# What factors are most likely to affect Toronto apartment ratings? An exploratory study of Toronto apartment data*

Yiming Tang

January 25, 2024

This study conducts an exploratory analysis of Toronto apartment ratings, leveraging open data from the City of Toronto's datasets spanning 2017 to 2023. The comprehensive data cleaning process involves variable selection, type conversion, and handling missing values and outliers. Visualizations, including histograms, boxplots, and scatter plots, provide insights into the dataset. The analysis identifies significant variables correlated with apartment scores and highlights the potential impact of property types on ratings. Recommendations for future work include statistical modeling to assess variable influences on scores and the exploration of geographical information through maps.

## Table of contents



*Code and data supporting this analysis is available at: https://github.com/Yiming1220/STA302

# 1 Introduction

Housing research has a long history in sociology, it is not only a commodity but also a kind of right (Pattillo, 2013). In this context, we see increasing attention to housing issues, which is not only about the actual needs of housing, but also involves the pursuit of social justice and equality. This evolving emphasis underscores housing research as an increasingly pivotal and substantive subject.

Meanwhile, Toronto's Apartment Building Standards program, initiated in 2017, stands as a pivotal bylaw enforcement initiative aimed at ensuring that proprietors and operators of apartment buildings comply with rigorous maintenance standards. This program, introduced to buildings with three or more storeys or 10 or more units, places a significant emphasis on maintaining the overall integrity of these structures (City of Toronto, 2024). This study delves into the complexity of the data, with a particular focus on establishing assessment scores. Code enforcement officers carefully examine all aspects during the assessment process, including common areas, mechanical and security systems, parking facilities and exterior grounds (City of Toronto data, 2024).

This report endeavors to conduct an exploratory analysis of the rating data for Toronto apartments before the year 2023. The research aims to unveil the factors most likely to influence the ratings of apartments in Toronto.

# 2 The Data

## 2.1 Data Description

The data is open data from the City of Toronto website (City of Toronto data, 2024), which records pertinent information of registered apartments in Toronto from the establishment of the Apartment Building Standards (RentSafeTO) until 2023, namely, the years from 2017 to 2023. Notably, there are 3,072 missing values within this dataset. The dataset comprises 11,760 rows and 40 columns, encompassing variables such as apartment identifiers, age, geographical location, feature details, among others.

## 2.2 Data Cleaning

Data cleaning can be delineated into four principal steps: variable selection, conversion of variable types, handling missing values, and addressing outliers.

Except for the property type column, all other character variables, including site address, ward name, etc., are excluded from the dataset. Subsequently, numerical variables lacking a discernible relationship with the rating are also removed, encompassing house identifiers, evaluation years, longitude, and latitude. Consequently, the number of data columns undergoes a reduction from 40 to 27. Due to the presence of the character string "N/A" in certain variables that should inherently be numerical, these variables are converted into numeric format. The instances of "N/A" are transformed into missing values upon conversion, leading to an increase in the count of missing values to 33,664. If all these missing values are removed, the dataset retains a mere 5.6% of the original samples. However, given the specific nature of certain variables, such as the year and the quantity of elevators, the application of mean imputation may not be entirely appropriate. Therefore, further variable selection is undertaken.

If at least 85% of the sample is to be retained, the data cannot be removed beyond 1764 rows. Taking into account that the variables with missing values may vary across different samples, vthose containing more than 1200 missing values will be excluded. Table 1 below shows the variables that will be further removed. After excluding the following variables, and subsequently removing all rows containing missing values, the sample size of the data amounts to 85.8% of the original dataset.

Table 1: Variables with more than 1200 missing values

| Variables | NAs |
|---|---|
| GARBAGE_CHUTE_ROOMS | 6658 |
| ELEVATORS | 4863 |
| STORAGE_AREAS_LOCKERS | 6987 |
| BALCONY_GUARDS | 3787 |
| OTHER_FACILITIES | 9506 |

Figure 1 presents a boxplot of the target variable, Score, revealing the presence of some outliers at the lower end of the distribution. Upon calculation, a total of 17 outliers are identified and subsequently removed.

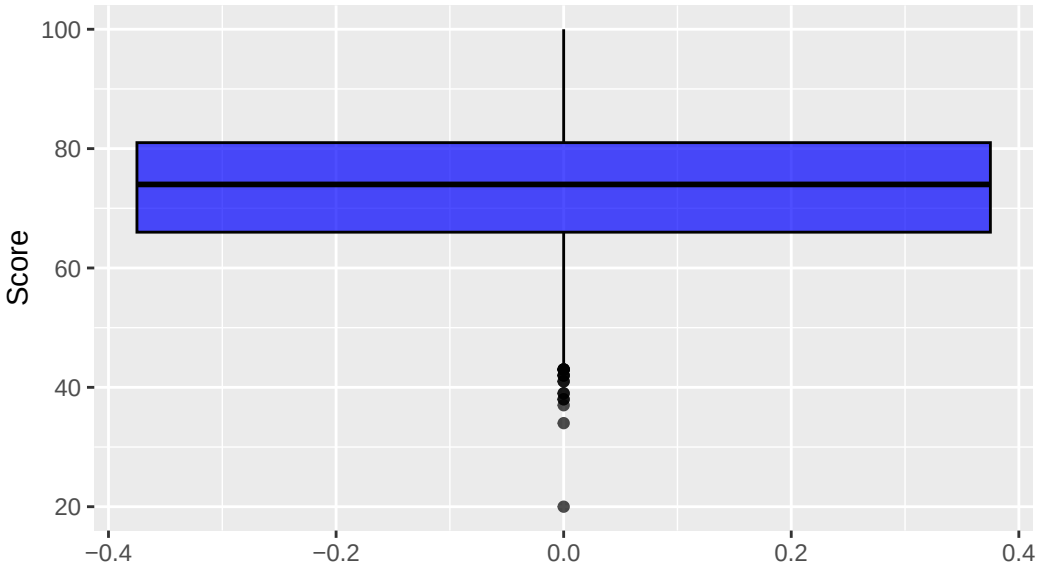


Figure 1: Boxplot of score

# 3 Results

## 3.1 Distribution of Score

Figure 2 illustrates the distribution of Score. The majority of apartments in the sample exhibit ratings concentrated between 60 and 85, with the highest prevalence observed for apartments with a score of 76.

## 3.2 The Relationship between House Type and Score

As previously mentioned, post variable selection, the dataset retains a non-numeric variable, property type. In order to comprehend its impact on the Score, it is imperative to isolate and conduct a dedicated analysis of this variable.

According to Figure 3, the property type is categorized into three distinct classes: private, social housing, and TCHC (Toronto Community Housing Corporation). Among these, the social housing category exhibits the highest median, while TCHC has the lowest median. The private category displays the most extensive distribution, generally positioned between the two.

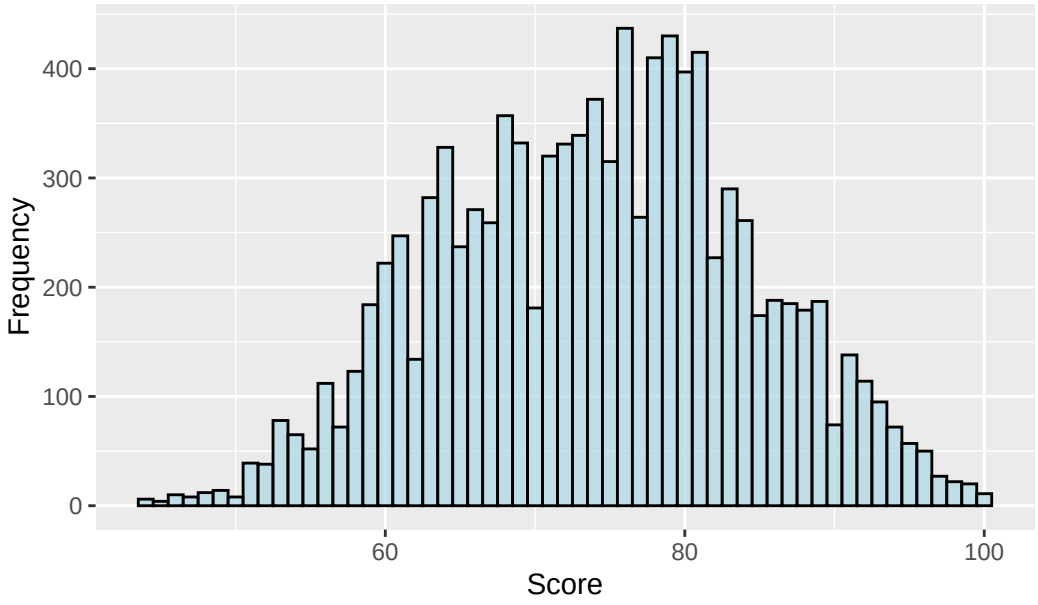


Figure 2: Distribution of Score

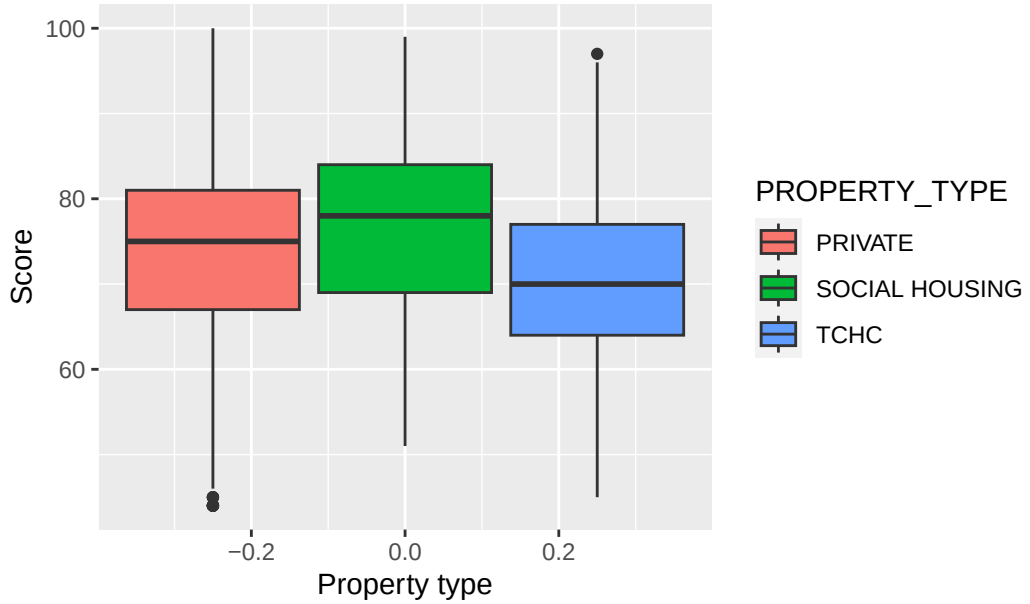


Figure 3: Boxplot of Scores by Property type

### 3.3 Correlation between Other Variables and Score

Excluding the column representing the property type, correlation coefficients are computed between the remaining variables and the Score. Variables exhibiting a strong correlation with the Score (absolute correlation coefficient greater than or equal to 0.7) are listed in Table 2. The ranges of these seven variables all span from 1 to 5, and each exhibits a positive correlation with the Score. Their increments are associated with an increase in the Score to a certain extent.

Table 2: Variables that are strongly correlated with Score

| Variables | Correlation.with.Score |
|---|---|
| ENTRANCE_LOBBY | 0.7594 |
| EXTERIOR_GROUNDS | 0.7282 |
| ENTRANCE_DOORS_WINDOWS | 0.7249 |
| STAIRWELLS | 0.7209 |
| LAUNDRY_ROOMS | 0.7160 |
| INTERIOR_LIGHTING_LEVELS | 0.7080 |
| INTERIOR_WALL_CEILING_FLOOR | 0.7007 |

Figure 4 depicts a scatter plot of the Score against the entrance lobby, with an added trend line for assessment. It illustrates a discernible positive correlation.

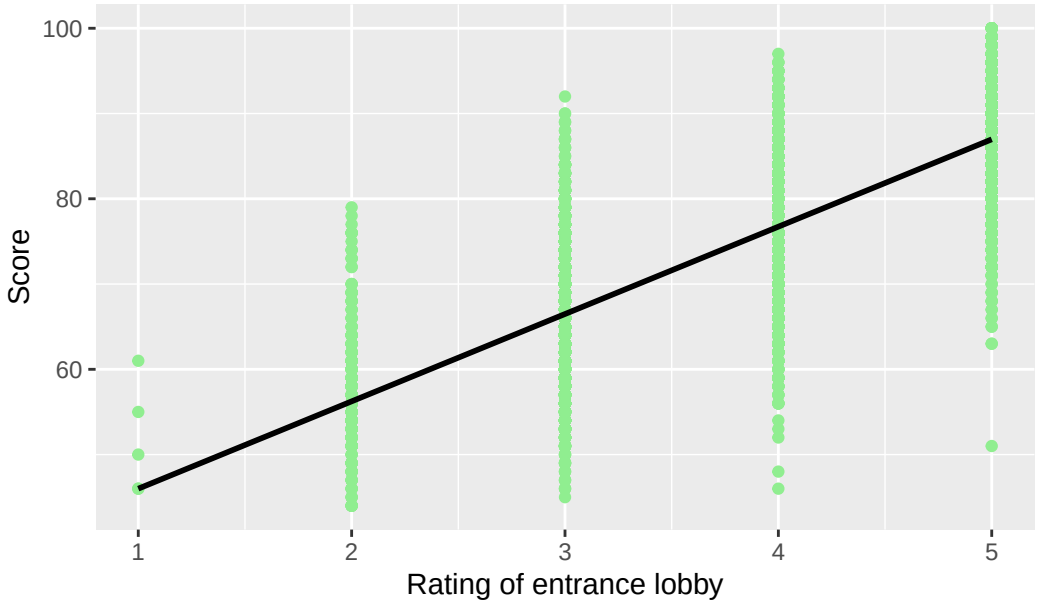


Figure 4: Scatter Plot for score vs entrance lobby

## 4 Discussion and Conclusion

This report provides insights into the factors influencing the ratings of apartments in Toronto through an exploratory analysis of open data from the Toronto government. The data cleaning process involved rigorous steps, including variable selection, type conversion, and addressing missing values and outliers. Data analysis was facilitated through the visualization of histograms, box plots, and scatter plots.

The property type characterized by private, social housing, and TCHC exhibits impacts on apartment ratings. Among these, social housing attains the highest median score, while private housing demonstrates the widest distribution. Furthermore, the identification of seven variables closely correlated with the scores provides actionable insights for apartment owners and operators. These variables, listed in descending order of correlation strength, include entrance lobby, external ground, entrance doors windows, stairwells, laundry room, indoor lighting levels, and interior wall ceiling floor ratings.

## 5 Limitation and Future Work

The utilization of scatter plots in the report may not be entirely appropriate for variables with only five ratings. For such variables, visualizations through boxplots would be more suitable. In addation, the variety of visualizations in the report seems somewhat limited, more types of plot and table can be introduced for data visualization.

In future work, statistical modeling can be employed to further analyze the impact of independent variables on the Score. Additionally, geographical information maps can be created using latitude and longitude, as well as Ward identifiers.

## 6 Reference

City of Toronto. (2024) "RentSafeTO: Evaluation Tool – City of Toronto." https://www.toronto.ca/

City of Toronto data (2024) "City of Toronto Open Data Portal". https://open.toronto.ca/dataset/apartment-building-evaluation/

Pattillo, M. (2013) 'Housing: Commodity versus right', Annual Review of Sociology, 39(1), pp. 509–531. doi:10.1146/annurev-soc-071312-145611.