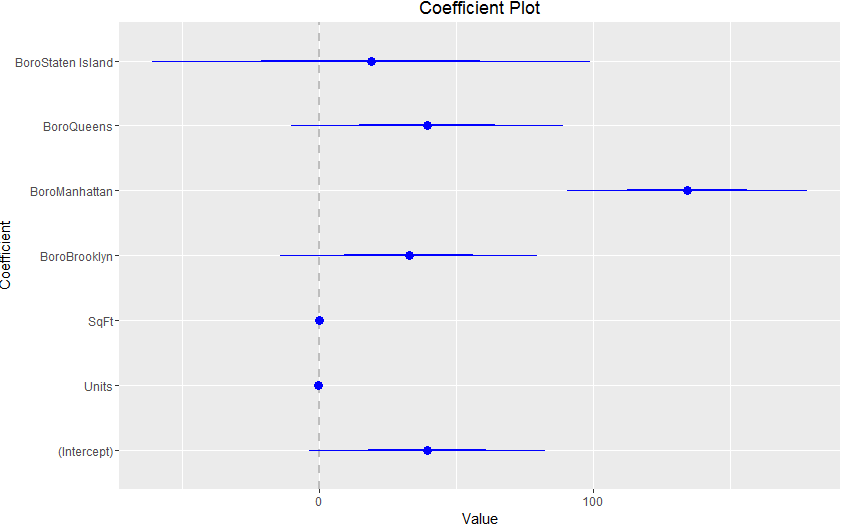
We then fit the model using the formula interface in lm. Since there are multiple predictors, we separated them on the right side of the formula using a plus sign (+). Fig 1.7 below shows the output result based on our model, as we can see, there was no coefficient for the bronx because that is the baseline level for Boro, and all other Boro coefficients are relative to that baseline.

Fig 1.7



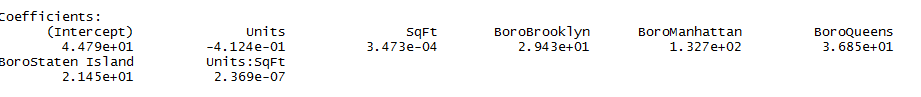
**Model 1**

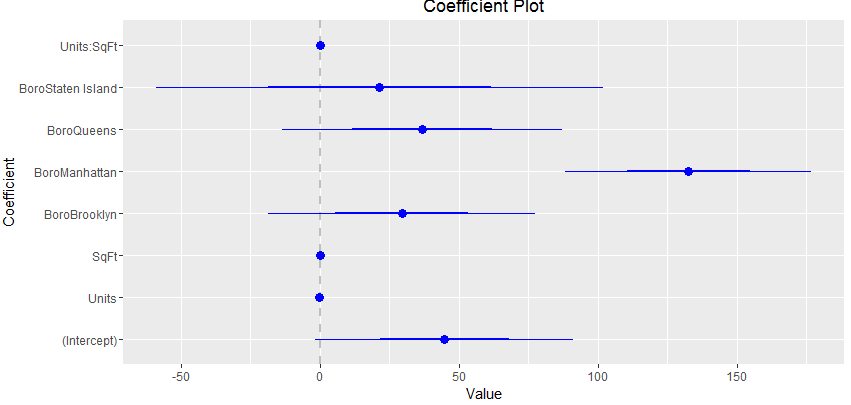
Since the coefficient represents the effect of the predictors on the response, we plotted a coefficient plot for our model 1, Fig 1.7, the plot showed us being located in Manhattan has the largest effect on value per square foot, but the number of units or square feet in a building has little effect on value.



**Model 2**

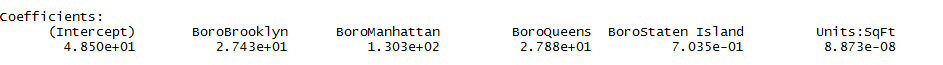
In this model, we added an interaction term to the variable.

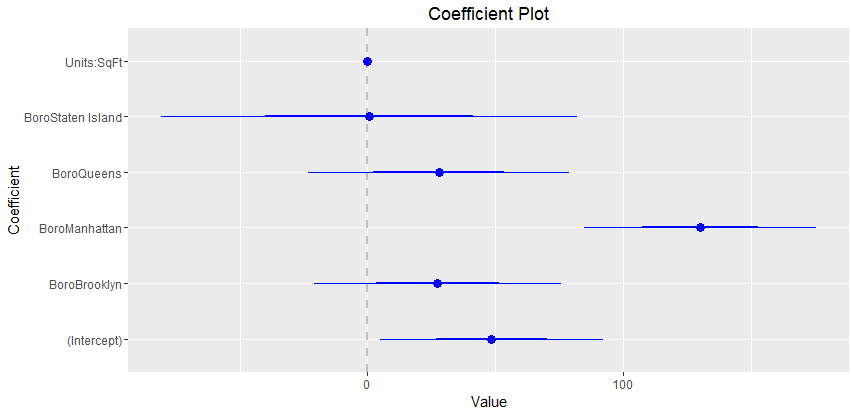




**Model 3**

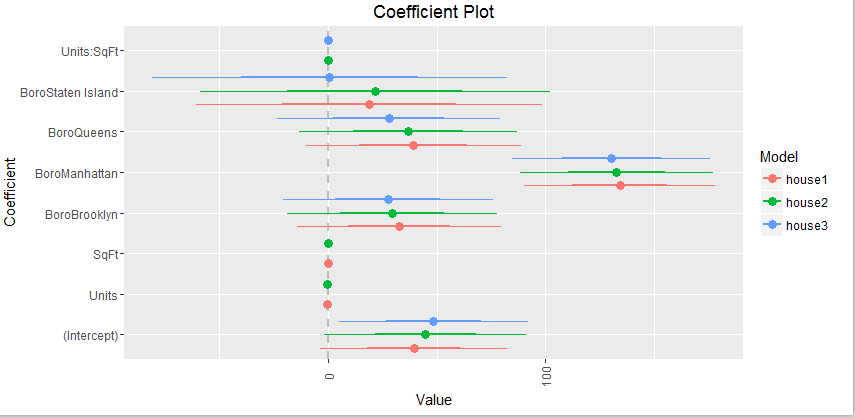
For the model 3, we plotted just the interaction term against the Boros as seen in Fig 1.8 below





**Coefficient Plot for the 3 models**

In Fig 1.9 below, we plotted a coefficient plot for multiple condo models. The coefficients are plotted in the same spot on the y-axis for each model.



**5. Regression Diagnostic**

**Analysis of Residuals**

To better improve the quality of our model, we went about analyzing the residuals, which is the difference between the actual response and the fitted values. The basic idea is that if the model is appropriately fitted to the data, the residual should be normally distributed.

Fit Model 1 – Looking at Fig 1.10 below, we did a plot of residuals versus fitted values and the pattern shows the residuals are not as randomly dispersed as desired, however we found out that it was due to the structure that Boro gives the data as seen in the 2nd plot to the right below, we then plotted the plot colored with Boro as seen in the 2nd plot to the left, the pattern in the residual is revealed to be the result of the effect of Boro on the model. To further find out if our Model 1 is the best, we plotted a Basic QQ plot and ggplot as seen in the last row in the Fig 1.10 below, the tails drift away from the line, indicating that it’s not the best fit.

|  |  |
| --- | --- |
|  |  |
|  |  |

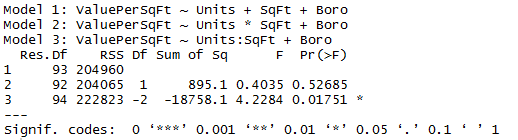
Fig 1.10

|  |  |
| --- | --- |
|  |  |

**Model 1, Model 2 and Model 3 comparison**

We made use of ANOVA procedure to compare all our three models because we do believe it serves the purpose in testing all relative merits of different models, and by using ANOVA, it will return a table of results including the residual sum of squares, which is a measure of error, the lower the better. Fig 1.11 shows that our Model 2 has the lowest RSS meaning it is the best model, but we know that RSS always improve when an additional variable is added to the model, and this can lead to excessive model complexity and over fitting. Based on this knowledge, we decided to use the AIC and BIC procedure to find the best model.

Fig 1.11.



Looking at the output in Fig 1.12, both the AIC and the BIC shows us that Model 1 is the best based on the lowest number in both methods.

Fig 1.12

|  |  |
| --- | --- |
| **AIC Method** | **BIC Method** |
|  |  |
|  |  |

**Cross Validation Method**

We decided to use one more method to know which is our best model by using cross validation method also called k-fold cross validation. For the cross validation, we used a glm function, and a comparison was done with the lm function we used earlier to confirm they both have the same output.

Fig 1.13

|  |  |
| --- | --- |
| Cross Validation (GLM) | **Comparison between glm and lm to ensure we get the same out** |
|  |  |

Looking at Fig 1.14 below, we also did a cross validation for all the 3 models, and Model 1 has the lowest AIC

|  |  |
| --- | --- |
| Model 2 | Model 3 |
|  |  |

**Conclusion**

We made use of 4 different methods to find our best model to fit our data, they include ANOVA, AIC, BIC, and Cross Validation. Except for the ANOVA method, all other methods indicated that our Model 2 best fit our data.

Appendix

housing <- read.csv("housingNew.csv") #code to bring in housing dataset#

summary(housing)

ggplot2(housing, aes(x=ValuePerSqFt)) +

geom\_histogram (binwidth=10) + labs(x="Value per Square Foot")

library(ggplot)

install.packages("ggplot")

install.packages('ggplot', repos='http://cran.rstudio.com', type='source')

#Model 1#

house1 <- lm(ValuePerSqFt ~ Units + SqFt + Boro, data = housing)

summary(house1)

library(coefplot)

coefplot(house1)

coef(house1)

install.packages("coefplot")

#Model 2 with Interacion and individual variables#

house2 <- lm(ValuePerSqFt ~ Units \* SqFt + Boro, data = housing)

coefplot(house2)

#Model 3 with Interaction without individual variables#

house3 <- lm(ValuePerSqFt ~ Units: SqFt + Boro, data = housing)

coefplot(house3)

#coefficient plot for all 3 models#

multiplot(house1, house2, house3)

#predict#

housepredict <- predict(house1, newdata =housingNew, se.fit =TRU, interval ="prediction", level =.95)

#Model 1 Diagnostics#

house1 <- lm(ValuePerSqFt ~ Units + SqFt + Boro, data = housing)

summary(house1)

coefplot(house1)

head(fortify(house1))

h1 <- ggplot(aes(x=.fitted, y =.resid),data= house1) + geom\_point() +

geom\_hline(yintercept = 0) +

geom\_smooth(se = FALSE)+

labs(x ="Fitted Values", y ="Residuals")

h1 + geom\_point(aes(color =Boro))

plot(house1, which =1)

plot(house1, which =2)

ggplot(house1, aes(sample =.stdresid)) + stat\_qq() + geom\_abline()

#Model1, 2 and 3 comparison#

anova(house1, house2, house3)

AIC(house1, house2, house3)

BIC(house1, house2, house3)

require(boot)

houseG1 <- glm(ValuePerSqFt ~ Units + SqFt + Boro, data = housing, family = gaussian(link ="identity"))#REFIT LM WITH GLM#

#to compare the results with lm

identical(coef(house1),coef(houseG1))

#crossvalidation of all model#

houseG2 <- glm(ValuePerSqFt ~ Units \* SqFt + Boro, data = housing)

houseG3 <- glm(ValuePerSqFt ~ Units + SqFt \* Boro + Class, data = housing)