The analysis of the sale price of single-family, detached houses in two regions in the Greater Toronto Area

Linxia Li (1005715488)

December 22, 2020

Repo: Code and data supporting this analysis is available at: https://github.com/Helen1024/STA304FinalProject

Abstract

In this report, I am interested in analyzing how physical condition, the list price and taxes and the geographical location of the single-family, detached house can affect the response variable the sale price as the sale price is a good indicator to reflect the economic situations of a city or a country and assist people who would like to purchase a house to predict the housing price in the future. By making assumptions and fitting a multiple linear regression model, it can be found that there exist positive correlations between the explanatory variables selected and the response variable. To illustrate, in the financial aspect (the list price and taxes), along with the increase in the list price and taxes of the property, the sale price of tend to increase. Similarly, in the geographical location aspect, a good geographical location of a house enhances the sale price.

Keywords

Sale Price, Single-Family, Detached Houses, Predictors: The number of bathrooms, List Price, Property Taxes, Geographical Location, Multiple Linear Regression Model

Introduction

Nowadays, house price is one of the major concerns for almost everyone and first-time buying is currently a major federal issue. We have known that house price is an appropriate indicator of both the overall market condition and the economic health of a city and a country. Besides, the house price might have been high during the current COVID-19 period. Therefore, anyone who would like to buy a house should estimate the price of the house in advance by some key features such as the geographical location of the house, the property taxes of the house, the size of the house and so on.

In this report, I focus on the housing price (sale price) of single-family, detached houses in two regions in the Greater Toronto Area which was recorded in a dataset called "HousingPrice_data" which is an observational data. The goal of this report is to analyze and predict the house price in two regions which are Toronto and Mississauga in the Greater Toronto Area by using a number of factors (predictors). The prediction of house price is quite useful and it is expected to assist people who plan to buy a house so that they should know the price range in the future and then they can plan their finance well. In addition, house price predictions are also quite beneficial for investors to master the trend of housing prices in a certain location. Therefore, choosing this variable is consistent with the goal mentioned above.

I will explore how the house price can be affected by three of the major aspects: the physical conditions, the last list price and the tax of the property and the location. The descriptions are as follows:

1. **Physical conditions:** Physical conditions are properties possessed by a house that can be observed by human senses, including the size of the house, the number of bedrooms (bathrooms), the availability of kitchen and garage (the number of parking spots), the area of the land and buildings and the age of the house (old, medium and new).
2. **List price and the tax of the property:** The last list price of the property is the suggested gross sale price of real estate property when it is put on the market. In other words, it is the price at which the seller determines to market the property. It is quite different from the actual sale price which will be analyzed and predicted in this report. The actual sale price is the price at which the seller and buyer agree and it is typically based on the sale price of comparable properties in the area. Property tax is a tax paid on property owned by an individual or other legal corporation which is calculated by a local government where the property is located and paid by the owner of the property.
3. **Geographical location:** Geographical location is another important factor in shaping the price of a house since the location determines the prevailing land price. In addition, the location also determines the ease of access to public facilities, such as schools, campus, hospitals and health centers, as well as some family recreation facilities such as malls and the fitness rooms.

Besides, I identified several features for each aspect. For the physical conditions, I consider the number of bedrooms, bathrooms and parking spots, the maximum square footage of the property, age and garden (with a garden and without a garden) as the representative parts. In case of geographical location, the region (Toronto and Mississauga) and the variable location (Downtown, Urban, Near the urban and suburb) are the chosen features.

We use linear regression to analyze our data and made the following hypotheses:

1. In physical conditions aspect, if the size of a house is bigger and it has more rooms and more facilities, the house price tend to be higher. In other words, the better the physical conditions of a house, the higher the actual sale price.
2. In the list price and the tax of the property aspect, the higher the list price and the tax of the house, the higher the actual sale price tend to be.
3. In geographical location aspect, if a house is in a good and convenient geographical location, the actual sale price of it tend to be higher.

Here is a brief description about each part of the report:

1. **Methodology:** This part includes data section and model section. I will discuss the data, the strengths and weaknesses, the variables of the data and the plot and tables of the raw data in the data section. In addition, I will establish the variables and present the relationships between the variables being used. I will also choose the model and explain it, discuss the features and the final model and do diagnostic checks.
2. **Results:** The results will be presented and explained.
3. **Discussion:** I will summarize what was done earlier, make conclusions and address the weakness and the next steps of the analysis.

Methodology

Data

Introduction of the Data (related to the sale price of the single-family, detached houses)

The data used in this report is called "HousingPrice_data.csv". The goal of this report is to analyze and predict the sale price of single-family, detached house in two areas in Greater Toronto Area which are the cities Toronto and Mississauga so that this data contains the response variable which is the actual sale price and a number of predictors which are used to fit a linear regression model. The data is recorded myself according to some information related to the house price online, especially from the Toronto Real Estate Board (TREB) which is the online information source for comprehensive coverage of real estate listings and services in the Greater Toronto Area. In this dataset, it contains 800 observations and 12 variables. Besides, this data is an observational data since it observes certain variables and would like to explore and determine if there is any correlation between the response variable and the predictors.

The target population (The set of all the units covered by the main objective of the study): All the single-family, detached houses in two regions (Toronto and Mississauga) in the Greater Toronto Area.

Sampling frame (A source material or device/list from which a sample is drawn): The list containing the information of every single-family, detached house in two regions in the Greater Toronto Area which provides sampling units, or access to sampling units.

The frame population: (The set of all units covered by the sampling frame) The frame population is all the targeted population that can be accessible by the sampling frame listed above.

Sample: (The population represented by the survey sample) The randomly selected single-family, detached house in those two areas in the Greater Toronto Area.

Sampling method: In order to obtain the survey sample, the method we can use is called Simple Random Sampling. This method can ensure the equally likely probability of being selected from the population.

Strengths and Weaknesses of the data and the study

Strengths: This data contains the response variable which is the actual sale price of detached house and a number of predictors such as the physical conditions, the property taxes and the geographical location. Based on those variables, we are able to analyze the relationship among different variables and have a relatively accurate inference to support the assumptions. In addition, this data involves many aspects of factors which can influence the sale price of a detached house so that people can have a better understanding of the real estate market in a specific region by analyzing the previous data related to the sale price.

Weaknesses: The predictors in this data can be used to analyze and predict the sale price of the detached house. However, some external factors might have a greater impact on the house price such as the economic fluctuations caused by COVID-19 or financial crisis so that those several basic information (the physical conditions, geographical location) of the house can not be used to predict the selling price of the house in those situations. Besides, this data was recorded myself and the sample size is small so that some data might be inaccurate and can cause bias.

Introduction to the variables in the data

The goal of the analysis is to use a number of factors which can have an influence on the sale price to predict it so that the data contains a response variable which is the sale price of the single-family, detached house in two regions (Toronto and Mississauga) in the Greater Toronto Area. The sale price of the detached house is a discrete variable with the unit Canadian dollars. Besides, there are a number of predictors which are used to analyze and predict the sale price. The variable list represents the last list price of the property in

Canadian dollars and it is a discrete variable. The list price is the suggested gross sale price of real estate property when it is put on the market. The variable bedroom, bathroom, parking represents the number of bedrooms, the number of bathrooms and the number of parking spots in a detached house respectively and they are all discrete variables. The variable maxsqfoot is the maximum square footage of the property representing the size of the detached house which is also a discrete variable. The variable taxes is a tax paid on property owned by an individual or other legal corporation which is calculated by a local government where the property is located and paid by the owner of the property and it is also a discrete variable. The remaining variables are all categorical variables. For the variable Region, T represents the city Toronto and M represents the city Mississauga. For the variable Location: "Downtown" means that the detached house is located in the city center, "Urban" represents that the house is located in the city, "Near Urban" means that the house is located outside the city but close to it and "Suburb" represents that the house is far from the city and located in the suburbs. For the variable Age: "Old" means that the house has been built for quite a long time (about 20 years or more), "Medium" means that the house has been built for a few years and "New" means that the house has just been built or completed less than five years. For the variable Garden: "Yes" represents the house has a garden and "No" represents the house does not have a garden.

Presenting the data

Here is the first few rows of the data I will use. (Table 1)

Table 1: Table 1 - First few rows of data

ID	sale	list	bedroom	bathroom	parking	maxsqfoot	taxes	Region	Location	Age	Garden
1	1265000	999900	2	2	1	3300	4732	T	Near urban	Medium	Yes
2	2200000	1999900	5	3	3	3600	7712	T	Urban	New	No
3	1225000	1169000	5	3	2	3000	4448	T	Suburb	Old	No
4	1900000	1995000	5	4	2	3100	6783	T	Urban	New	Yes
5	1622000	1450000	3	2	0	2800	7436	T	Near urban	Medium	Yes
6	1180000	979000	3	2	1	3300	3650	T	Suburb	Old	No

Here are two summary table about the numerical (Table 2) and categorical (Table 3) data that we will use.

Table 2: Table 2 - Summary of numerical data

ID	sale	list	bedroom	bathroom	parking	maxsqfoot	taxes
Min. : 1.0	Min. : 672000	Min. : 649000	Min. :1.000	Min. :1.000	Min. : 0.000	Min. : 1400	Min. : 2178
1st Qu.:200.8	1st Qu.:1140750	1st Qu.:1127750	1st Qu.:3.000	1st Qu.:3.000	1st Qu.: 3.000	1st Qu.: 2300	1st Qu.: 4493
Median :400.5	Median :1415500	Median :1413500	Median :4.000	Median :3.000	Median : 4.000	Median : 2800	Median : 5541
Mean :400.5	Mean :1781622	Mean :1775831	Mean :3.976	Mean :3.326	Mean : 3.978	Mean : 2876	Mean : 6892
3rd Qu.:600.2	3rd Qu.:2107000	3rd Qu.:2098000	3rd Qu.:5.000	3rd Qu.:4.000	3rd Qu.: 5.000	3rd Qu.: 3200	3rd Qu.: 7589
Max. :800.0	Max. :5156000	Max. :5499000	Max. :7.000	Max. :6.000	Max. :10.000	Max. :19000	Max. :25575

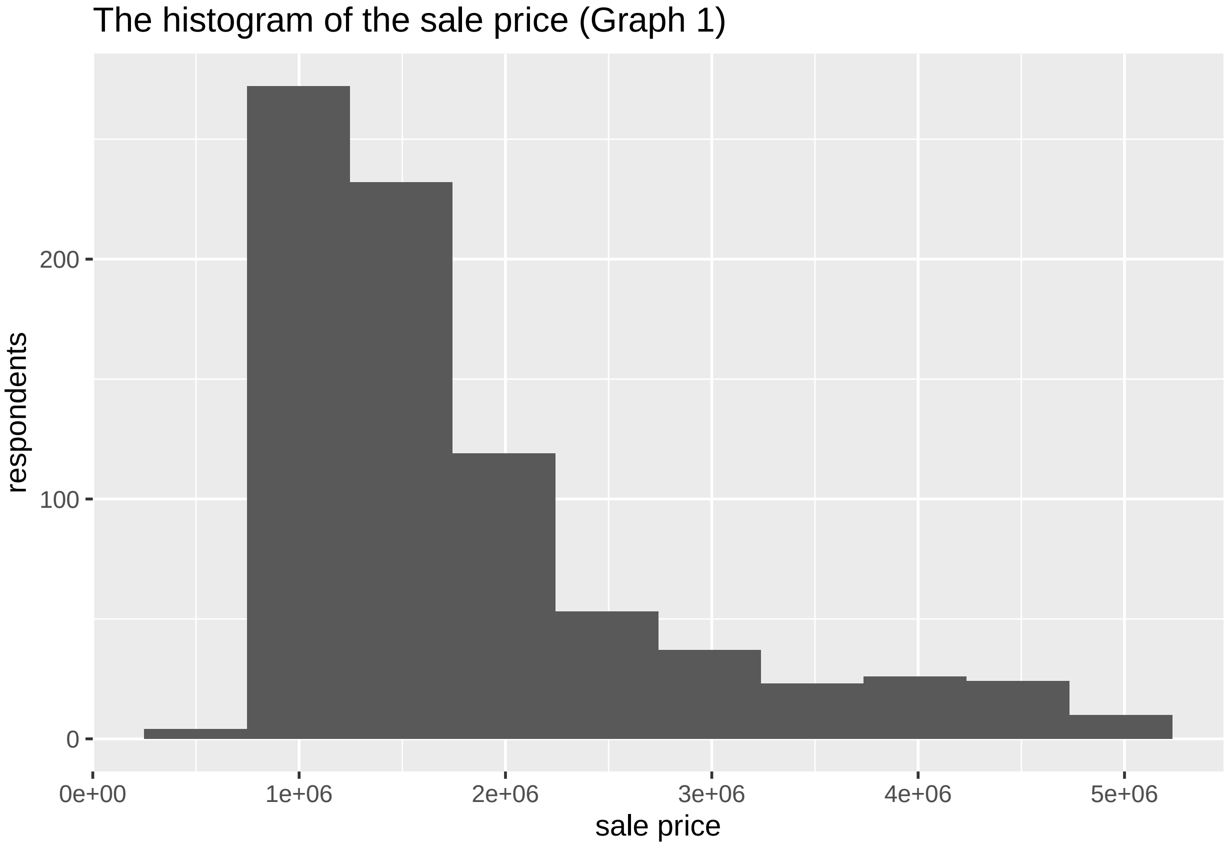
Table 3 - Counting categorical data

Region	M	T		
Quantity of each type	316	484		
Location	Downtown	Near urban	Suburb	Urban
Quantity of each type	202	200	206	192
Age	Medium	New	Old	
Quantity of each type	352	148	300	
Garden	No	Yes		
Quantity of each type	441	359		

Ploting the data

Variable of interest

Here is the histogram for the variable of interest - Sale Price of the detached house (Graph 1)



Inference: From the above histogram of the sale price of the detached house:

1. The distribution is right-skewed, single-peaked. The center of it might be 1500000 Canadian dollars and the sale price of the single-family, detached houses is mostly concentrated between 1000000 Canadian dollars and 2000000 Canadian dollars.

2. There are some outliers in this histogram which are representative for those small number of single-family, detached houses which had a lower selling price or those which had a higher selling price.

Model

Feature selection

Here are the features selected into the dataset.

Response variable: Sale price of the single-family, detached house.

Features: (See Appendix for more about how to select features)

1. Basic information / Physical condition: the number of bathrooms
2. Financial aspect: list, taxes
3. Geographical location: Region, Location

We select features based on the following criterion:

1. Too many NAs: If there are too many NAs in this feature, it should be removed for the convenience of the analysis. Rejecting these features can also avoid biases caused by non-response.
2. Replication: For the multiple linear regression model, each features should be independent. Therefore, the most representative ones are chosen among a certain type.
3. Not correlated with response variable: For the simplicity of our model, I tend to choose the features that are most correlated with the model.

(See Appendix for more about how to select features)

Relationships between variables

Physical condition of the house & Sale price

Below is the summary table for the relationship between sale price and the number of bathrooms (Table 4)

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

Table 4: The summary table for the sale price and the number of bathrooms (Table 4)

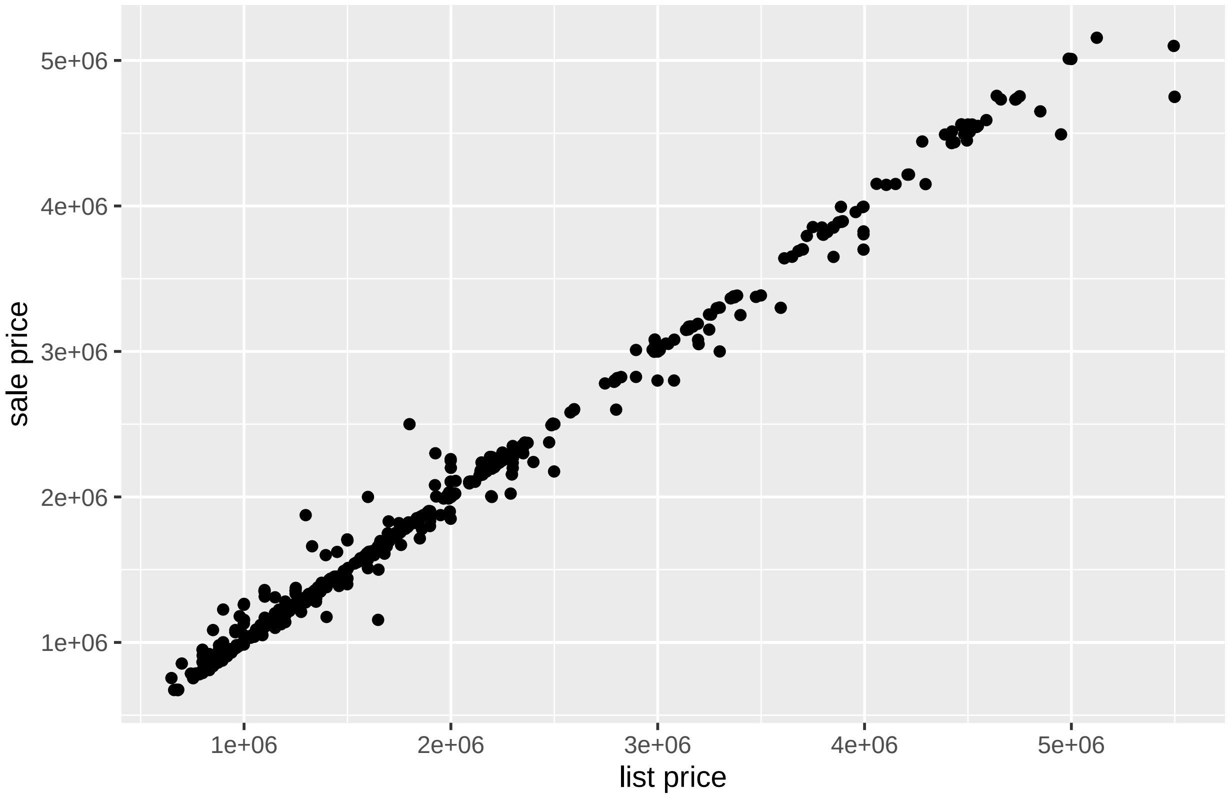
bathroom	mean_sale_price	sd_sale_price	min_sale_price	max_sale_price	median_sale_price
6	3621333.3	1034153.1	2305000	5012000	3338500
5	3045364.4	1050324.5	1423000	5156000	2801000
4	2082912.2	958458.5	854000	4738000	1754000
3	1538783.9	733529.0	787000	4754000	1318000
2	1120223.8	323743.6	672000	2300000	1061000
1	990961.8	335935.7	675000	1875000	825000

Inference:

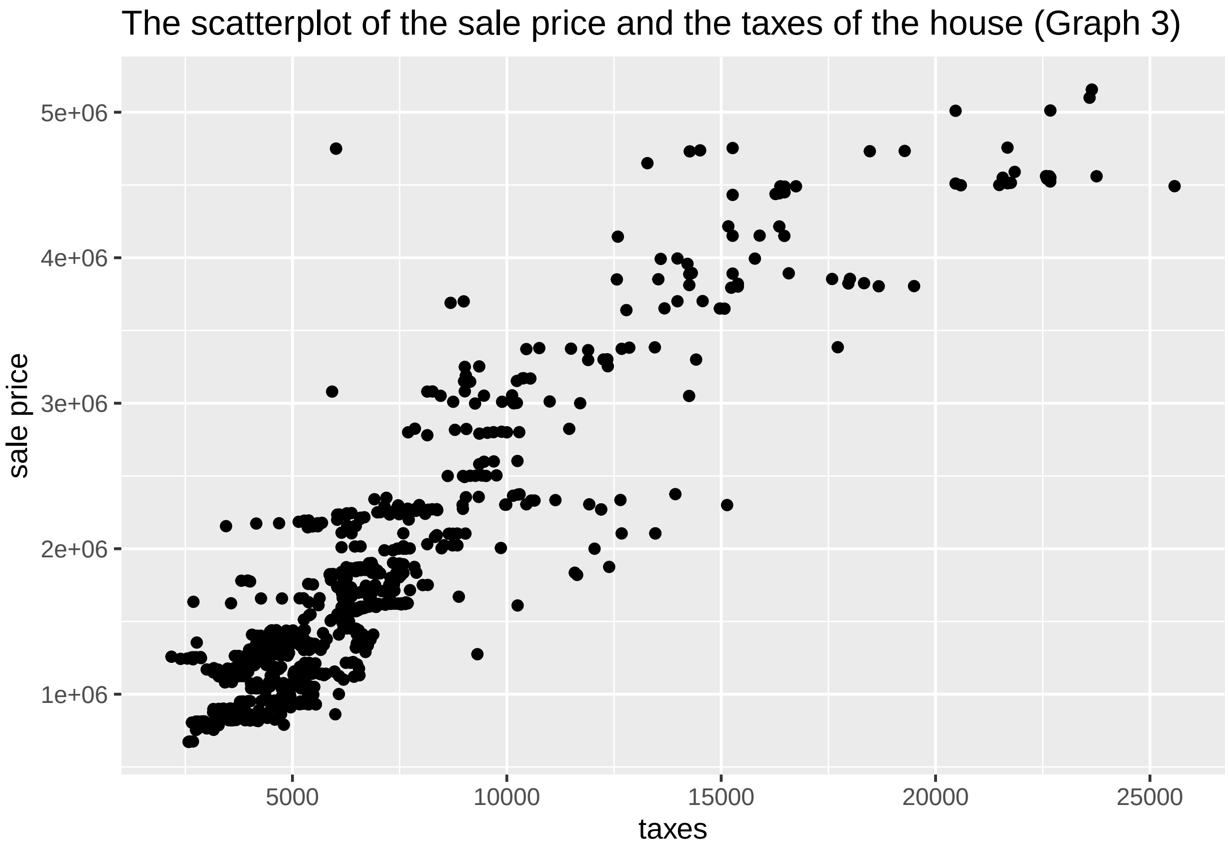
The number of bathrooms: The summary table above shows the mean, median, minimum, maximum and standard deviation of the sale price from different numbers of bathrooms in the house. We have known that mean or median is a measurement of the center of a variable so that I use arrange in R to arrange the mean of sale price of the detached house for different numbers of bathrooms. We can find out that the detached houses which have 6 bathrooms had the highest mean of sale price and those which have only 1 bathroom had the lowest mean of sale price.

Financial aspect of the house & Sale price

Below are the scatterplots for the relationships between sale price and the list price and the property taxes of the detached house



The scatterplot of the sale price and the list price of the house (Graph 2)

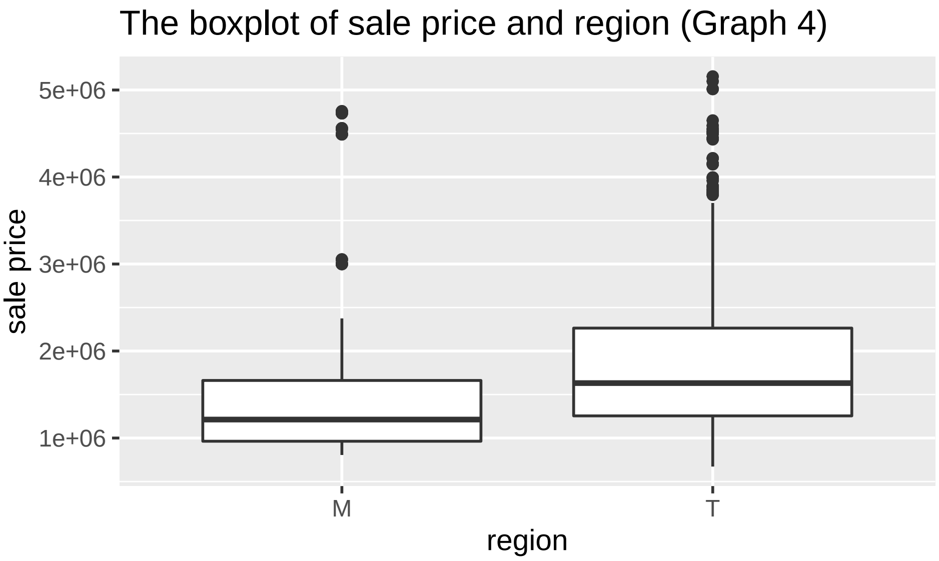


Inference:

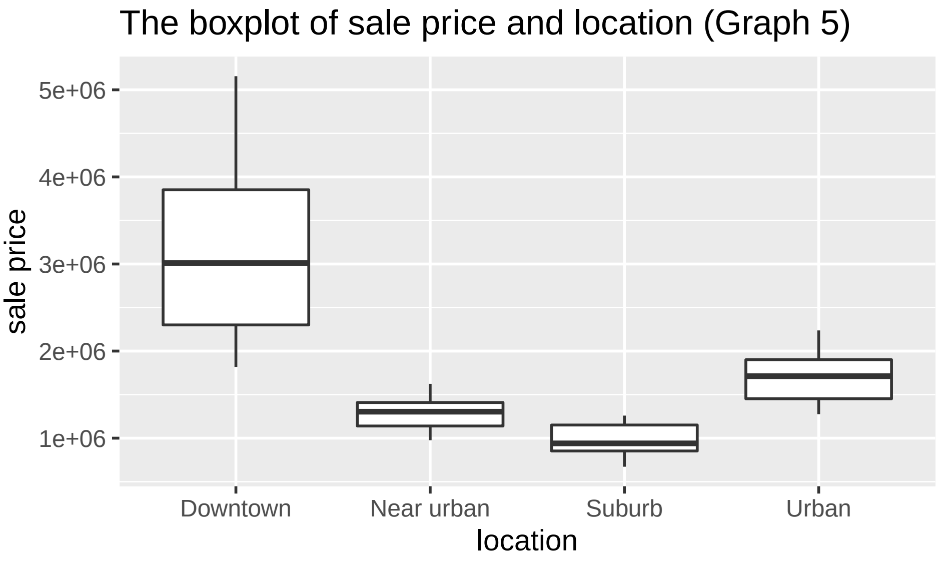
1. List price: From the above scatterplot, we can find out that the scatterplot fitted is approximately a straight line so that there is a strong linear trend between sale price and list price.
2. Taxes: From the above scatterplot, we can find out that there is a relativaly strong linear trend between sale price and taxes.

Geographical location & Sale price

Below are the boxplots between geographical location and the sale price (Graph 4 and Graph 5)



The boxplot of sale price and region (Graph 4)



The boxplot of sale price and location (Graph 5)

Inference:

1. Region:

a. The medians of the boxplots (Mississauga and Toronto) are below the middle so that the distributions of those two boxplots are right-skewed.

b. The IQR (Interquartile range) of the boxplot (for the region Toronto) is the largest. The IQR of the boxplot (for the region Mississauga) is the smallest. However, both of them have some outliers which are representative for those small number of single-family, detached houses which had a higher sale price.

2. Location:

a. The median of the boxplot (Downtown) is in the middle so that the distribution of the boxplot is symmetric. The medians of the boxplots (Near urban and Urban) are above the middle so that the distributions of those boxplots are left-skewed. The median of the boxplot (Surburb) is below the middle so that the distribution of the boxplot is right-skewed.

b. The IQR (Interquartile range) of the boxplot (for the location Downtown) is the largest. The IQR of the boxplot (for the location Urban) is the second largest. The IQR of the boxplots (for the location Near urban and the location Suburb) are approximately the same and the smallest.

Model Identification

We first use five variables such as bathroom, list, taxes, Region and Location to fit a multiple linear regression model with sale price as the response variable.

Discussions of the five independent variables (features)

bathroom: This is a numerical variable taking on any value within 1 and 8. It can not be treated as categorical variable in the model since different numbers of bathroom can have various influence on the sale price of the house. If it is treated as a categorical variable, the predictors and the analysis of the model might be inaccurate.

list and taxes: These are two numerical variables that represent the listed price and taxes of the house. The list price is the suggested gross sale price of real estate property when it is put on the market and taxes is a tax paid on property owned by an individual or other legal corporation which is calculated by a local government where the property is located and paid by the owner of the property.

Region and Location: These are two categorical variables which measure the geographical location of the house. The variables region and location have two and four types respectively. We use region and location rather than region-groups and location-groups since different types of regions and locations can have different influences on the sale price of the house. From the above boxplots, we can find out that different types of regions and locations have various range of the sale prices of the detached houses so that we should use region and location instead of the groups of them.

Overall, we choose these five independent variables and use them in the model to predict the sale price since all of them can influence the sale price of the house. We choose the sale price as the response variable since it can reflect the overall economic situation of a city or a country and the overall standard living of people. Besides, we can use the previous sale price of the house in a specific geographical location to predict the sale price of it in the future. Therefore, we can have a better understanding of the real estate market and economic situation of a city or a country by using it as the response variable and analyzing it.

Reasons to choose multiple linear regression model

1. Multiple linear regression is a regression model that estimates the relationship between a quantitative dependent variable and two or more independent variables using a straight line. Here, bathroom, parking, list, taxes, Region and Location are six independent variables and all of them can influence the sale price which is a quantitative dependent variable. Therefore, we can use multiple linear regression model to estimate the relationship between them.
2. For the Bayesian model, it assumes the parameters we would like to explore follow some distributions. However, we did not explore the distributions of those six variables so that we can not use the Bayesian model.
3. For the logistic regression model, it is a statistical model that in its basic form uses a logistic function to model a binary response variable. We use sale price as the response variable which is a numerical variable. If we change our response variable to the binary form 0 and 1, it can only show whether the price of the house is high or low instead of showing the exact and accurate value of it. Therefore, we are not willing to change the response variable to a binary form so that we can not use the logistic regression model.

Our Linear Model:

$$\begin{aligned}\hat{sale} = {} & \beta_0 + \beta_1 bathroom + \beta_2 list + \beta_3 taxes \\ & + \beta_4 RegionT + \beta_5 LocationNearurban + \beta_6 LocationSuburb \\ & + \beta_7 LocationUrban\end{aligned}$$

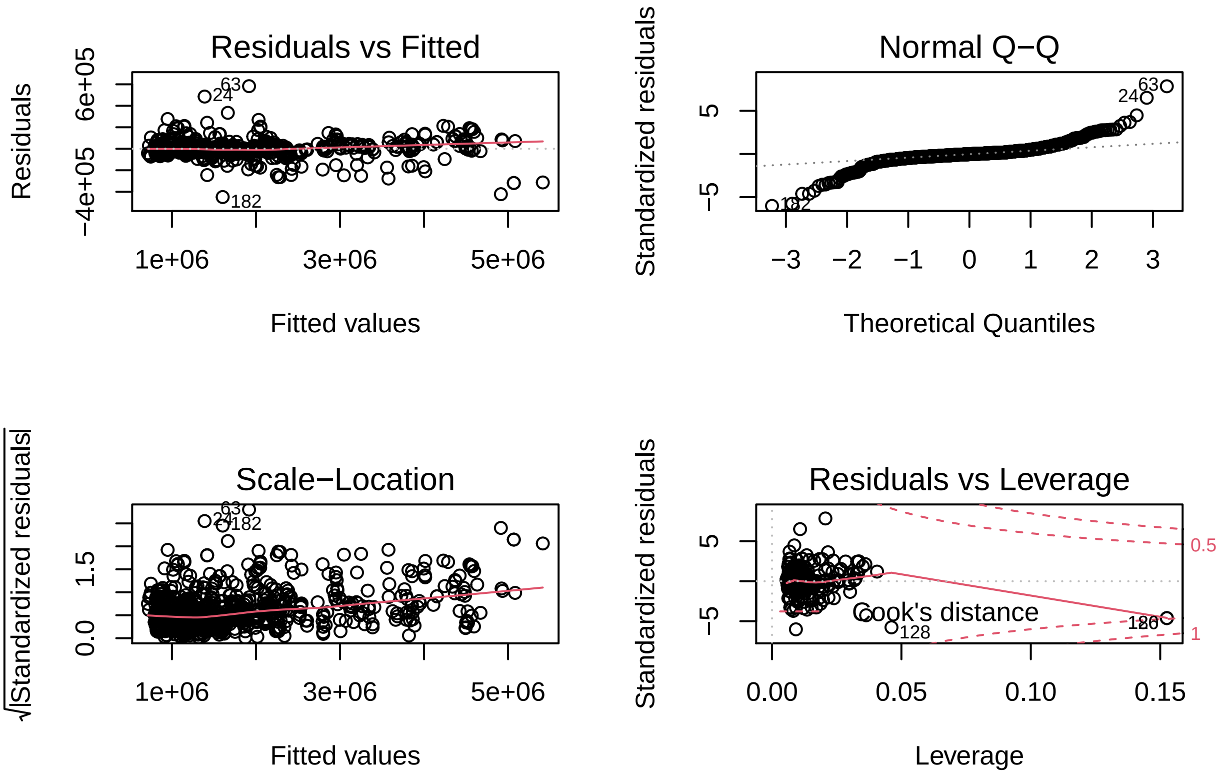
Where:

- β_0: the intercept of the model.
- β_1: holding all other explanatory variables in the model fixed, for a one-unit increase in the number of bathroom, on average sale price changes by β_1 dollars.
- β_2: holding all other explanatory variables in the model fixed, for every one dollar increase in the list price of the house, on average sale price changes by β_2 dollars.
- β_3: holding all other explanatory variables in the model fixed, for every one dollar increase in the taxes of the house, on average sale price changes by β_3 dollars.
- β_4: holding all other explanatory variables in the model fixed, the sale price for the region Toronto is β_4 larger/smaller than that for the region Mississauga.
- $\beta_5, \beta_6, \beta_7$: similar to the interpretation of β_4, but corresponding to different variable which is location.

Model Diagnostics

Below are the linear model diagnostics plots (Graph 6)

Graph - Linear model diagnostics plots



Residuals vs Fitted

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 those five predictors: the number of bathroom, the list price and taxes of the detached house, region and location.

Normal Q-Q plot

The normal QQ plot shows if residuals are normally distributed. We have known that if the residuals are lined well on the straight dashed line, the model we can fit is pretty good. From the normal QQ plot we made above, we can find out that the residuals which are on the bottom left and the top right of this plot are not on the straight dashed line. However, 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.

Scale-Location plot

The scale-location plot shows if residuals are spread equally along the ranges of predictors and can check the assumption of constant variance (homoscedasticity). From the Scale-Location plot we made above, we can find out that the red line is approximately horizontal. Besides, the spread around the red line does not vary with the fitted values which means that the residuals appear randomly spread. Therefore, it is good to see that a horizontal line with equally (randomly) spread points so that the assumption of constant variance holds.

Residuals vs Leverage plot

The residuals vs leverage plot can help us find the influential cases. We have known that not all outliers are influential in the linear regression analysis. Even though they have extreme values , they might not be important to determine a regression line which means that the results (model we would like to fit) would not be much different if we include or exclude them from analysis. However, some outliers can have an influence on the results. If there are some outliers at the upper right corner or at the lower right corner (outside of the dashed line, Cook's distance), the results can be influenced. However, in the residuals vs leverage plot we made, we did not find any points which are at the upper right corner or at the lower right corner (outside of the dashed line, Cook's distance) so that there are no outliers and the model we fit is good.

Results

Model Summary

Here is the summary table for our multiple linear regression model. (Summary table 5)

Summary table 5

Term	Estimate	p-value
(intercept)	173100	$< 2e-16$
bathroom	-13970	0.000139
list	0.8876	$< 2e-16$
taxes	16.33	$< 2e-16$
RegionT	47540	5.36e-15
LocationNear urban	-87530	6.51e-12
LocationSuburb	-107700	7.57e-14
LocationUrban	-54110	5.64e-07

We fit a linear regression model from sale price, the number of bathroom, list price, taxes, region and location and we can express it as:

$$\begin{aligned}\hat{sale} &= 173100 - 13970 bathroom + 0.8876 list + 16.33 taxes \\ &+ 47540 RegionT - 87530 LocationNearurban - 107700 LocationSuburb \\ &- 54110 LocationUrban\end{aligned}$$

Hypothesis test

In order to test whenever there is a significant linear relationship between the sale price of the house and the number of bathroom, the list price and the taxes of the house, the region and location. We need to do hypothesis testing for the estimates of the regression parameters.

For each slope estimates β_n:

$H_0 : \beta_n = 0$
$H_a : \beta_n \neq 0$

The null hypothesis states that β_n is equal to zero, while the alternative hypothesis states that β_n is not equal to zero. In this case, we use a benchmark significance level of 5%. From the summary table 5, we can find out that all of the p-values for all the predictors are smaller than the significance level which is 5%. Therefore, the null hypothesis should be rejected which indicates that there are significant correlations between the response variable sale price and those five predictors.

Here is the result of the test:

The number of bathroom, the list price and taxes of the house, region and location with the sale price of the house are all significant features.

In addition, the multiple R-squared gives percentage of variation in the response variable explained by the regression line. When we use the function summary to summarize the fitted multiple linear regression model, we can find out that the multiple R-squared is 0.994 which means that 99.4% of variation in the response variable which is the sale price of the house can be explained by the regression line fitted. Besides, the overall p-value of the global F-test which is $< 2.2e - 16$ is quite small so that we can conclude that there is a linear regression between the response variable sale price and those five predictors.

Model interpretation:

In terms of hypothesis testing of the estimated parameters in the multiple linear regression model, we can conclude that:

The number of bathroom

From the result of the hypothesis test of the number of bathroom, we can conclude that we have strong evidence to support that there is a correlation between sale price and the number of bathroom since the p-value is 0.000139 which is smaller than the significance level (0.05). In the multiple linear regression line, holding all other explanatory variables constant, as the number of bathroom increases by one unit, on average the sale price of the corresponding detached house might decrease 13970 Canadian dollars.

List price and taxes

From the result of the hypothesis tests of the list price and taxes, we can conclude that we have strong evidence to support that there are correlations between sale price and the list price and taxes since the p-values of them are both smaller than the significance level (0.05). In the multiple linear regression line, holding all other explanatory variables constant, as the list price of the detached house increases by one dollar, on average the sale price of the corresponding detached house might increase 0.8876 dollars. Holding all other explanatory variables constant, as the tax of the detached house increases by one dollar, on average the sale price of the corresponding detached house might increase 16.33 dollars.

Region and Location

We have evidence to support that the sale price of the detached house has correlations with the predictor region since the p-value is smaller than the significance level. In general, the region of a detached house can affect the actual sale price of it since different economic situations in different regions/cities can lead to different sale price. From the summary table above, we can find out that holding all other explanatory variables constant, the average sale price of the detached house in Toronto is higher than the average sale price of the detached house in Mississauga.
Besides, we also have evidence to support that there is a correlation between the sale price of the detached house and the predictor location since the p-value is smaller than the significance level. The location of a detached house determines the ease of access to the public facilities and some family recreation facilities so that it can affect the sale price of a detached house. From the summary table above, we can find out that holding all other explanatory variables constant, the average sale price of the detached house in Downtown is higher than the average sale prices of the detached houses in suburb, urban or near urban.

Discussion

Summary

The goal of this report is to explore and analyze the sale price of the single-family, detached houses in two regions (Toronto and Mississauga) in the Greater Toronto Area and the factors which can have influences on it. The factors can be classified into three categories which are the physical conditions (basic information) of the house, financial aspect, and the geographical location of the house. Physical conditions including the size of the house, the number of rooms (bedrooms and bathrooms), the number of parking spots and the age of the house are properties possessed by a house. Financial aspect includes the last list price and the taxes of a house which can represent the previous value of the house. Geographical location including the region and location (whether the house is located in downtown or suburb) determines the prevailing land price and the ease of access to public facilities and some family recreation facilities. In this report, some key features, strengths and weaknesses and the variables of the data have been discussed firstly. Then, the first few rows, the numerical and categorical variables of the data and the variable of interest which is the response variable sale price have been presented by summary tables and plot. Next, by selecting three main features of the house (physical conditions, financial aspect and the geographical location), a multiple linear regression model was fitted to estimate the response variable (sale price) chosen. In the model section, I also used model diagnostics to show that whether the fitted multiple linear regression is a good model or not. The feature and model selection were presented in the appendix. Then, the results of the analysis and the explanation were presented. Eventually, the conclusions, weakness and next steps can be made.

Conclusions

In terms of the results and the hypothesis testing of the estimated parameters in the multiple linear regression model, we can make the following conclusions. In the physical condition (basic information of the house) aspect, the sale price and the number of bathroom selected tend to have a negative correlation since the estimated coefficient for the number of bathroom is a negative number. In addition, the sale price of the detached house has positive correlations with the predictors list price and taxes as they increase one unit, the sale price will also increase. Specifically, holding all other explanatory variables constant, as the list price of the detached house increases by one dollar, on average the sale price of the corresponding detached house might increase 0.8876 dollars. Holding all other explanatory variables constant, as the tax of the detached house increases by one dollar, on average the sale price of the corresponding detached house might increase 16.33 dollars. In the geographical location aspect, we can find out that the overall sale price of the detached house in Toronto is higher than that of the detached house in Mississauga. Besides, the overall sale price of the detached houses which are located in downtown is higher than the sale prices of the detached houses located in suburb, urban or near urban. From the hypothesis testing of those estimated parameters, we can find out that all of the p-values for all the predictors are smaller than the significance level which is

5%. Therefore, the null hypothesis should be rejected which indicates that there are significant correlations between the response variable sale price and those five predictors.

Overall, we can find out that a higher list price and a higher tax of the single-family, detached house will increase the actual sale price of that house. However, in the physical condition aspect, the sale price of the house tend to decrease when the number of bathrooms increases which is something surprise us in the result. Besides, the single-family, detached houses which are located in a good geographical location tend to have a higher price. Specifically, the overall sale price of the detached house in Toronto tends to be higher than that of the detached house in Mississauga. The overall sale price of the detached houses which are located in downtown is higher than the sale prices of the detached houses located in suburb, urban or near urban. In addition, the sale price of the house in a specific region and location can reflect the overall economic situation and development in a city or a country. It can also be influenced by the economic fluctuations caused by COVID-19 or financial crisis.

Weaknesses

The model fitted in this report is a multiple linear regression model which uses five independent variables to estimate one response variable which is the sale price of the detached house, However, we can hardly measure linearity since I just use model selection (See Appendix for more about how to select features) to select the most representative features which can influence the sale price of the house instead of measuring the linearity. Therefore, one of the weaknesses of the analysis is that the linear trend is hard to estimate. Besides, the data used in this report was recorded myself and the sample size is small so that some data might be inaccurate and can cause bias. The predictors selected in the feature selection can be used to analyze and predict the sale price of the detached house. However, expect several basic information (the physical conditions and geographical location) of the house, some external factors can also influence the house price such as the economic fluctuations caused by COVID-19 or financial crisis so that the prediction of the sale price by only using the explanatory variables (predictors) selected might be inaccurate.

Next steps

We can apply more analysis: First, we can try model models for a better demonstration of the relationships. For instance, the generalized additive model might be helpful for exploring non-linear relationships and the function term in this model might have a positive effect in our model accuracy. Then, we can also add more features to the model and penalize overfitting by methods such as ridge regression and LASSO regression. This step can improve the feature selection and at the same time identifying the most significant features. In addition, we can train models and do some predictions, which is useful not only in knowing more about the performance about the fitted model, but also in making predictions when we get more data and assess the sale price of the detached houses. In order to train a more accurate model, machine learning techniques such as nueral network might be a good choice.

References

1. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

2. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

3. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

4. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

5. Toronto Regional Real Estate Board. https://trreb.ca
6. Will Kenton (2020). "Multiple Linear Regression (MLR)", https://www.investopedia.com/terms/m/mlr.asp
7. "Multivariable Methods", https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable7.html
8. "Akaike Information Criterion", https://www.sciencedirect.com/topics/medicine-and-dentistry/akaike-information-criterion
9. "Difference Between AIC and BIC", 12 Oct. 2010, http://www.differencebetween.net/miscellaneous/difference-between-aic-and-bic/
10. "How to interpret p-value with COVID-19 data", https://towardsdatascience.com/how-to-interpret-p-value-with-covid-19-data-edc19e8483b
11. "Statistical Learning (I): Hypothesis Testing on House Price Dataset", https://towardsdatascience.com/practical-practice-of-hypothesis-testing-on-house-price-dataset-1fb169bc04ee

Appendix

Feature selection

First of all, we chose a response variable that we are interested in: the sale price of the single-family, detached house

Next, we identified several parts of possible features:

1. Basic information (physical conditions): the number of bedroom, bathroom and parking spots (the variables bedroom, bathroom and parking), the maximum square footage of the property (maxsqfoot) , Age and Garden (make sure that the survey is evenly distributed between these categories, making the inference representative)
2. Financial aspect: the list price and the taxes of the house (list, taxes)
3. Geographical location: Region, Location

We clean the data selecting these features above into a single data frame.

We then implemented backward BIC (Bayesian Information Criterion) to select the features. This method involves choosing the optimum model by deleting one most insignificant feature a time. BIC is a penalized cost function for the model. After several steps, when the value of BIC reaches its minimum, we get our optimum model.

Table 6: Table 6 - BIC

Step	Df	Deviance	Resid. Df	Resid. Dev	AIC
	NA	NA	786	4.383733e+12	18033.03
- Age	2	5871358298	788	4.389604e+12	18020.73
- maxsqfoot	1	33569601	789	4.389637e+12	18014.06
- Garden	1	6193585710	790	4.395831e+12	18008.50
- bedroom	1	29569819263	791	4.425401e+12	18007.18
- parking	1	32689630707	792	4.458090e+12	18006.38

From the table above (Table 6), we can see that five features are deleted from the model. However, we are looking for some features that are most influential to the sale price of single-family, detached house so that we need to make further feature selection.

We then turn to choose from the remaining model by assessing the p-values of the coefficients. We tend to choose features with smaller p-value, which means more likely to reject the null hypothesis of coefficients being zero. The summary table of our model after using BIC is below.

Table 7: Table 7 - p-values

names	pvalue
(Intercept)	<2e-16
list	<2e-16
bathroom	0.000139
taxes	<2e-16
RegionT	5.36e-15
LocationNear urban	6.51e-12
LocationSuburb	7.57e-14
LocationUrban	5.64e-07

It can be seen from the above table (Table 7) that the p-values are smallest in terms `list, bathroom, taxes, Region, Location`. Thus we choose these features for our final linear model.