A Simple Linear Model for Toronto and Mississauga House Prices

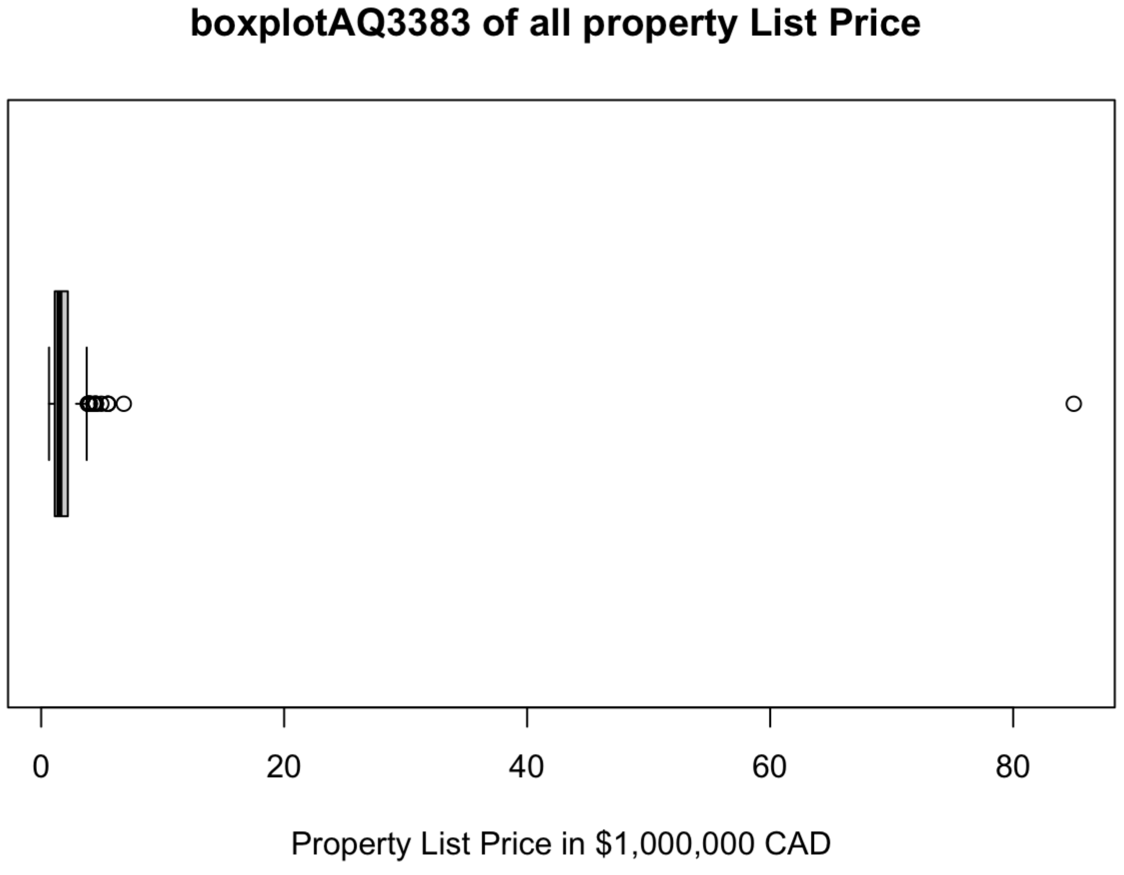
AQ3383

October 20, 2020

I. Exploratory Data Analysis

```
set.seed(3383)
data_AQ3383 = original_data_AQ3383[sample(nrow(original_data_AQ3383), 200), ]
```

Part 1 The boxplot with no unusual points removed is shown below.



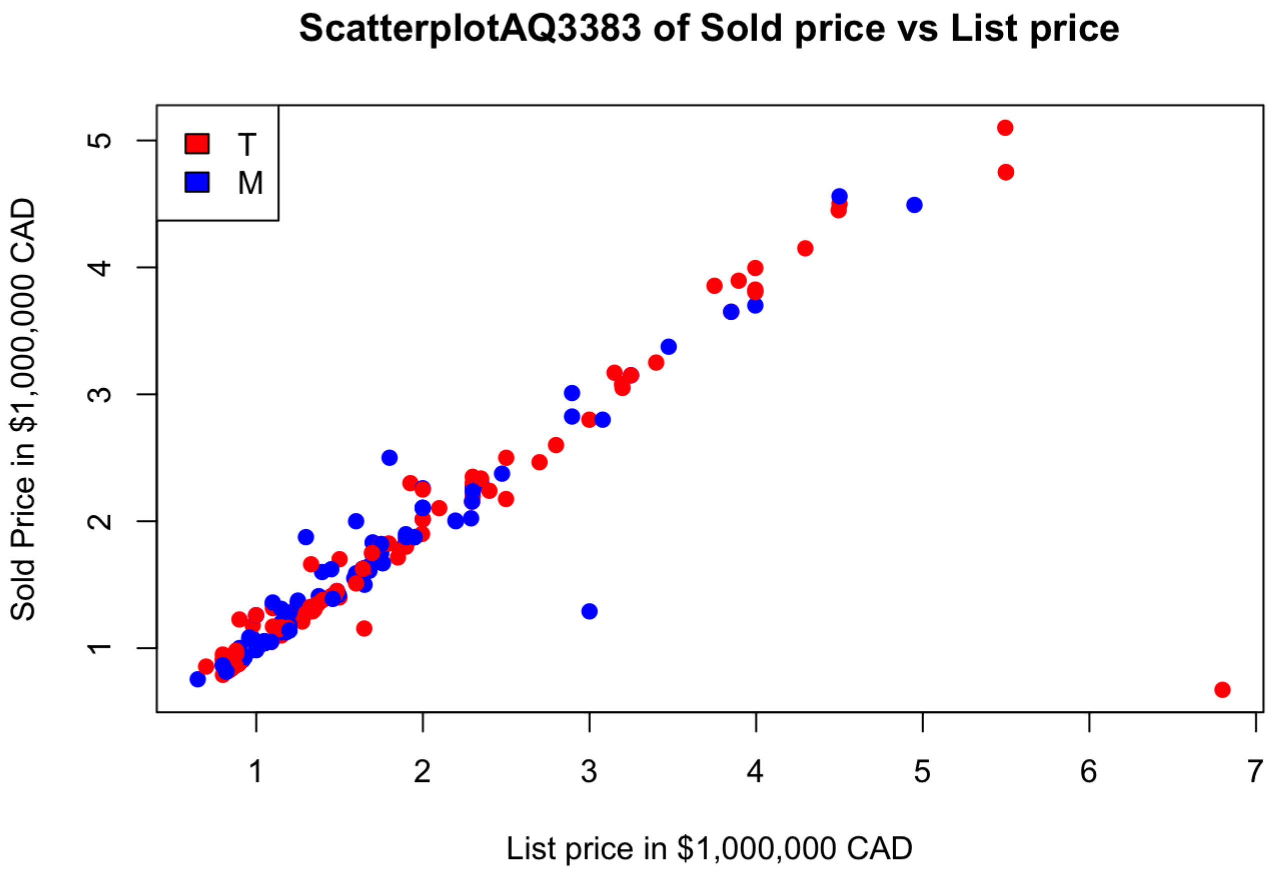
I use this plot because we can easily see if any points are outliers from graph plotted. In our case, the outlier is the maximum value. Therefore when creating a subset of what we currently have, I remove this particular point.

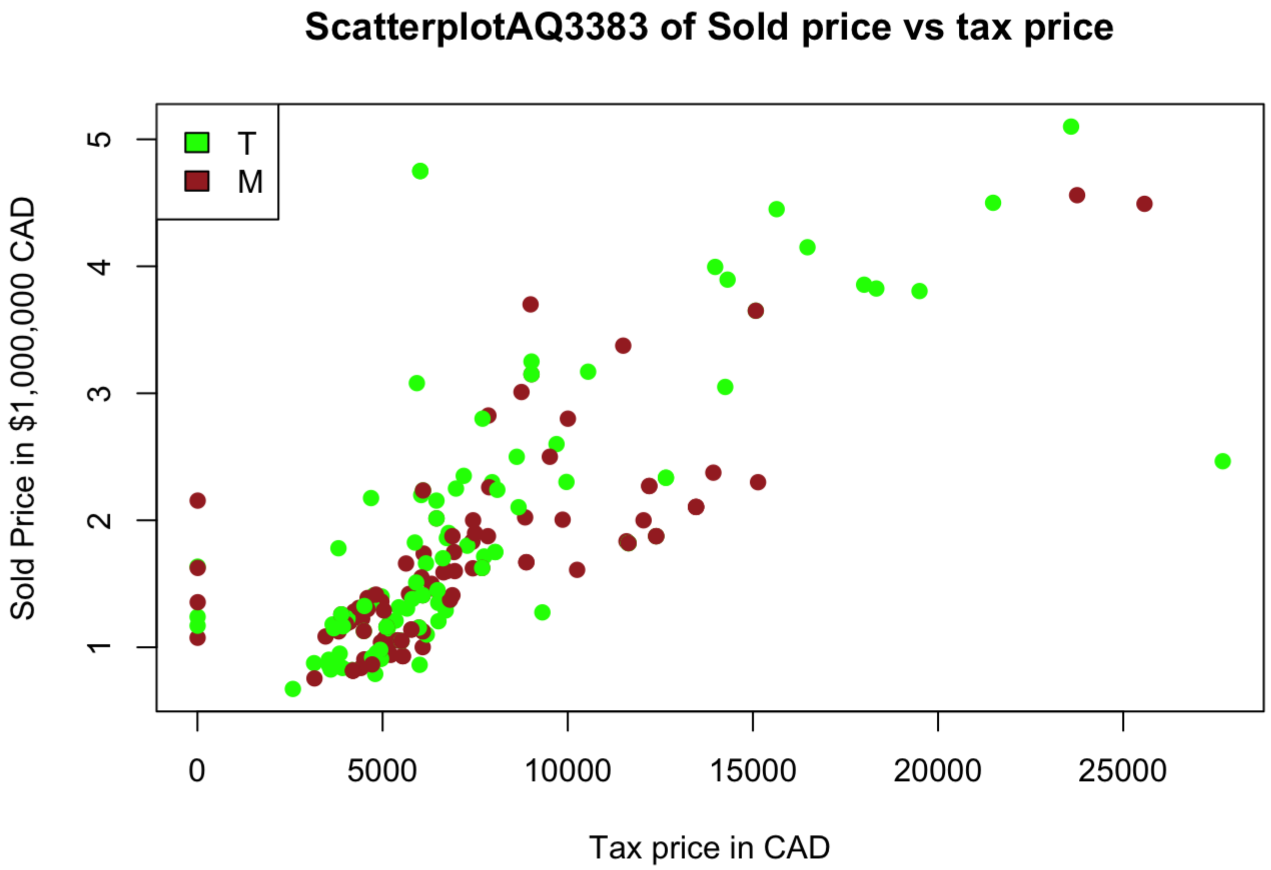
```
#find the ID of the outlier
data_AQ3383[which.max(data_AQ3383$list),]
```

```
##      ID  sold  list taxes location
## 112 112 1.085 84.99  4457        T
```

```
#remove the outlier
df_AQ3383 <- data_AQ3383[data_AQ3383$ID != 112,]
```

Then we use data frame df_AQ3383 for the rest of the assignment.





Interpretation: Both graphs above shows a somehow linear relationship between, therefore we can use linear regression for both graphs. Sold price vs list price has a more significant linear relationship than sold price vs tax price. Heteroscedasticity occurs in sold price vs tax price graph. In the boxplot with all data of listing price included, there exists a significant outlier, which is about 85 million CAD. This might be a error occurred when measuring data.

II. Methods and Model

Three simple linear regressions (SLR) for sale price from list price and its corresponding graph shown below:

```
##
## Call:
## lm(formula = dfy_AQ3383 ~ dfx1_AQ3383)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9821 -0.1073 -0.0242  0.1184  0.7429
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.35438    0.05900   6.006 9.01e-09 ***
## dfx1_AQ3383  0.77949    0.02795  27.887  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4173 on 197 degrees of freedom
## Multiple R-squared:  0.7979, Adjusted R-squared:  0.7969
## F-statistic: 777.7 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = dfy_M_AQ3383 ~ dfx_M_AQ3383)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45148 -0.04358 -0.01940  0.05841  0.41395
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.13901    0.02145   6.481 5.48e-09 ***
## dfx_M_AQ3383  0.89045    0.01188  74.941  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1031 on 86 degrees of freedom
## Multiple R-squared:  0.9849, Adjusted R-squared:  0.9847
## F-statistic:  5616 on 1 and 86 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = dfy_T_AQ3383 ~ dfx_T_AQ3383)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.6577 -0.1544 -0.0038  0.1853  0.7889 
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.54309    0.10714   5.069 1.65e-06 ***
## dfx_T_AQ3383  0.70401    0.04609  15.276  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5367 on 109 degrees of freedom
## Multiple R-squared:  0.6816, Adjusted R-squared:  0.6787 
## F-statistic: 233.4 on 1 and 109 DF,  p-value: < 2.2e-16
```

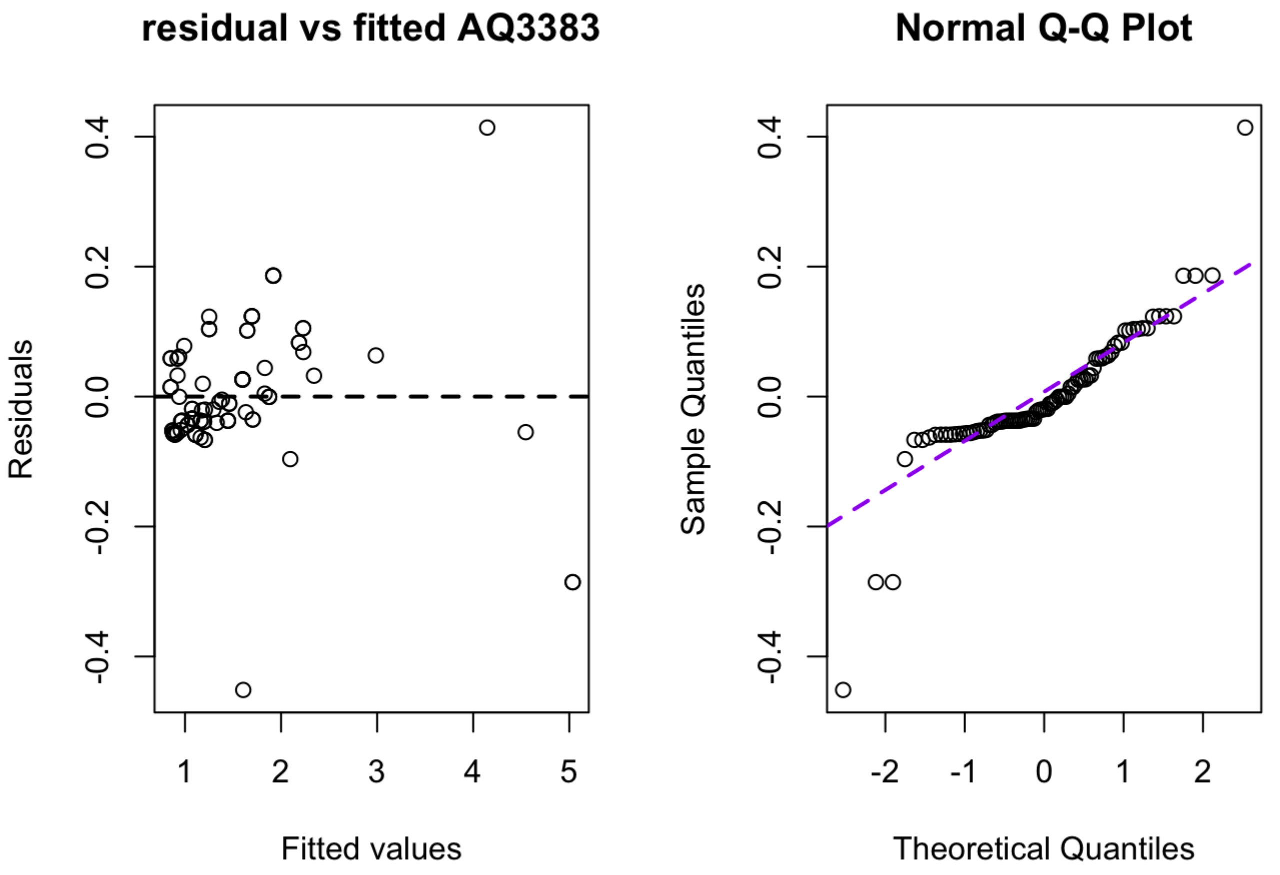
Table:

Regression	R^2	**estimated intercept** β_0	**estimated slope** β_1	**estimate of the variance of the error**	**p-value for** $H_0 : \beta_1 = 0$	**95% CI for** β_1
All	0.7979	0.3544	0.7795	0.1741	p-value: < 2.2e-16	(0.7244, 0.8346)
Mississauga Neighborhood	0.9849	0.1390	0.8905	0.0106	p-value: < 2.2e-16	(0.8668, 0.9141)
Toronto Neighborhood	0.6816	0.5431	0.7040	0.2880	p-value: < 2.2e-16	(0.6127, 0.7954)

Interpret and compare: We see the difference between R^2 is quite different. Mississauga neighborhood has a 0.98 R^2, which is the highest among all. However Toronto Neighborhood has a 0.68 R^2. As a result, R^2 based on all data is 0.80, which is in between other two sets of data. This shows us it is necessary to evaluate two neighborhoods respectively. This is normal because the distance between data and fitted regression line for all neighborhoods is usually not as well as how data fits the two neighborhoods respectively, since the list and sold price are often more similar in one neighborhoods.

A pooled two-sample t-test is not the best to be used when determine if there is a statistically significant difference between the slopes of the simple linear models for the two neighborhoods. Since we are dealing with housing price in two different cities, therefore it is reasonable to assume the two sets of data are independent. However, according to data shown above, they do not have the same variance. Therefore we should not use pooled two-sample t-test here.

III. Discussions and Limitations



According to data summary shown above, I picked Mississauga neighborhood to do the following evaluation. This is because it has a R^2 of 0.9849, which means it has the highest portion of explained response variable variation by this model.

Violations: According to fitted vs residual graph, we can tell there are a few irregulars. The graph is mostly linear, which is good. There are also a few outliers in Normal Q-Q plot. However thee residuals follow a mostly straight line, which concluded there are not much violations, except for a few outliers.

Possible predictors to to fit a multiple linear regression for sale price could be size of the house, and how old this house is. These are factors directly related and can cause effect to the sale price