The location of an apartment has an impact on its quality*

An analysis on the relationship between features and score of apartments in Toronto

Zihao Liu

24 April 2022

Abstract

This report explores the relationship between the quality and the features of apartments in Toronto. The year when the apartments were built ranges from 1800 to 2000s. And it was found that the wards where apartments locate and property type of apartments are influential to the quality of apartments. The findings in this report can help companies in real estate industy, and the results can help people who are renting to find appropriate housings.

Keywords: Multiple linear regression; Toronto; Apartments; Housing; Applied statistics

1 Introduction

Housing is an important part of our daily life. A high quality living space benefits people who live in there. Apartment, as one of the housing types, is common in big cities, like Toronto. The apartments are diversified, some of them are located downtown, and some of them are located in less crowded areas. Some of the apartments are modern high-rise, and some of the apartments are low-rise and have a long history. This report is going to explore different features of apartments in Toronto, and their relationship with the quality of apartments will be explored as well.

This report is going to have a brief overview of the features and quality of apartments in Toronto. And a multiple linear regression model is used to analyze the relationship between the quality of apartments and the features of apartments. Other summary statistics are used to analyze the impacts of the features as well. After the analysis, it was found that the property types and wards where the apartments are located have impacts on the quality(score) of apartments, while the number of storeys and number of units of an apartment have only slight influences on the quality of apartments. The findings can be helpful for companies in real estate industry. And some of the results can also assist people who are renting apartments in Toronto.

Data section includes the source of the data, the cleaning process, and an overview of the variables in the data. Model section contains the mathematical equation for the model, the explanation of the equation and other evaluation of the performance of the model. Results section includes the result of the model and some other summary statistics associated with the features of apartments. Discussion section contains the summary of the findings, the weaknesses, and the further steps of this report. This report and the analysis in this report is created using R programming language (R Core Team 2020). R packages `tidyverse`(Wickham et al. 2019), `knitr`(Xie 2014), `dplyr`(Wickham et al. 2022), `ggplot2`(Wickham 2016),`here`(Müller 2020), `gridExtra`(Auguie 2017), `tidymodels`(Kuhn and Wickham 2020), and `modelsummary`(Arel-Bundock 2022) were used in this report.

2 Data

The data used for this report is from Open Data Toronto Portal. And it is imported using the R package `opendatatoronto` (Gelfand 2020). The data contains the features and scores of apartments registered with

*Code and data are available at: https://github.com/ZihaoLiu2/Toronto_apartments_statistics

Table 1: First 10 rows of the dataset of apartment data in Toronto

Year Registered	Year Evaluated	Year Built	Property Type	Ward	Ward Name	Address
2017	2021	1959	PRIVATE	14	Toronto-Danforth	190 COSBURN
2017	2021	1990	SOCIAL HOUSING	14	Toronto-Danforth	1480 QUEEN S
2017	2021	1960	PRIVATE	14	Toronto-Danforth	15 BATER AV
2018	2021	1927	PRIVATE	14	Toronto-Danforth	12 BATER AV
2017	2021	1965	PRIVATE	14	Toronto-Danforth	100 GAMBLE
2017	2021	1962	PRIVATE	14	Toronto-Danforth	150 COSBURN
2017	2021	1956	PRIVATE	14	Toronto-Danforth	130 COSBURN
2017	2021	1950	PRIVATE	14	Toronto-Danforth	165 COSBURN
2017	2021	1965	PRIVATE	14	Toronto-Danforth	175 COSBURN
2017	2021	1968	PRIVATE	20	Scarborough Southwest	3207 KINGSTO

RentSafeTO in Toronto, and the scores are evaluated by "Bylaw Enforment Officers" (they inspect different components of a apartment building, and give an averaged score for the building) (toronto 2022). For data cleaning, the features and score of apartments were selected, and the format of some variables was corrected. The rows with missing values were removed. The cleaned data contains 9512 observations and 10 variables.

Table 1 contains the first 10 rows of the cleaned data. Year Registered is the year a apartment is registered with RentSafeTO. Year Evaluated is the time that the score was given to an apartment. Year Built is the year that the building was constructed. Property type represents the type of housing, it can be private, social housing, and TCHC(Toronto community housing). Ward is the numer id for the ward where the apartment is located. Ward Name is the specific name of the ward where the apartment is located. Storey is the number of storeys of a apartment and Unit represents the number of units of a apartment. Score represents the score evaluated for a apartment.

Figure 1:distribution of scores of apartments

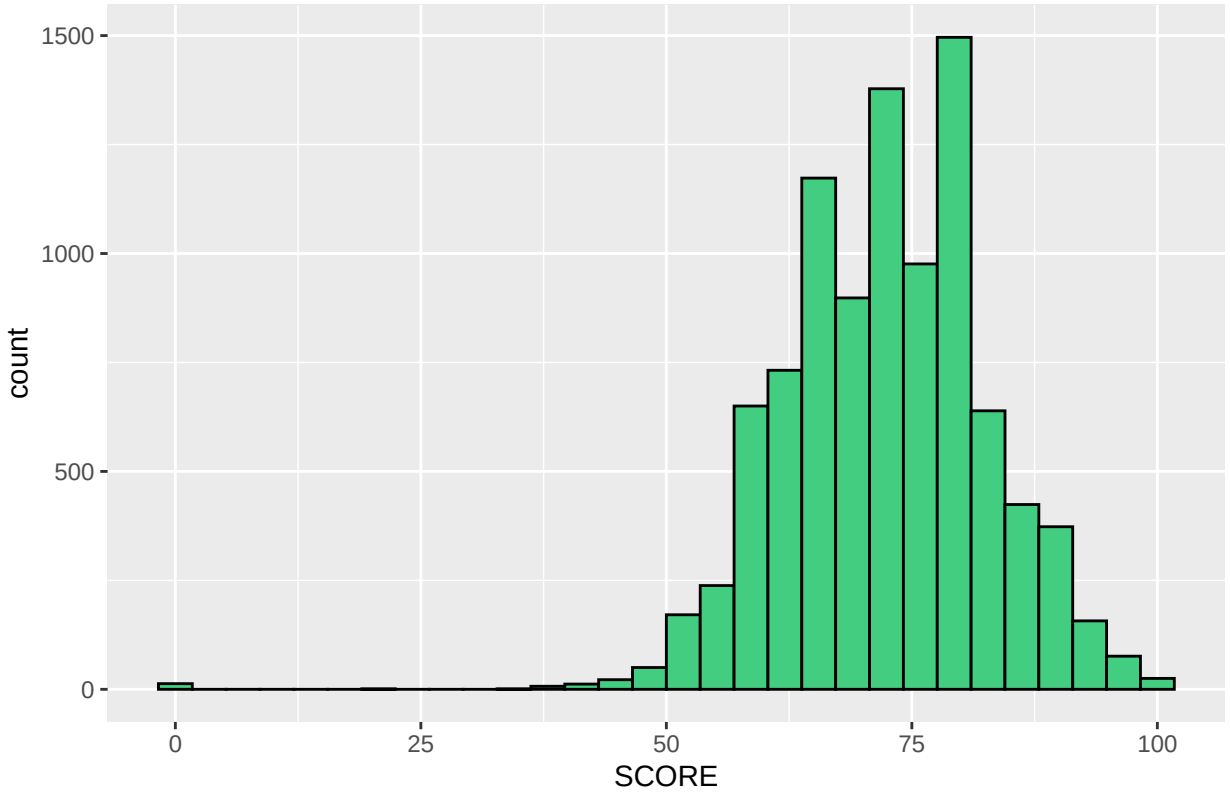

Figure 1 exhibits the distribution of the scores of apartments in Toronto. It ranges from 0 to 100. The distribution peaks at around 80 and it concentrates around 70. So, a lot of the apartments have a score of 60 to 85, and a few of the apartments received a score of zero and a full score of 100.

Figure 2: Distribution of the Ward where the apartments are located

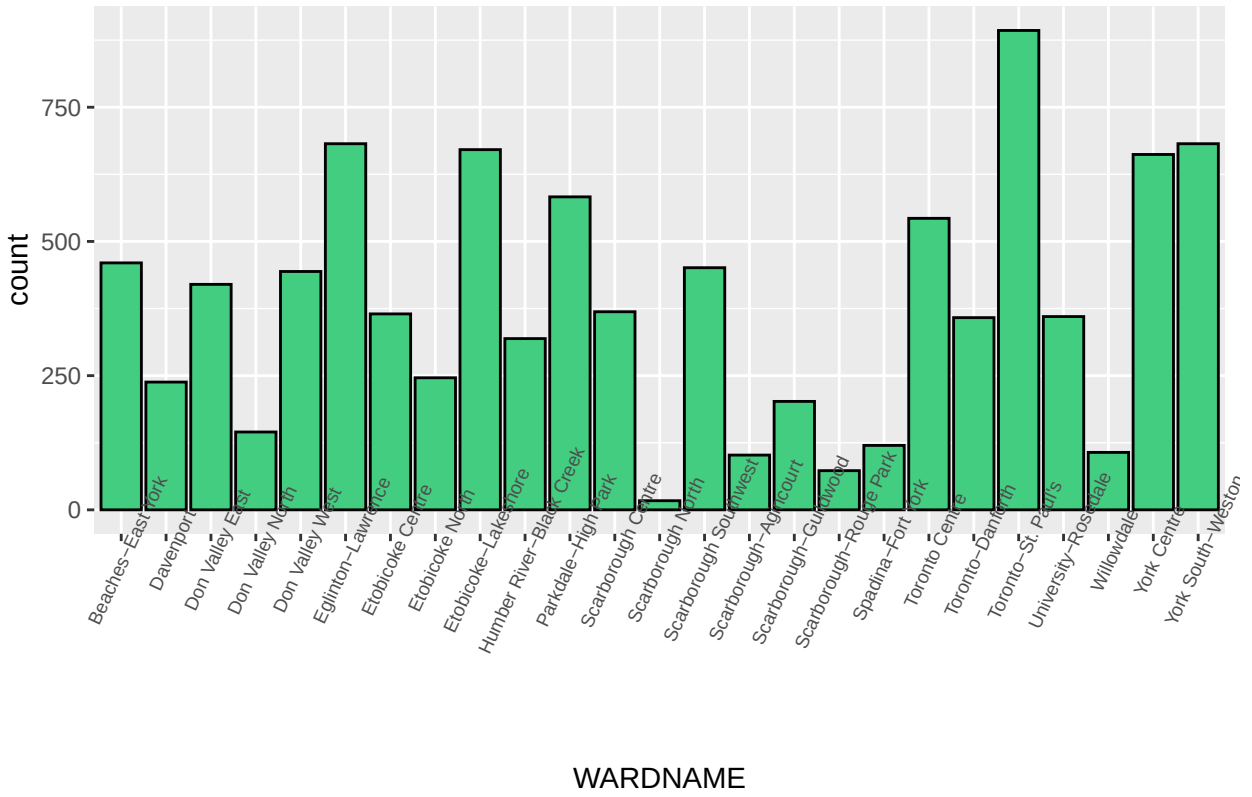

Figure 2 shows the distribution of wards where the apartments are located. Wards are like neighboorhoods. Toronto-St. Paul's has the highest number, 893 apartments are located in this ward, which is around 10% of the total. Scarborough North has the lowest number, only 17 apartments are located in this ward. There are 3 types of wards based on the number of apartments located in each ward. The first type are the wards with the number of apartments less than 250, the second type are the wards with number of apartments around 500, and the third type are the wards with the number higher than 650. The number of apartments in each ward varies a lot.

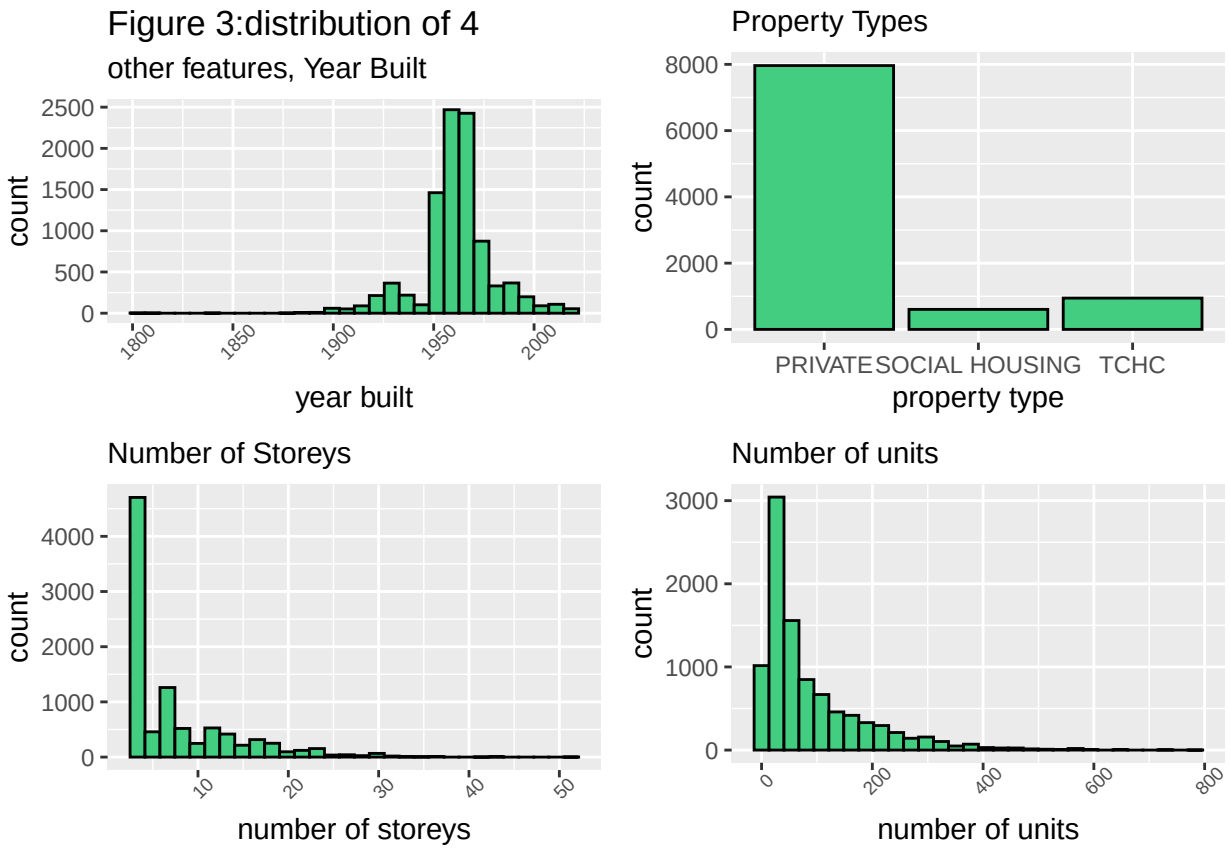

Figure 3 exhibits the distributions of 4 other features of apartments in Toronto. The plot on the top left shows the distribution of the year the apartments were built. It ranges from 1800 to 2000s. And the distribution cocnentrates around 1960s. A big portion of the apartments were built in the period from 1950 to 1970. The plot on the top right exhibits the distribution of property types. Private buildings are the majority of the apartments, while social housing and TCHC only have a fairly small portion of the apartments. The plot on the bottom left shows the distribution of the number of storeys of apartments. Most of the apartments have a number of storeys less than 10, and the highest apartment has a number of storeys more than 50. The plot on the bottom right shows the distribution of number of units. It concentrates around 50 units, and the apartment with the highest number of units has more than 800 units.

3 Model

A multiple linear regression model is used to explore the relationship between the score of an apartment and the features of an apartment. The dependent variable is score, which is a continuous variable, and a linear regression model is appropriate for predicting continuous variable. For the explanatory variables, there are categorical and numeric variables. After comparing a few potential models, I decided a final model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 I(X_4 = SocialHousing) + \beta_5 I(X_4 = TCHC) + \beta_6 I(X_5 = Davenport) + ...$$

$$+ \beta_{29} I(X_5 = YorkSouthWeston) + \epsilon$$

In the model, Y represents the score of an apartment. X_1 represents the year when the apartment was built. X_2 represents the number of storeys of the apartment. X_3 represents the number of units of the apartment. $I(X = \alpha)$ is an indication function, it is used to model the categorical variables. It returns 1 when X is α, and returns 0 otherwise. X_4 is a categorical variable, and it represents the type of an apartment.X_5 represents the ward where the apartment locates. β_0 is the intercept, it is the value of score when other features are zero. $\beta_1, ..., \beta_{29}$ are the coefficients of the features. Each coefficient represents how much the score would increase if the feature increases by 1 and other features hold constant.

The apartment data is split into training and testing data. Training data is used to train the model, and testing data is used to evaluate the performance of the model trained. RMSE(Root-mean-square deviation) was used to determine the performance. The RMSE of the model is around 9.487, it means the score predicted by the model differs from the actual scores by 9.487 on average. It shows that the model performs moderately. It reveals some relationship, but not extremely precise.

4 Results

Table 2 is the results from the regression. For instance, intercept is -141.64, it means the score of an apartment is predicted to be -141.64 when other features are zero. The coefficient for year built is 0.11, it means the score of an apartment is increased by 0.11, with year built increasing by one, and the other features holds constant. The number of units has a coefficient of 0.0016(rounded to 0.00), it means that the score of an apartment is 0.16 higher with 100 more unit, with other features hold constant. The coefficient for property type TCHC is -5, it means that the score of an apartment is decreased by 5 when its type of housing is TCHC(Toronto community housing). The coefficient for ward name Davenport is -2.87, which means the score of an apartment is decreased by 2.87 when the apartment is located in Davenport and other features hold constant.

In table 3, the average score of apartments in 25 wards of Toronto are exhibited. Scarborough North has the highest average score, around 80.88. And Davenport has the lowest score, around 69.22.

The average score of the 3 types of apartments are shown in table 4. Social housing has the highest average score. TCHC (Toronto community housing) has the lowest average score.

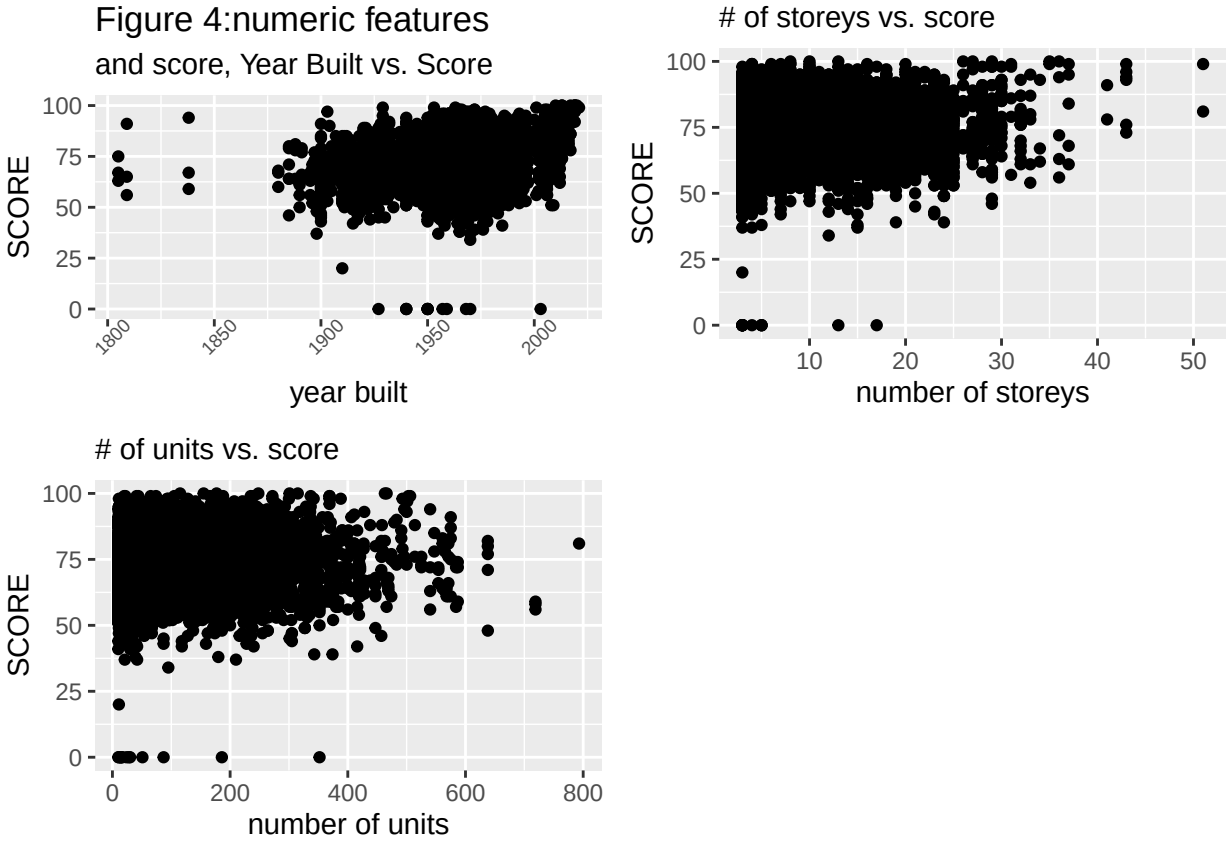

Figure 4 exhibits the relationship between the numeric variables and the score of apartments. For the plot on the top left, the scatterplot shows the relationship between year built and score, and there seems to be a positive trending, which means the score increases as the year increases. For the plot on the top right, it shows the relationship between number of storeys and the score. It seems to have a slightly positive trending.

Table 2: Relationship between score of an apartments and its features

	Model 1
(Intercept)	−141.64
	(14.72)
YEAR_BUILT	0.11
	(0.01)
CONFIRMED_STOREYS	0.16
	(0.04)
CONFIRMED_UNITS	0.00
	(0.00)
PROPERTY_TYPESOCIAL HOUSING	0.04
	(0.53)
PROPERTY_TYPETCHC	−5.00
	(0.42)
WARDNAMEDavenport	−2.87
	(0.89)
WARDNAMEDon Valley East	2.36
	(0.75)
WARDNAMEDon Valley North	3.79
	(1.07)
WARDNAMEDon Valley West	3.66
	(0.74)
WARDNAMEEglinton-Lawrence	−0.66
	(0.66)
WARDNAMEEtobicoke Centre	−1.76
	(0.78)
WARDNAMEEtobicoke North	−4.52
	(0.87)
WARDNAMEEtobicoke-Lakeshore	−1.01
	(0.67)
WARDNAMEHumber River-Black Creek	−5.69
	(0.81)
WARDNAMEParkdale-High Park	−3.04
	(0.69)
WARDNAMEScarborough Centre	0.81
	(0.78)
WARDNAMEScarborough North	5.94
	(2.79)
WARDNAMEScarborough Southwest	−1.90
	(0.74)
WARDNAMEScarborough-Agincourt	2.73
	(1.19)
WARDNAMEScarborough-Guildwood	−1.85
	(0.94)
WARDNAMEScarborough-Rouge Park	1.36
	(1.42)
WARDNAMESpadina-Fort York	1.19
	(1.14)
WARDNAMEToronto Centre	−1.00
	(0.71)
WARDNAMEToronto-Danforth	−0.55
	(0.78)
WARDNAMEToronto-St. Paul's	1.42
	(0.64)
WARDNAMEUniversity-Rosedale	−0.64
	(0.78)
WARDNAMEWillowdale	3.08
	(1.21)

Table 3: The averaged score of apartments in different wards

WARDNAME	average_score
Beaches-East York	72.56739
Davenport	69.21849
Don Valley East	76.37619
Don Valley North	79.19310
Don Valley West	76.70045
Eglinton-Lawrence	71.97507
Etobicoke Centre	72.13425
Etobicoke North	69.23984
Etobicoke-Lakeshore	71.69151
Humber River-Black Creek	68.80564
Parkdale-High Park	69.28130
Scarborough Centre	74.41192
Scarborough North	80.88235
Scarborough Southwest	71.19512
Scarborough-Agincourt	77.31373
Scarborough-Guildwood	71.56436
Scarborough-Rouge Park	75.05479
Spadina-Fort York	75.23333
Toronto Centre	71.71271
Toronto-Danforth	73.26257
Toronto-St. Paul's	73.83763
University-Rosedale	71.70556
Willowdale	77.47664
York Centre	71.57704
York South-Weston	70.43402

Table 4: The average score of different types of apartments

PROPERTY_TYPE	average_score
PRIVATE	72.56223
SOCIAL HOUSING	75.07285
TCHC	69.44762

For the plot on the bottom left, it shows the relationship between number of units and score. The points in the plot seem to scatter randomly, and there is no obvious trending.

5 Discussion

5.1 Summary

In the data section, it was found that the scores of the apartments were concentrated around 70, which shows that a large portion of apartments have a moderate or good quality. After that, the distribution of the wards where the apartments are located shows that the number of apartment varies in different wards. Another interesting fact found in data secion is that the apartments have a long history in Toronto, some of the apartments were built in 1800s. Most of the apartments were built in the last century and concentrates around 1960s. In addition, most of the apartments are private, most of the apartments have the number of storeys under 10, and most of the apartments have the number of units under 100. It shows that most of the apartments are not big in size and most of them are low-rise buildings.

The model and performance of the model suggests that number of storeys, number of units, year apartments were built, property type and ward apartment locates have influences on the score of apartments. However, the influence of variables are fairly different, and some of them have small amount of impacts. The coefficient for number of units in the model is 0.0016, which means the score would only increase by 0.16 if the number of units increases by 100. And the random pattern in the plot of number of units vs. score suggests that the number of units has almost no relation with the score of apartments. So it can be concluded that the number of units has nearly no impact on the score. The coefficient for the year built is 0.11, which means the apartments that are built more recently have higher scores, with other features hold constant. The coefficient for the number of storeys is 0.16, which shows that the score of apartments would only increase by 1.6 if the number of storeys increases by 10. Also, the trending between it and the score suggests that the relationship between them is weak. So, the relationship between the score of apartments and the number of storeys is not strong.

For the categorical variables, the coefficient for property type being TCHC(Toronto community housing) is -5, which means when other features of apartments are the same, the apartments that are TCHC are 5 scores less than the other apartments. For the wards of apartments, 14 of the wards have negative coefficients, which means the scores of the apartments located in these wards are negatively influenced by their locations. The magnitude of the coefficients for wards is much higher, which suggests that the wards are more influential than the number of storeys and the number of units. Lastly, the table of average scores in different ward shows that the apartments in Scarborough North have the highest average score, and the apartments in Davenport have the lowest average score.

5.2 Weaknesses

For weaknesses, the rows with missing values were removed from the original data. Although the removed observations are a very small portion of the data, the accuracy of the results can be decreased. On the other hand, the data used for the analysis contains only the apartments registered with RentSafeTO, so the data might not be a comprehensive representation of the apartments in Toronto. So, the results and model generated from the data might not reflect the full image of the relationship between the features and quality of apartments. In addition, the outliers in the data might impact the accuracy of the model. The performance of the model is moderate, which means the relationship reflected by the model might not be highly accurate.

5.3 Application and next steps

The findings in this report can be used to assist decision makings in real estate industry. And it can also be used to assist policy making in housing related departments in government. The influence of wards can be used to help construction planning. The impacts of different property types can be used to plan the social housing and community housing in the future. Moreover, the findings can help the people who are renting

apartment to make decisions on apartments with different features. However, the findings in this report should only be used as suggestions, decisions should be made with the consideration of other resources as well. For instance, the people who are renting should not exclude the apartments from the pool just because their locations are in wards with low average score. For the next steps, the model used in this report can be improved by transforming the variables. In addition, more features of apartments in Toronto can be used to explain the quality of apartments in further studies.

Appendix

Data Sheet

Extract of the questions from Gebru et al. (2021)

The datasheets is created using R programming language and a R Markdown file (R Core Team 2020).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
    - The dataset contains the features and score of apartments registered with RentSafeTO in Toronto. The dataset is from Open Data Toronto Portal and can be accessed with R package `opendatatoronto` (Gelfand 2020).

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?* -The data is created by RentSafeTO in Toronto.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
    - The provider of the funding is unknown. The data is available on Open Data Toronto Portal (toronto 2022).

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - The dataset contains the scores and features of apartments in Toronto.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There is one type of instances, the apartments that are contained in the dataset.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - The dataset is a sample of the apartments in Toronto. The largeer set is all the apartments in Toronto. The dataset might be representative of the larger set, because it contains the information of more than 9000 apartments.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
    - The data is in the form of a dataset, and can be accessed using R package `opendatatoronto` (Gelfand 2020).
5. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is available on Open Data Toronto Portal.
6. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No

7. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
   - No
8. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
   - No
9. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
   - Not possible

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*
   - The features of apartments are observable, and the score of apartments are evaluated by Bylaw Enforcement officers (toronto 2022).
2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - The data was collected by RentSafeTO.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - The apartments in the data are registered with RentSafeTO.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - Bylaw Enforcement officers (toronto 2022).
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - The data is updating and it was first created in 2017.
6. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - The data is from Open Data Toronto Portal.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - The rows with missing values are removed from the original data.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - Yes.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - R programming language was used to clean and preprocess the data (R Core Team 2020).

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - The dataset is used in the report about relationship between features and quality of apartments in Toronto.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - https://github.com/ZihaoLiu2/Toronto_apartments_statistics

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The dataset will be viewable by public on GitHub.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - GitHub

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Zihao Liu (the original data is still updating on Open Data Toronto Portal)
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - zihaohans.liu@mail.utoronto.ca
    - RentSafeTO@toronto.ca (Contact for the original data)

References

Arel-Bundock, Vincent. 2022. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready.* https://CRAN.R-project.org/package=modelsummary.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics.* https://CRAN.R-project.org/package=gridExtra.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal.* https://CRAN.R-project.org/package=opendatatoronto.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files.* https://CRAN.R-project.org/package=here.

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

toronto, open data. 2022. "Apartment Building Evaluation." 2022. https://open.toronto.ca/dataset/apartment-building-evaluation/.

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, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.