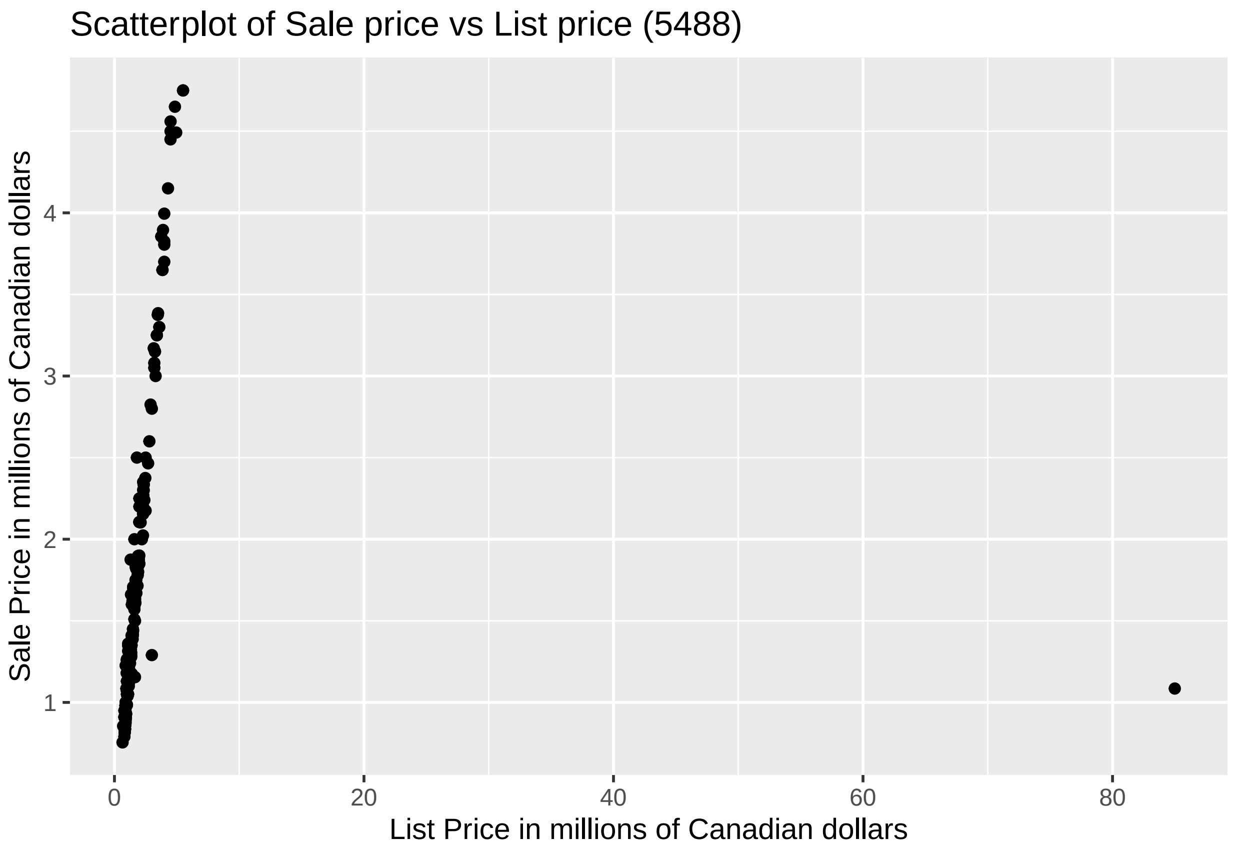
A simple linear model to determine the expected sale price of detached single family homes in two neighbourhoods in the Great Toronto Area

LL5488

October 24, 2020

I. Exploratory Data Analysis

A scatterplot of Sale price vs List price to describe the data



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

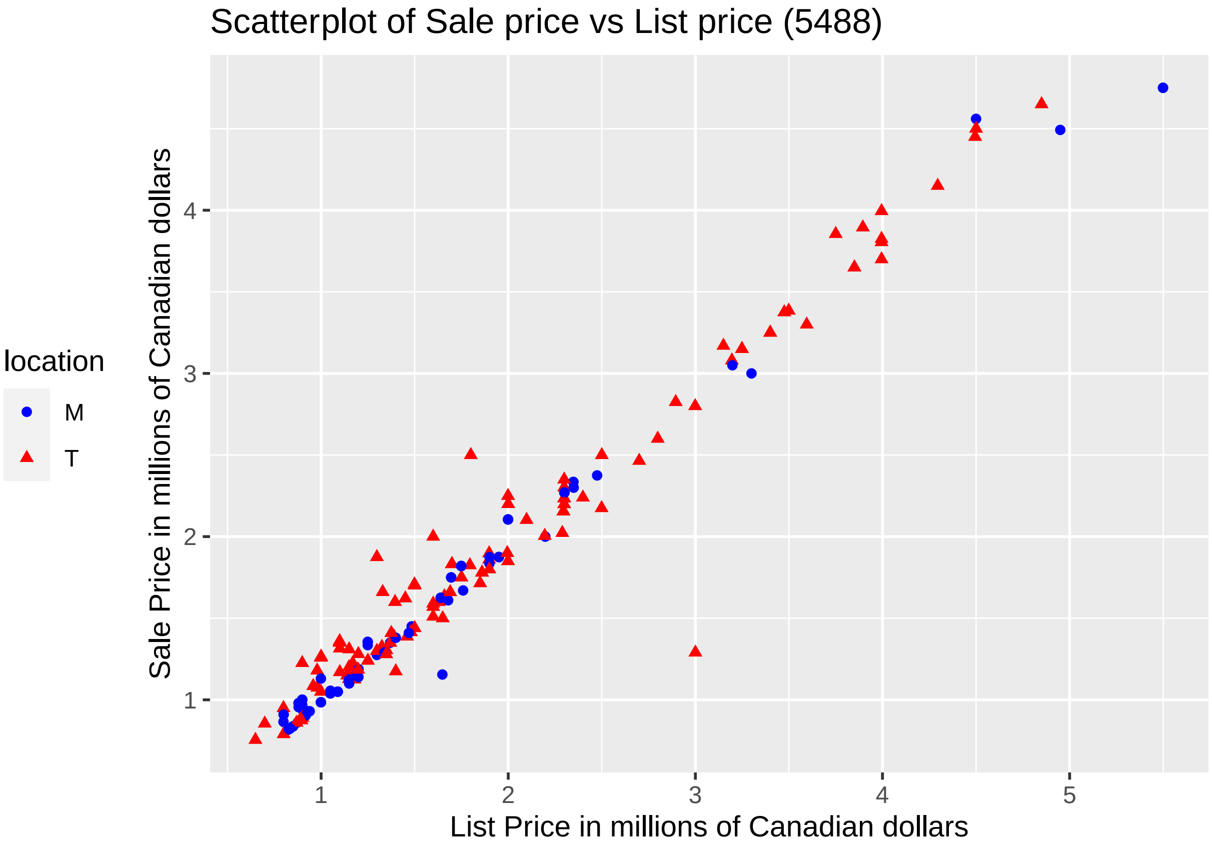
I made a scatterplot to describe the data we randomly selected. In this scatterplot, I chose the sale prices as the response variable (Y) and the list prices as the explanatory variable (X).

The description of the scatterplot

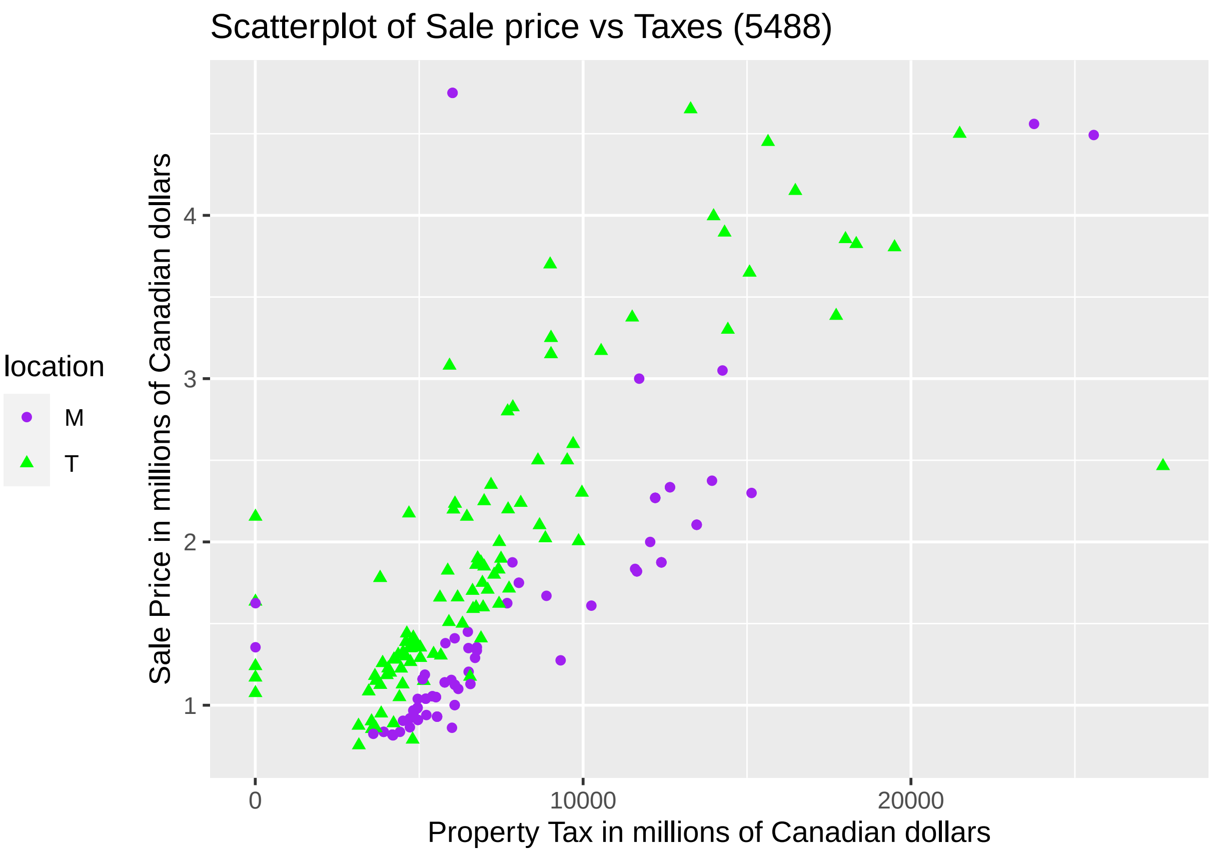
We can see from the scatterplot that there is a strong linear trend between sale price and list price. Besides, this scatterplot has several outliers which can represent that some list prices are higher and some sold prices are lower.

Explanation: Choice of the variable may vary. However, from the above scatterplot of sale price vs list price, we can find out that there is an influential point at the right bottom of the plot. The x and y value of this point are both abnormal so that it is a 'bad' leverage point. Therefore, it is reasonable to remove this case (with ID 112) that corresponds to this point and create a subset data by removing this case.

The Scatterplots and their interpretations



Interpretation: The Scatterplot of Sale price vs List price has a strong linear trend. Besides, we can find out that there is an outlier in the plot since the y value of this outlier is abnormal.



Interpretation: The scatterplot of Sale price vs Taxes also has a linear trend. We can find out that the values of the property tax mainly concentrate around 5000. Besides, there are also some outliers in this scatterplot.

The comparison between two scatterplots

Both scatterplots are more or less linear. Therefore, it is quite appropriate to use linear regression. However, the linear trend between sale price and list price is stronger than that between sale price and taxes.

II. Methods and Model

```
##
## Call:
## lm(formula = sy ~ sx)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57629 -0.07217 -0.02077  0.06135  0.71897
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15397    0.02553   6.032 7.88e-09 ***
## sx           0.90411    0.01233  73.300  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.177 on 197 degrees of freedom
## Multiple R-squared:  0.9646, Adjusted R-squared:  0.9645
## F-statistic:  5373 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = sym ~ sxm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45176 -0.05231 -0.01992  0.05879  0.41507
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14010    0.02123    6.60 3.22e-09 ***
## sxm          0.88996    0.01163   76.54  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1039 on 86 degrees of freedom
## Multiple R-squared:  0.9855, Adjusted R-squared:  0.9854
## F-statistic:  5858 on 1 and 86 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = syt ~ sxt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60252 -0.09512 -0.01615  0.08961  0.69008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18684    0.04463   4.187 5.76e-05 ***
## sxt          0.90190    0.01989  45.352  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2146 on 109 degrees of freedom
## Multiple R-squared:  0.9497, Adjusted R-squared:  0.9492
## F-statistic:  2057 on 1 and 109 DF,  p-value: < 2.2e-16
```

Three simple linear regressions (SLR) for sale price from list price were fit using: (1) All data (from the subset made in part I) (2) Properties of neighbourhood M (3) Properties of neighbourhood T

The table below gives several model summary values

Type	(1) All	(2) Neighbourhood M	(3) Neighbourhood T
R^2	0.9646	0.9855	0.9497
Estimated intercept, $\hat{\beta_0}$	0.1540	0.1401	0.1868
Estimated slope, $\hat{\beta_1}$	0.9041	0.8900	0.9019
Estimated variance of errors, $\hat{\sigma^2}$	0.0313	0.0108	0.0461
p-value ($H_0 : \beta 1 = 0$)	$p < 0.05$	$p < 0.05$	$p < 0.05$
95% CI of $\beta 1$	(0.8798, 0.9284)	(0.8668, 0.9131)	(0.8625, 0.9413)

Interpret and compare the three R^2 values

We have known that R-squared (R^2) is a statistical measure that represents the percentage of variation for a dependent variable explained by the regression line. From the three simple linear regressions (SLR) for sale price from list price, we can see that all three R^2 values are high so that the percentages of variation for the sale prices explained by those three regression lines are higher. To 1 decimal place, they are all equal to 0.9. We can notice that the R^2 based on all the data is not the largest since the sale prices and list prices are similar across neighbourhoods and the model summaries such as $\hat{\beta_0}$, $\hat{\beta}1$ and $\hat{\sigma^2}$ are similar.

Discussion of the pooled two-sample t-test

In order to use a pooled two-sample t-test to determine if there is a statistically significant difference between the slopes of the simple linear models for the two neighbourhoods, four conditions should be met.

Since properties in neighbourhood M are separate from those in neighbourhood T, it is reasonable to assume that these two populations are independent. Therefore, the condition “The two samples are independent” us met.

Based on Normal error SLRs ($e_i^M \overset{\text{iid}}{\sim} N(0, \sigma_M^2)$, i = 1, . . . , 88, and $e_i^T \overset{\text{iid}}{\sim} N(0, \sigma_T^2)$, i = 1, . . . , 111), it follows that:

$$b_1^M \sim N(\beta_1^M, \frac{\sigma_M^2}{SXX_M})$$

$$b_1^T \sim N(\beta_1^T, \frac{\sigma_T^2}{SXX_T})$$

where b_1^M and b_1^T are least squares estimators of β_1^M and β_1^T respectively. Therefore, b_1^M and b_1^T are both follow a normal distribution which satisfies the conditions.

Besides, from the above table, we can see that σ^2 is 0.0313, σ_M^2 is 0.0108 and σ_T^2 is 0.0461. Those three values are approximately equal so that it satisfies that the two populations have the same variance.

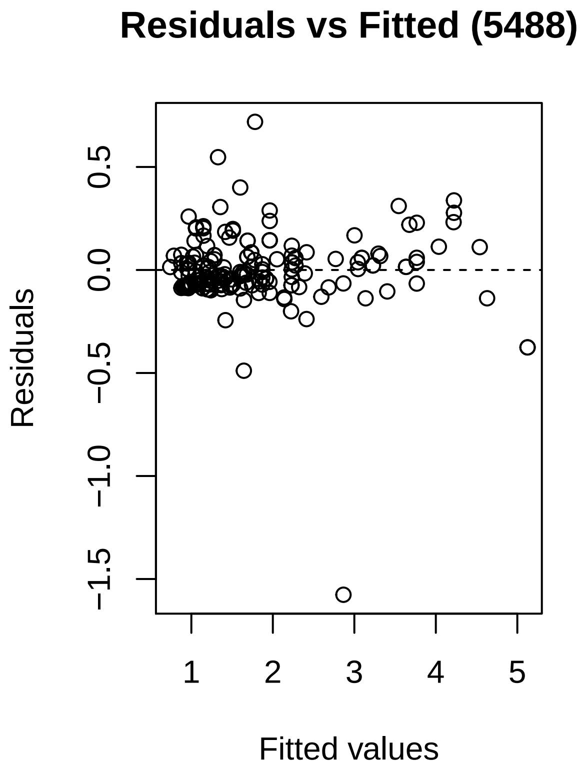
Overall, we can use a pooled two-sample t-test to determine if there is a statistically significant difference between the slopes of the simple linear models for the two neighbourhoods since those four conditions can be met.

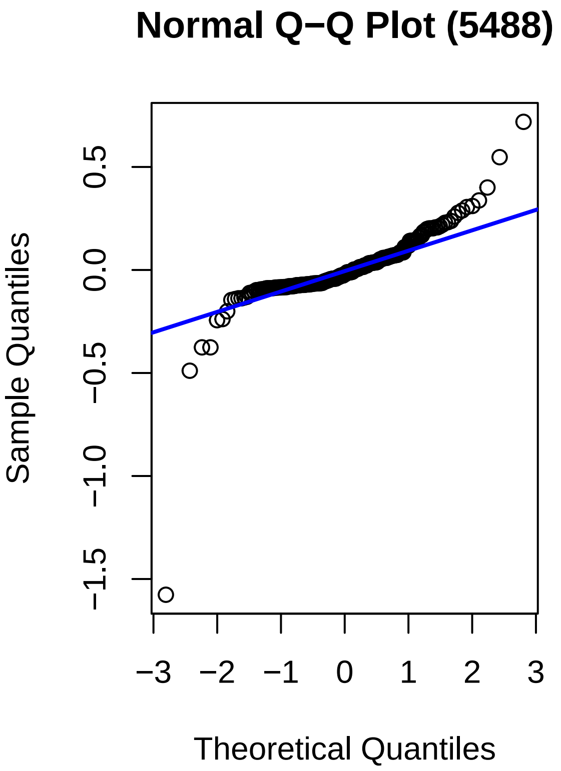
Since we can use the pooled two-sample t-test, we can start by stating appropriate null and alternative hypotheses where $H_0 : \beta_1^M - \beta_1^T = 0$ and $H_a : \beta_1^M - \beta_1^T \neq 0$ where β_1^M and β_1^T represent the slopes of the SLR for neighbourhood M and neighbourhood T respectively.

Then, we can calculate the test statistic and then p-value. If the p-value we calculated is smaller than 0.05, then we can conclude that the difference is quite significant and the data provides evidence that the coefficients are not the same. If the p-value we calculated is larger than 0.05, then we can conclude that the difference is not significant and the data can provide evidence that the coefficients are the same.

III. Discussions and Limitations

Since the R^2 we calculated in Part II of the two neighbourhoods are approximately the same and their estimated slopes are also approximately the same, we can find out that there is not a subdivision into 2 different models by neighbourhoods. Therefore, we use the overall model for the combined neighbourhoods and asses the residuals by using the Residuals vs Fitted graph and the Normal Q-Q Plot.





Residual plot

From the residual plot we made above, we can find out that the residuals are equally spread around a horizontal line which indicates that the model provides an adequate summary of the data. Besides, those residuals are correlated non-linearly with the fitted values which suggests that there is a linear trend between sale price and list price based on the combined neighbourhoods.

Normal Q-Q plot

The normal QQ plot shows if residuals are normally distributed. From the normal QQ plot we made above, we can find out that most of the residuals are lined well on the straight dashed line which means that the model we fitted is good and the residuals are normally distributed. Besides, from the plot, we can also find out that the response is right-skewed.

Two potential numeric predictors

1. The size of the house: As we all known that, the larger the house, the higher the sale price.
2. The number of rooms of the house: If a house has lots of rooms, the sale price might be higher.

Therefore, we can use the size and the number of rooms of the house to fit a multiple linear regressions for sale price.