SLR Analysis of List price and Sale price of properties in GTA (.)

JW0935 (.)

October 24, 2020 (.)

I. Exploratory Data Analysis Part 1 (.)

Part 1 (.)

In 2020, people are facing a huge global pandemic, COVID-19, where Toronto is one of the most affected cities in Canada. The real estate in GTA has been seriously affected, and the price of detached houses is at the highest level in history. First-time home buying has become the main problem for the federal government. This leads us to investigate into the sale price of detached houses. Using the dataset from the Toronto Real Estate Board (TREB) on detached houses, we build SLR models to predict the sale price of properties by list price in two separate neighbourhoods, Toronto and Mississauga, to help buyers predict the expected sale price of detached houses.

Our data is obtained by randomly selection from the dataset of TREB, including 200 observations and 5 variables: id, sold, list, taxes and location.

We create 3 scatter plots of sold price and two explanatory variables:

Fig.1.

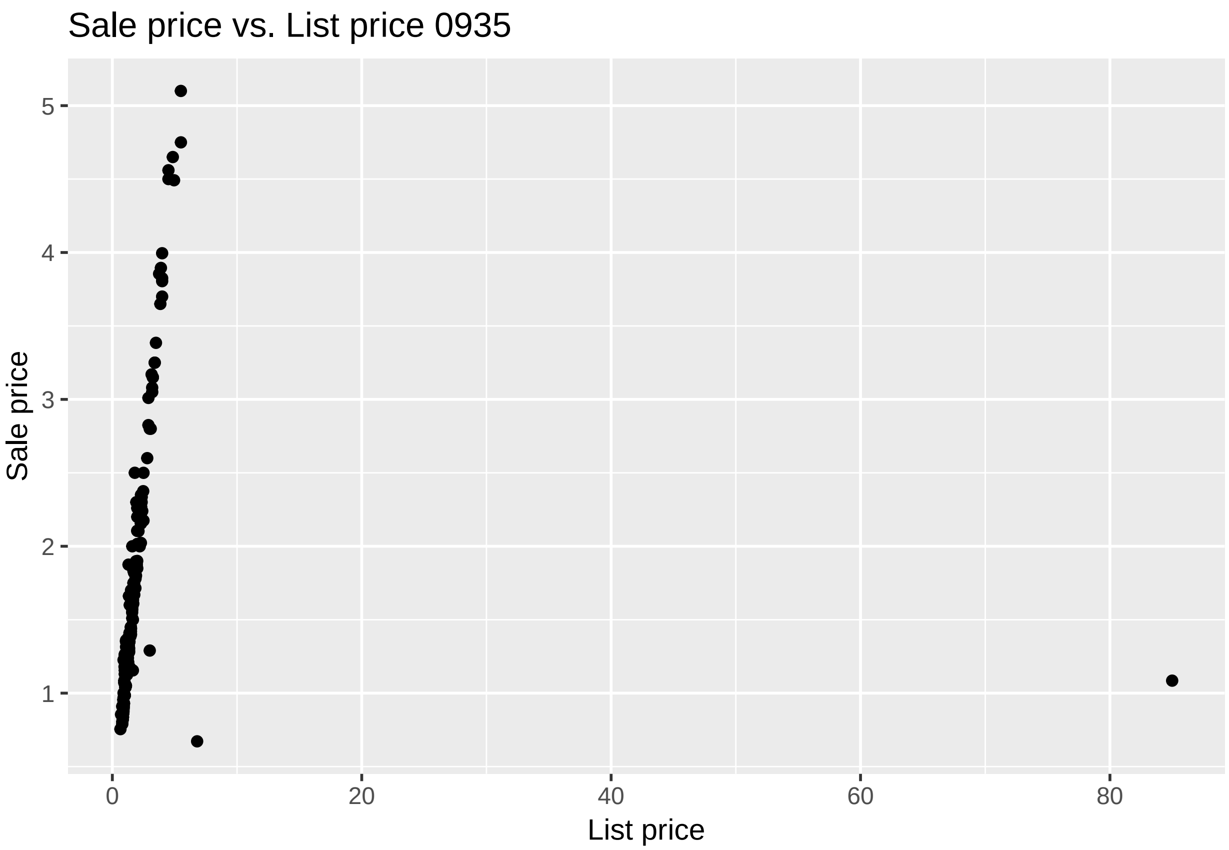


Figure 1 shows the relationship between list price and sale price of properties. We can clearly see a linear relationship between list price and sale price. We also observe some outliers and leverage points. There is a point with high residual and high leverage at approximately list price of 83 millions of dollars, and another high leverage and residual point with around list price of 7 millions of dollars. These two points are removed from our sample data because they are considered as influential points that may drastically influence the slope or intercepts of our SLR model.

Fig.2.

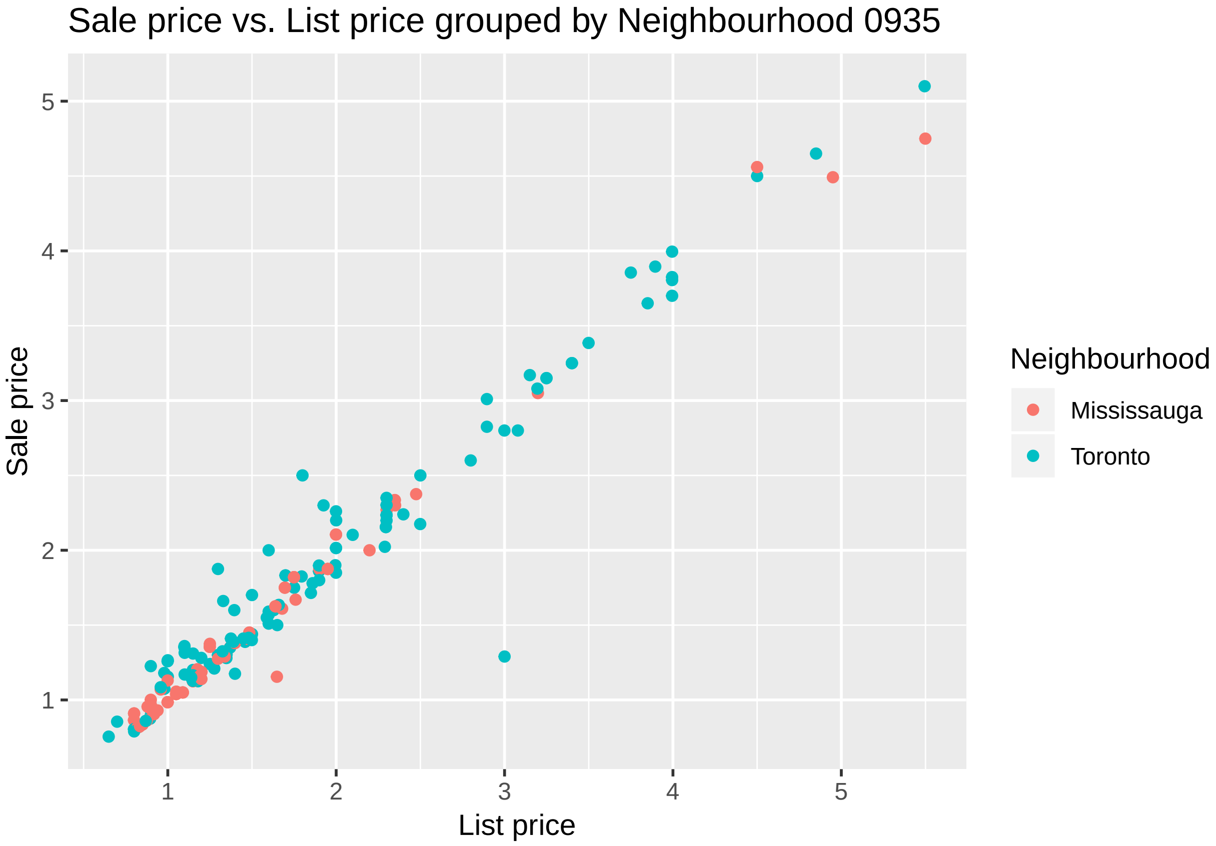
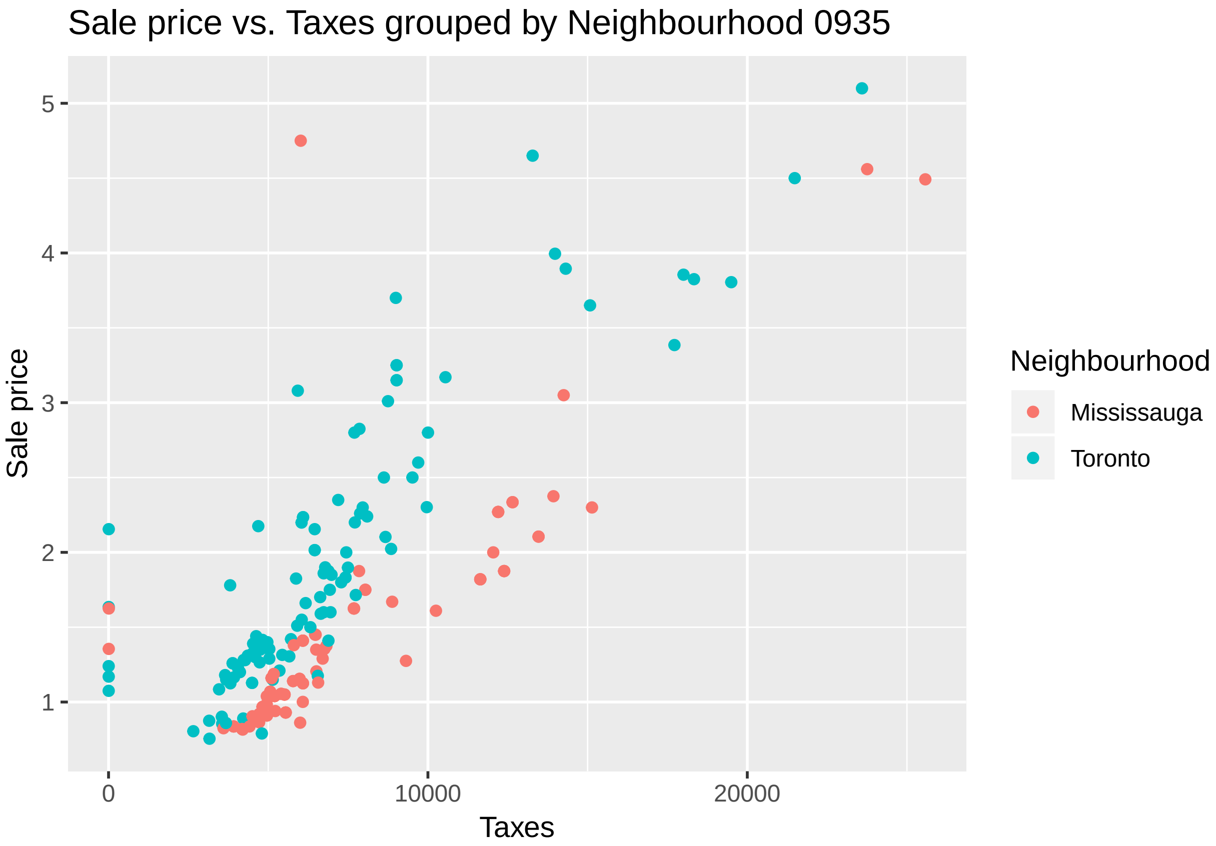


Figure 2 shows the relationship between list price and sale price of properties grouped by location after the deletion of two outliers. We see that properties in neighbourhood Mississauga and Neighbourhood Toronto are almost distributed together, which means the relationship between list price and sale price of properties has little to do with location. Besides, properties in neighbourhood Toronto has a large variability because they distributed more apart than properties in Mississauga. There is still an outlier of list price of 3 millions of dollars in Toronto, but it may not substantially affect the slope because it has high residual but low leverage.

Fig.3.



From figure 3, we see the relationship between taxes and sale price of properties grouped by neighbourhood. Properties in Mississauga are almost distributed below properties in Toronto, meaning for the same taxes, sale price of properties in Toronto is consistently higher than properties in Mississauga. Thus, the relationship between taxes and sale price of properties is different in each neighbourhood.

II. Methods and Model

Part 2 (.)

We continue our study by building SLR models, predicting sale price of properties using list price as predictor for all neighbourhoods with formula:

$Sale\hat{P}rice = 0.1422 + 0.9111 ListPrice$, where $Sale\hat{P}rice$ is the expected sale price of properties for all neighbourhoods, $ListPrice$ is the list price of properties for all neighbourhoods.

SLR model for neighbourhood Mississauga with formula:

$Sale\hat{P}riceM = 0.1127 + 0.9096 ListPriceM$, where $Sale\hat{P}riceM$ is the expected sale price of properties for neighbourhood Mississauga, $ListPriceM$ is the list price of properties for neighbourhood Mississauga

SLR model for neighbourhood Toronto with formula:

$Sale\hat{P}riceT = 0.1873 + 0.9007 ListPriceT$, where $Sale\hat{P}riceT$ is the expected sale price of properties for neighbourhood Toronto, $ListPriceT$ is the list price of properties for neighbourhood Toronto

We create a table for the results of 3 models (All rounded to 4 decimal places):

Table (.)

Value\Location	All neighbourhoods	Mississauga	Toronto
R^2	0.9609	0.9843	0.9459
Estimated intercept	0.1422	0.1127	0.1873
Estimated slope	0.9111	0.9096	0.9007
Estimated variance of the error	0.0311	0.0093	0.0471
P-value of testing H_0: slope is 0	6.9463e-140	1.9834e-78	7.1317e-71
95% CI for slope	(0.8852, 0.9370)	(0.8848, 0.9344)	(0.8599, 0.9416)

Table includs R^2, estimated intercept, estimated slope, estimated variance of error, p-value for testing the null hypothesis that slope is 0, and 95% confidence interval for slope for all of our 3 models.

We find that about 96.09% of total variation in sale price for properties in all neighbourhoods can be explained by list price, about 98.43% of total variation in sale price in neighbourhood Mississauga can be explained by list price, and about 94.59% of total variation in sale price in neighbourhood Toronto can be explained. R^2 of 3 models are all large and similar because figure 1 & 2 show the strong linear relationship between sale price of properties and list price for no matter in all neighbourhoods or Mississauga or Toronto, and the pattern is also very close, where only R^2 of model for neighbourhood Toronto is slightly smaller because there is still an outlier and a relatively larger variability that we discussed.

When considering whether conduct a pooled two-sampled t-test, we need to first satisfy the assumption of independency between two samples. We find that Mississauga is close to Toronto and in the range of GTA, so that the sale price of properties in Toronto may affect sale price in Mississauga, so two groups are not independent. Besides, we need to check whether two populations have the same variance. Since population variance is unknow, we can use the sample variance to estimate the population variance, where the estimated variance of Mississauga is 0.0093 and the estimated variance of Toronto is 0.0471, and are not the same, so the assumption of same population variance for two-sampled test is invalid. Thus, we cannot use the pooled two-sampled test in this case.

III. Discussions and Limitations

Part 3 (.)

We choose the model of predicting sale price of properties by list price for all neighbourhoods as our final model, and with formula: $Sale\hat{P}rice = 0.1422 + 0.9111 ListPrice$

Based on figure 1 & 2 and our fitted models, we can see the pattern of model is similar whether our sample is grouped or not, meaning there is no such difference between Mississauga and Toronto in the relationship of list price and sale price, so the model of two neighbourhoods together can explain the expected sale price well.

When testing SLR assumptions, we assume errors have normal distribution with mean 0 and variance σ^2, which is in the true model and we cannot observe, so we use residuals to estimate errors, and make inference of our model by observing residuals. We assume the errors are uncorrelated because our sample items are drawn randomly and independently.

We create residual plots:

Fig.4.

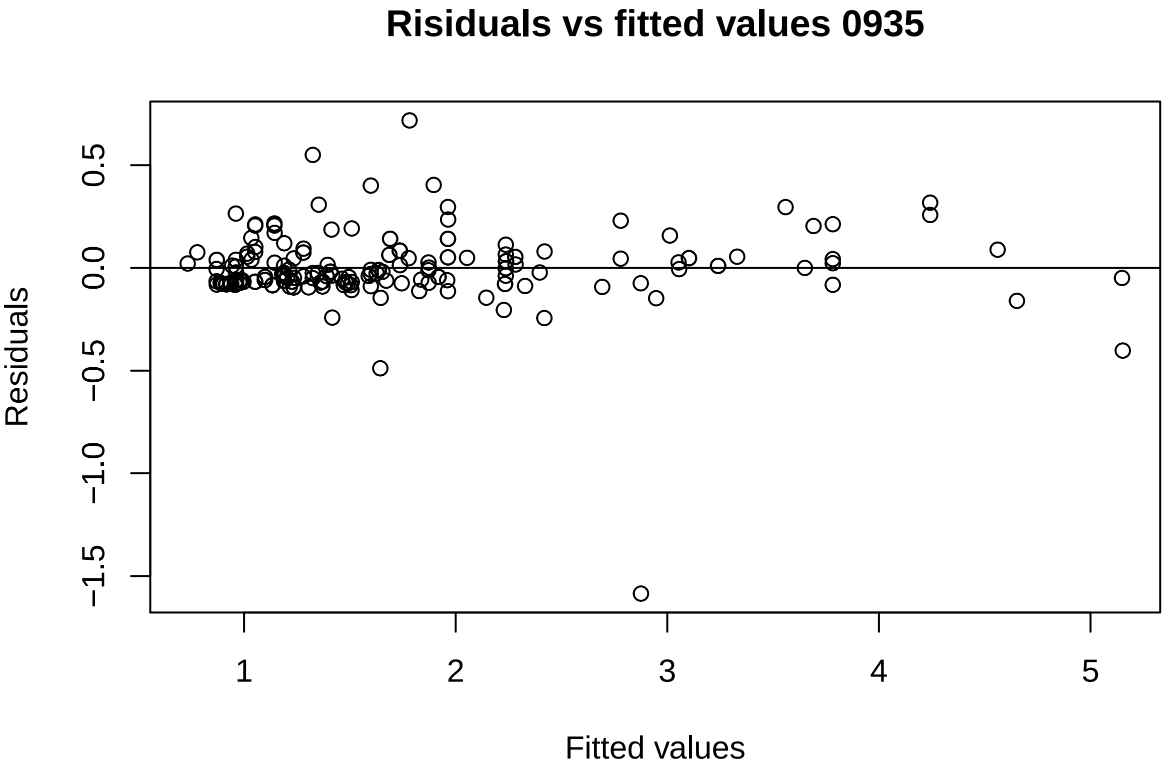
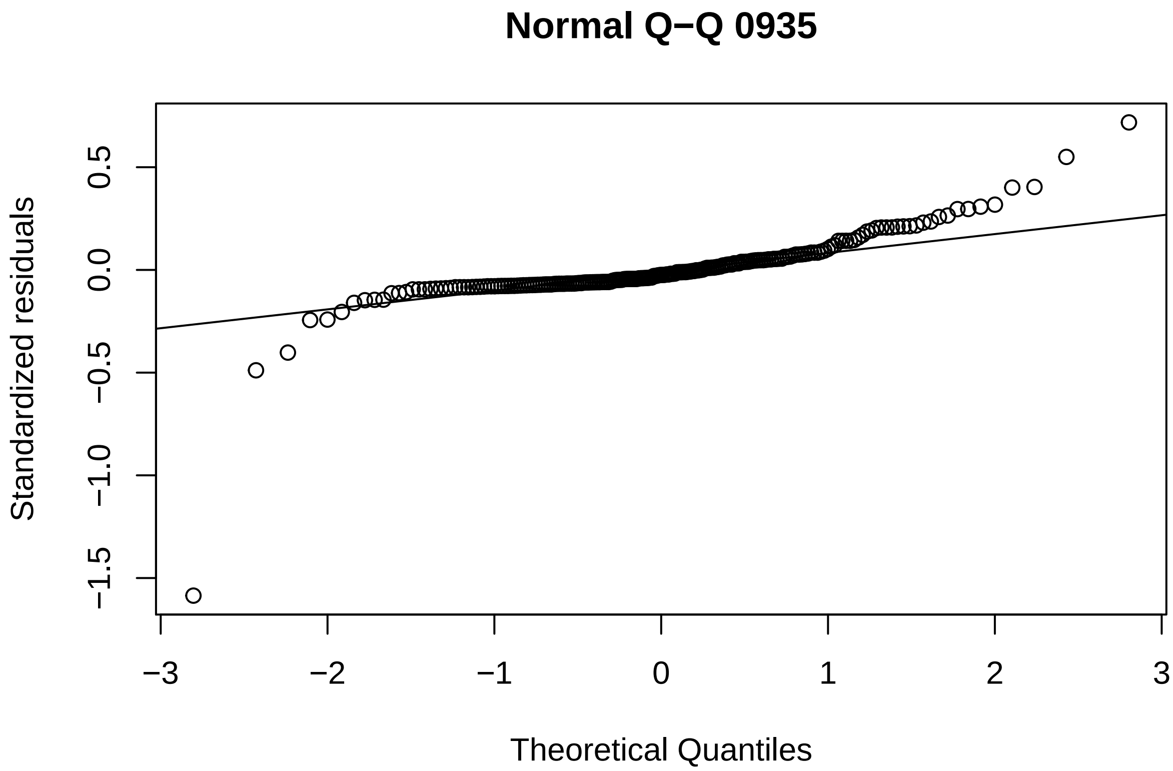


Figure 4 is the residual scatterplot, showing the relationship between fitted values and residuals. We can see residuals have mean 0, and there is no pattern between fitted values and residuals, so our SLR model is appropriate. Besides, we see that the all of points are almost randomly distributed along the line, so the assumption of homoscedasticity holds.

Fig.5.



From the normal qq-plot showed in figure 5, we see it is heavy-tailed, and residuals are not evenly distributed along the line, so it is not normally distributed, violating the model assumption of normality.

```
##
##  Shapiro-Wilk normality test
##
## data:  residual
## W = 0.68629, p-value < 2.2e-16
```

Besides, we conduct a Shapiro-Wilk test, with H_0: residuals are normal, where the p-value is less than 2.2e-16 and less than significance level of 5%, so we reject the null hypothesis, which means the residuals are not normal, proving our finding.

By our study, we know the influence of list price of properties on sale price. However, there are other predictors to predict sale price. For example, the size of the house (in sq ft), generally, larger houses have higher sale price. Besides, the age of the houses can also be used to predict the sale price, as we know the younger the house, the higher the price is.