Analysis of Linear Regression Models for Toronto Home Prices[*]

Yetao Guo

18 April 2024

Abstract

Accurately predicting house prices is crucial for economic planning and individual investment decisions in the real estate market. This report uses a linear regression model based on 2011 Toronto house price data, aiming to predict house prices in the future. However, the study also reveals several weaknesses of the model and proposes some improvement measures, including collecting more recent data to enhance the timeliness of the model, re-evaluating and optimizing the selection of variables in the model, and exploring the introduction of other potential influences to improve the model's predictive power. We expect to construct a more accurate and comprehensive house price forecasting model through these measures, which will provide strong support for relevant decision-making.

Contents

Introduction **2**

Data **3**

Dataset 3

Variables 3

Missing Data 4

Plots 4

Model **13**

Features 13

Model Concerns 15

Results **15**

Discussion **16**

Data and Model Findings 16

Weaknesses and next steps 17

Appendix **18**

Datasheet 18

[*] https://github.com/yetaoguo/Toronto-house-price.git

References **22**

Introduction

As the global economy grows and urbanization accelerates, the real estate market remains a crucial component of national economies. It occupies a central role in the global economy, representing individual wealth and a substantial portion of national wealth, and is key to economic stability and growth (Jones 2021). Additionally, the real estate market significantly impacts financial stability; fluctuations in property prices can lead to economic booms and busts, such as the real estate boom and subsequent banking crisis that triggered the devastating financial crisis of 2008 (Jones, Cowe, and Trevillion 2018). Price volatility in real estate is of particular concern in economically developed and densely populated areas like Toronto. As Canada's largest real estate market, price fluctuations in Toronto affect consumer confidence and behavior since changes in real estate values influence household wealth and economic expectations. High housing costs in Toronto also impact affordability and living standards, posing critical challenges for urban planning and social policy. This report aims to explore Toronto's house prices in-depth and forecast them using linear regression models to provide a reference basis for relevant decision-making.

While exploring Toronto's house prices, we have selected several factors that may affect house prices as initial variables.

For data collection, we used the 2011 Well-being Toronto Housing, which is authoritative, accurate, and representative and can provide us with a reliable basis for analysis. We can explore the reasons behind Toronto's housing prices by digging deeper into these data. In the data processing and analysis stage, we first explored the distribution of the variables and revealed the quantitative characteristics and degree of dispersion among them through descriptive statistical analysis. Second, we used correlation analysis to examine the degree of correlation between the variables and house prices and determine which factors significantly impact house prices. The results of these analyses provide strong support for the subsequent modeling work. In constructing the linear regression model, we built a mathematical model that could reflect the changes in Toronto's house prices by choosing appropriate independent and dependent variables. We verified the reliability and validity of the model by testing the fit, significance level, and other indicators of the model.

We obtained the prediction model of Toronto house prices by predicting the linear regression model. The prediction results can be obtained for the given conditions, and these prediction results can not only help the government departments formulate reasonable real estate policies and guide the healthy development of the market but also provide decision-making references for investors and help them seize the market opportunities.

It is important to point out that although this report adopts a linear regression model to forecast Toronto house prices, the real estate market is a complex and changing system that is affected by various factors. Therefore, in practical application, we also need to combine other methods and models for comprehensive analysis to improve the accuracy and reliability of the forecast.

In summary, this report provides a vital reference for relevant decision-making through the in-depth discussion of Toronto house prices and the application of linear regression models. The Toronto real estate market will achieve a smoother and healthier development in the future under the joint effect of policy guidance and market mechanisms.

Data

Dataset

In order to achieve the goals set out in this paper, this paper provides data related to housing in different neighbourhoods in Toronto, based on 2011 Census data, as well as some summary statistics on these homes. This paper uses the R statistical programming language(R Core Team 2020), along with several packages: using tidyverse(Wickham et al. 2019), readxl(Wickham and Bryan 2023), haven(Wickham and Miller 2022)for data manipulation and preparation, the tables were created using knitr(Xie 2021) and kableExtra(Zhu 2021), DiagrammeR(Iannone 2023), ggplot2(Wickham 2016)build graphical diagrams,corrgram(Wright et al. 2021) was used for producing correlograms, car (Fox and Weisberg 2019)package provides advanced regression models, modelsummary(Arel-Bundock 2022) facilitated the summarization of model outputs,tinytext (Xie 2024)simplify the management of the LaTeX, statistical analyses were supported by MASS(Venables and Ripley 2002).

Variables

The dataset consists of 11 variables.

1. Neighbourhood: Name of Neighbourhood.
2. Neighbourhood Id: Specific IDs for different neighbourhoods.
3. Home Prices: Real Estate Sale Prices.
4. Social Housing Waiting List: Social Housing Waiting List.
5. Social Housing Turnover: Social Housing Turnover Rates.
6. Social Housing Units: Social Housing Total Units in the city.
7. Percent Mid-Century Highrise Population: Percent of Population Living in Mid-Century Highrises.
8. Rent Bank Applicants: Rent Bank Applicants.
9. Mid-Century Highrise Population: Mid-Century Highrise Total Population.
10. Mid-Century Highrise Households: Mid-Century Highrise Households.
11. Percent Mid-Century Highrise Households: Percent of Private Households in Mid-Century Highrises

    Of the above variables, Home prices is the response variable, except for the variables Neighbourhood and Neighbourhood Id which are potential predictors.

Home_Prices	Mid-Century_Highrise_Households	Mid-Century_Highrise_Population
317508	690	1810
251119	4110	13395
414216	430	1260
392271	600	1050
233832	870	2305
292861	4465	12445

Missing Data

The data is from 2011, which is a long time ago. The real estate market changes rapidly and is affected by various factors, such as economic conditions, policy adjustments, and population shifts. Therefore, using 2011 data to predict current or future housing prices may be highly biased. The data for each neighborhood are based on the mathematical aggregation of smaller sub-areas (in this case, Census Tracts) that, when combined, define the entire neighborhood. While I have listed multiple potential independent variables, other important variables have yet to be considered. For example, factors such as a community's infrastructure, educational resources, and accessibility may have a significant impact on home prices.

To sum up, although we used a linear regression model to forecast the 2011 Toronto housing price data, these data have inherent limitations. In practical applications, we need to collect and process data more comprehensively and accurately to improve the accuracy and reliability of our predictions.

Plots

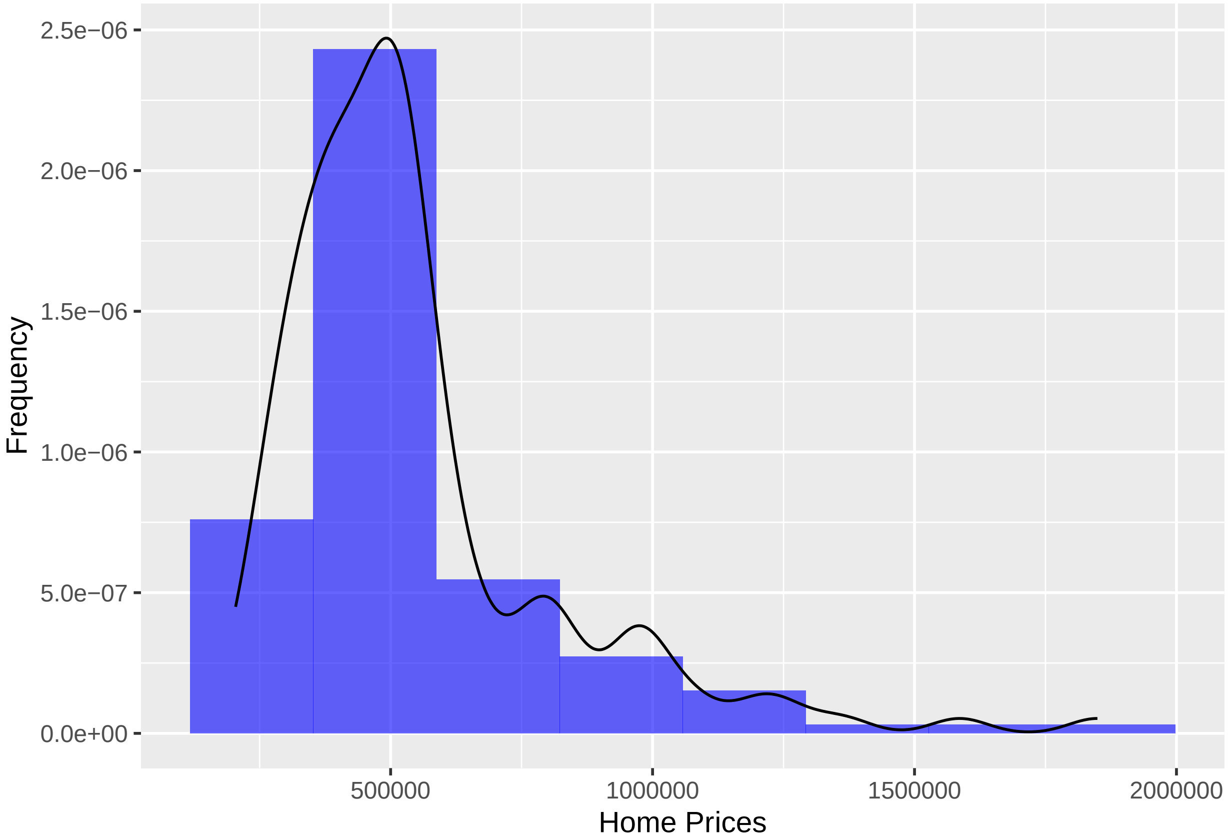

Figure 1: Distribution of Home Prices

Figure1 is the Home Prices histogram. It reflects the distribution of real estate sales prices in different communities. Looking at the histogram, it shows a clear right-skewed distribution, meaning that home prices in most communities are concentrated in a relatively low range, while home prices in a few communities are very high. This distribution may reflect differences in the level of economic development, geographic location, educational resources, infrastructure, and other market factors between communities.

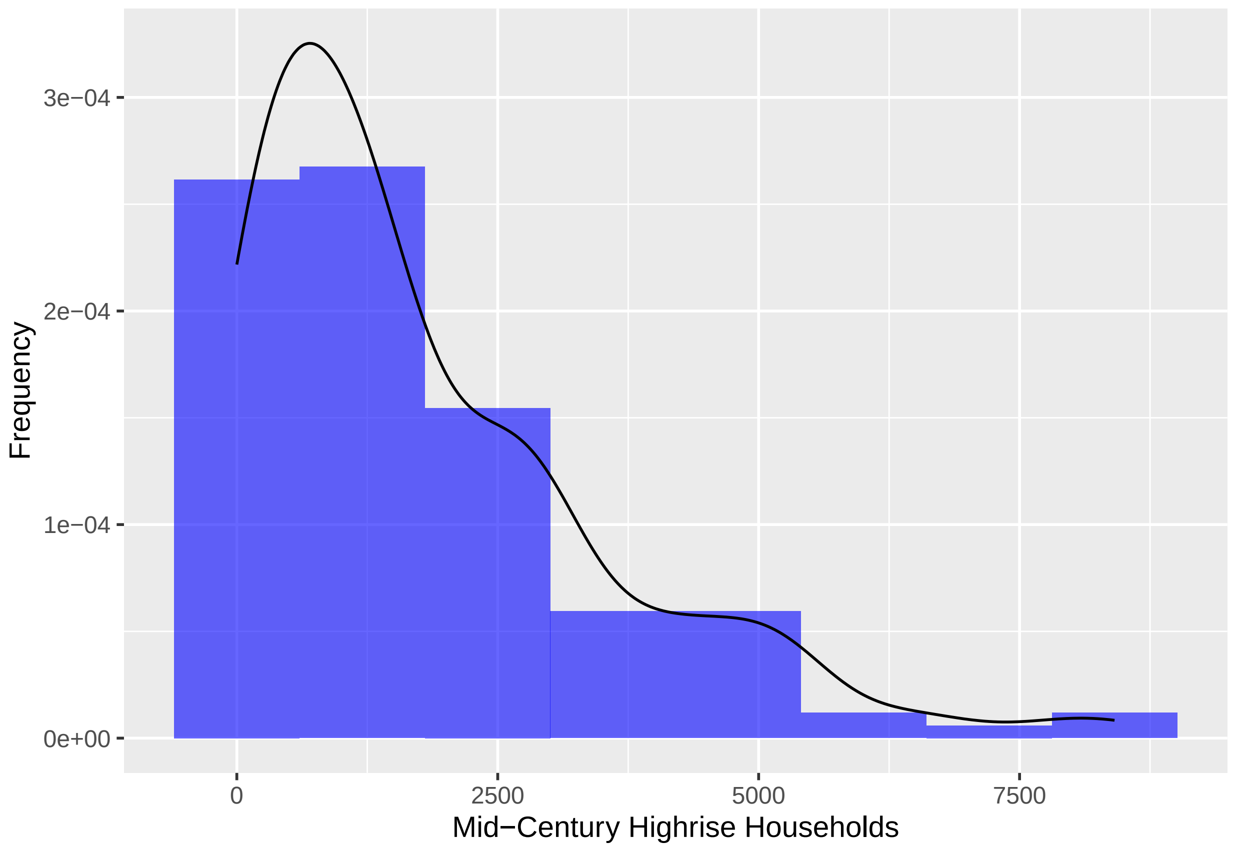

Figure 2: Distribution of Mid-Century Highrise Households

Figure2 is the histogram of Mid-Century Highrise Households. It similarly exhibits a right-skewed characteristic, suggesting that most neighbourhoods have a limited number of Mid-Century Highrise Households, while a few neighbourhoods have a higher number of such households. This reflects differences in the type of homes and composition of residents in different communities.

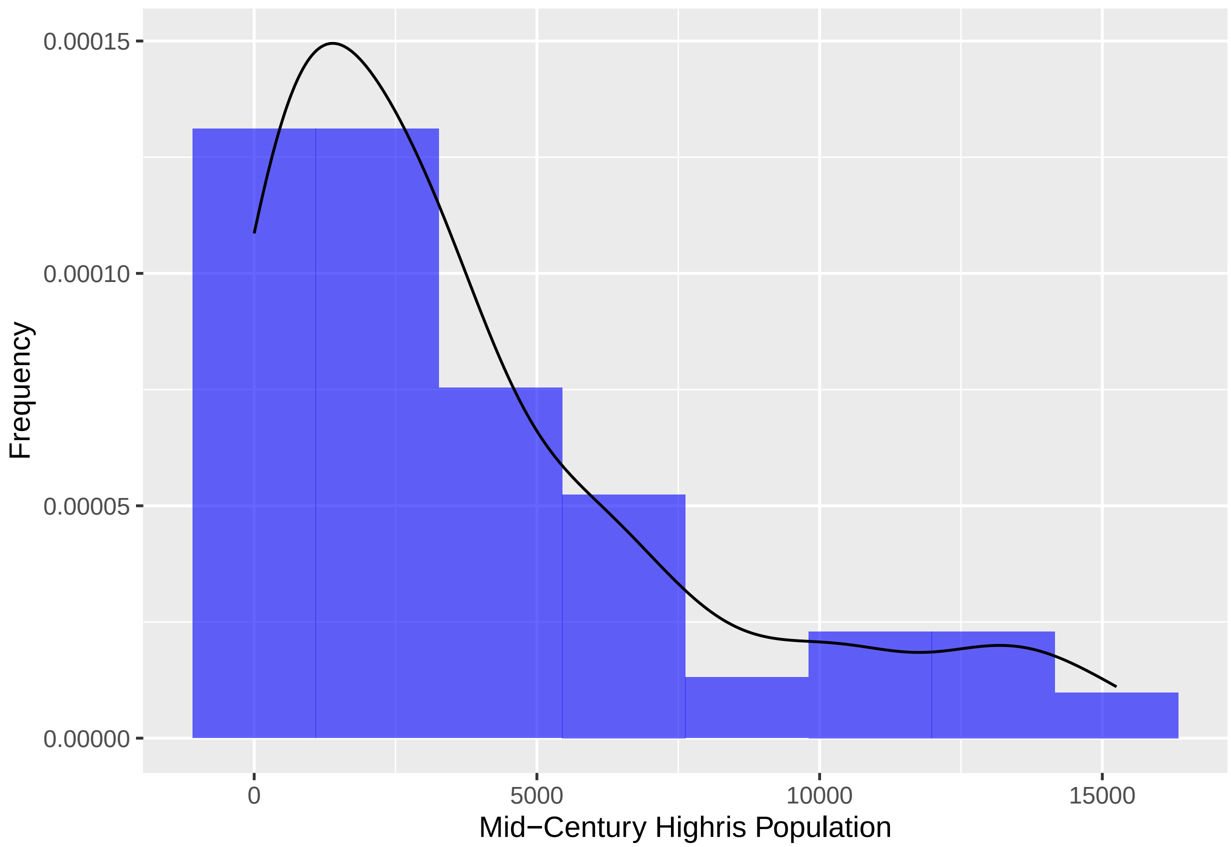

Figure 3: Distribution of Mid-Century Highris Population

Figure3 is the histogram of Mid-Century Highrise Population. It is also right-skewed, indicating that most communities have a low Mid-Century Highrise population, while a few communities have a high concentration of Mid-Century Highrise residents. This reflects the historical development and demographic composition of different communities with respect to high-rise buildings.

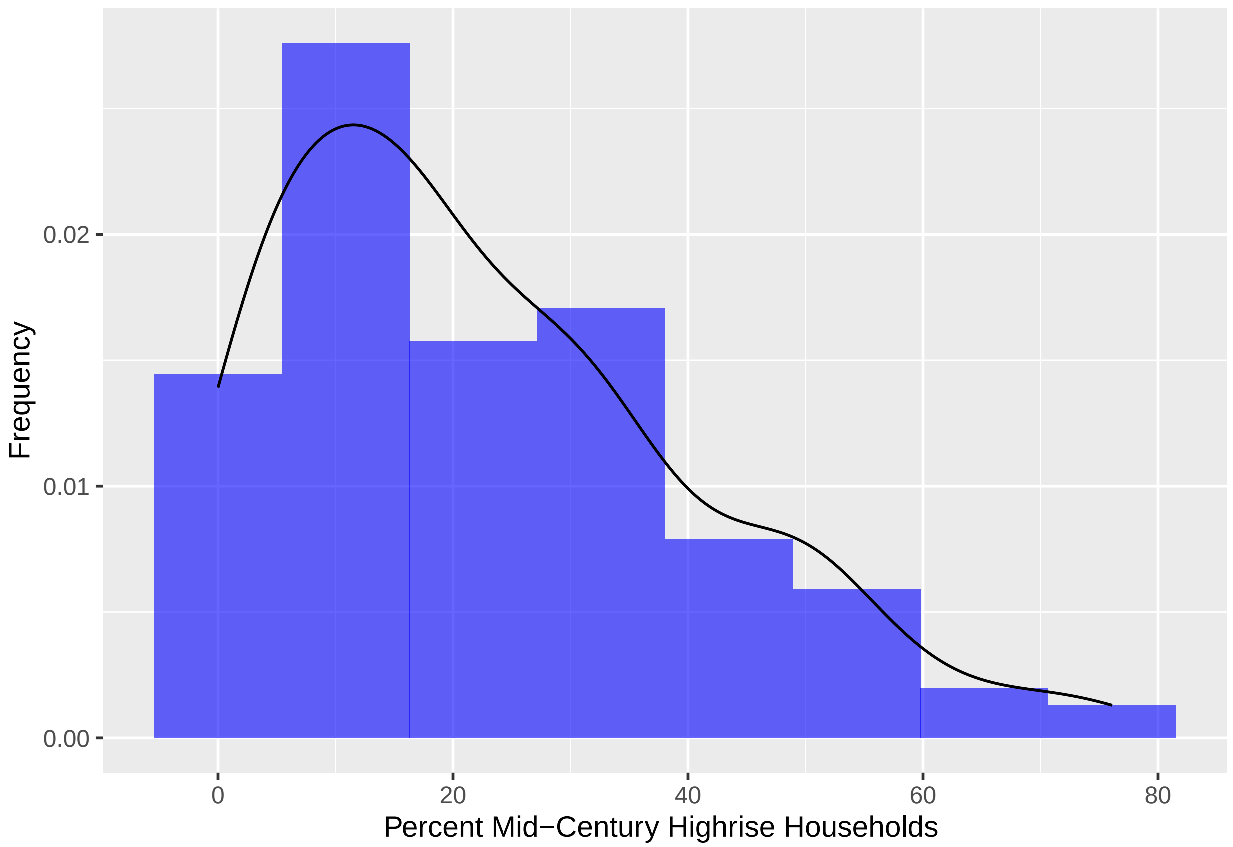

Figure 4: Distribution of Percent Mid-Century Highrise Households

Figure4 is the histogram of Percent Mid-Century Highrise Households.It also shows a right-skewed distribution. Most communities have a low percentage of private households in mid-century highrise housing, while a few communities have a high percentage. This is related to housing policies, resident demand, and housing market conditions in different communities.

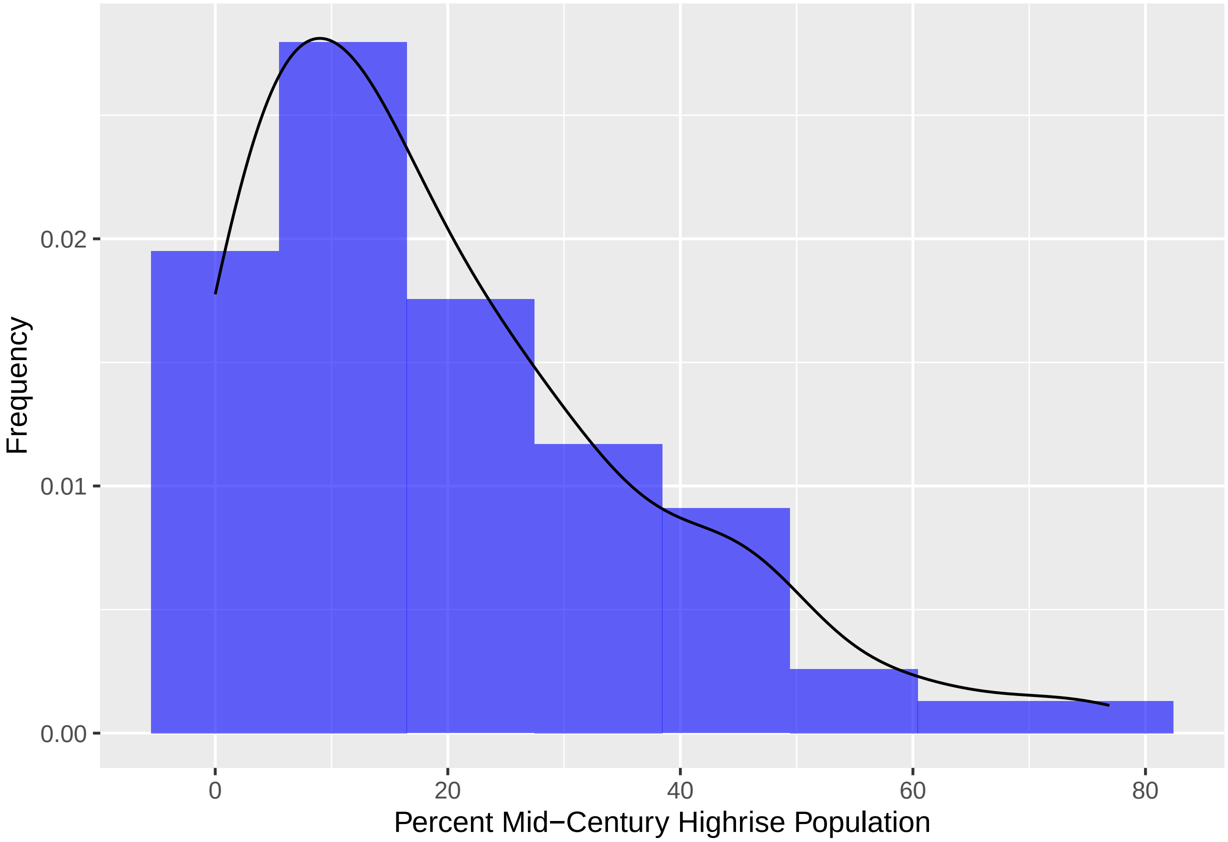

Figure 5: Distribution of Percent Mid-Century Highrise Population

Figure5 is the Percent Mid-Century Highrise Population histogram. It also shows a right-skewed pattern. This indicates that most neighbourhoods have a low percentage of population living in mid-century highrise buildings, while a few neighbourhoods have a higher percentage of mid-century highrise residents. This reflects patterns of urban growth and population movement, as well as differences in how different communities retain or update their high-rise buildings.

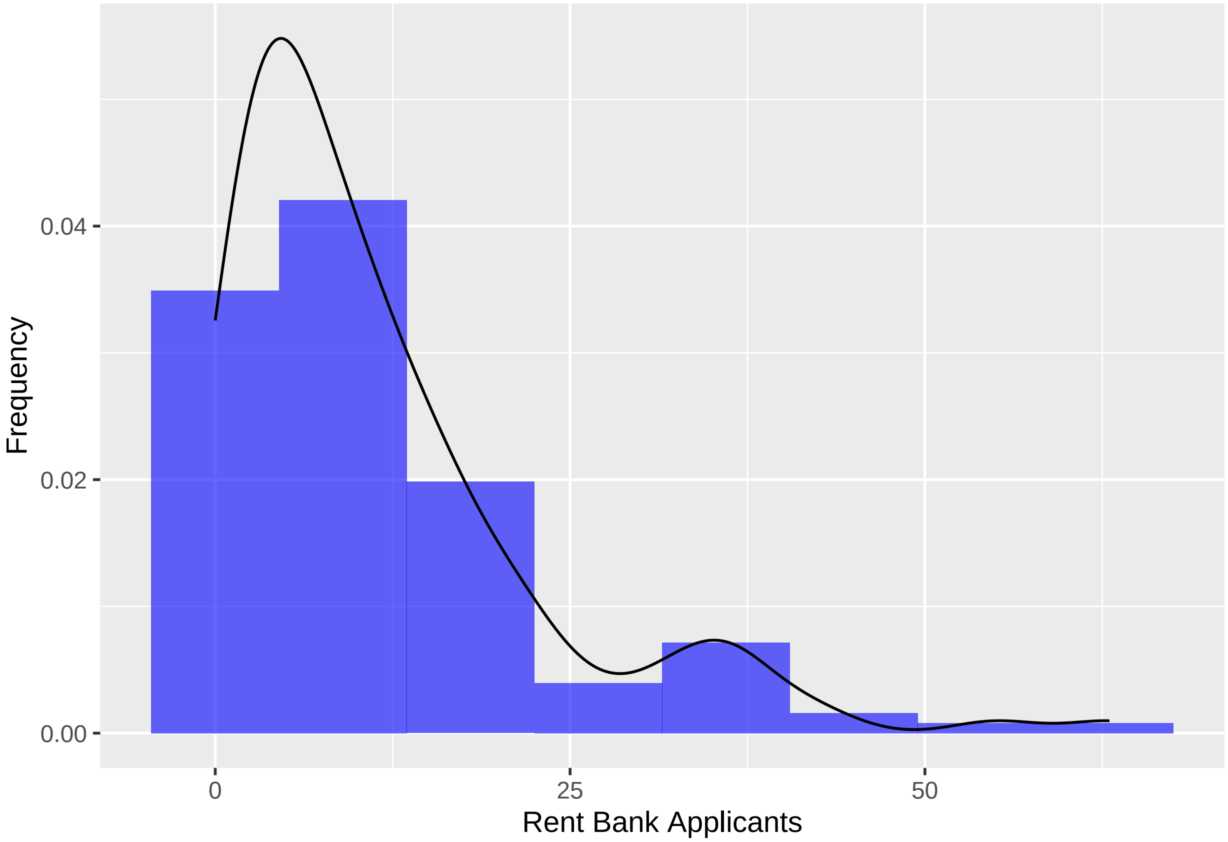

Figure 6: Distribution of Rent Bank Applicants

Figure6 is the histogram of Rent Bank Applicants ,which similarly shows a right-skewed distribution. Most neighbourhoods have a low number of Rent Bank Applicants, while a few neighbourhoods face a higher number of rent arrears, which is related to the economic conditions of the neighbourhood, the income level of the residents and the implementation of the rent policy.

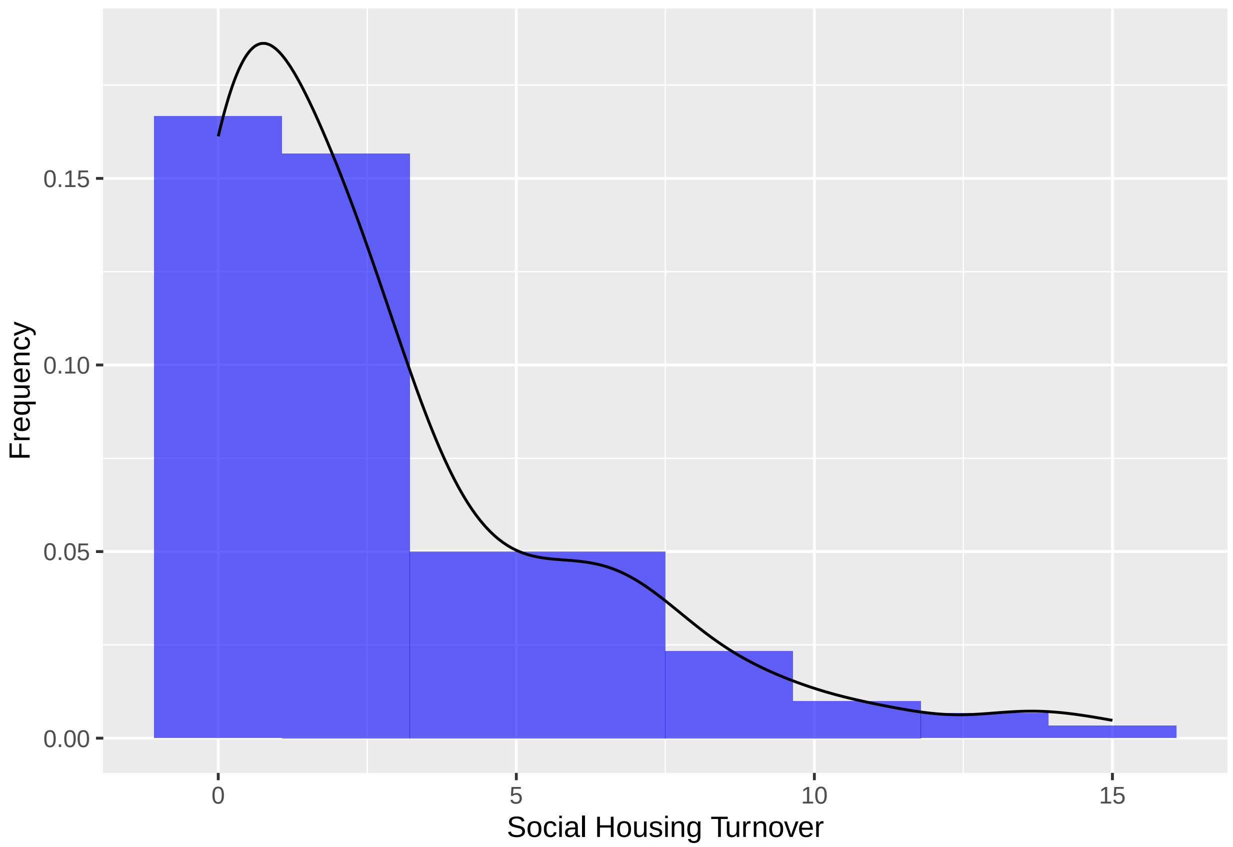

Figure 7: Distribution of Social Housing Turnover

Figure7 is the histogram of social housing turnover rates ,which also shows a right-skewed character. This means that in the majority of communities the turnover rate of social housing is at a low level, indicating a high level of housing stability. However, there are a few communities with high turnover rates, which indicates frequent turnover or higher mobility of social housing in these communities.

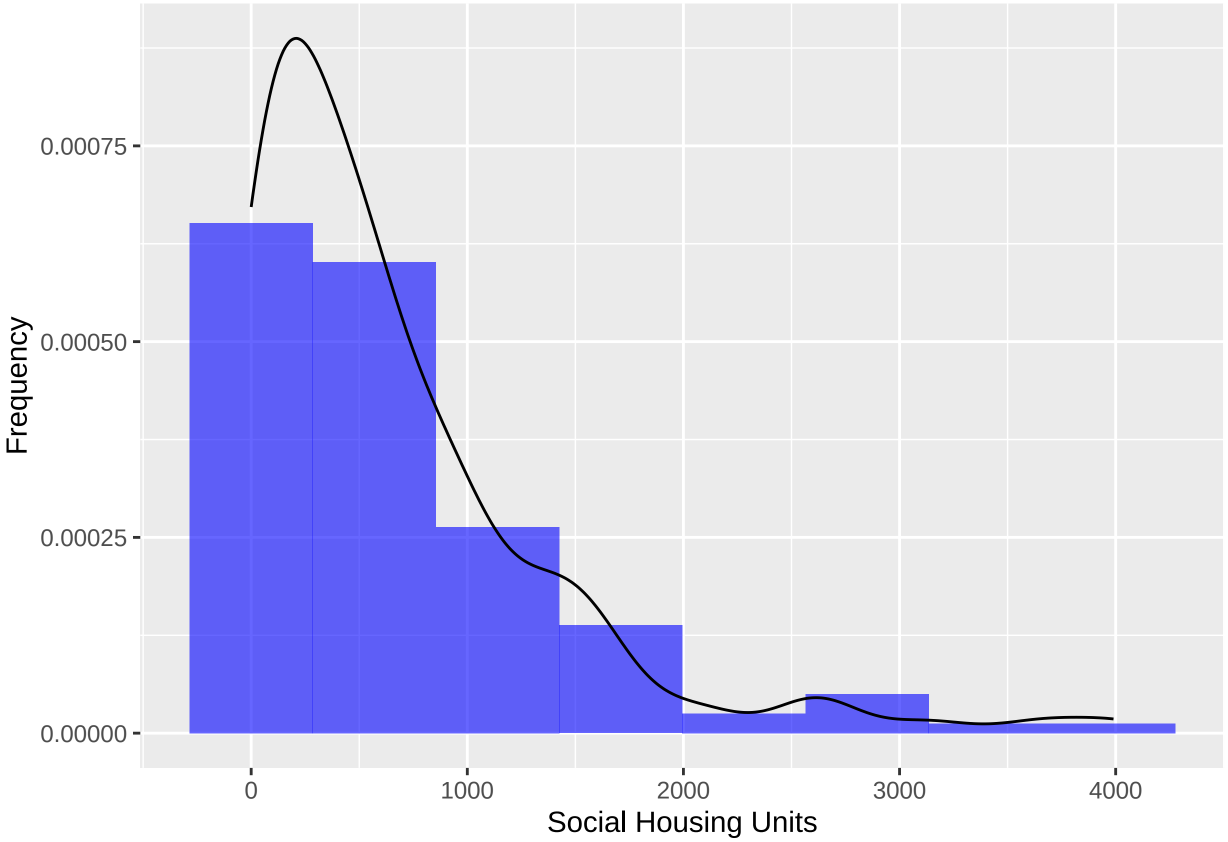

Figure 8: Distribution of Social Housing Units

Figure8 is the histogram of Social Housing Units. It shows a right-skewed distribution, suggesting that most communities have a relatively limited amount of social housing, while a few communities have a large amount of social housing. This is associated with the planning, policy and economic conditions of different communities, reflecting the uneven distribution of social housing between communities.

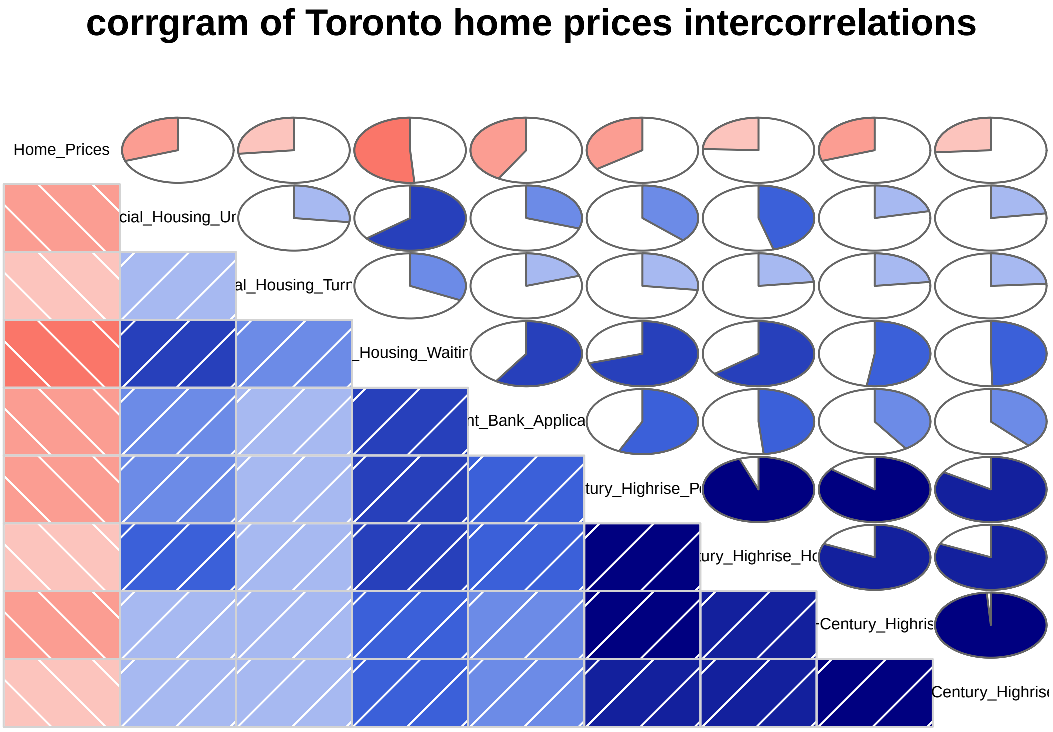

Figure 9: correlation matrix

Figure9 is a correlation coefficient plot ,it is a tool to visualise the strength of correlation between multiple variables. The strongest correlation is between Social_Housing_Waiting_List and Home Prices. This indicates that the length of the social housing waiting list has a strong negative correlation with home prices. In other words, the more applicants waiting for social housing, the lower the home prices are likely to be. This is because social housing usually exists to meet the needs of low-income households who is not be able to afford high-priced homes, and therefore the length of social housing waiting lists can be used as an indicator of the lower socio-economic conditions and relatively low level of house prices in the area.

Other variables, such as Social_Housing_Turnover, Social_Housing_Units, and Percent_Mid-Century_Highrise_Population, while also negatively correlated with housing prices, the specific strength of the correlation lies between Social_ Housing_Waiting_List and Mid-Century_Highrise_Households. These variables may reflect a number of aspects of a community's socio-economic status, housing policies, historical development, and the availability of specific types of dwellings, and their impact on housing prices may be multifaceted, both directly and indirectly. Correlation analysis only initially reveals the relationship between variables and does not directly indicate causality. In addition, linear correlation coefficients can only describe the linear relationship between the variables and cannot reveal the possible non-linear relationship. Therefore, I will consider other factors and conduct a more in-depth study during further data analysis and modelling.

Model

This paper predicts that Home prices depend on the community's Social Housing Waiting List, Social Housing Turnover, Social Housing Units, Percent Mid-Century Highrise Population, Rent Bank Applicants, Mid-Century Highrise Population, Mid-Century Highrise Households, and Percent Mid-Century Highrise Households. To try and model house prices I created the following multiple regression model:

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \epsilon$$

β_1 indicates how house prices change when Mid-Century_Highrise_Households change. If β_1 is positive, an increase in the number of Mid-Century_Highrise_Households leads to an increase in the price of houses; if it is negative, it leads to a decrease in the price of houses.

β_2 reflects the effect of Mid-Century_Highrise_Population on house prices. Its positivity, negativity, and magnitude also determine the direction and extent of change in house prices when population size changes.

β_3 shows the effect of Percent_Mid-Century_Highrise_Households on house prices. If β_3 is positive, a rise in percentage leads to a rise in house prices; and vice versa.

β_4 measures the effect of Percent_Mid-Century_Highrise_Population on house prices. Its interpretation is similar to β_3, but focuses on the percentage of population rather than the percentage of households.

The β_5 reflects the effect of Rent_Bank_Applicants on house prices. If the number of Rent_Bank_Applicants increases, how the price of a home will change depends on how positive or negative β_5 is.

β_6 and β_7 reflect the impact of Social_Housing_Turnover and Social_Housing_Waiting_List on house prices respectively. Social housing turnover may reflect activity in the housing market and changes in demand, while waiting list length may reflect pressure on housing demand and lack of supply.

Features

The final model obtained using backward stepwise regression contains four independent variables: Mid-Century_Highrise_Households, Percent_Mid-Century_Highrise_Households, Rent_Bank_Applicants and Social_Housing_Waiting_List. Bank_Applicants (number of rent bank applicants), and Social_Housing_Waiting_List (length of social housing waiting list). All four variables show a significant effect on Home_Prices at a significance level of 0.1.

Through backward stepwise regression, the model automatically eliminated variables that were highly co-variate with other variables and retained the most representative independent variables. This has the effect of reducing the impact of extraneous variables on model stability and accuracy. Compared to the initial model, the final model is more concise and contains only the necessary independent variables. This makes the model easier to interpret and apply, while also improving computational efficiency.

The independent variables in the model have clear economic significance. For example, the number of households in mid-century high-rise dwellings and their percentage share reflect the supply and demand for particular types of dwellings; the number of rent bank applicants reflect the level of activity in the rental market; and the length of waiting lists for social housing reveal pressures on housing demand and the effectiveness of the implementation of social housing policies. The model is applicable to similar areas or situations with similar property market characteristics.

However, it should be noted that due to the complexity and regional variability of the property market, the model requires further adaptation and validation when applied to other regions. Overall, this final model obtained after backward stepwise regression is a concise, reliable and economically meaningful multiple linear regression model that can be used to analyse and predict changes in housing prices.

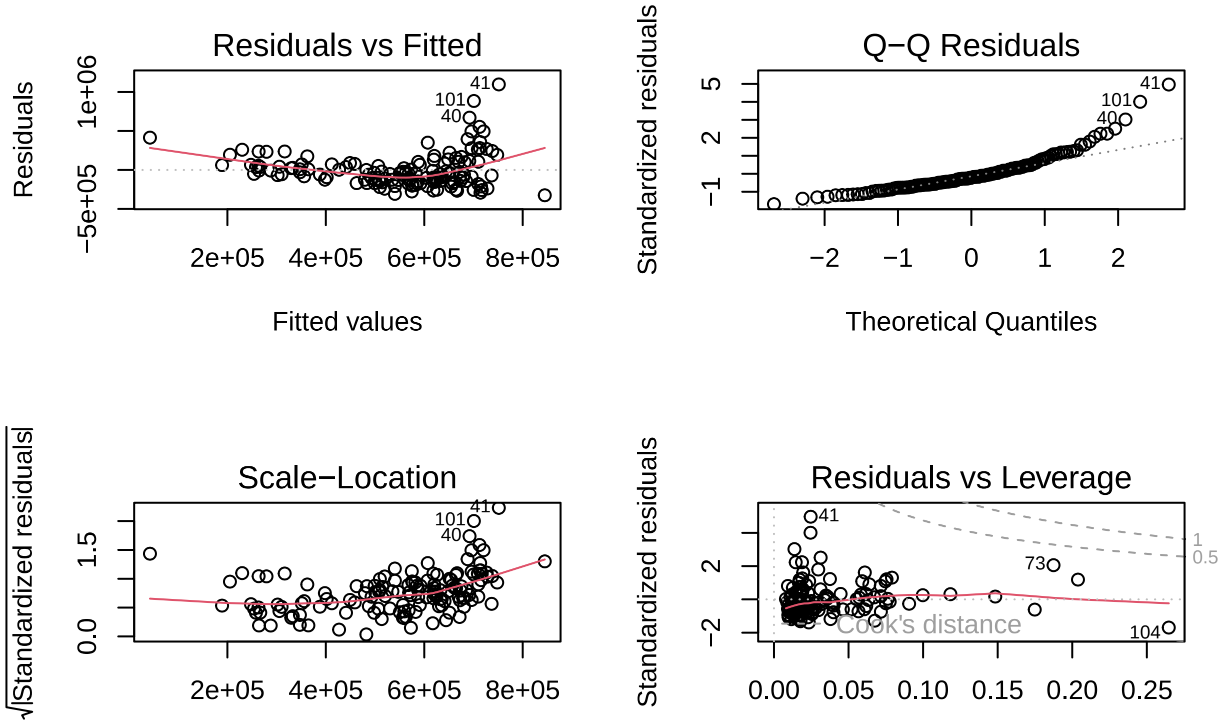

Figure 10: Residual Diagnostic Plots

Model Concerns

Figure10 is a backward stepwise regression yielded an optimal model with multiple significant independent variables, still has some areas of concern, especially when considering the data sources and prediction objectives.

Firstly, the issue of data timeliness is a notable concern. The model uses data from 2011, which is now more than ten old. During this time, the Toronto real estate market have experienced dramatic changes in a number of areas, including home prices, population distribution, and economic conditions.Therefore, models constructed based on outdated data is not accurately reflect current market conditions, leading to inaccurate prediction results.

Secondly, the homoskedasticity of the model residuals is not well satisfied, which may lead to instability of the estimated coefficients of the model and affect the predictive accuracy of the model. The residual analysis reveals the existence of three outliers, which is caused by data errors, atypical events or extreme conditions in the market. However, if we remove the outliers when doing data analysis, new outliers appear, so these three outliers are usualy ignored.

In addition, although the model selects a significant set of independent variables through backward stepwise regression, this does not mean that these variables still have the same predictive power in the current market environment. In particular, when market conditions change, some factors that were significant become less important, while some new factors come into play a role.

Finally, the linear regression model itself has its limitation. It assumes that there is a linear relationship between the independent and dependent variables, but in the real world this relationship is not viable. Especially when it comes to complex economic and social phenomena, a simple linear model may not be able to capture all the influencing factors and their interactions.

In short, especially when applying the model to property markets in different regions or over different time periods. The stability and forecasting accuracy of the model is affected by a variety of factors such as market environment and policy changes.

Results

I explored a model above and explained it by analysing the values of the predictive coefficients and their impact on the overall model. I will continue to share the results of this model.

Mid-Century_Highrise_Households: The coefficient estimate for this variable is 60.95, implying that an increase in the number of mid-century high-rise households by one unit is associated with an average increase in predicted house prices of 60.95 units (. This coefficient is statistically significant because the t-value is greater than 2 and the p-value is less than 0.01. Percent_Mid-Century_Highrise_Households: The coefficient estimate for this variable is -3695.21, indicating that a 1% increase in the percentage of mid-century high-rise households is associated with an average decrease in predicted home prices of 3695.21 units. However, the p-value for this coefficient is 0.05162, which is slightly higher than the usual significance level of 0.05, making this coefficient slightly less significant and is need to be analysed further on a case-by-case basis. Rent_Bank_Applicants: the coefficient estimate for the number of Rent_Bank applicants is -4845.12, indicating that when the number of Rent_Bank applicants increases by 1 unit, the predicted home price decreases by an average of 4845.12 units. This coefficient is statistically significant as the p-value is less than 0.05. Social_Housing_Waiting_List: The coefficient estimate for the length of the social housing waiting list is -444.49, indicating that when the length of the social housing waiting list increases by 1 unit, the predicted house price falls by an average of 444.49 units. This coefficient is statistically significant due to the very small p-value (much less than 0.01).

The Multiple R-squared is 0.3213, which indicates that the independent variables in the model explain 32.13% of the total variation in house prices. While this percentage is not particularly high, it does indicate that these independent variables have some degree of influence on house prices. There is also nearly 68%

	model1	model2	model3	Final model
(Intercept)	716 237.771	723 259.212	772 333.833	762 149.494
	(40 688.955)	(39 760.926)	(36 482.650)	(35 808.813)
Mid-Century_Highrise_Households	98.059	56.846	59.058	60.946
	(40.215)	(21.487)	(22.001)	(22.024)
Mid-Century_Highrise_Population	−21.019			
	(21.164)			
Percent_Mid-Century_Highrise_Households	15 966.168	16 071.097	−3410.584	−3695.209
	(7586.588)	(7263.366)	(1887.696)	(1881.735)
Percent_Mid-Century_Highrise_Population	−20 594.070	−20 708.369		
	(8199.639)	(7468.236)		
Rent_Bank_Applicants	−4047.954	−4271.715	−4825.895	−4845.122
	(2129.752)	(2045.014)	(2085.402)	(2091.779)
Social_Housing_Turnover	−8855.436	−10 481.578	−8539.593	
	(6276.900)	(6195.057)	(6307.096)	
Social_Housing_Units	−59.761			
	(39.867)			
Social_Housing_Waiting_List	−255.936	−370.261	−418.193	−444.493
	(114.394)	(88.490)	(88.925)	(87.045)
Num.Obs.	140	140	140	140
R2	0.379	0.367	0.330	0.321
R2 Adj.	0.341	0.338	0.305	0.301
AIC	3849.0	3847.6	3853.4	3853.3
BIC	3878.4	3871.1	3874.0	3871.0
Log.Lik.	−1914.476	−1915.785	−1919.719	−1920.670
F	9.984	12.854	13.227	15.977
RMSE	210 215.17	212 190.88	218 238.10	219 725.86

of the variation that is not explained by the variables in the model, which could mean that there are other important variables that are not being considered or that there is random error.

Discussion

Data and Model Findings

In this research, we utilize a dataset containing a number of key variables to explore the factors that influence house prices. The data covers a number of areas, such as the number of households in mid-century high-rise dwellings, the percentage of households in mid-century high-rise dwellings, the number of rent bank applicants, and the length of social housing waiting lists. These variables have been chosen because they are generally considered to be important factors influencing the property market. However, it is important to note that these variables appear to be comprehensive; it is possible that other potentially important influences, such as economic cycles, policy changes, or geographic location, have been missed.

Although some of the coefficients in the model are statistically significant, we also note that the explanatory power of the model is relatively limited. The Multiple R-squared value is only 0.3213, which implies a significant amount of variation not explained by the variables in the model. This may be due to the ignorance of important influencing factors in the model or due to the complex and non-linear relationship between the variables. Therefore, in future studies, we can consider incorporating more variables or using more complex models to improve the explanatory power and predictive accuracy of the models.

In summary, we can draw some preliminary conclusions and recommendations through the discussion of the data and model findings. The number of households in medieval high-rise dwellings has a positive effect on house prices. In contrast, the percentage of households in medieval high-rise dwellings, the number of rent bank applicants, and the length of social housing waiting lists have a negative effect on house prices. However, it is also important to note that the explanatory power of the model is limited and future research needs to further explore and consider additional influencing factors and model optimization methods.

Weaknesses and next steps

As mentioned earlier, the Multiple R-squared value of the model is 0.3213, indicating that the model explains only 32.13 percent of the variation in house prices. This implies that a significant amount of information needs to be included in the model and that all the important factors that affect house prices need to be adequately captured. This may result in the model's prediction accuracy not being high enough to provide an adequate basis for decision making. In constructing the model, although we have considered several variables, some important variables have not been included, such as market supply and demand, location superiority, school quality, and accessibility. These variables may have a significant impact on housing prices, but we did not include them in the model due to data access limitations or other reasons.

In future studies, we can try to incorporate more variables, especially those that have a significant impact on housing prices. By expanding the range of variables, we can improve the explanatory power of the model and make it more accurately reflect the actual situation of housing prices. To more accurately describe the relationship between variables and house prices, we can consider methods such as non-linear modeling or introducing interaction terms. This can better capture the complex relationship between variables and improve the prediction accuracy of the model.

Appendix

Datasheet

Motivation

1. `*For what purpose was the dataset created?Was there a specific task in mind?Was there a specific g`

   -The dataset was created to conduct a linear regression model based on 2011 Toronto home price data. The purpose is to predict home prices through a number of dimensions.

2. `*Who created the dataset (for example, which team, research group) and on behalf of which entity (`

   -The dataset was created by the author of this paper.

3. `*Who funded the creation of the dataset?If there is an associated grant, please provide the name o`

   -No monetary cost was required. All works are based on the programming software R (R Core Team 2020) and other useful packages.

4. `*Any other comments?   *`

   -The dataset can be a little old.

Composition

1. `*What do the instances that comprise the dataset represent (for example, documents, photos, people`

   -The instances in the dataset represent various housing-related data points for neighborhoods in Toronto. It includes information about housing prices, household demographics in mid-century high-rises, and social housing statistics. There do not appear to be multiple types of instances; all data points are related to housing and demographics.

2. `*How many instances are there in total (of each type, if appropriate)?  *`

   -141 instances are in total.

3. `*Does the dataset contain all possible instances or is it a sample (not necessarily random) of ins`

   -Typically, datasets provided by Open Toronto aim to be comprehensive but may sometimes represent a sample due to practical constraints. If it is a sample, the representativeness would need to be validated by the dataset providers.

4. `*What data does each instance consist of?   "Raw" data (for example, unprocessed text or images) o`

   -Each instance consists of “raw” data points, including numerical values for housing statistics like prices, highrise population percentages, and social housing waitlist numbers.

5. `*Is there a label or target associated with each instance? If so, please provide a description.*`

   -The first table is labeled with Percent of Private Households in Mid-Century High Rises.

6. `*Is any information missing from individual instances?   If so, please provide a description, expl`

   -N/A

7. `*Are relationships between individual instances made explicit (for example, users’ movie ratings,`

   -Not mentioned.

8. `*Are there recommended data splits (for example, training, development/validation, testing)?   If`

   -The dataset does not explicitly outline relationships like a social network or user ratings system.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a

   -It is a good quality dataset.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for exa

    -The dataset includes URLs, suggesting it is not entirely self-contained and linked to external resources such as housing agencies and statistical databases. There are no guarantees provided that these resources will remain constant over time.

11. *Does the dataset contain data that might be considered confidential (for example, data that is pro

    -No. Typically, datasets from public sources like this are anonymized and aggregated to avoid disclosing personal data.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening

    -No.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describ

    -The dataset appears to identify sub-populations in terms of household types, such as those living in mid-century highrise buildings. Distributions are given as percentages or counts per neighborhood but do not detail age, gender, or other demographic variables.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or

    -The dataset does not clearly indicate that individuals can be directly or indirectly identified. The data seems to be aggregated at the neighborhood level, which generally protects individual privacy.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data tha

    -There does not appear to be sensitive data

16. *Any other comments?*

    -N/A

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for ex

   -The dataset does not provide information about the individual instances of data acquisition methods, such as whether the data was directly observed, reported by subjects, or inferred from other data.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or

   -The data came from various sources, including Toronto Community Housing Corporation, the City of Toronto's Shelter, Support and Housing Administration, the City of Toronto Affordable Housing Office, and Statistics Canada. Average Home Price data was taken from Realosophy.com.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, determ

   -Not mentioned.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors

   -Not mentioned.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of t

   -The dataset indicates that the data was updated in 2011. Still, there is no clear information about the exact duration of data collection or whether the data collection timeframe matches the creation timeframe of the instances.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If s

   -Not mentioned.

7. *Did you collect the data from the individuals in question directly, or obtain it via third partie
   -Open Toronto website.
8. *Were the individuals in question notified about the data collection? If so, please describe (or s
   -No.
9. *Did the individuals in question consent to the collection and use of their data? If so, please de
   -Open Toronto is a public website that everyone can access.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their
    -N/A
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example,
    -None
12. *Any other comments?* -None

Preeprocessing/cleaning/labeling

1. *Was any preeprocessing/cleaning/labeling of the data done (for example, discretization or bucketin
   -Highlight some rows in column SHORT_NAME with yellow color
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to sup
   -The “raw” data is contained in the GitHub repository in the paper.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide
   -The software used to clean the data consists of R packages, which have been cited and credited.
4. *Any other comments?*
   -None

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   -No
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so,
   -Yes.They are available at: https://github.com/yetaoguo/Toronto-house-price.git
3. *What (other) tasks could the dataset be used for?*
   -Other tasks might consist of more analysis on the prediction of Toronto house prices and factor that caused increase and decrease in house price
4. *Is there anything about the composition of the dataset or the way it was collected and preprocess
   -no effect
5. *Are there tasks for which the dataset should not be used?If so, please provide a description.*
   -The dataset should not be used for any illegal activity.
6. *Any other comments? *
   -None

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, in

   -The dataset will not be distributed to third parties but will be shown on GitHub.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does th

   -The dataset will be distributed on GitHub via this paper's repository.

3. *When will the dataset be distributed?*

   -The dataset will be distributed on April 18, 2024.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, a

   -N/A

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the in

   -N/A

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual in

   -N/A

7. *Any other comments?*

   -N/A

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   -The author of the paper will host the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   -The dataset manager can be found on the GitHub website.

3. *Is there an erratum?If so, please provide a link or other access point.*

   -None

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete

   -The dataset will not be updated.

5. *If the dataset relates to people, are there applicable limits on the retention of the data asso

   -There are no limits.

6. *Will older versions of the dataset continue to be supported/hosted/maintained?If so, please des

   -The author of the paper may host older versions of the dataset,but data will not be maintained.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for t

   -Others who want to build on to the data set through the GitHub website should be verified by the author.

8. *Any other comments?*

   -N/A

References

Jones, C. (2022). Urban economy. : real estate economics and public policy. Routledge.

Jones, C., Cowe, S., & Trevillion, E. (2018). Property boom and banking bust : the role of commercial lending in the bankruptcy of banks. John Wiley & Sons, Inc.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, H. (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. Retrieved from http://haozhu233.github.io/kableExtra/ and https://github.com/haozhu233/kableExtra

Xie, Y. (2021). Knitr: A general-purpose package for dynamic report generation in R (R package version 1.37). Retrieved from https://yihui.org/knitr/

Arel-Bundock, V. (2022). Modelsummary: Summary tables and plots for statistical models and data: Beautiful, customizable, and publication-ready (R package version 0.10.0). Retrieved from https://vincentarelbundock.github.io/modelsummary/

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. ISBN: 978-3-319-24277-4. Available at https://ggplot2.tidyverse.org

Xie, Y. (2024). Tinytex: Helper functions to install and maintain TeX Live, and compile LaTeX documents (R package version 0.50). Retrieved from https://github.com/rstudio/tinytex

Wickham, H., & Bryan, J. (2023). Readxl: Read Excel files. Retrieved from https://readxl.tidyverse.org and https://github.com/tidyverse/readxl

Fox, J., & Weisberg, S. (2019). An R companion to applied regression (3rd ed.). Sage. Thousand Oaks, CA. Retrieved from https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S (4th ed.). Springer. New York. ISBN 0-387-95457-0. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/

Wright, K., et al. (2021). Corrgram: Plot a correlogram (R package version 1.14). Retrieved from https://kwstat.github.io/corrgram/

Wickham, H., & Miller, E. (2022). Haven: Import and export 'SPSS', 'Stata' and 'SAS' files (R package version x.x.x). Retrieved from https://CRAN.R-project.org/package=haven

Iannone, R. (2023). DiagrammeR: Create graph diagrams and flowcharts using R (R package version 1.0.6.1). Retrieved from https://CRAN.R-project.org/package=DiagrammeR

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://vincentarelbundock.github.io/modelsummary/.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression.* Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Iannone, Richard. 2023. *DiagrammeR: Create Graph Diagrams and Flowcharts Using r.* https://CRAN.R-project.org/package=DiagrammeR.

Jones, Colin. 2021. *Urban Economy: Real Estate Economics and Public Policy.* Routledge.

Jones, Colin, Stewart Cowe, and Edward Trevillion. 2018. *Property Boom and Banking Bust: The Role of Commercial Lending in the Bankruptcy of Banks.* John Wiley & Sons.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s.* Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Jennifer Bryan. 2023. *Readxl: Read Excel Files.*

Wickham, Hadley, and Evan Miller. 2022. *Haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files.* https://CRAN.R-project.org/package=haven.

Wright, Kevin et al. 2021. *Corrgram: Plot a Correlogram.* https://kwstat.github.io/corrgram/.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

———. 2024. *Tinytex: Helper Functions to Install and Maintain TeX Live, and Compile LaTeX Documents.* https://github.com/rstudio/tinytex.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/ https://github.com/haozhu233/kableExtra.