# Toronto Home Prices Linear Regression Model

Assignment 2 October 25, 2021

Yutong Lu - 1005738356

## Introduction

The dataset used in this report is based on the 2011 data of two datasets "Wellbeing Toronto – economics" [28] and "Wellbeing Toronto – demographics" [29]. Both datasets are accessed using City of Toronto Open Data [10]. We are interested in looking at some factors that a family may concern about when purchasing a home, and how they could be associated to home prices in Toronto.

Children and the availability of child care may be associated with home prices in the neighbourhood. According to Ost [22], it appears that the potential cost of being a houseowner is very important to the childbearing decisions of young people, suggesting a close relationship between children and housing. Also, for families with children, the availability of children care and other relevant services could be one of their main concerns. Daysal et al. [9] reported that as the home price increases, it generally leads to healthier children at birth. As a result, the number of child care spaces in the neighbourhood as one of the social amenities, may be related to the local home prices.

On the other hand, higher housing prices may put more stress on families that are already struggling with debt payments. According to Toronto Housing Market Analysis by Canadian Centre for Economic Analysis and Canadian Urban Institute [7], the average home ownership costs increased by 60% from 2006 to 2018, but almost simultaneously, there was a 15% decrease in the median household income from 2006 to 2016. This outpacing increase of average home prices may lead to a heavier debt stress. Also, Press [23] reported that after the pandemic started, there is a 4% increase in the total household debt. As a result, it infers that if a family is already facing debt stress, they may choose not to purchase a home with higher price that could increase their debt by a great amount, and they may choose to settle in an area with lower home prices.

In terms of employment and home prices, Agnew and Lyons [2] reported that after the financial crisis 2007-2009 in Ireland, employment factors show a greater impact on the housing sale prices. Specifically, the variable job creation is less significantly related to the sale prices than the variable job destruction, and thus the authors concluded that higher house payment increases the accessibility to labour market [2]. The results of Agnew and Lyons [2] are very important because it suggests that there may be a positive linear relationship between the number of business or employment rates and home prices. On the other hand, in a study on Ontario housing prices, the authors reported that as the unemployment rate increases, we expect the housing prices to decrease [21]. This corresponds to the significant variable job destruction reported by Agnew and Lyons [2].

In this research, we will explore if there is a statistically significant linear relationship between the factors we are interested in and the home prices in Toronto. This significance means that based on our method in the following sections, the results we get from the data collected are not purely by chance [17]. Linear relationship means that we can fit a straight line through the data points and that straight line can explain the relationship between variables [15].

**The research question of this report is what factors are linearly related to home prices in Toronto, and how are they linearly related to the home prices in Toronto.**

In our data, home price is the variable that we want to study, and the other variables, risk category, employment per area, business per area, child care spaces per area, are the variables that change on themselves and in turn may affect the home prices. Thus, we are trying to what combinations of these factors best explains the variations in home prices, and how they affect the home prices in the presence of each other.

Based on the information provided by Daysal et al. [9], Agnew and Lyons [2], and Canadian Centre for Economic Analysis [7], we hypothesize that the number of local employments, business establishments, available child care spaces and the risk category of missing debt payments are all linearly related to the home prices in Toronto. More specifically, when all other factors are the same, neighbourhoods with the highest risk of missing loan payments will have the lowest home prices when compared to medium and high-risk neighbourhoods. When all other factors are the same, the more child care spaces a neighbourhood has, the higher home prices it will also have. When all other factors are the same, when there are more local employments or jobs available, then the home prices will also be higher. When all other factors are the same, if there are more business establishments in the area, the home prices of the area will also be higher.

Thus, this research is important because it not only may provide potential guidance for families finding a home but could also give some insights about what a particular region can improve on in terms of social care to make this region be more appealing to people that wish to settle down.

There are seven sections in our report, including this introduction and bibliography. In the Data section, we will introduce the data collection and cleaning process of our dataset, provide a summary of the data, including the important variables and their numerical and plot summaries. Then in the Methods section, we will introduce the process of choosing independent variables in our model, and report our results and the final model in the Results section. Based on our results, we will draw conclusions, discuss weaknesses of this research, and propose future works that can be done in the Conclusion section. Any extra plots and numerical summaries that are relevant to our research will be included in the Appendix.

# Data

## Data Collection Process

"Wellbeing Toronto – economics" [28] contains data from Toronto City Planning, Economic Development Culture & Tourism, Children's Services, Employment & Social Services, Social Policy Analysis & Research, Municipal Licensing & Standards and Economic Development Culture & Tourism, as well as other two resources, TransUnion and Realosophy.com. "Wellbeing Toronto – demographics" [29] contains data from Statistics Canada Census. The data from City of Toronto are collected from 2011 census and the data for each neighbourhood is aggregated from the information of smaller regions [28][29]. Community Data Program provided the data for TransUnion based on postal code, and about 92% of Canadians with credit files are covered by TransUnion, but the debt information of areas with fewer than 15 credit files are not provided due to privacy considerations [28]. Home prices data were collected between 2011 to 2012 in CAD, collected by Realosophy [28]. Based on the data section on the Realosophy website, the real estate data on Realosophy is from Toronto Real Estate Board [1].

It is also important to note that there are some limitations in the datasets. The census data is collected by adding up the data from smaller areas, which means that the total value for a variable in one neighbourhood may be smaller than reality due to rounding or suppressing of the number in the subdivisions [28][29]. This may affect our model in terms of fitting a line that is different from the line with not rounded or suppressed data, or in other words, we may have biased estimates for the coefficients of the line. Another limitation is that not all Canadians with credit files are covered by TransUnion and there are inaccessible data for postal codes with fewer credit files, and thus we may be missing data during the collection and calculated for each neighbourhood without those missing individuals' information. This may lead to less generalizable models in our research because we may not be able to generalize to an area with fewer credit files.

## Data Summary

"Wellness Toronto – economics" [28] is a dataset with 140 observations and 8 variables. Two of variables are character variables Neighbourhood Id and names, and the other 6 variables are all numerical, which are the number of licenced businesses establishments, number of child care spaces, debt risk score, home prices, the number of local employment, and the number of social assistance recipients. Based on our research question, we will only be using the number of child care spaces, debt risk score, and home prices in our research. Note that the 140 observations in the dataset represent the 140 neighbourhoods in Toronto. Similarly, for "Wellness Toronto – demographics" [29], it is also with 140 observations representing 140 neighbourhoods, arranged in their IDs, and this is why we will be able to combine them in the following cleaning process. Different from the economics dataset, there are much more variables in the demographic dataset, including total area of the neighbourhood in square kilometer, total population of the neighbourhood, different features of income, and more. However, we will not be using the other variables expect for total area in our research.

## Data Cleaning

To begin with, "Wellbeing Toronto – economics" [28] and "Wellbeing Toronto – demographics" [29] are excel files imported directly into R using the package opendatatoronto [10] by first searching and storing the id of the specific dataset we wanted on the Toronto Open Data Portal and then getting the dataset and the specific worksheet in the excel file that we wanted [11].

Firstly, no missing value was found in the dataset. Because of a format problem of excel, the first row was in the raw dataset is the variable names, and the actual variable names displayed in R were numbers. As a result, in the economics dataset, we first renamed the columns using their actual names that are stored in the first row, which are "neighbourhood", "neighbourhood id", "businesses", "child care spaces", "debt risk score", "home prices", "local employment", and "social assistance recipients", and then we removed the first row. For the demographic dataset, because we only needed the variable for neighbourhood areas, so we first selected the column "Demographics", which stands for total area, and then changed its name to "area in square kilometer". After renaming all the variables we wanted, we then combined the this column from the demographic dataset to the economics dataset. After inspection, the variable types in this dataset are all characters, so we then changed all of them into numeric because they are all numeric in nature.

Because we wanted to eliminate the effect of area of each neighbourhood on the number of child care spaces that they are geographically capable to have, we created a new variable child care spaces per area by dividing the variable child care spaces by area in square kilometer. Similarly, we created two more variables "employment per area" and business per area by dividing "local employment" and "businesses" by "area in square kilometer", respectively. Also, to make the numbers more readable, we created a new variable "home prices in 10K CAD" by dividing the variable "home prices" by 10000. Then based on the metadata provided by Wellness Toronto [28], a Debt Risk Score lower than 707 means the risk of missing loan payments 3 consecutively is high, while a score higher than 769 means the risk is low. As a result, we created a new variable "risk category" indicating their risk and separated the observations into either "high", "middle" or "low" based on the observed Debt Risk Score. Observations with "high" have scores lower than 707, observations with "middle" have scores greater or equal to 707 but smaller than 769, and observations with "low" have scores higher or equal to 769.

Finally, we selected "home prices in 10K CAD", "risk category", "employment per area", "businesses per area", and "child care spaces per area" for the linear regression model that will be built based on our research question, and also "neighbourhood id" for future reference of the actual neighbourhood. Now, the cleaned dataset has 140 observations, each for one neighbourhood in Toronto, and 6 variables.

## Important Variables

The five important variables in this dataset are "home prices in 10K CAD", "risk category", "employment per area", "businesses per area" and "child care spaces per area". "Home prices in 10K CAD" is a numerical variable of the average home price in the neighbourhood between 2011 and 2012, in ten thousand (10K) CAD, which will be the study variable in our linear regression model. “Risk category” is a categorical variable that indicates the risk of missing 3 consecutive loan payments, with levels low, middle and high. “Employment per area” is the number of local jobs for individuals over 15 years old per square kilometer. “Businesses per area” is the number of licenced business establishments per square kilometer. “Child care spaces per area” is a numerical variable of the count of licenced spaces that provide children service per square kilometer. The variables “risk category”, “employment per area”, “businesses per area” and “child care spaces per area” are the potential independent variables that we want to include in the model to explain the study variable “home prices in 10K CAD”.

## Numerical Summaries

The following Table 1 is a summary for all numerical variables in our dataset. We can see that for every variable, its mean is much greater than the median.

The variable “home prices in 10K CAD” has a mean of 54.819, a median value of 49.121, and a standard deviation of 26.767. We can see that the average home prices for all neighbourhoods in Toronto area (548.193K CAD) is much higher than the median 491.21K CAD. From this information, we may infer that the home prices is right skewed, meaning that there may be outliers on the far right, bringing the average of home prices up.

The variable “employment per area” has a mean of 3077.367, which is much greater than its median at 1354.332. This indicates that there may be very large outliers in our data. Because we already divided by local area, it is not due to the effect of purely larger area having more jobs, and thus we may infer that some neighbourhoods may have much more employments per area than others. Similarly, for the variable “businesses per area”, its mean 175.046 is greater than its median 90.185, following that there may be some neighbourhood having much more business establishments per area than other neighbourhoods.

The variable “child care spaces per area” has a mean of 42.031, a median value of 29.583, and a standard deviation of 39.883. The median is lower than the mean, so again, it means that there may be neighbourhoods with much more child care spaces per area than other neighbourhoods, and the mean is brought up by these large outliers.

Table 1: Summaries for numerical variables

| Variable | Mean | Median | Standard Deviation |
|---|---|---|---|
| Home prices in 10K CAD | 54.819 | 49.121 | 26.767 |
| Employment per area | 3077.367 | 1354.332 | 9558.117 |
| Business per area | 175.046 | 90.185 | 287.226 |
| Child care spaces per area | 42.031 | 29.583 | 39.883 |

Table 2 is a count for every category in the categorical variable “risk category” and tells us about average “home prices in 10K CAD” in each category. From this table, we can see that the number of neighbourhoods with middle risk is the greatest and there are slightly more low-risk neighbourhoods than high-risk ones. Simultaneously, the neighbourhoods with middle risk of missing debt payments also have the middle mean home prices among the three groups. We observe that the high risk neighbourhoods have the lowest mean home prices and the low risk neighbourhoods have the highest mean home prices in our data. This makes sense because according to Canadian Centre for Economic Analysis [7], the debt increase has outpaced the household income increase, and thus people with more debt may not be financially capable of purchasing expensive homes, resulting in them living in neighbourhoods with low average home prices.

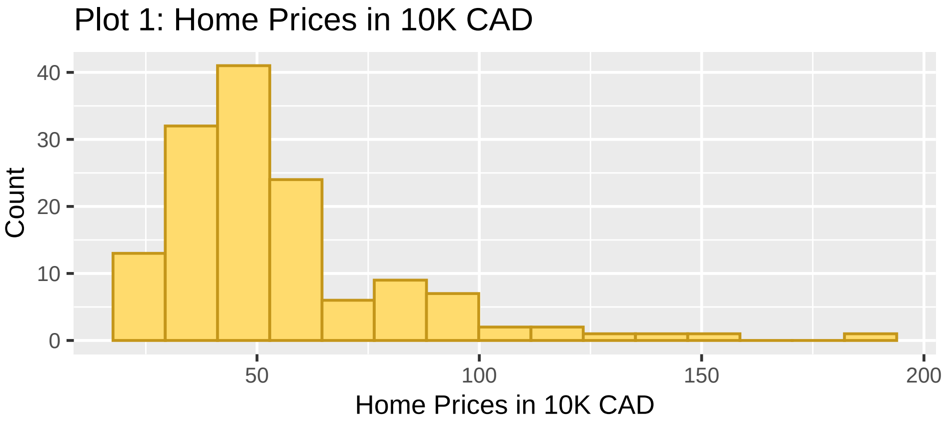
Table 2: Average Home prices with each risk level

| Risk Category | Number of Neighbourhoods | Mean Home Prices in 10K CAD |
|---|---|---|
| high | 21 | 35.62331 |
| low | 24 | 88.30566 |
| middle | 95 | 50.60297 |

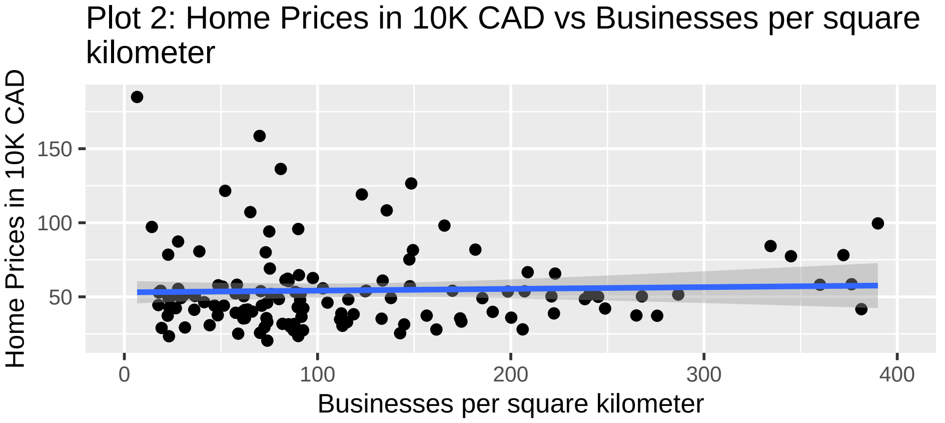
All analysis for this report was programmed using `R version 4.1.1` [24]. In this section of this report, the tables are created using R packages Tidyverse [31], knitr [33], and kableExtra [35].

## Plot Summaries

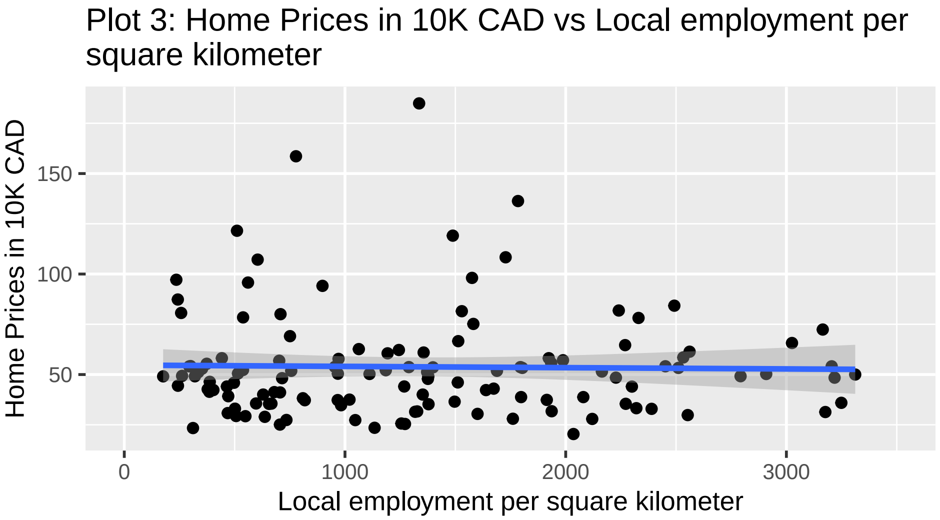
Plot 1 is a histogram of the variable "home prices in 10K CAD". From the histogram, we can see that the distribution is right skewed and there are some large outliers in this variable, which corresponds to the larger mean than median in the numerical summaries.



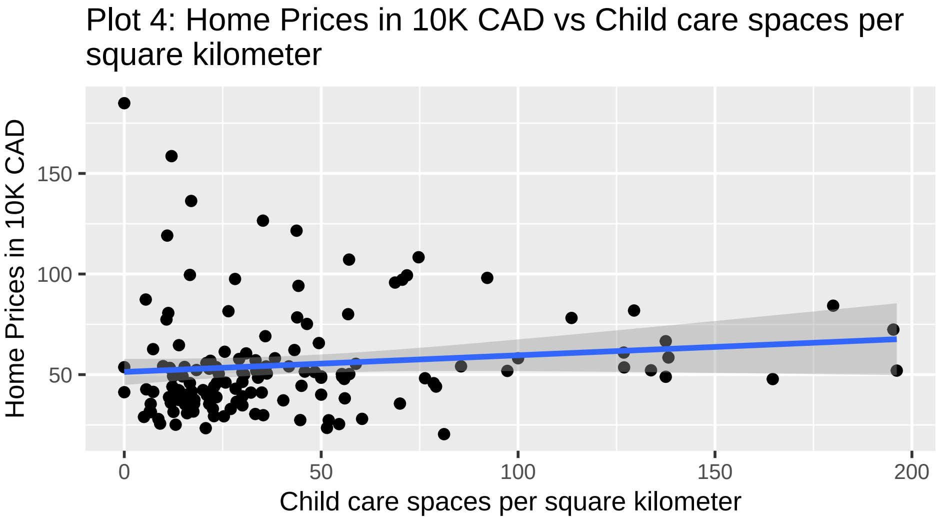
Plot 2 is a scatterplot of businesses per area and home prices, and we fitted a blue smooth line to the points. Because there exist very large outliers that will scale the plot axis to the point where other points will just appear as an aggregation, we limit the x-axis to be between 0 to 400, which is a reasonable measure of the non-outliers in the data. According to the plot, there appears to be a very weak to no linear relationship between these two variables, because the line appears to be flat for all values of businesses per area. This indicates that maybe businesses per area alone cannot explain home prices very well, but we still do not know how it will perform in the presence of other predictors.



Plot 3 is a scatterplot of local employment per area and home prices, and we fitted a blue smooth line to the points. Similarly, we have limited the x-axis to be in the interval (0, 3500), and this is because otherwise the outliers in the dataset will increase the scale of x-axis to the point where the relationship of other points are not visible. According to the plot, there appears to be a very weak to no relationship between these two variables. This indicates that on its own, employment per area does not seem to explain home prices very well, but we still do not know how if it will be able to explain more in the presence of other predictors.



Plot 4 is a scatterplot of child care spaces and home prices, and we fitted a blue smooth line to the points. According to the plot, there appears to be a very weak linear relationship between these two variables, however, we still do not know if child care spaces can explain more in the presence of other variables. Also, when comparing to the other numerical variables, “employment per area” and “businesses per area”, it shows a slightly stronger positive relationship between child care spaces and home prices. Furthermore, notice that at each value of child care spaces, the variance of home prices around the line is not the same, and the variance is generally very large.



All analysis for this report was programmed using `R version 4.1.1` [24]. In this sections of this report, the plots are created using R packages Tidyverse [31] and stringr [32].

# Methods

In this research, the method that we will use to model the relationship in our research question is a frequentist linear regression model, which means that we believe the parameters in our model is fixed but unknown in the population [4]. Firstly, there are two types of variables involved in this model, which are study (dependent, or response) variable, the variable we want to study, and auxiliary (independent, or predictor) variable, the variable that we think may explain the changes in the study variable, and we should have the value of them for all the observations in our data [3]. Linear regression is a way to model the relationship between one study variable and one or multiple independent variables by fitting a straight line and thus a linear equation, to the data we observed [18]. Also, in this model, the study variable needs to be continuous numerical, but the independent variables can be either numerical or categorical. However, if the line is too flat, then the outcome variable may not be a significant linear relationship between the study variable and the auxiliary variable(s) [19].

When we have a categorical variable in the linear regression model, we create dummy variables for all categories, except for one category, of a categorical variable [30]. In other words, if there are $n$ categories in the variable, then we create $n-1$ dummy variables. The category without a dummy variable will then be the reference category and will not appear in the model itself. Depending on the category an observation takes, at most one of the dummy variables will take a value of one, where one means that the observation is in this category, and zero otherwise [30]. If an observation takes zero in all dummy variables, then this means that this observation is in the reference category.

Because we want to have more than one independent variable in this linear regression model, we are using a multiple linear regression instead of a simple one, which involves multiple predictors in the model to explain our study variable [14]. In this research, we will include one numerical independent variable and one categorical independent variable with three categories in our model, so we need a total of $1+(3-1)=3$ $x_i$ terms in the model.

Therefore, the linear regression model we will use is:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

where $\beta_0$ represents the intercept of the regression line, and we can interpret it as the value of $y_i$ when all $x_{ji}=0$. Notice that because sometimes it is meaningless to let $x_{ji}=0$, we should be careful when interpreting $\beta_0$. $\beta_1$ represents the slope of the line or coefficient of the numerical variable $x_{1i}$ [18]. We can interpret $\beta_1$ as the expected amount of increase in $y_i$ when there is a one-unit increase in $x_{1i}$ when all the other predictors are the same. Because we want to introduce a 3-level categorical variable into the model, we then have two dummy variables in our model, $x_{2i}$ and $x_{3i}$, it means that the categories given by our categorical variable have different intercepts but the same slope for the numerical variable. For example, for an observation in the $x_{2i}$ category, then the new intercept of this observation is $\beta_0+\beta_2$ because the $x_{2i}=1$ and $x_{3i}=0$ for this observation. $\beta_2$ represents that the observations in the category $x_{2i}$ have an expected $y_i$ outcome that is $\beta_2$ different from the reference category when the values of $x_{1i}$ are the same. Similarly, $\beta_3$ represents that the observations in the category $x_{3i}$ are expected to have a $y_i$ outcome that is $\beta_3$ different from the reference category when the values of $x_{1i}$ are the same. Finally, $\epsilon$ is a random error term, which represents the natural variation to fully represent data points in the actual population [13].

Because our research question is what factors linearly relate to the home prices of Toronto, our dependent variable in the model will be "home prices in 10K CAD". Also, we want at least one numerical variable and one categorical variable as predictors in the model. This frequentist linear regression model is appropriate because firstly, our dependent variable is a continous numerical variable, and we want to include at least one numerical variable and one categorical variable as the independent variables, which satisfies the requirements for variable type of linear regression model. Also, as seen in the three scatterplots between the independent variables and the dependent variable, most points accumulate around the line. This means that our data may satisfy the Normality assumption of linear regression [25], which assumes the study variable home prices Normally distributes and displays a "bell-shape" at each predictor value around the line we fitted.

Our dependent variable is home prices, and we will then select the independent variables in this model using a step-wise backward selection method [8], and then we will use coefficient of determination and adjusted coefficient of determination [5] to decide which variables to include in our model.

The coefficient of determination $R^2$ is a value that measures how much of the variation in the study variable can be explained by a specific model [6]. $R^2$ can take a value between 0 and 1, and if the value is closer to 1, that means more variation in the study variable can be explained by the independent variable [6]. As a result, we can then get an idea about how the independent variables in a model may explain the variation in the dependent variable in a linear model.

The formula for $R^2$ is

$$R^2 = 1 - \frac{RSS}{SST} = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2}$$

where $RSS$ represents the residual sum of square, which is the variation that is not explained by our model, and $SST$ represents the total sum of square, which is the total variation in our data [5]. After subtracting the proportion of variation that is not explained by our model from 1, it is natural that the left over portion, which is $R^2$, represents the proportion of variation that is explained by our model [6].

However, we cannot use $R^2$ to compare two models with different numbers of predictors because $R^2$ always tends to increase when we add more independent variables since more predictors naturally explain the response a little bit more, even if the added predictor is not a particularly useful independent variable to be included the model [27]. In other words, $RSS$ on the same data always decreases when there are more predictors while $SST$ stays the same for the same data. Thus, we will use adjusted coefficient of determination $R^2$ instead, which is a value that gives us the ability to compare two models with different numbers of independent variables because it adjusts for the number of independent variables and only will increase for a lot when the newly added predictor enhances the our model not simply by predictor addition or by chance. We can also see this in the formula for adjusted $R^2$ [5], which is

$$R^2_{adj} = 1 - \frac{RSS/n - p - 1}{SST/n - 1}$$

where $n$ is the number of observations in the data, $p$ is the number of predictors in the model, 1 in the numerator represents the intercept of the model. However, because of the division in both numerator and denominator of $R^2_{adj}$, we cannot interpret this as the proportion of variation explained by our model like $R^2$ anymore, but we will use this to determine whether to leave a variable in our model.

In this research, we will do a step-wise backward selection method [8], which means that we will start with a model with all the predictors we have in the dataset. Then, we will calculate the $R^2$ and $R^2_{adj}$ value for our model, remove predictors in our model once at a time, and compare the $R^2_{adj}$ value with the last, bigger model with more predictors. If the $R^2_{adj}$ of the model with more predictors is lower than the $R^2_{adj}$ of the model with fewer predictors, then we will choose the smaller model. Otherwise, we will choose the bigger model with more predictors [8]. We want a minimum of two predictors in the model, one numerical and one categorical, so we will stop after eliminating the least two important predictors in the full model.

After obtaining the predictors we want using step-wise backward selection [8], we will also report the $R^2$ and $R^2_{adj}$ value for another model with the interaction terms [16] between the two predictors in the model. As mentioned above, the way we added the dummy variables $x_{2i}$ and $x_{3i}$ means that the categories given by our categorical variable have different intercepts but the same slope for the numerical variable. However, if we add interaction terms in the model, then we are saying that the different levels in our categorical level not only have different intercepts but different slopes [16]. Then, our model will be

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{1i} x_{2i} + \beta_5 x_{1i} x_{3i} + \epsilon_i$$

For example, for an observation in the $x_{2i}$ category, we know that it has $x_{2i} = 1, x_{3i} = 0$. Then we can combine the terms and thus the new intercept of this observation is $\beta_0 + \beta_2$ and the new slope for $x_{1i}$ is $\beta_1 + \beta_4$.

As a result, this longer version of our model is called the full model, and the shorter version without interaction terms is the reduced model, which contains a subset of predictor terms in the full model [34].

We then will then use a partial F test to determine if we can remove the additional interaction terms in the model all at once, or in other words, whether the full model is statistically significantly different from the reduced model [34]. We choose a partial F test instead of looking at $R^2_{adj}$ again because we cannot remove the interaction terms between the a numerical variable and a multi-level categorical variable step-wise, and it does not make sense to remove the interaction term for some categories but not others. Also, because $R^2, R^2_{adj}$ are only statistical measurements instead of formal tests [6], there are no clear cut-off values to tell us when the model is good enough. Therefore, we will use a partial F test to decide if all the additional interaction terms in the full model are necessary [34]. The test statistic [34] is

$$F = \frac{RSS_{(drop)}/k}{RSS_{(full)}/n - p - 1}$$

which follows a F distribution with degrees of freedom of $k, n - p - 1$.

$RSS_{(drop)} = RSS_{(reduced)} - RSS_{(full)}$ is the amount of variation that becomes unexplained if we use the smaller, reduced model, $k$ is the number of predictors we dropped from the full model, $n$ is the number of observations in the data, $p$ is the number of predictors in the full model, 1 stands for the one parameter that is the intercept of the model. We already know from the $R^2$ that $RSS$ always decreases with the addition of more predictors when using the same data, so $RSS_{(full)}$ is smaller than $RSS_{(reduced)}$. The null hypothesis and the alternative hypothesis are

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

$$H_0 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

where $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and $\boldsymbol{\beta}_2$ represents the set of k predictors you want to remove from the full model, and $\beta_0, \boldsymbol{\beta}_1$ are the intercept and the p predictors in the full model, respectively [34]. We will use a significance level of $\alpha = 0.05$ and calculate one-sided p-values. If $p - value = P(F_{df=(k,n-p-1)} > F) < \alpha$, which means that the probability of observing our test statistic is very small under the null hypothesis, we will reject the null hypothesis and have evidence that at least one of the predictors removed should stay in the model. On the other hand, if $p - value = P(F_{df=(k,n-p-1)} > F) > \alpha$, then we fail to reject the null hypothesis and we have evidence that we can remove the dropped predictors from the full model and use the reduced model instead [34]. In our case, we will look at the full model with interaction terms, and the reduced model with no interaction terms, and decide which one to use based on the results of this partial F test.

After deciding the predictors to include in the model, we will then use the appropriate linear model with or without interaction terms depending on our partial F test result to perform linear regression, get an estimate for the coefficient of each predictor, and use a Student's t test [26] to calculate test statistics and p-values for each coefficient in our model. We will compute our test statistic using the Student's T-test formula [26]

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{var[\hat{\beta}_i]}}$$

The null hypothesis and the alternative hypothesis are

$$H_0 : \beta_i = 0$$

$$H_a : \beta_i \neq 0$$

$\hat{\beta}_i$ represents the estimated coefficient, $\beta_i$ represents the $\beta_i = 0$ in the null hypothesis. This means that in the null hypothesis, there is no relationship between this predictor and the response variable in the presence of other predictors. We then calculate a two-sided p-value using $p - value = P(|t_{df=n-p-1}| > T)$ where $t_{df=n-p-1}$ represents a $T$ distribution, with a degree of freedom of number of observation subtracted by the number of predictors and the one intercept in the model. This p-value represents the probability of observing a value that is as extreme or more extreme than our test statistics, and thus the smaller the p-value, the less likely that our test statistic is purely by chance under the null hypothesis. In this case, we will use a significant level of $\alpha = 0.05$, meaning that if our p-value is smaller than 0.05, we will reject the null hypothesis that there is no relationship between this predictor and the study variable in the presence of other predictors.

# Results

The following Table 3 is a table for the different possible models in our research and their corresponding $R^2$ and $R^2_{adj}$ values. We can see from the table that the full model, which is the model with four predictors, indeed have a high $R^2$ value 0.3858623, but the least $R^2_{adj}$ value 0.3629467. This is because in our full model, the highest $R^2$ value is only due to the highest number of predictors, and these predictors all can explain a little bit of the study variable [27]. Based on our graphical summaries, we noticed that employment per area and business per area have very weak to no relationship to home prices, so we decided to remove them from the models first and observed the change in $R^2_{adj}$. According to the table, after we removed "employment per area" and "businesses per area" from the model one by one, there is even a gradual increase in the $R^2_{adj}$ value, which confirms with our choice of removing these two variables from our full model. This may be explained by Agnew and Lyons [2], where they reported that job creation is not as significant as job destruction in relationship to home prices, and we only have the variables employment and business per area, but we do not have the unemployment rate as a predictor.

Table 3: $R^2$ and $R^2_{adj}$ for different models of study variable, Home Prices in 10K CAD

| Predictors in the model | $R^2$ | $R^2_{adj}$ |
|---|---|---|
| Model 1: risk category, child care spaces per area, employment per area, businesses per area | 0.3858623 | 0.3629467 |
| Model 2: risk category, child care spaces per area, businesses per area | 0.3824011 | 0.3641019 |
| Model 3: risk category, child care spaces per area | 0.3819576 | 0.3683243 |
| Model 4: risk category, child care spaces per area, risk category*child care spaces per area (interaction terms included) | 0.4034744 | 0.381216 |

From Table 3, we can also see that the $R^2_{adj}$ value for the model with interaction terms (0.381216) is greater than the reduced model without interaction terms (0.3683243). Then, we used a partial F test to determine whether we can remove the interaction terms in the reduced model all at once. The resulted p-value is 0.093, which is greater than our significance level 0.05. Then, we fail to reject the null hypothesis that all removed predictors have coefficients of zero, which means that we have evidence that we can remove all the interaction terms at once and use the reduced model without interaction terms.

As a result, in this research, we will include the numerical variable "child care spaces per area" and the categorical variable "risk category" as the predictors, and "home prices in 10K CAD" is the study variable, and we will not include interaction terms between the two independent variables. Thus, we are trying to find a linear equation that explains the relationship between home prices, number of child care spaces and risk level of missing debt payments. Because there are three categories for risk, we will have one numerical independent variable and two dummy independent variables for the categorical variable in the model.

Table 4 is a report of the estimated coefficients in our model and their p-values. "$\hat{\beta}_i$" represents the parameters that we are estimating in the model. The column "Variable/Intercept" represents the variable that the coefficients associated with, or the intercept of the model. "Estimate" represents the estimate for each of the $\hat{\beta}_i$ in our model. The last column "p-Value" represents the p-value calculated using Student's T-test [26] under the null hypothesis $\hat{\beta}_i = 0$.

Table 4: $\hat{\beta}_i$ Estimate and Corresponding p-Value

| $\hat{\beta}_i$ | Variable/Intercept | Estimate | p-Value |
|---|---|---|---|
| $\hat{\beta}_0$ | (Intercept) | 32.296 | 1.239e-09 |
| $\hat{\beta}_1$ | child care spaces per area | 0.088 | 0.055 |
| $\hat{\beta}_2$ | risk category: low | 52.544 | 1.112e-13 |
| $\hat{\beta}_3$ | risk category: middle | 14.478 | 0.006 |

In other words, our model from this data is

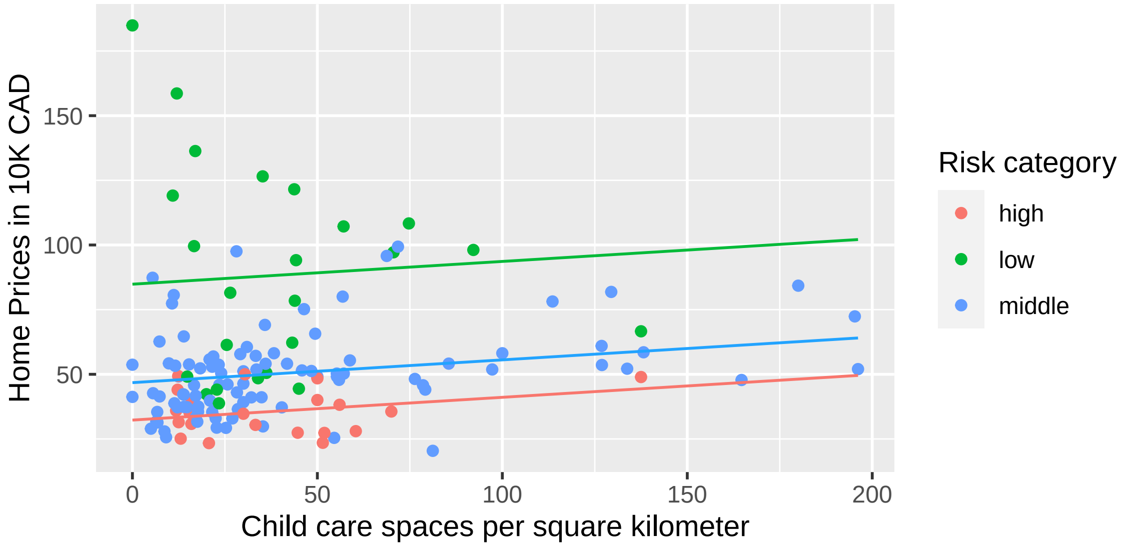
$$\hat{y_i} = 32.296 + 0.088x_{1i} + 52.544x_{2i} + 14.478x_{3i} + e_i$$

First of all, $\hat{y_i}, x_{1i}, x_{2i}, x_{3i}$ are our study variable and the three auxiliary variables, which are child care spaces and two of the three risk categories, which are low and middle based on alphabetical order. $\hat{\beta_0} = 32.296$ is the estimated intercept of our straight line. We can interpret it as when there is no child care spaces in the neighbourhood, the default expected home prices in 10k is 32.296 for the neighbourhood with high debt risk. $\hat{\beta_1} = 0.088$ represents the estimated coefficient of child care spaces, and we can interpret it as when the risk category is the same, then for every one more child care space built in the neighbourhood, the average home prices of the neighbourhood is expected to increase by $0.088 \times 10k$ CAD. $x_{2i}$ and $x_{3i}$ are two of the risk categories "low" and "medium". Then, $\hat{\beta_2}$ is the estimated coefficient for low risk category, and $\hat{\beta_3}$ represents the estimated coefficient for medium risk category. We can interpret $\hat{\beta_2} = 52.544$ as that when the number of child care spaces is the same, the average home prices for low risk neighbourhoods is $52.544 \times 10k$ CAD higher than the high risk neighbourhoods. Similarly for the interpretation for $\hat{\beta_3} = 14.478$, when the number of child care spaces is the same, the average home prices for middle-risk neighbourhoods will be $14.478 \times 10k$ CAD higher the high risk neighbourhoods. $e_i$ represents the residual, which is the difference between the actual $y_i$ and our estimated $\hat{y_i}$.

As a result, we can also express our model as the following:
For high-debt-risk neighbourhoods, $\hat{y_i} = 32.296 + 0.088x_{1i} + e_i$,
For medium-debt-risk neighbourhoods, $\hat{y_i} = (32.296 + 14.478) + 0.088x_{1i} + e_i = 46.774 + 0.088x_{1i} + e_i$,
For low-debt-risk neighbourhoods, $\hat{y_i} = (32.296 + 52.544) + 0.088x_{1i} + e_i = 84.84 + 0.088x_{1i} + e_i$.

Plot 5 is an visualization for our three linear models above [20]. The top green line represents the linear model for low-debt-risk neighbourhoods, the blue line represents the linear model for middle-debt-risk neighbourhoods, and the red line at the bottom represents the linear model for high-debt-risk neighbourhoods. It demonstrates that the neighbourhoods with different debt risk levels have the same slope for child care spaces per area, but have different intercepts with y-axis. This assumption is based on our $R^2_{adj}$ values when picking the predictors, and the insignificant F test statistic in the partial F test when comparing the two models with or without interaction terms. This means that when the number of child care spaces in the neighbourhood is the same, low-debt-risk neighbourhoods have the highest expected home prices among the neighbourhoods in all categories. Conversely, low-debt-risk neighbourhoods have the lowest expected home prices among all categories when the number of child care spaces in the neighbourhood is the same. This makes sense because people in neighbourhoods with low debt risk may be able to afford the more expensive housing, and thus the home prices in the neighbourhoods may be higher when the number of child care spaces is the same. On the other hand, people that are already struggling with debt may not choose to buy expensive housing, so the neighbourhoods they live in may have lower home prices when the number of child care spaces is the same.

Plot 5: Home Prices in 10K CAD vs Child care spaces per square kilometer, categorized by debt risk level



We have also reported the p-value for each coefficients in Table 4. Out of all of our p-values, only the p-value for the coefficient $\hat{\beta}_1$ is bigger than our significance level. Because $\hat{\beta}_0$ has a p-value of 1.239e-09, which is much smaller than our significance level 0.05, we reject the null hypothesis that the intercept is zero. In other words, we reject that the default expected home prices is zero for high-debt-risk neighbourhoods when there are no child care spaces in the presence of other predictors. However, the p-value of $\hat{\beta}_1$ is 0.055, which is greater than our significance level 0.05, so we fail to reject the null hypothesis that there is no significant linear relationship between the child care spaces and home prices, in the presence of other predictors. This seems reasonable based on the report of Harkness et al. [12], where they reported no association between housing prices and the wellbeing of children living in poverty for a long time.

The p-value for $\hat{\beta}_2$ is 1.112e-13, which is much smaller than the significance level 0.05, so we reject the null hypothesis of $\beta_2 = 0$, which means that when the number of child care spaces is the same, the home prices in the neighbourhoods with low debt risk is the same as the home prices in high-risk-neighbourhoods in the presence of other predictors. Last but not least, the p-value for our $\hat{\beta}_3$ is 0.006, which is smaller than 0.05, so we reject the null hypothesis $\beta_3 = 0$, meaning that when the number of child care spaces is the same, the home price in the medium-risk-neighbourhoods is the same as the home price of high-risk-neighbourhoods in the presence of other predictors. This significant linear relationship between debt risk category seems reasonable based on the report by Canadian Centre for Economic Analysis and Canadian Urban Institute [7], where they reported the outpaced increase in average housing cost comparing to the decreasing median income.

Therefore, based on our results, we reject our hypothesis that the number of local employments, business establishments, available child care spaces and the risk category of missing debt payments are linearly related to home prices in Toronto. We have evidence that only the debt risk category of Toronto neighbourhoods is significantly linearly related to the home prices. Also, we fail to reject our hypothesis that high-debt-risk neighbourhoods will have the lowest home prices when compared to medium and high-risk neighbourhoods with all other predictors have the same values.

All analysis for this report was programmed using `R version 4.1.1` [24]. We used the `lm()` function in base `R` to derive the estimates of a frequentist linear regression in this section.

# Conclusions

Our research question is what and how the different factors are linearly related to home prices in Toronto. Our hypothesis is that the number of local employments, business establishments, available child care spaces and the risk level of missing three debt payments consecutively are all linearly related to home prices in Toronto. We also hypothesized a positive linear relationship between the independent variables local employments, business establishments, available child care spaces, and the study variable home prices when conditioning on other predictors, and when all other factors are the same, high-debt-risk neighbourhoods will have the lowest home prices when compared to medium and high-risk neighbourhoods.

We decided to build a frequentist [4] linear regression model for the study variable home prices with one numerical and one categorical variable, based on our research question. In order to decide which predictors to include in the model, we used step-wise backward [8] selection based on $R^2_{adj}$ values [5] of the models with different numbers of predictors. When we were left with two predictors, we then came up with two models. One full model is with the interaction terms [16] of the numerical and categorical variables, and the other reduced model has no interaction terms [34]. Then, we used a partial F test [34] to decide whether we can remove the interaction terms in the full model all at once. After deciding whether to use the full or reduced model, we then used the resulting model for our data and calculate the estimates for the coefficients or the intercept in the model. We then used a Student's T test [26] to calculate the p-value to determine if this predictor is significantly linearly related to the response.

In our results, we found out that the model with the two independent variables, child care spaces and risk category, has the highest $R^2_{adj}$ value, whereas the models with the variables employments and businesses have lower $R^2_{adj}$ values. As a result, we decided to include these two independent variables in the model.

Then, we added the interaction terms between these two variables in another model, an even higher $R^2_{adj}$ is reported. We then performed a partial F test on these two models. Our reported p-value indicates that we could remove the interaction terms in the bigger model and use the reduced model without interaction terms instead.

Using this model, we then have the linear regression model $\hat{y_i} = 32.296 + 0.088x_{1i} + 52.544x_{2i} + 14.478x_{3i} + e_i$, or in an alternative form, three models with three different intercepts for each of the risk category:
For high-debt-risk neighbourhoods, $\hat{y_i} = 32.296 + 0.088x_{1i} + e_i$,
For medium-debt-risk neighbourhoods, $\hat{y_i} = (32.296 + 14.478) + 0.088x_{1i} + e_i = 46.774 + 0.088x_{1i} + e_i$,
For low-debt-risk neighbourhoods, $\hat{y_i} = (32.296 + 52.544) + 0.088x_{1i} + e_i = 84.84 + 0.088x_{1i} + e_i$.

This means that when the number of child care spaces is the same, the high-debt-risk neighbourhoods have the lowest expected home prices, the middle-debt-risk neighbourhoods have the middle expected home prices, and the low-debt-risk neighbourhoods have the highest expected home prices.

The reported p-value shows that the number of child care spaces is not significantly linearly related to home prices in the presence of other predictors. All the dummy variables from the categorical variable, debt risk category, are significantly linearly related to home prices in the presence of other predictors. We find out that when there are no child care spaces in the neighbourhoods, the neighbourhoods with low debt risk have the highest expected home prices, and the neighbourhoods with high debt risk have the lowest expected home prices. Our results could be explained by the lack of significant relationship between children wellbeing and housing prices [12], and the outpacing increase in the homeownership cost when compared to the decreasing median income [7]. Therefore, we have evidence to reject our hypothesis that employments, businesses, child care spaces and debt risk level are all linearly related to home prices in Toronto. Based on our results, out of all the factors in the hypothesis, only debt risk level is linearly related to the home prices in the presence of child care spaces, but child care spaces is not linearly related to the home prices in the presence of debt risk level. However, we fail to reject the hypothesis that high-debt-risk neighbourhoods will have the lowest home prices when all other factors are the same.

## Weaknesses

There are some weaknesses and limitations in this report. Firstly, according to the scatterplot of child care spaces and home prices (Plot 4), as well as the scatterplot including all predictors (Plot 5), we can see that the variance of home prices at different values of child care spaces is not constant. Thus, we may have violated the homoscedasticity of errors, or the constant error variance assumption of linear regression [25]. This may cause our model to perform better at some predictor values than others.

Also, it is possible that there may be other predictors in the true, population-level relationship that are linearly related to our response variable. As a result, we may have violated the linearity assumption or zero mean error assumption of linear regression, which means that we assume there are no left-out or extra predictors in our model, and the error has a mean of zero in the model [25]. This means that resulting in the potential biased estimates of the coefficients in our linear regression model.

## Next Steps

Firstly, we have mentioned some potential violations of the linear regression assumptions in the limitation, and we also see that there are outliers in both dependent and independent variables in our data. As a result, we may use some techniques in the future research to adjust for these violations, skewness and outliers in our data.

Also, in future research, we could switch the variable types of child care spaces and debt risk scores in the model. Because both child care spaces and debt risk scores are numerical variables in the original dataset, it is possible that turning the number of child care spaces into a categorical variable during the data cleaning process may be a better choice. We choose to turn the debt risk score into a categorical variable based on the defined levels in the metadata of our dataset. By doing this, we are essentially saying that the actual value for debt risk score does not matter, rather the category matters more. As a result, in future researches, we can turn the number of child care spaces into a categorical variable. Then, we are saying that the actual number of child care spaces does not matter, and the categorized child care spaces may do a better job in explaining the variation in the home prices. Simultaneously, we could leave the debt risk score as a numerical variable, and explore the relationship between the categorical independent variable child care spaces, the numerical independent variable debt risk score, and the study variable home prices in Toronto.

Moreover, in this report, we divide employments, businesses and child care spaces by the total area of each neighbourhood. This is because we want to "normalize" these variables and eliminate the effect of differences in the neighbourhood area. However, in future research, we could try another way to normalize these variables. We could try to divide these variables by the total population of each neighbourhood, and investigate the relationship between these variables per individual and the home prices.

## Discussion

In conclusion, we are saying that when considering the home prices in Toronto neighbourhoods, the number of child care spaces is not significantly linearly related to it, when the neighbourhoods are categorized based on debt risk level. On the other hand, the risk level of missing debt payments is significantly linearly related to the home prices, and neighbourhoods with low debt risk could have the highest average home prices among all risk levels when there are no child care spaces in the neighbourhoods. Also, the number of employments and business establishments may not explain much variation in the home prices in Toronto.

Eventually, this may bring potential guidance for people trying to purchase a home. If a family with children is seeking an area with more child care spaces, it is not necessary for them to purchase a home in a neighbourhood with high home prices. Also, if a family is trying to find a more expensive home, then it is likely that this family will find it in an area with low risk level of missing three debt payments consecutively. On the other hand, if a family is trying to purchase a home in a neighbourhood with more job opportunities in Toronto, then it may not matter whether this family chooses to settle in an area with higher or lower home prices.

# Bibliography

1. *About Our Data.* (n.d.). Realosophy. https://www.realosophy.com/about-our-data. (Last Accessed: October 16, 2021)

2. Agnew, K., & Lyons, R. C. (2018). *The impact of employment on housing prices: Detailed evidence from FDI in Ireland. Regional Science and Urban Economics, 70*, 174–189. https://doi.org/10.1016/j.regsciurbeco.2018.01.011

3. *Auxiliary variable.* (2018, May 8). CROS - European Commission. https://ec.europa.eu/eurostat/cros/content/auxiliary-variable_en#:~:text=In%20statistical%20research%2C%20an%20Auxiliary,all%20units%20of%20the%20population. (Last Accessed: October 15, 2021)

4. Bartolucci, F., & Scrucca, L. (2010). Point Estimation Methods with Applications to Item Response Theory Models. In *International Encyclopedia of Education* (Vol. 7, pp. 366–373).

5. Bhalla, D. (n.d.). *Difference between adjusted R-squared and R-squared.* ListenData. https://www.listendata.com/2014/08/adjusted-r-squared.html. (Last Accessed: October 19, 2021)

6. Bloomenthal, A. (2021, October 10). *Coefficient of Determination.* Investopedia. https://www.investopedia.com/terms/c/coefficient-of-determination.asp. (Last Accessed: October 18, 2021)

7. Canadian Centre For Economic Analysis, & Canadian Urban Institute. (2019, January). *Toronto housing market analysis.* https://www.toronto.ca/legdocs/mmis/2019/ph/bgrd/backgroundfile-140633.pdf. (Last Accessed: October 16, 2021)

8. Choueiry, G. (n.d.). *Understand Forward and Backward Stepwise Regression.* Quantifying Health. https://quantifyinghealth.com/stepwise-selection/. (Last Accessed: October 19, 2021)

9. Daysal, N. M., Lovenheim, M., Siersbæk, N., & Wasser, D. N. (2021). *Home prices, fertility, and early-life health outcomes.* Journal of Public Economics, 198, 104366. https://doi.org/10.1016/j.jpubeco.2021.104366

10. Gelfand, S. (2020). *opendatatoronto: Access the City of Toronto Open Data Portal.* R package version 0.1.4. https://CRAN.R-project.org/package=opendatatoronto

11. Gibbs, A., & Stringer, A. (2021, January 20). *Chapter 16 Short tutorial on pulling data for Assignment 1.* Probability, Statistics, and Data Analysis. https://awstringer1.github.io/sta238-book/section-short-tutorial-on-pulling-data-for-assignment-1.html#section-toronto-open-data-portal. (Last Accessed: October 16, 2021)

12. Harkness, J., Newman, S., & Holupka, C. S. (2009). Geographic Differences in Housing Prices and the Well-Being of Children and Parents. *Journal of Urban Affairs, 31*(2), 123–146. https://doi.org/10.1111/j.1467-9906.2009.00448.x

13. Hayes, A. (2021, September 24). *Error term.* Investopedia. https://www.investopedia.com/terms/e/errorterm.asp. (Last Accessed: October 16, 2021)

14. Hayes, A. (2021, March 30). *Multiple linear regression (MLR).* Investopedia. https://www.investopedia.com/terms/m/mlr.asp. (Last Accessed: October 16, 2021)

15. Hayes, A. (2021, January 30). *Linear relationships.* Investopedia. https://www.investopedia.com/terms/l/linearrelationship.asp. (Last Accessed: October 16, 2021)

16. *Interaction effects in regression.* (n.d.). Stat Trek. https://stattrek.com/multiple-regression/interaction.aspx. (Last Accessed: October 23, 2021)

17. Kenton, W. (2021, April 18). *Statistical significance definition.* Investopedia. https://www.investopedia.com/terms/s/statistically_significant.asp. (Last Accessed: October 16, 2021)

18. *Linear regression.*(n.d.). http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm. (Last Accessed: October 15, 2021)

19. *Lecture 17 Simple Regression.* (n.d.). http://web.pdx.edu/~newsomj/pa551/lectur17.htm. (Last Accessed: October 15, 2021)

20. Moon, K.W. (2020, October 6). *ggPredict() - Visualize multiple regression model.* The Comprehensive R Archive Network. https://cran.r-project.org/web/packages/ggiraphExtra/vignettes/ggPredict.html. (Last Accessed: October 23, 2021)

21. Nistor, A., & Reianu, D. (2018). Determinants of housing prices: evidence from Ontario cities, 2001-2011. *International Journal of Housing Markets and Analysis, 11*(3), 541–556. https://doi.org/10.1108/IJHMA-08-2017-0078

22. Ost, C.E. (2012). Housing and children: simultaneous decisions?— a cohort study of young adults' housing and family formation decision. *Journal of Population Economics, 25*(1), 349–366. https://doi.org/10.1007/s00148-010-0345-5

23. Press, J. (2021, May 20). *Bank of Canada warns of risks from household debt and housing market.* Advisor's Edge. https://www.advisor.ca/news/economic/bank-of-canada-warns-of-risks-from-household-debt-and-housing-market/. (Last Accessed: October 15, 2021)

24. R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/. (Last Accessed: October 15, 2021)

25. *Regression model assumptions.* (n.d.). JMP. https://www.jmp.com/en_ca/statistics-knowledge-portal/what-is-regression/simple-linear-regression-assumptions.html. (Last Accessed: October 23, 2021)

26. Student. (1908). The probable error of a mean. *Biometrika*, 1–25.

27. The Investopedia Team. (2021, April 30). *R-squared vs. adjusted R-squared: What's the difference?* Investopedia. https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp. (Last Accessed: October 19, 2021)

28. Toronto Open Data Portal. (2014). *Wellbeing Toronto – economics* [Data set]. Social Development, Finance & Administration. https://open.toronto.ca/dataset/wellbeing-toronto-economics/

29. Toronto Open Data Portal. (2017). *Wellbeing Toronto – demographics* [Data set]. Social Development, Finance & Administration. https://open.toronto.ca/dataset/wellbeing-toronto-demographics/

30. Trochim, W.M.K. (n.d.). *Dummy variables.* Conjoint.ly. https://conjointly.com/kb/dummy-variables/. (Last Accessed: October 16, 2021)

31. Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686. (Last Accessed: October 23, 2021)

32. Wickham, H. (2019). *stringr: Simple, consistent wrappers for common string operations.* R package version 1.4.0. https://CRAN.R-project.org/package=stringr. (Last Accessed: October 23, 2021)

33. Xie, Y. (2021). *knitr: A general-purpose package for dynamic report generation in R.* R package version 1.31.

34. Zach. (2020, December 6). *What is a partial F-test?* Statology. https://www.statology.org/partial-f-test/. (Last Accessed: October 23, 2021)

35. Zhu, H. (2021). *kableExtra: construct complex table with 'kable' and pipe syntax.* R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra. (Last Accessed: October 23, 2021)

# Appendix

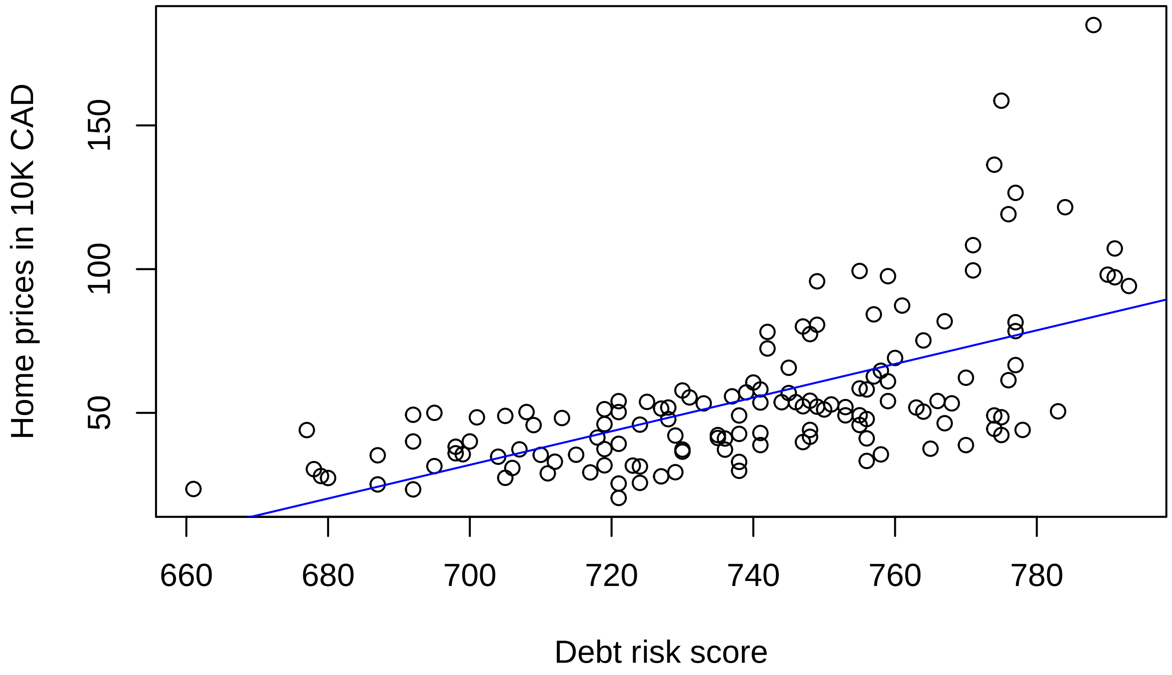
Here is a glimpse of the merged dataset with all the predictors in the hypothesis:

```
## Rows: 140
## Columns: 6
## $ Neighbourhood_id           <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9"~
## $ home_prices_10k            <dbl> 31.7508, 25.1119, 41.4216, 39.2271, 23.3832~
## $ risk_cate                  <chr> "middle", "high", "middle", "middle", "high~
## $ employment_per_area        <dbl> 1936.5570, 705.2174, 385.5882, 471.2000, 31~
## $ businesses_per_area        <dbl> 81.85444, 58.91304, 63.82353, 57.60000, 23.~
## $ child_care_spaces_per_area <dbl> 6.480558, 13.043478, 7.352941, 30.000000, 2~
```

The numerical variable debt risk score has a mean of 739.157, a median of 741 and a standard deviation of 28.626. The close mean and median values may suggest that there are no large outliers in this variable.

Plot 6 is the scatterplot of home prices and this numerical debt risk score. We can see that in the scatterplot, there appears to be a moderate positive linear association between the home prices and debt risk score. The blue line is a fitted frequentist linear regression line between these two variables. Notice that the variance of home prices across different values of debt risk score is approximately the same, except when the debt risk score is relatively high.

Plot 6: Home prices vs Debt risk score


Home prices in 10K CAD
Debt risk score