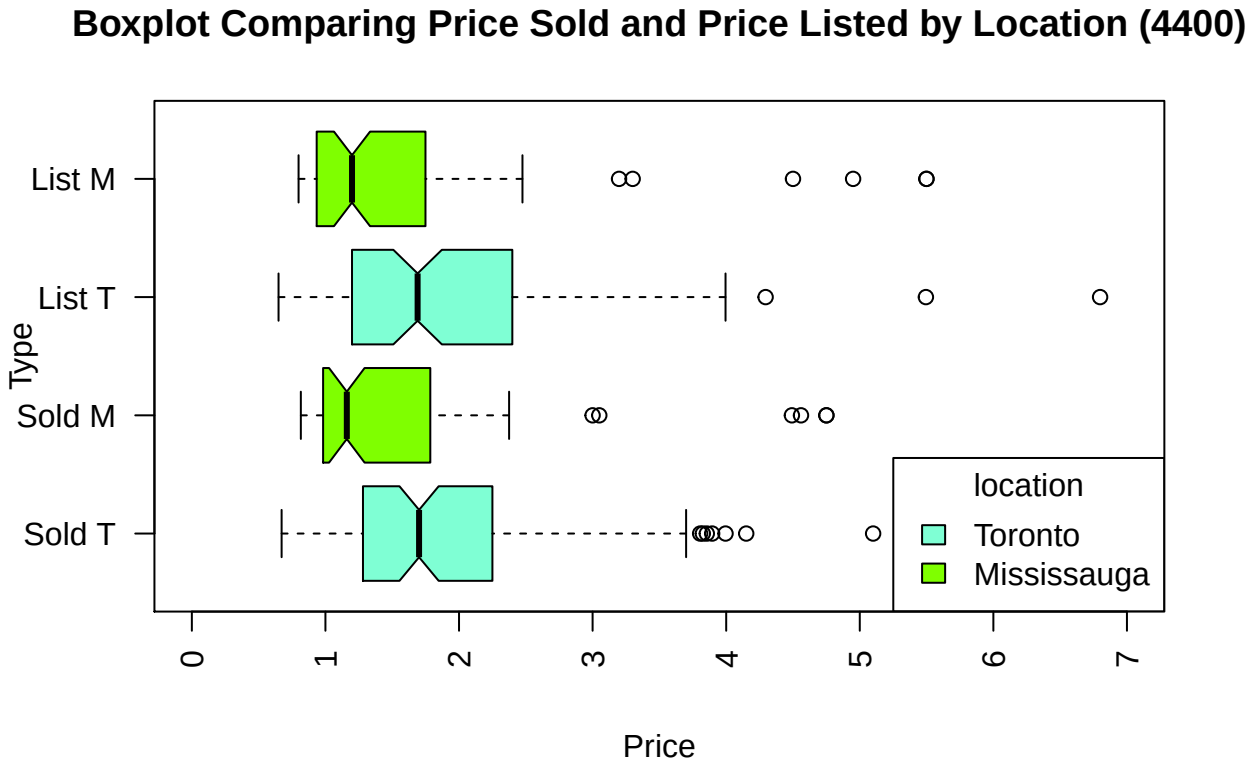
Toronto and Missauga Housing Sale Data Analysis

JS4400

October 24, 2020

I. Exploratory Data Analysis



Description of the Data I created a box and whisker plot to describe the housing data. The data consists of the sale price, list price, last years property taxes and the location(M = Mississauga, T = Toronto) of 200 houses sold in the Mississauga and Toronto area after I did a sample. The data was obtained from the Toronto Real Estate Board (TREB) on detached houses.

The box and whisker plot show that the houses in Toronto are listed and sold at higher prices than the houses in Mississauga on average. There is a considerable amount of data outside the range of whiskers meaning there could be outliers which I will examine further in the variable removal section. All the points outside the whiskers were to the right of the data meaning that houses were likely to be listed and sold for high prices far from the median rather than low prices far from the median. There is one piece of data that is not in the range of the graph that has a listed price of almost 85 million that can not be seen in the box and whisker plot. When looking at the data the majority of listed prices are slightly greater than the sold price which

makes since as people will negotiate and find price agreeable for the buyer and seller but some had lower list price than sold price.

Variable Removal Reasoning Point id 112 sold for 1.085000 million, was listed for 84.99000 million, paid 4457.000 in taxes and was in Toronto

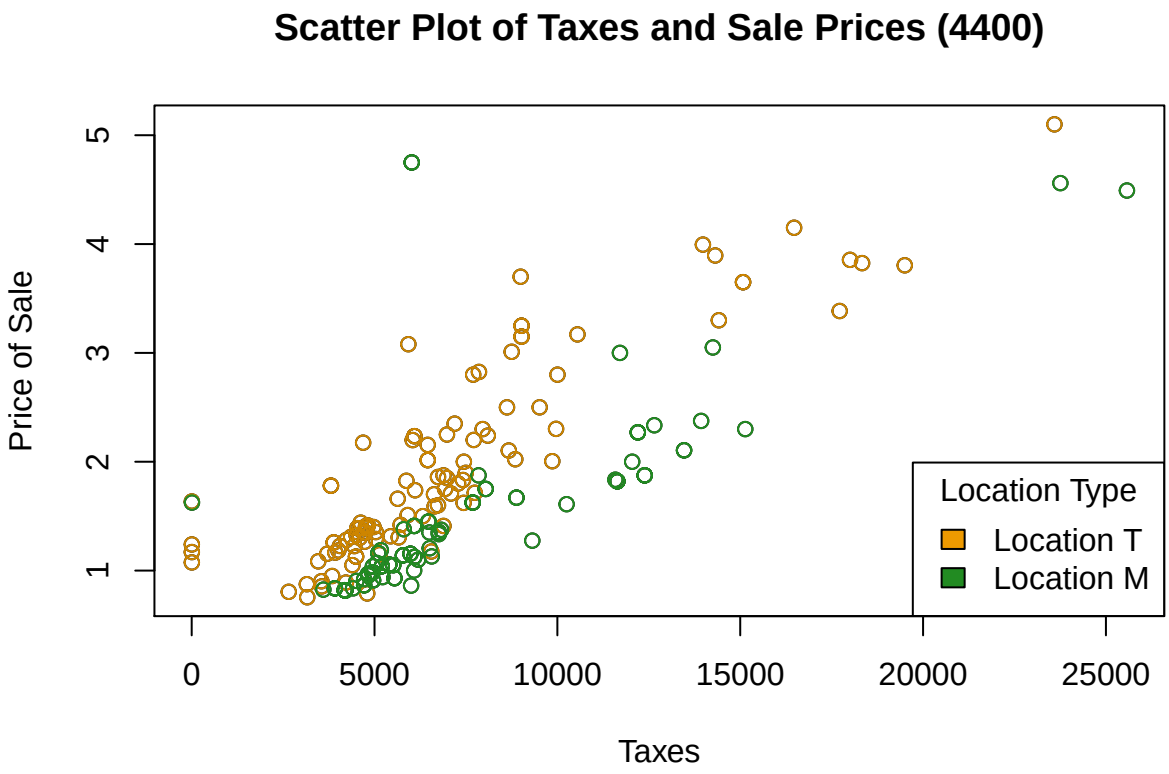
leverage of point 112: 0.9694499

standardizes residual value of point 112: -12.5213714

There are 200 pieces of data in the model so I can verify it's high leverage using the formula leverage $> 4/n$, in this case $4/n = 0.02$ meaning its high leverage. Now to determine if the point has good or bad leverage we use the standardized residual to see if it's greater than 4. In this case the value is over 12 clearly resulting in an influential point. Of course I can not remove the point just because its influential, there needs to be a reason to remove it. I'm removing point 112 because the listed price was clearly overvalued at almost 85 million as the house sold for 1.085 million, almost all points had reasonable list prices.

Point id 95 sold for 0.672000 million, was listed for 6.79900 million, paid 2577.000 in taxes and was in Toronto

The reason I choose to remove point 95 from the data set is that the list price was the second highest overall yet it sold for the lowest amount of the 200 points, a list price of almost 7 million but only selling for under 700k. This point will have a negative effect on the linear model and does not fit the rest of the data, its the only other point other than 112 to have a drastic list price difference from sold price.



Major Highlight's from each plot The scatter plot of list prices and sale prices shows that there looks to be a strong linear trend with list price and sale price. There are only a few points on the graph that look like they could be outliers after I removed the two points that did not fit with the rest of the model. This will likely result in a high value of correlation coefficient.

The scatter plot on the amount of taxes paid and the sale price shows more variability in its plot but still looks to have a linear relationship. This would make sense as property taxes are based on current value of the house determined by the local government. The one big difference between the people in Toronto and Mississauga would be that it looks like people in Toronto pay more in property taxes compared to people in Mississauga.

The graph of the box and whisker plot showcasing the data shows that Toronto seems to have similar sized upper and lower quartile ranges which is a measure of variability in the data while Mississauga looks to have a larger upper quartile compared to the lower quartile. The interquartile range of Mississauga is smaller than Toronto and both have similar looking quartiles for their list price and sold price.

II. Methods and Model

Result of Regression Table

	R^2	Intercept	Slope	Variance of Error	p-value	95% Slope Confidence Interval
All Data	0.9625168	0.1612844	0.8975438	0.0308872	9.988e-142	(0.8725932 0.9224943)
M Location	0.9859434	0.1406437	0.8874720	0.0100391	3.233e-84	(0.8651534, 0.9097906)

	R^2	Intercept	Slope	Variance of Error	p-value	95% Slope Confidence Interval
T Location	0.9440773	0.2030901	0.8919042	0.0469431	1.4156e-67	(0.8498997, 0.9339088)

R^2 Interpretaion The R^2 value represents the coefficient of determination of the linear model. The linear model with all of the data has a R^2 value of 0.9625, the model with just Mississauga has an R^2 value of 0.9859 and the Toronto model has an R^2 value of 0.9441. All of the models have a similar R^2 value. The coefficient of determination represents the percentage of variation in Y's explained by the regression line. They appear similar as all of data sets have data that is close to the linear regression line.

Pooled Two Sample t-test discusion Normal distribution of the estimates, independence and the same variance are the three conditions required to conduct a pooled two sample t-test on the slope parameters of Toronto and Mississauga. If these conditions hold a two sample t-test could be completed. By assumption the data sets are independent and follow a normal distribution, this is Because the residuals and the slope estimators data sets are linear functions of Yi, which are assumed to be normally distributed. Now I just have to show they have the same variance.

The variance of the Mississauga Only model slope is 0.0001261676

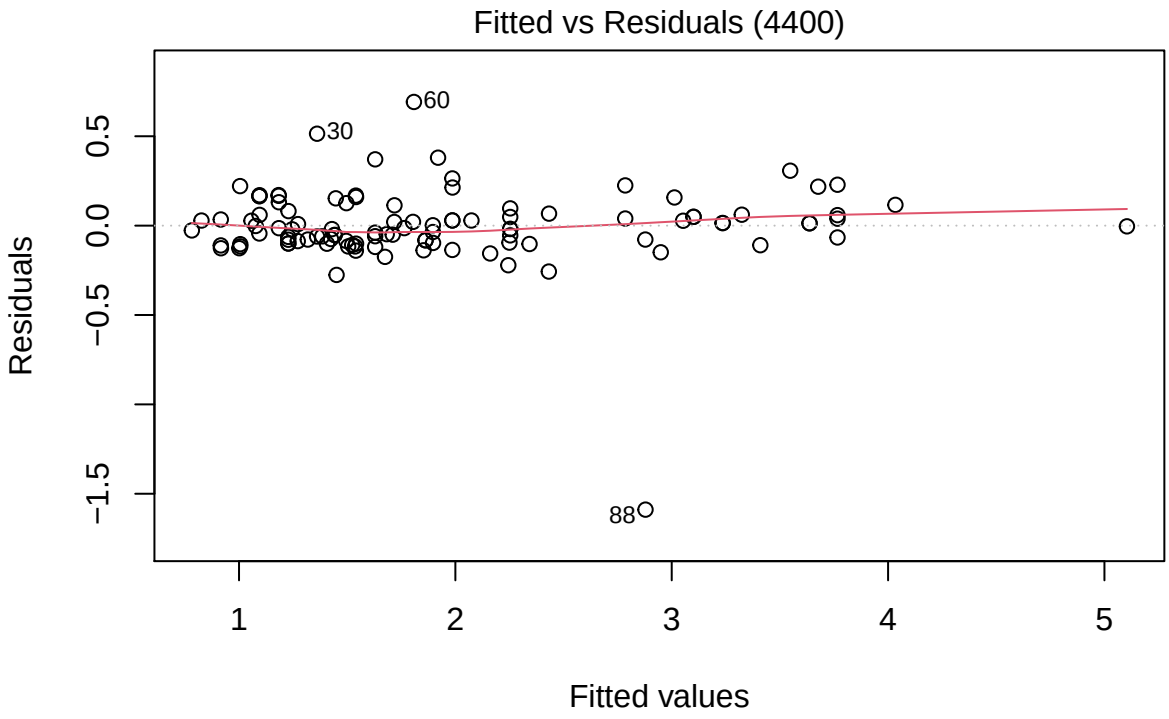
The slope variance of the Toronto model is 0.0004487741

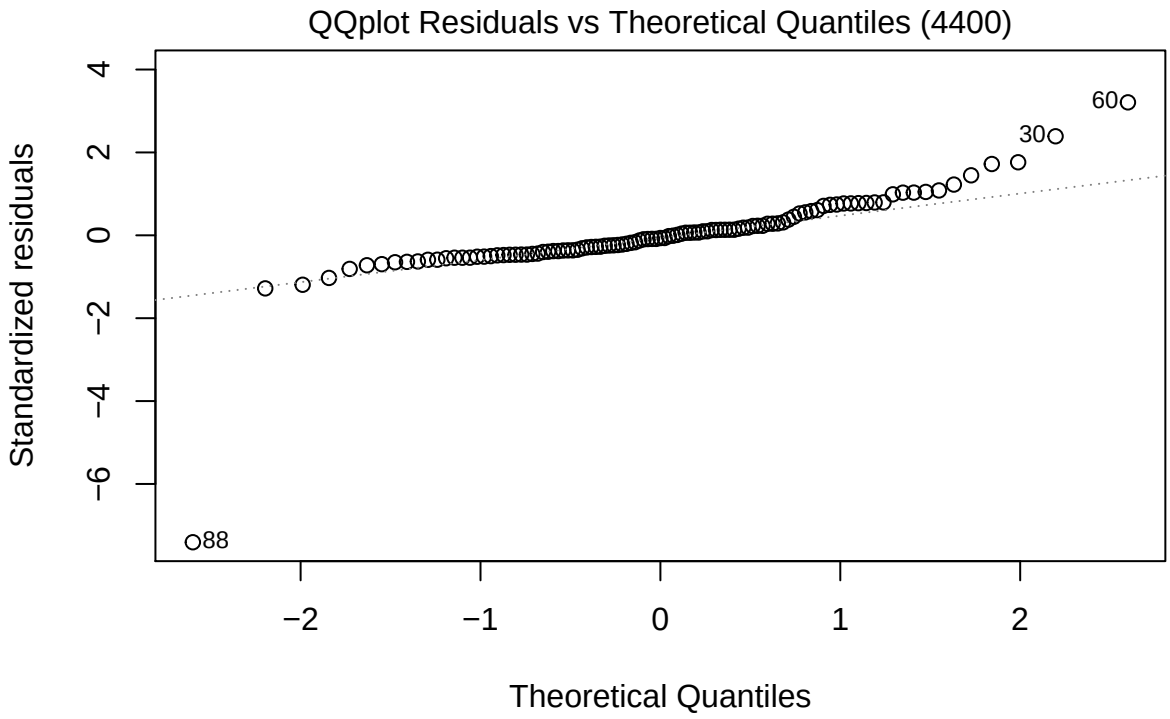
In the sample that I was given the variances of the slope estimators are not exactly equal to each other as Mississauga has a slope variance of 0.0001261676 and Toronto has a 0.0004487741 slope variance but this is a difference of only 0.000322607. The variances are close enough that the same variance condition holds, thus we can conduct the two pooled t-test as it meets the required conditions.

III. Discussions and Limitations

Explanation of my choice The fitted model that I'm choosing is the Toronto data model. I'm choosing the Toronto model as it has a high R^2 value, a low variance of the error term and looks like it will fit the linear model assumptions. The Toronto model looks to have the low variation in the scatter plot shown in part 1 as well so it might be the most likely to fit a linear model. After removing the extreme outliers from the set at the beginning I believe that the Toronto model will fit the assumptions of need for simple linear regression.

Violation of normal SLR assumption disscusion To verify if there are any violations of the normal error SLR assumptions the model needs to have a normally distributed error term with constant variance. Plotting the data against a normal probability plot will show the goodness of fit as a normal distributions and the fitted vs residuals plot will show if the model error term has constant variance.


Fitted vs Residuals (4400)
Residuals
Fitted values
60
30
88
0.5
0.0
−0.5
−1.5
1
2
3
4
5



In the normal qqplot graph there exists small deviation from the normal line at the tails of the plot but the majority of the points fall near the fitted line. With the majority of the data following the normal line and only small amounts of deviation I can conclude the error terms follow a normal curve. Examining the fitted vs residual plot there looks to be constant variance as no clear trend exists in the graph. The normal curve of the error term is centered at zero and has a mean of zero, thus it follows the normal slr assumptions.

Two potential Numeric Predictors on the Model Two other potential numeric predictors that could be used to model sold price are the square footage of the house and the distance to the nearest public transportation (ex. ttc). Square footage is really valuable as it determines the size of the house and the larger the home I theorize that the more money it will cost, size of living space will always be a factor in the price of a home. As for the distance to the nearest public transportation stop, this is important as people want to be close for convenience purposes and to lower the time of their commute.