Evaluation of Housing Datasets

I choose to analyze the Excel file – "rollingsales_queens.xls", because, compared with other two files, the "Neighborhood" has no missing values and the "Sale Price" has not too much zero. The information provided in Queens is good enough to do some insights by using models.

After drawing the distribution of "Sale Price", "Gross Square Feet" and "Land Square Feet", I find their trends are highly right skewed. In order to normalize the target variable, I will use log10 function to transform them. Besides, the zero in "Sale Price", "Gross Square Feet" and "Land Square Feet" is meaningless. I conduct a new dataset which leaves out those zeros.

The right graph in Figure 1. Shows the relationship between log transformation in "Gross Square Feet" and "Sale Price" in new dataset. We can find that the adjusted space is concentrated between 6 and 10, and the adjusted sale price is mainly in 12 and 15. The general trend between Sale Price and Space is positive, except some outliers.

I also want to find out what kind of relationship between Sale Price and "Building Class Category" or "Neighborhood". Based on figure 2., I choose five different neighborhoods with three kinds of family homes to analyze the distribution of Sale Price. Interestingly, the median Sale Price is based on the number of family home, the more family homes the higher the price. And the median sale price of each building class in "Bayside" is the highest while in "Arverne" is the lowest, we can guess the living condition and community management may vary in those neighborhoods.

In addition, the year built is also an important factor on Sale Price. Therefore, I create a new dataset which contains three types of family homes and year-built data in 2006, 2008, 2010 and 2012. When the timeline is from 2006-2012, the sale price of one family homes didn't change a lot, however, during the 2010-2012, it increased quite a lot. The median price changed from lower than $500,000 to around $750,000, and the maximum price was higher than the other two types of homes.

```
# plot scatter
plot(queens$`GROSS SQUARE FEET`, queens$`SALE\nPRICE`, xlab="Space", ylab="SalePrice")
abline(lm(queens$`SALE\nPRICE` ~ queens$`GROSS SQUARE FEET`))

newq2 <- queens[which(queens$`GROSS SQUARE FEET`>0 &
                      queens$`LAND SQUARE FEET`>0 & queens$`SALE\nPRICE` >0),]

plot(log(newq2$`GROSS SQUARE FEET`), log(newq2$`SALE\nPRICE`), xlab = "Adjusted Space", ylab="Adjusted SalePrice")
abline(lm(log(newq2$`SALE\nPRICE`) ~ log(newq2$`GROSS SQUARE FEET`)))
```

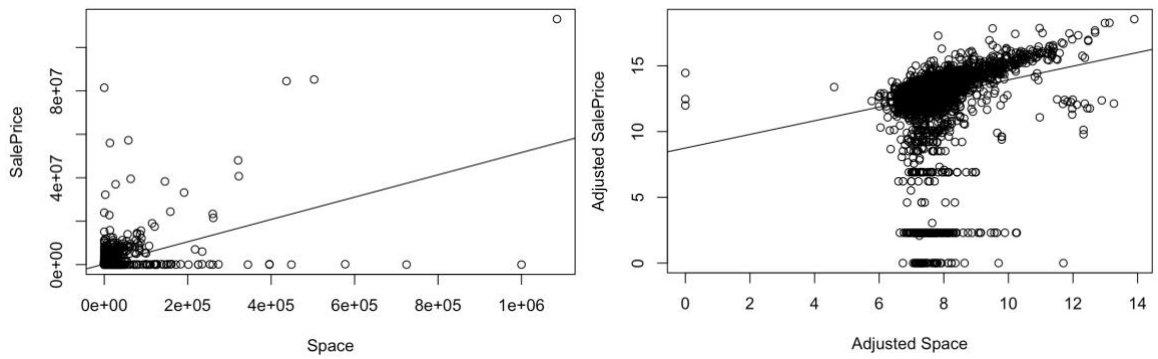

Figure 1. Comparison between raw data and adjusted data

```
# find 1-, 2- and 3- family homes
family <- subset(newq2, (`BUILDING CLASS CATEGORY`=="01  ONE FAMILY HOMES" | `BUILDING CLASS CATEGORY`=="02  TWO FAMILY HOMES"|
                  `BUILDING CLASS CATEGORY`=="03  THREE FAMILY HOMES")&(NEIGHBORHOOD == "ARVERNE"|NEIGHBORHOOD =="ASTORIA"|
                 NEIGHBORHOOD =="BAYSIDE"|NEIGHBORHOOD =="BELLEROSE"|NEIGHBORHOOD =="COLLEGE POINT"))
require(ggplot2)
ggplot(data=family, aes(x=family$NEIGHBORHOOD, y=family$`SALE\nPRICE`))+
  geom_boxplot(aes(fill=family$`BUILDING CLASS CATEGORY`))

#building year:2000-2013
budyear <-subset(newq2, (`BUILDING CLASS CATEGORY`=="01  ONE FAMILY HOMES" | `BUILDING CLASS CATEGORY`=="02  TWO FAMILY HOMES"|
                         `BUILDING CLASS CATEGORY`=="03  THREE FAMILY HOMES")&(`YEAR BUILT` == "2006"|`YEAR BUILT` == "2008"|
                          `YEAR BUILT` == "2010"|`YEAR BUILT` == "2012"))
```

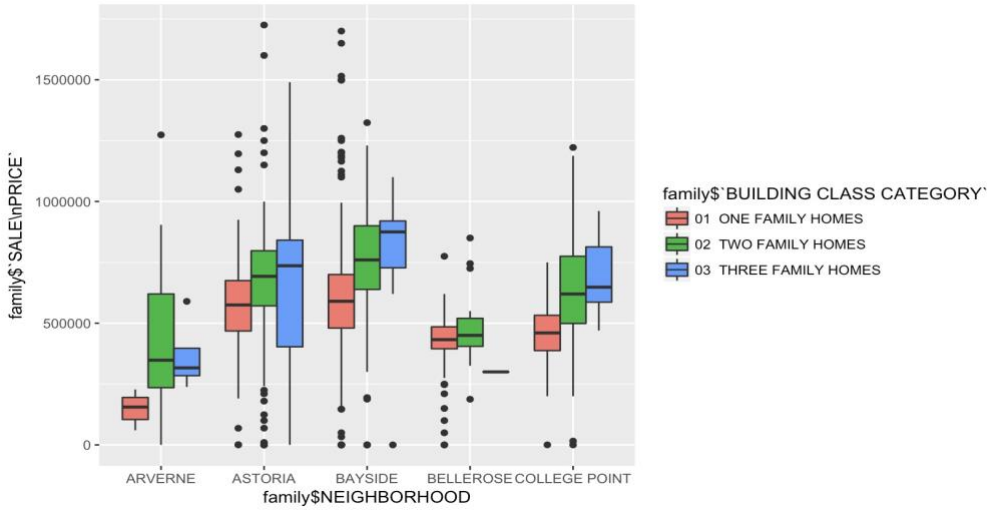

Figure 2. Sale Price on Neighborhood and Building Class

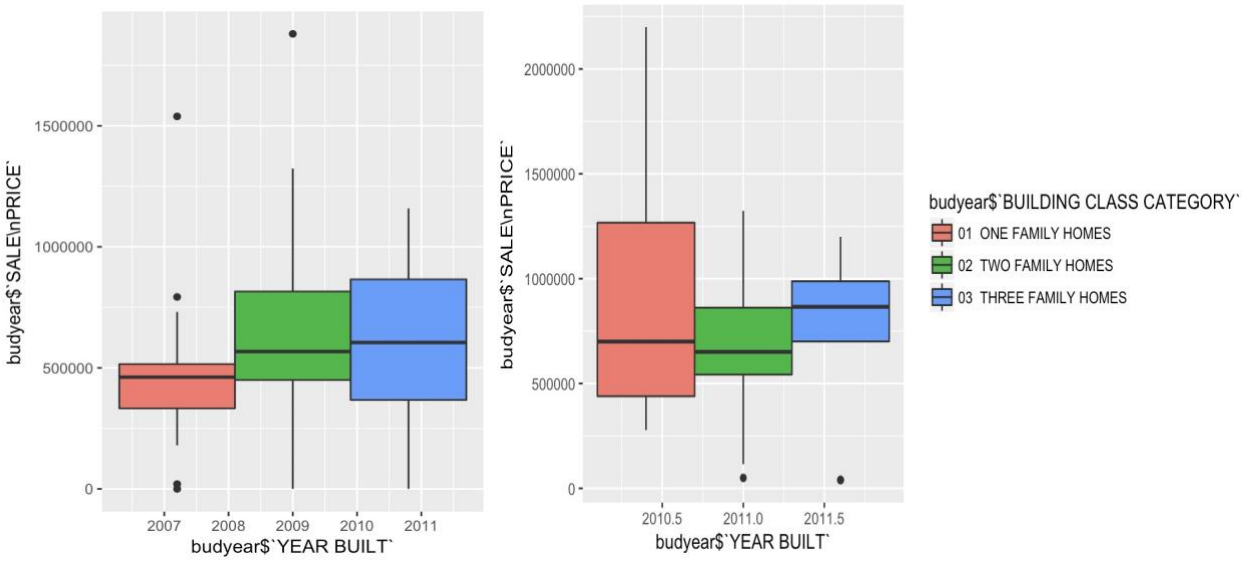

Figure 3. Sale Price on Year Built and Building Class

1b. Pick one or more models (these need not be restricted to the models you've learned so far [multivariate regression, KNN, K-Means]) to explore the chosen data. Interpret the model fits and indicate significance. Describe any cleaning you had to do and why.

I choose the multivariate regression model to explore the data. Before applying to the model, I create a new dataset which "gross square feet", "land square feet" and "sale price" are more than zero. Because the zero exist in "square feet" is meaningless in the dataset.

I choose four independent variables which are "Gross Square Feet", "Land Square Feet", "Neighborhood" and "Building Class Category" to predict sale price. Because the sale price, "Gross Square Feet" and "Land Square Feet" are left skewed, I apply log() function on them. I also use factor() function to "Neighborhood" and "Building Class Category" which are string type.

The model that I choose can well explain the Sale Price, because the p-value is less than 0.05, and R2 is 0.975 which close to 1. Generally, the higher the R2 the better the model explains sale price. And the smaller the p-value is, the more significant the factor is. Therefore, "Gross Square Feet" is more significant than "Land Square Feet"(0.02), all types of "Neighborhood" are significant and "Elevator Apartments" in "Building Class Category" is the most significant one.

The possible model function would be: Log(Sale Price) = 0.37*log(Gross Square Feet)+0.13*log(Land Square Feet)+9.83*(Woodhaven)+0.18*(COOPS - ELEVATOR APARTMENTS).

Based on Figure 4., most of the residual points are concentrated on zero, which means that the difference between observed sale price and predict sale price followed by linear regression model is very small.

```
m3 <-
  lm(log(newq2$`SALE\nPRICE`) ~ 0+log(newq2$`GROSS SQUARE FEET`)+log(newq2$`LAND SQUARE FEET`)+
       factor(newq2$NEIGHBORHOOD)+factor(newq2$`BUILDING CLASS CATEGORY`), data=newq2)
summary(m3)
ploy(m3)
plot(resid(m3))
```

```
Call:
lm(formula = log(newq2$`SALE\nPRICE`) ~ 0 + log(newq2$`GROSS SQUARE FEET`) +
    log(newq2$`LAND SQUARE FEET`) + factor(newq2$NEIGHBORHOOD) +
    factor(newq2$`BUILDING CLASS CATEGORY`), data = newq2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3201   0.1252   0.4202   0.6616   3.8048

Residual standard error: 2.042 on 9312 degrees of freedom
Multiple R-squared:  0.975,     Adjusted R-squared:  0.9748
F-statistic:  4275 on 85 and 9312 DF,  p-value: < 2.2e-16
```

```
log(newq2$`GROSS SQUARE FEET`)                                  ***
log(newq2$`LAND SQUARE FEET`)                                   *
factor(newq2$NEIGHBORHOOD)AIRPORT LA GUARDIA                    ***
factor(newq2$NEIGHBORHOOD)ARVERNE                               ***
factor(newq2$NEIGHBORHOOD)ASTORIA                               ***
factor(newq2$`BUILDING CLASS CATEGORY`)09  COOPS - WALKUP APARTMENTS          **
factor(newq2$`BUILDING CLASS CATEGORY`)10  COOPS - ELEVATOR APARTMENTS       ***
factor(newq2$`BUILDING CLASS CATEGORY`)11  SPECIAL CONDO BILLING LOTS         *
```

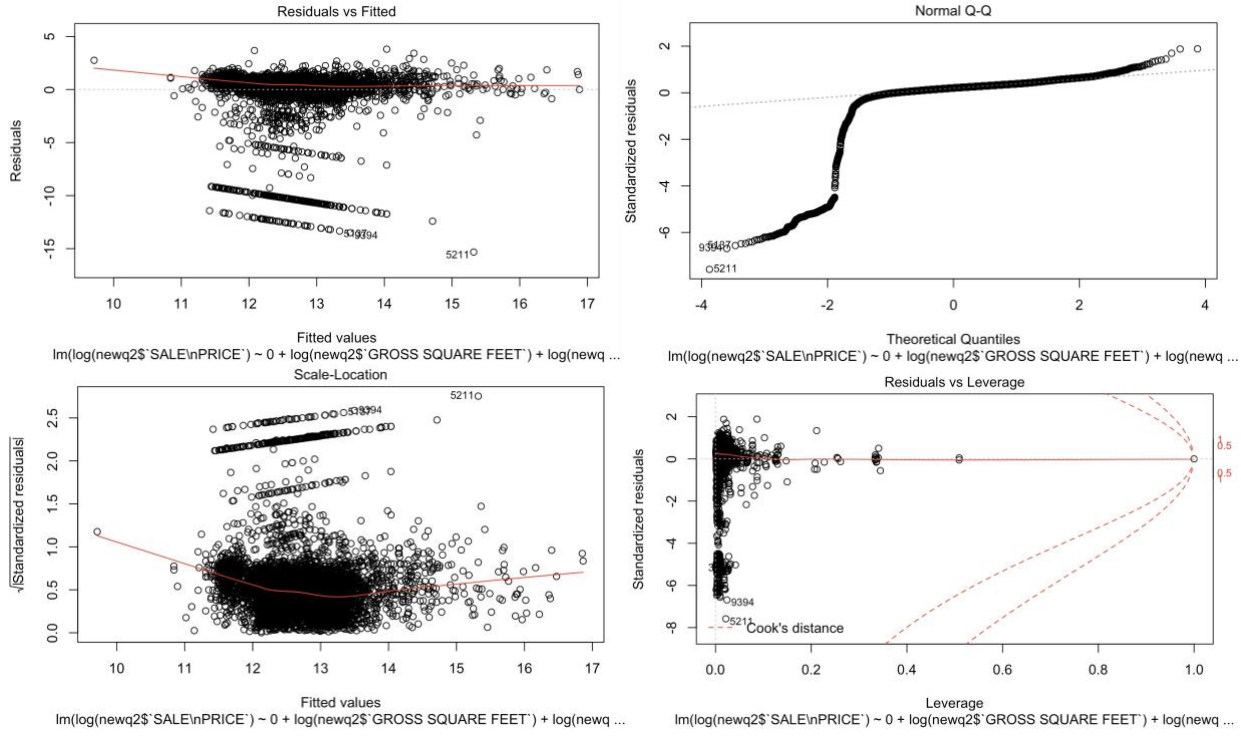

Figure 4. Model Performance and Summary Results

Apply the model(s) to predict quantities of interest. Describe (contingency table) and plot the predictions.

The quantities of interest that I want to predict is Sale Price.
In order to examine the efficiency of my model, I separate the dataset into training and testing part, the sample size is 50% in training set and the rest in test set. The dimension in training set contains 500 observations and 6 variables. To start, I hypothesis that “Gross Square Feet”, “Land Square Feet”, “Building Class Category”, Neighborhood” and “Year Built” are potential predicators. Therefore, I construct a new data frame consisting solely of these variables. Before using linear regression model, I want to investigate how those variables influence each other. Based on the following figure 5, we can see the “Gross Square Feet” and “Land Square Feet” are highly skewed. We can have a better fitting result, if those variables are normal distribution. The Sale Price and Year Built are almost normal distribution. “Neighborhood” is not an important factor in influencing price, based on the dataset that I just choose five random neighborhood.
So, I log those variables with highly skewed. The significant indicator show that all the variables are significant (except “Year Built”). The Figure 6 shows part of the final predicted result of Sale Price.

```
#Split the data into training and testing set
set.seed(700)
split <- sample(seq_len(nrow(family)), size = floor(0.5 * nrow(family)))
train <- family[split, ]
test <- family[-split, ]
dim(train)

#Constracted a new dataframe contain some variables
train <- subset(train, select=c(`SALE\nPRICE`,`GROSS SQUARE FEET`,
                                `LAND SQUARE FEET`, NEIGHBORHOOD,
                                `BUILDING CLASS CATEGORY`,`YEAR BUILT`))
head(train)
summary(train)
pairs.panels(train, col="red")

fit <- lm(`SALE\nPRICE`~0+log(`GROSS SQUARE FEET`)+log(`LAND SQUARE FEET`)+
  factor(`BUILDING CLASS CATEGORY`)+`YEAR BUILT`, data=train)
summary(fit)

confint(fit, conf.level=0.95)

test <- subset(test, select=c(`SALE\nPRICE`,`GROSS SQUARE FEET`,
                              `LAND SQUARE FEET`, NEIGHBORHOOD,
                              `BUILDING CLASS CATEGORY`,`YEAR BUILT`))

prediction <- predict(fit, newdata=test)
head(prediction)
```

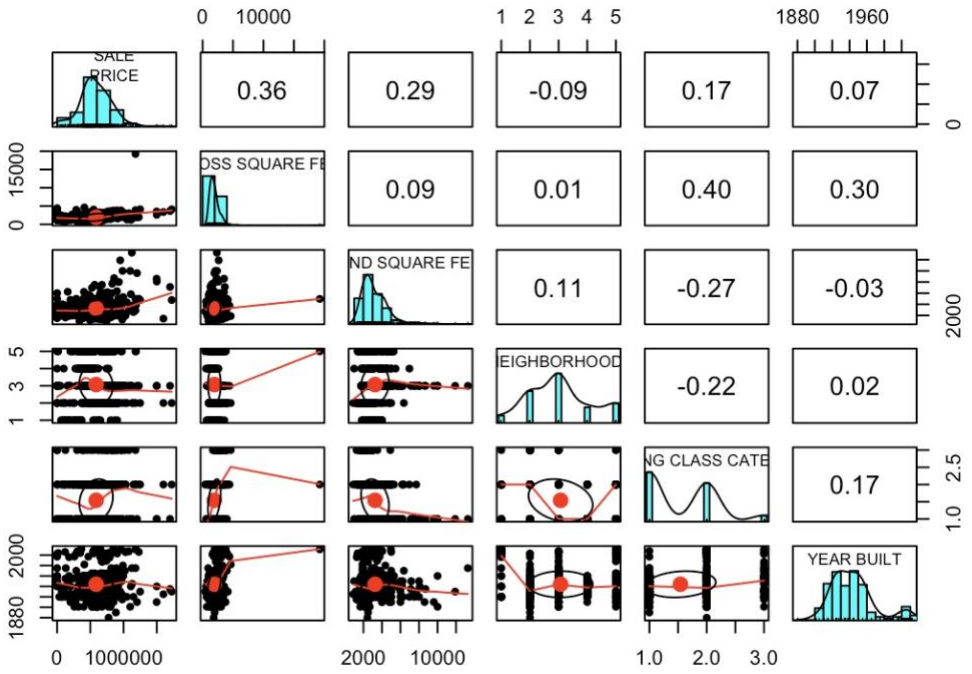

Figure 5. Fitting pattern between variables (Contingency Table)

```
Call:
lm(formula = `SALE\nPRICE` ~ 0 + log(`GROSS SQUARE FEET`) + log(`LAND SQUARE FEET`)
+
    factor(`BUILDING CLASS CATEGORY`) + `YEAR BUILT`, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-850457 -105659   21225  132775  911575

Coefficients:
                                                    Estimate Std. Error
log(`GROSS SQUARE FEET`)                            284513.1    33512.6
log(`LAND SQUARE FEET`)                             128910.3    26758.3
factor(`BUILDING CLASS CATEGORY`)01  ONE FAMILY HOMES   -1499278.8   803209.7
factor(`BUILDING CLASS CATEGORY`)02  TWO FAMILY HOMES   -1491404.5   804622.7
factor(`BUILDING CLASS CATEGORY`)03  THREE FAMILY HOMES -1506977.0   806866.5
`YEAR BUILT`                                           -554.3      415.8
                                                    t value Pr(>|t|)
log(`GROSS SQUARE FEET`)                              8.490 2.44e-16 ***
log(`LAND SQUARE FEET`)                               4.818 1.94e-06 ***
factor(`BUILDING CLASS CATEGORY`)01  ONE FAMILY HOMES    -1.867   0.0625 .
factor(`BUILDING CLASS CATEGORY`)02  TWO FAMILY HOMES    -1.854   0.0644 .
factor(`BUILDING CLASS CATEGORY`)03  THREE FAMILY HOMES  -1.868   0.0624 .
`YEAR BUILT`                                          -1.333   0.1831
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 222800 on 494 degrees of freedom
Multiple R-squared:  0.8801,    Adjusted R-squared:  0.8786
F-statistic: 604.2 on 6 and 494 DF,  p-value: < 2.2e-16
```

Figure 6. Statistic Summary of Fitting model based on training set

```
> prediction <- predict(fit, newdata=test)
> head(prediction)
       1        2        3        4        5        6
592022.4 585627.2 352472.5 699835.0 685540.7 647313.8
```

Figure 7. The head of Predicted Result

Examine the fit(s). Perform a significance test that is suitable for the variables that are investigating and describe the results

What I use for performing a significance test is by calculating the R-square value for the fitting model on the test set. In general, the R^2 is the metric for evaluating the goodness of fitting my model. The higher the better.

The final result is (1-SSE/SST), 20.17%, which indicate that my model is not perfectly match with observed sale price. The error between observed one and predicted one are quite big. Therefore, I think that may be caused by three reasons, the first one is not enough samples for prediction; the independent variables should add more; the function of linear model should be revised to improve the model performance.

```
SSE = sum((test$`SALE\nPRICE` - prediction)^2)
SST = sum((test$`SALE\nPRICE`-mean(train$`SALE\nPRICE`))^2)
SSE
SST
1 - (SSE/SST)
```

```
> SSE
[1] 2.737729e+13
> SST
[1] 3.429629e+13
> 1 - (SSE/SST)
[1] 0.2017418
```

Figure 8. The Result of Significance

Based on what I did above, I found that “Year Built” is not an important factor in my model. Two square feet variables are highly skewed, so we should transform them into normal distribution first before conducting models.

And in the long run, the sale price of “One family homes” doesn’t change quite a lot, and the more families in a home the higher the sale price. However, during the 2010-2012 year, the price of “One family homes” raised quite a lot even higher than “Three family homes”. There might happen something on “One family homes”, such as new living policy or increased immigrates.

```
#Heatmap
new_matrix <- data.matrix(factor(family$NEIGHBORHOOD))
library(ggplot2)
ggplot(data=family, aes(x=NEIGHBORHOOD, y=`BUILDING CLASS CATEGORY`,fill=`SALE\nPRICE`))+geom_tile()
```

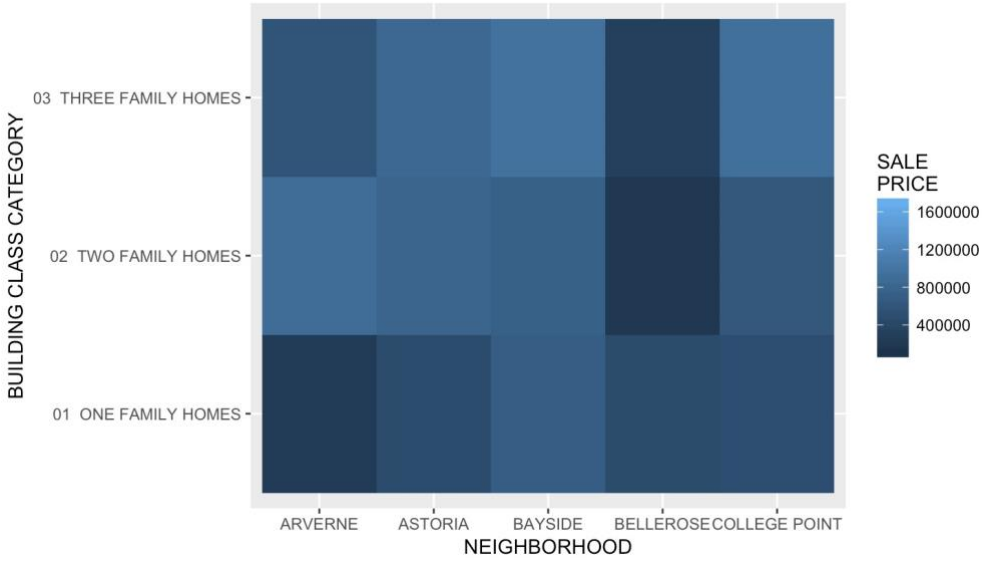

Figure 9. Heatmap of Sale Price in three Family and five Neighborhoods

Also, the HeatMap shows that the highest sale price is “Three family homes” in Bayside while the lowest sale price is “One family homes” in Arverne.

In conclusion, my model can well explain the training data set, however, when it applies to the test data set, the model performance is not as good as in the training data set. Based on significant test, the three kinds of “Building Class Category” are not very significant compared with other two space measurements. Therefore, I should reconsider about my model, maybe I should add more variables or filter my dataset in a different way.

The model can be used to give a general prediction on Queen’s sale price in some certain areas, due to the limitation of variables and a relatively low prediction rate. However, people can still use the predicted value to judge whether it is worthwhile to buy a house/apartment or not. If they want to determine a true value of a house/apartment, they need to do more investigation rather than just using this model to predict.

Reference:

1. https://stackoverflow.com/questions/14604439/plot-multiple-boxplot-in-one-graph
2. https://stackoverflow.com/questions/1686569/filter-data-frame-rows-by-a-logical-condition
3. https://stats.stackexchange.com/questions/5135/interpretation-of-rs-lm-output