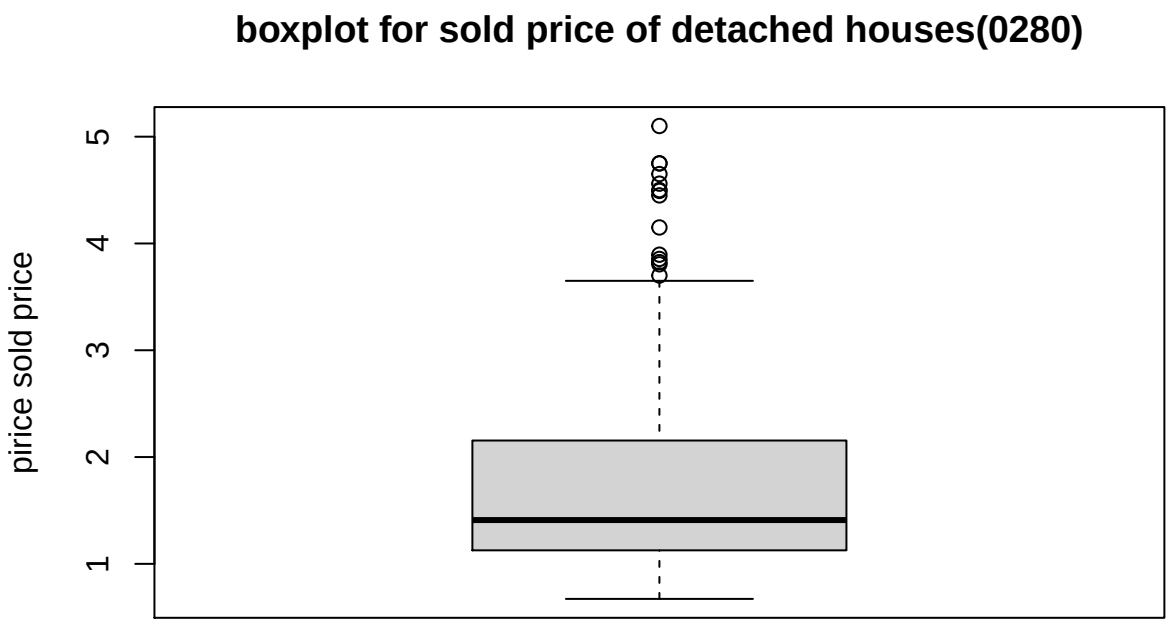
A research on the price and tax for detached house during COVID-19(0280)

Z.L0280

October 24, 2020 (.)

I. Exploratory Data Analysis




```
## [1] 84.99
```

```
## [1] 0.699
```

```
## [1] 5.1
```

```
## [1] 0.672
```

```
##     ID  sold  list taxes location
## 70 112 1.085 84.99  4457        T
```

```
##     ID  sold  list taxes location
## 99 110 0.855 0.699  3558        T
```

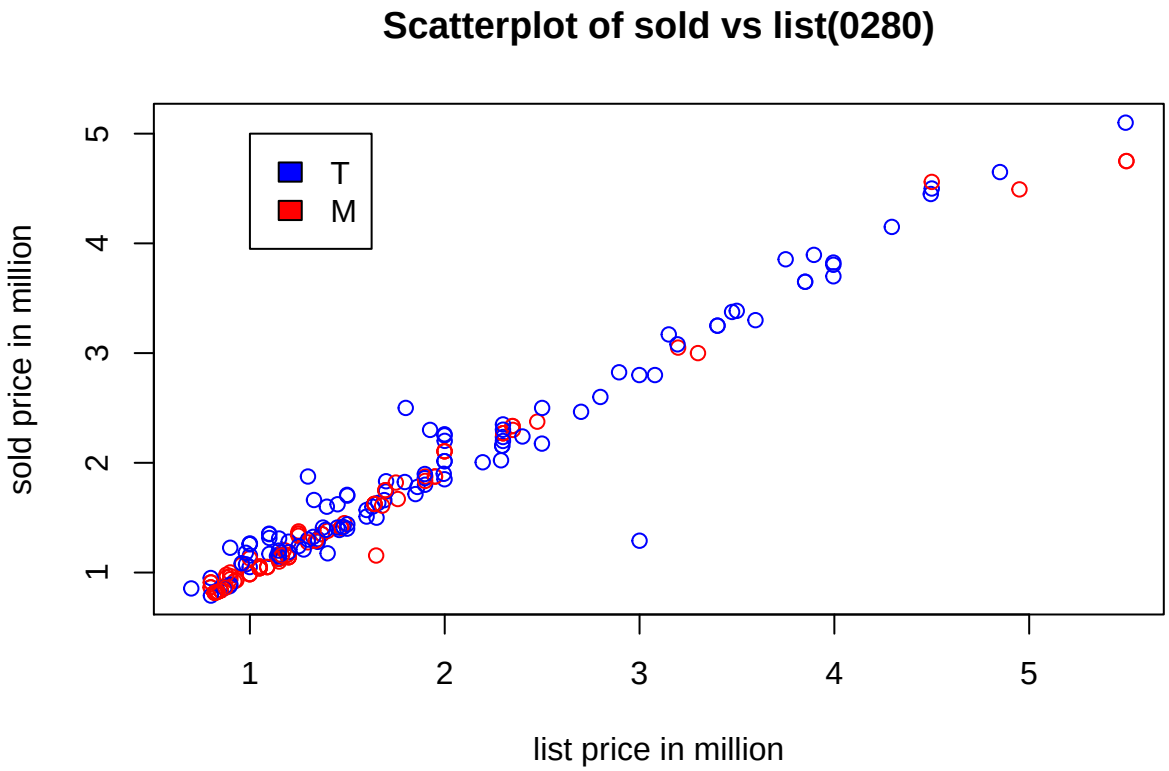
```
##     ID sold  list taxes location
## 135 96  5.1 5.495 23592        T
```

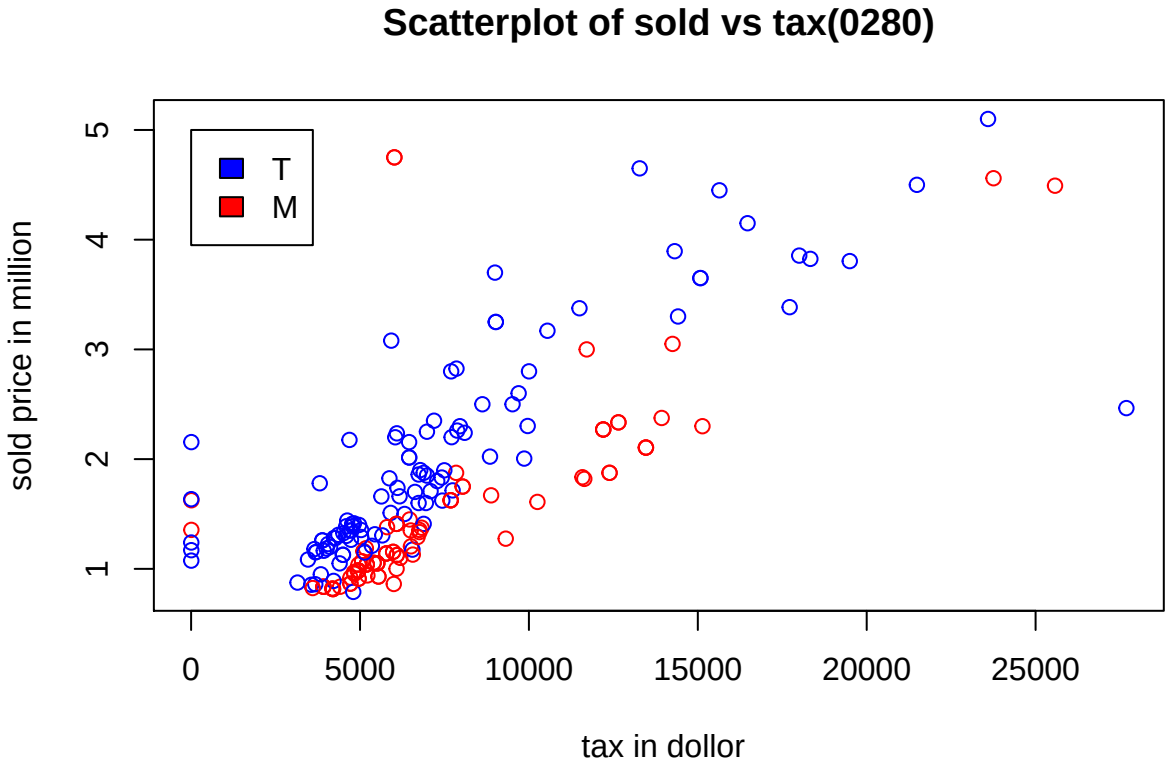
```
##   ID  sold  list taxes location
## 5 95 0.672 6.799  2577        T
```

a) I choose to remove the two cases which are listed at 84.99M and sold at 1.085M and listed at 6.799M and sold at 0.672M because the huge difference in the price reveals the fact that at least one of them

are not made seriously and would affect the accuracy of the research result.

b)





c) 1)Since I believe the sold price can better demonstrate the price level over a certain area, thus I choose to use the boxplot of sold price to discover the spread of price. In the boxplot, it's clear that the majority sold price of detached house during Covid-19 fall between 1 to 2 million. Several outliers can be seen in the area which the price is greater than approximately 3.8 million.

2)The scatterplot named “scatterplot of sold vs list” demonstrates that the sold price and the list price follows a strong linear relationship. Generally speaking, the sold price and list price in T-neighborhood are higher than that of M-neighborhood.Additionally, one obvious outlier with list price at 3M is worth notetaking.

3)The scatterplot named “scatterplot of sold vs tax” indicates that the sold price and tax follows a weak linear relationship.Note that with same amount of tax being charged, the sold price of detached house in T-neighborhood are usually higher than that of M-neighborhood.Generally, the higher the sold price is for a detached house, the greater the tax will be charged.Additionally, more outliers and leverage point are identified.

II. Methods and Model

```
##
## Call:
## lm(formula = datafinal0280$sold ~ datafinal0280$list)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57041 -0.07112 -0.02540  0.06250  0.72195
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.15533    0.02525   6.152 4.22e-09 ***
## datafinal0280$list  0.90169    0.01220  73.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1774 on 196 degrees of freedom
## Multiple R-squared:  0.9654, Adjusted R-squared:  0.9652
## F-statistic:  5465 on 1 and 196 DF,  p-value: < 2.2e-16
```

```
##                         2.5 %    97.5 %
## (Intercept)        0.1055404 0.2051250
## datafinal0280$list 0.8776368 0.9257463
```

```
## [1] 0.03147076
```

```
##
## Call:
## lm(formula = dataMZZZ$sold ~ dataMZZZ$list)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45070 -0.04953 -0.01945  0.05886  0.41676
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.13941    0.02044   6.821  1.1e-09 ***
## dataMZZZ$list  0.88974    0.01140  78.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1019 on 88 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9856
## F-statistic:  6093 on 1 and 88 DF,  p-value: < 2.2e-16
```

```
##                     2.5 %    97.5 %
## (Intercept)   0.09879638 0.1800276
## dataMZZZ$list 0.86708581 0.9123918
```

```
## [1] 0.01038361
```

```
##
## Call:
## lm(formula = dataT0280$sold ~ dataT0280$list)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.59431 -0.09671 -0.01853  0.09689  0.69248
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.19313    0.04536   4.257 4.49e-05 ***
## dataT0280$list  0.89706    0.01994  44.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2183 on 106 degrees of freedom
## Multiple R-squared:  0.9502, Adjusted R-squared:  0.9498
## F-statistic:  2024 on 1 and 106 DF,  p-value: < 2.2e-16
```

```
##                     2.5 %    97.5 %
## (Intercept)    0.1031910 0.2830702
## dataT0280$list 0.8575307 0.9365922
```

```
## [1] 0.04765489
```

Table (.)

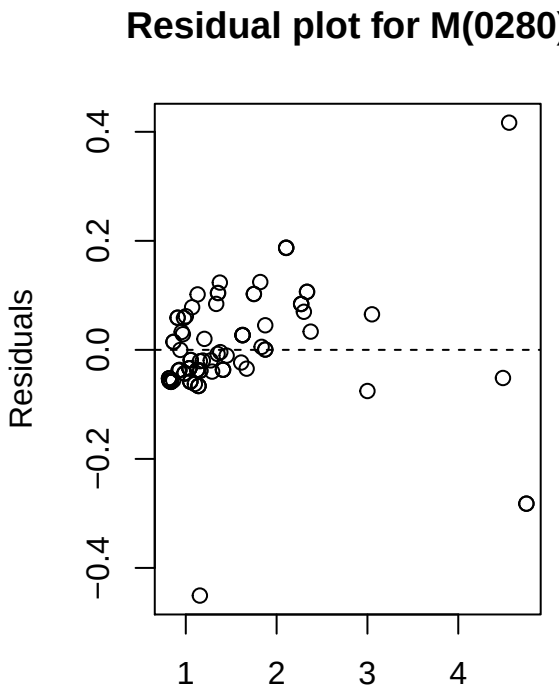
```
##                              All               M-Neighbour       T-Neighbour
## R^2                          0.9654            0.9858            0.9502
## estimated intercept          0.15533           0.13941           0.19313
## estimated slope              0.90169           0.88974           0.89706
## estimated variance of error 0.03147           0.01038           0.04765
## p-value                      2e-16             2e-16             2e-16
## 95% CI                       (0.87763,0.92575) (0.86708,0.91239) (0.8575,0.9365)
```

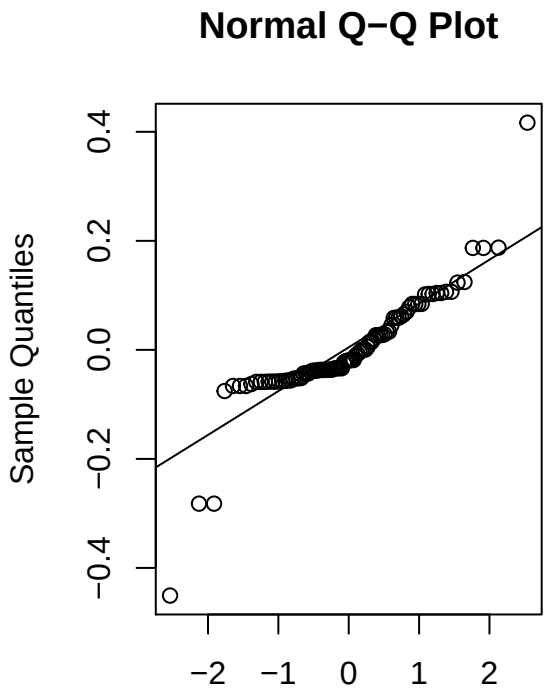
b) R^2 in three cases are a little bit different from each other. The reason for this to happen is that there are more outliers in the data T-Neighborhood than T-neighborhood. There are more data points in T-neighborhood one and the house price in Toronto are more dispersed thus leading to a less fit linear relationship.

c) In order to apply a two-sample t-test. It's statistical assumptions must be satisfied. 1)Data values must be independent.

2) Data in each group are normally distributed. 3)The variance from two independent data are equal. Then check it one by one.The price of detached house are not independent because Mississauga and Toronto are close to each other that it serves as a satellite city of Toronto thus the house price in Mississauga is greatly affected by the house price in Toronto. Secondly, as can be discovered from the scatterplot above, all the plots are right-skewed which is a violation of the trait, bell shape of a normal distribution. Thirdly, from the variance we calculated in the above model, the variance from T-neighborhood is 0.04765 while the variance for M-neighborhood is 0.01038 which are not equal. In conclusion, neither of the assumptions are followed thus we can't use the pooled two-sample t-test.

III. Discussions and Limitations





a)I will choose the fitted model for M-neighborhood since it has the greatest R^2 of all 3 indicates that it has the highest percentage of the dependent variable variation that a linear model explains. Additionally, it has the least estimated variance of error which further proves the regression model in M-neighborhood is the best fit.

b)Under normal error SLR assumptions,4 assumptions have to be satisfied. 1) A simple linear model is appropriate. 2)The errors are uncorrelated. 4)The errors have constant variance. 4)The errors are Normally distributed. Now I will check if they follows the assumption. By the residual plot for residuals vs fitted value, I conclude it's a linear model since there is no pattern. For the second one, due to limitation of plot I can't check it right here. Thirdly, still by the residual plot, the variance are not constant that it is either increasing or decreasing which the plot demonstrates a fanning pattern. Lastly, by QQplot, it's clear that the errors are not following the normal distribution which is another violation of the assumption.

c) Two potential numeric predictors can be the size of the detached house and the time period of the detached house since its constructed.Population in the city is also a considerable numeric predictors.