What factors can influence apartment buildings evaluation?

Wantong Qiu

27/04/2022

Abstract

When talking about people's happiness level, people will think about the quality of food, clothing, shelter, and transportation. Indeed, these four areas are the sum total of our lives. With the progress and development of society, people seem to have a higher pursuit of life quality. Today, we will discuss the theme of housing that can give people the most sense of life experience among the four directions. In the process of improving the quality of life, people seem to have more demand for housing. In the face of increasing demand, people need to know if they evaluate whether an apartment is suitable for living. Data from 'About Apartment Building Evaluation' on the Open Toronto website can help us learn from the evaluation of apartments, which is analyzed and studied by selecting some related factors from the two aspects of building internal facilities and external facilities.

Contents

Here is some body text.[1]

[1]This footnote will appear at the bottom of the page.:https://github.com/wantongqiu/Apartment-Evaluation-.git

1 Introduction

In fact, housing is an eternal topic, as long as there are people, then there will be housing demand. Of course, there are many different types of housing. When we discuss why we choose apartments, there will be many voices. For example, people who live in apartments can enjoy many different internal comprehensive public facilities, such as swimming pools, gyms, laundry facilities, and so on(Ovation 2015). Moreover, the house security system of the apartment is also a highlight, and apartment life helps social connections. Although there is a sense of community in rural or suburban areas, the distance of apartment life increases the possibility of establishing lifelong connections(Ovation 2015). So the next thing we're going to talk about is what are the factors that affect we choose an apartment?

With the development of society, there is more and more apartment. The government has also been increasing its intervention in these housing facilities, increasing the supply and affordability of housing(McAfee 2017). But when people are faced with choices, it is easy to get confused, because we cannot know which factors we need to consider when choosing an apartment, which are good or which are inadequate. On the other hand, through the data about "Apartment Building Evaluation" on the Open Toronto website, we can see the data About 40 variables of many different buildings. There are 20 of them for evaluation of the apartment's internal and external facilities on a scale of 0 to 5. In addition, we can also analyze and study the scoring criteria of different factors through the building year, number of floors, number of units, and apartment property type. The scores for these interior and exterior facilities give a deeper understanding of how the apartment is evaluated.

This paper is organized as follows: find the data about "About Apartment Evaluation" on the website of Open Toronto, and download the relevant data document. Before starting, load data into R Studio(R Core Team 2021) and download some packages available for use. In order to make the subsequent analysis and research more accurate, data cleaning was carried out and some incomplete observations were deleted. Then we introduce the cleaned data set and variables we need to use and use kableExtra(Zhu 2021) to organize the data we need to use into tables. Next, in the model section, I explained the regression model that I will use in this paper and the reasons for using it, as well as simple steps for establishing the model. Use ggplot2(Wickham 2016)in the result section to make figures that show the relationship between variables. Then I will give a more detailed explanation of the model mentioned above, including the selection of variables and the specific steps of establishing the model, as well as the best final model obtained through the validation of the model. Then, in the discussion section, some limitations of the data and model will be mentioned as well as the future directions for further study. Then, a summary of the whole research question and the introduction of what we really know in this paper is put in the first paragraph of the paper.

Table 1: Data about in the variables about Exterior facilities

Property_type	Year_built	Parking_Area
SOCIAL HOUSING	1996	4
SOCIAL HOUSING	1992	4
PRIVATE	1965	5
TCHC	1968	3
PRIVATE	1967	4

Table 2: Data about the first 5 rows in the variables about Interior facilities

Confirmed_storeys	Confirmed_units	Security	Elevator
4	29	5	4
6	102	5	4
28	214	5	5
17	352	4	2
21	275	5	5

2 Data section

2.1 Data cleaning

Before starting, 9790 observations in the dataset were collected with data cleaning. In order not to affect subsequent observation and analysis, we will use tidyverse(Wickham et al. 2019) to make the missing values in raw data were filtered. Then, data1 was formed, in which 493 observations were left after filtering. So in the following parts, we will analyze and study the factors for evaluating apartments by analyzing the 9,273 observations.

2.2 Variables

KableExtr(Zhu 2021) was used to create Table 1 and Table2 to show 7 variables that will affect the evaluation of apartments. Specifically, Table 1 mainly presents three variables about exterior facilities, including the year when the apartment is built, the score of the parking lot of the apartment, and the property types of the apartment. As can be seen in table1, property types are mainly divided into 3 categories, including owed privately, by Toronto Community Housing Corporation (TCHC) or another assisted, Social, or supportive housing Provider. Table2 mainly shows the data of four variables in exterior facilities that can affect the evaluation of apartments, including the number of stories, number of units, and the scores levels of security levels and elevators from 0 to 5 score. Then five lines of data were randomly selected for display, mainly to show the data form of each variable and what data types were recorded. In the following parts, the selected variables will be further explored and studied in terms of how they affect the evaluation of apartments.

3 Model

In this paper, I will explore my research question through linear Regression model. Linear Regression can help you in evaluating Apartment. Linear regression means that the graph of the relationship between two variables is or near a straight line. The linear regression model is used to help find whether the relationship between variables of Score and other variables is linear. Because the research question of this paper is to find different factors that affect scoring. Therefore, if the relationship between the variable of Score and the linear of other variables can be found in this model, it will be good.

Before starting building model, we need to divide our data into a train set and a test set. In the following steps, we will ue the train data set in model building. Firstly, we will select variables that may affect the response of Score to form the first model through our comment sense, literature reviews, or graphs. And then we're going to sort out the data for our model. The second step, then, is to compose Model2 by looking for the most significant level variables with the response of Score based on Model1. The third step is that we can use the stepwise selection method to screen out more preferred model3.Stepwise selection is both forward and backward methods that can be allowed in the process of variable selection on the same model. The process is to iterate between forwarding and backward selection until we are unable to add or delete further variables which owns the smallest AIC(Akaike's Information Criterion). The iteration allows the procedure to check whether, due to the addition or deletion of a predictor, the conditional relationship between Y and the predictors has changed so that a variable that might previously have been added or deleted now might explain a significant portion variation.

After finishing initial model selection, we needed to validate our model.Firstly, we need to check the situation of linearity and distribution of the model. Then check whether the residual plots and QQ plots meet the assumptions of the relationships. Uncorrelated Errors and constant variance. Then, the partial F test is used to compare whether the model we selected is the best model. To be specific, we randomly select several variables from the selected model to build a new model. Then we use the partial F test to check the value of the P-value if it is less than 0.05, we can select the originally selected model. So far we have used train data set to carry out. Next, we will do further validation by using the test data set. We use the test set to repeat some of the initial validation steps we just did for the train set, including whether the check meets the two conditions and the assumptions. Finally, check for actual points, including leverage points, outliers, or influential points. Then the above steps represent that our model is complete.

4 result

4.1 Exterior facilities

The first variable of the year built is presented by means of the histogram. It can be seen from figure1 that the number of apartments built from 1950 to 1970 showed a trend of rapid growth, and the number of apartments built in this period was also the largest. Since 1970, the number of apartments built has drastically decreased. What's interesting is that this diagram shows Normal Distribution. A detailed analysis of the Linear Regression Model will be given in the latter part of the model.

Figure 1: The built year of the apartments

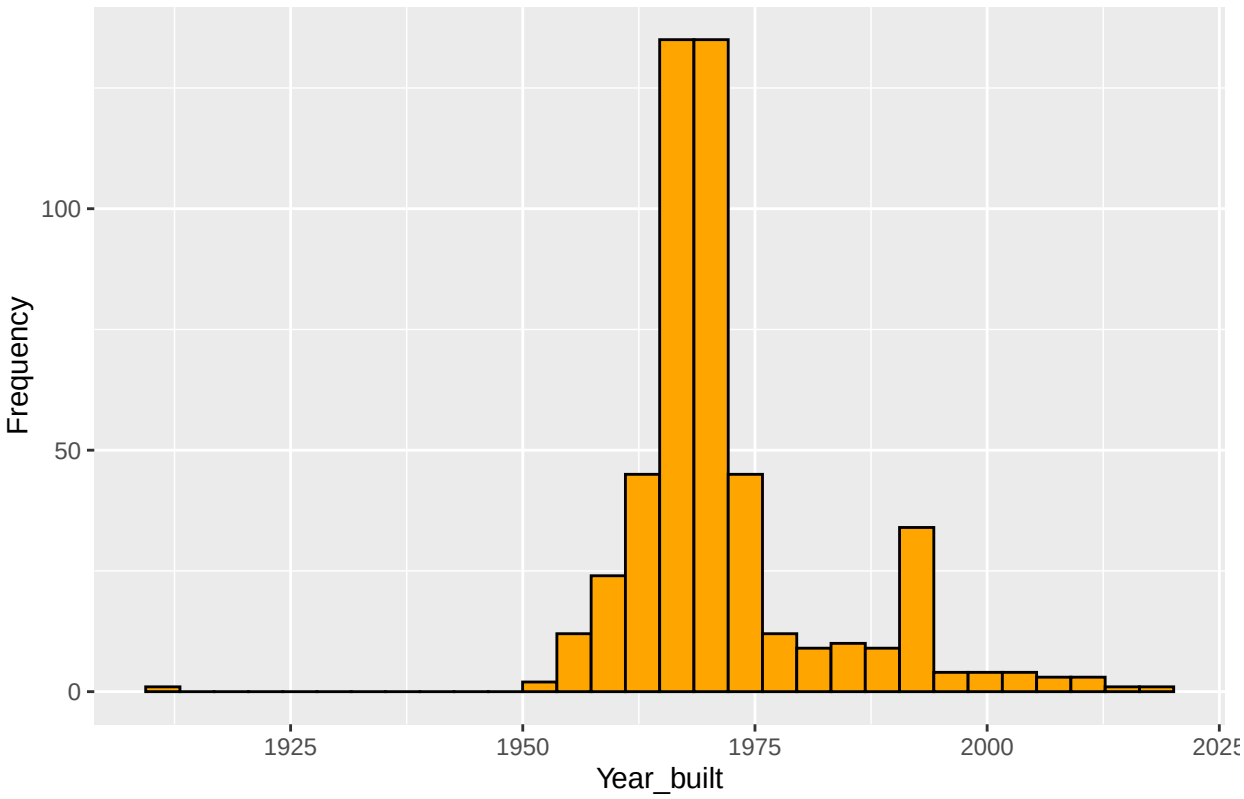
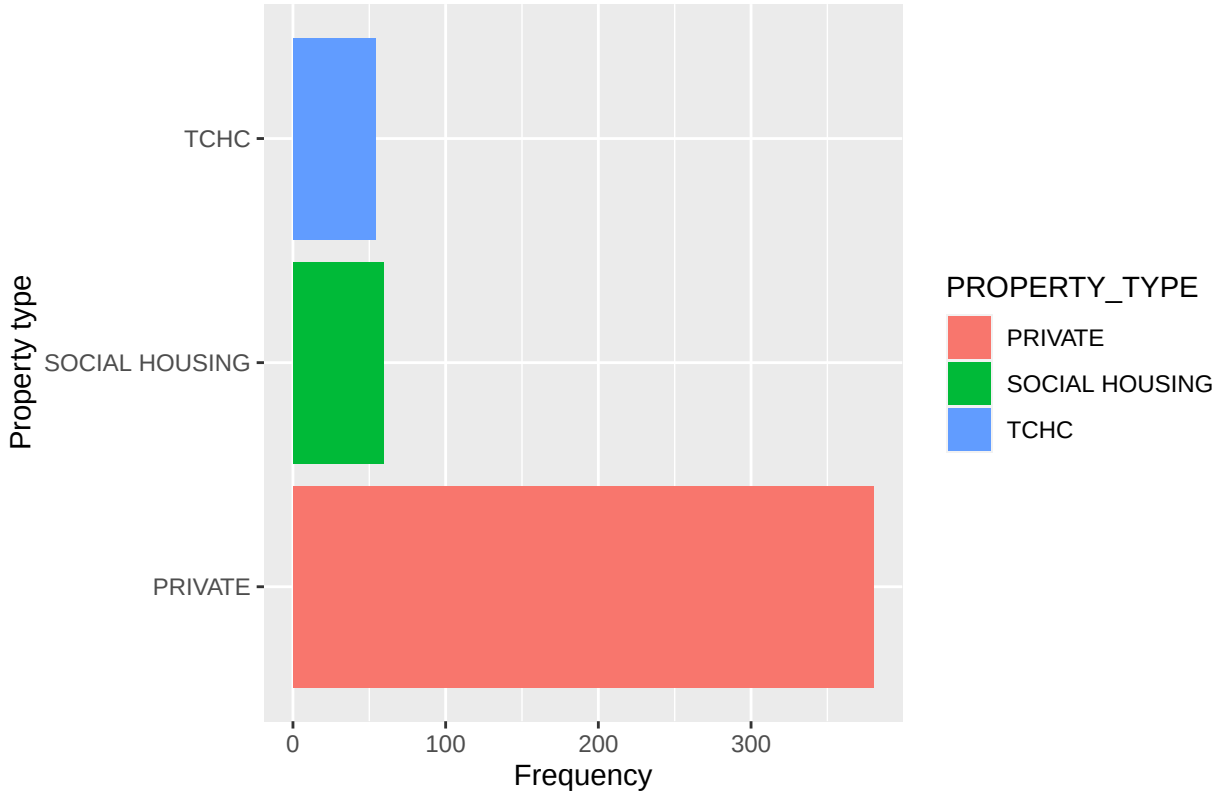


Figure 2 shows the parking area score. On a scale of 0 to 5, in this graph, we can see that the number of people rated as 4 is the largest. In other words, more than half of apartment parking lots are in the middle level. Even fewer apartments scored a perfect 5. Of course, the lowest number of apartments got a score of 1, which means that most of the apartments didn't give people a bad experience.

Figure3 presents data about variables of property type through a bar chart. Here we can see that there are only three property types in the whole data1, including TCHC (Toronto Community Housing Corporation), Social housing (Social or supportive housing Provider), or privately owned. Then in this figure, we can see that the least apartments are in social housing, which I have indicated in green. Observations, with a number of nearly 400, are private apartments marked in red. Finally, apartments for TCHC are highlighted in blue. This means that nearly 400 evaluations for different factors in the entire dataset are scored around the private apartments.

Figure 3: The property type of the apartments



4.2 Interior facilities

Here are some of the effects of interior facilities on apartment ratings. In Figure 4, the score of the number of Storeys and units is presented through Scatter Plot to reflect some similarities and differences between them. In this figure, we can see that the two variables have in common that most of the points are concentrated in a small number of storeys and units, so the points will become fewer and more scattered in the further part. Then in legend, you can see the change of score: darker color means lower score, lighter color means higher score. In the first half of the graph, the dots are lighter and therefore have higher scores, and in the second half of the graph, the number of dots becomes less. For example, those with fewer floors but more units are also welcomed with high scores. In addition, some units with multiple floors have high scores and are favored by many people. Therefore, we can see from this figure that most apartments do not include individual special cases but actually tend to have a smaller number of storeys and units that will be more popular.

Figure4: The relationship between the scores of units and storeys

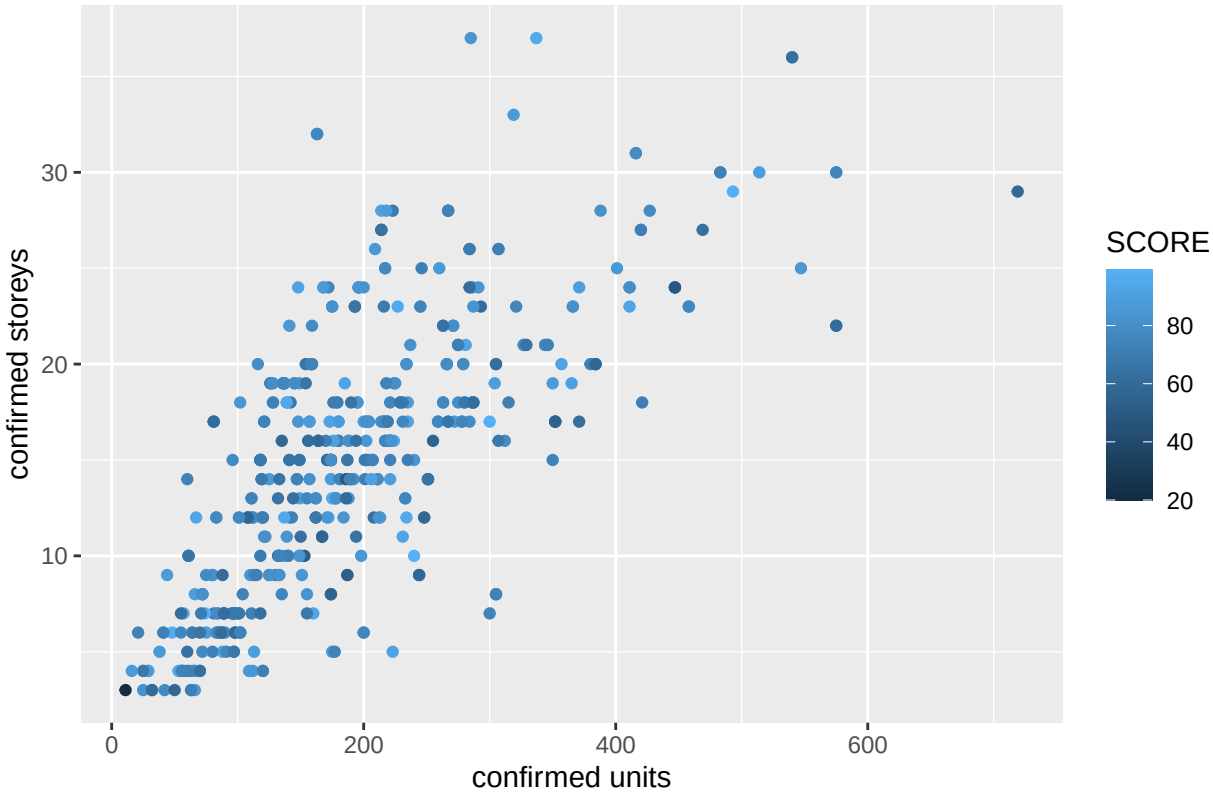
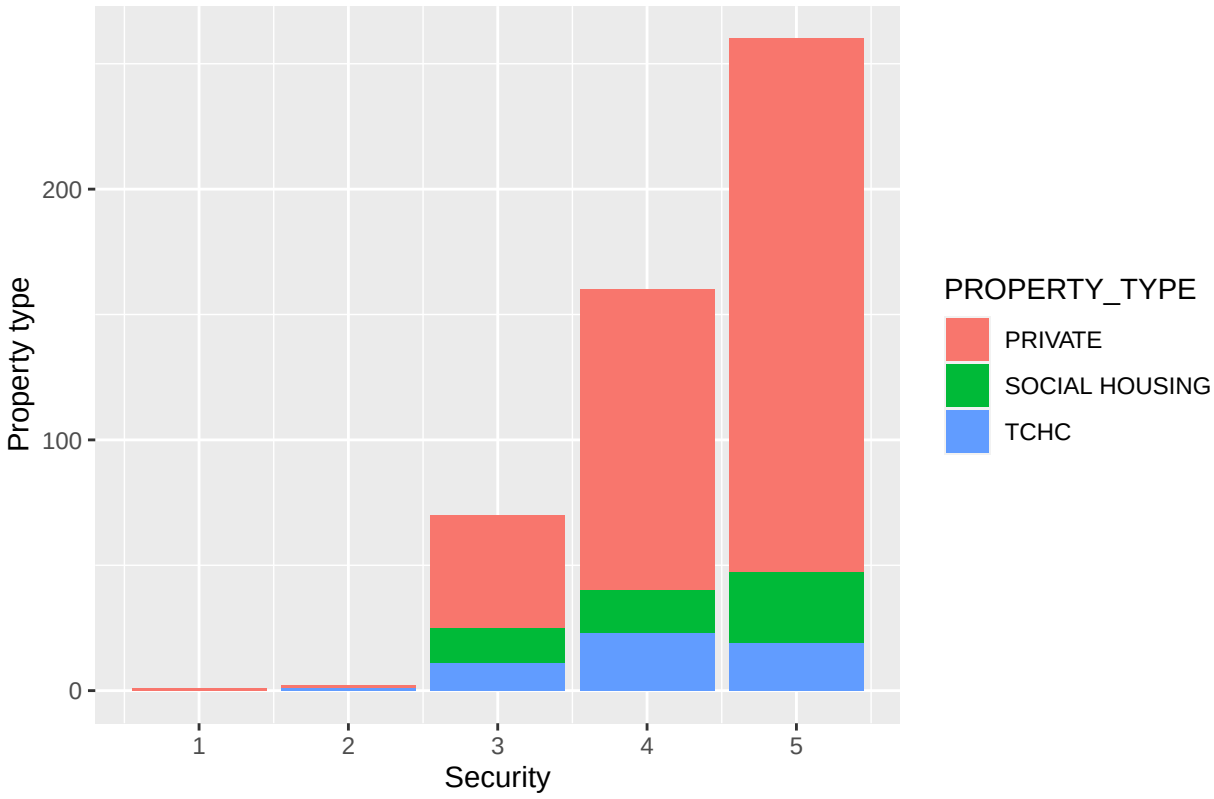


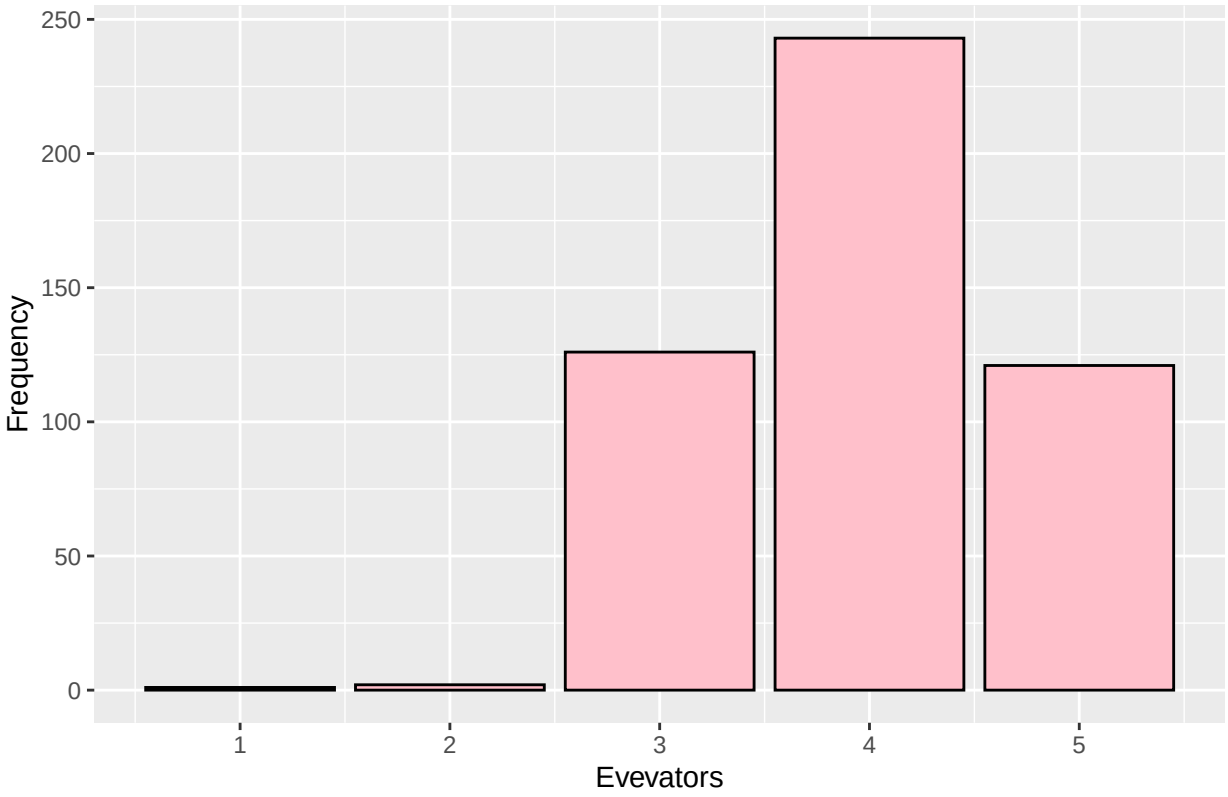
Figure 5 is the relationship between the score of security and the type of apartment building. Firstly, there are three types of apartments mentioned above, including private, social housing, and TCHC. The X-axis is the variable for security level, on a scale of 0 to 5. It can be seen that the number the score of 4 is the largest, and the proportion of private is the largest at the security level of 4. In other security levels private apartments also occupy the largest proportion, while social housing apartments for the least. There is no social housing in security level scores of 1 and 2. It means that the overall security level of social housing is relatively high. What's more, the security level of most apartments is very good, because houses with security levels above 3 apartments for more than 80% of the total. Therefore, it can be seen that the development of society has greatly improved the security level.

Figure 5: the relationship between sevurity level and property type



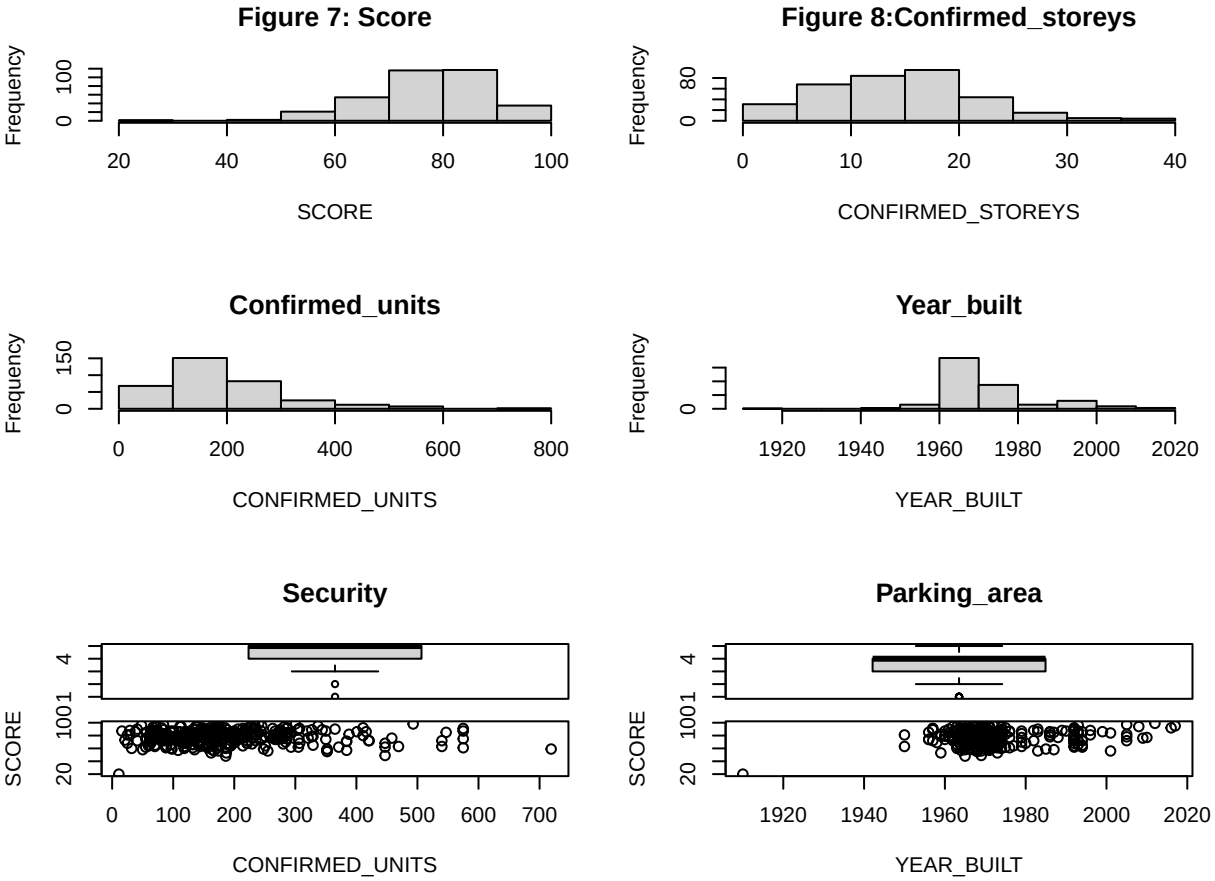
In Figure 6, we can see that elevator scores ranged from 0 to 5. As we can see the biggest proportion was 4 points. And then we have the lowest percentage of scores 1 and 2. The picture conveys the message that in today's society, people's awareness of safety has been enhanced, so the developers of apartment buildings have a good attitude towards the use of elevators in both old and new buildings, and for the convenience of people, the management of elevators becomes better, because most of the ratings are good.

Figure 6 : The evaluation of elevators



4.3 Model Result

Before we building the model we need to divide the valid data1 into train sets (346 observations) and test sets (147 observations). Then, the data of the train set is summarized first, and we can see that the numerical data will show the data about the minimum value, median, mean maximum value, and so on.



Histograms of variables of Score, Confirmed Storeys Confirmed Units, and year built are used to show data about these four numerical variables. The box plots of security and parking area are the data presentation of category variables. Finally, the scatter plots about the numerical variables of confirmed units and years were built to show the relationship between the two predictors and response. It could be seen that the overall trend of the two plots was going up.

4.3.1 Model building

When building a model1, we use common sense, literature, and figures to build. So model 1 includes 14 variables. The variable of SCORE is the response, The 14 variables are the predictors which include the variables of CONFIRMED UNITS, CONFIRMED STOREYS, YEAR BUILT, and NO of AREAS EVALUATED, SECURITY, concealed PARKING AREA, TAIRWELLS, ENTRANCE DOORS WINDOWS, ENTRANCE LOBBY, GARBAGE CHUTE ROOMS, STORAGE AREAS LOCKERS PROPERTY TYPE, WARD.

Then, on the basis of Model1, we further filter model2 in table 3 by looking at the significant levels of each Predictor. There are 12 variables that meet the significant levels. Including the variables of STOREYS, YEAR BUILT, SECURITY, PARKING AREA, TAIRWELLS, ENTRANCE DOORS, WINDOWS ENTRANCE LOBBY, GARBAGE CHUTE ROOMS, STORAGE AREAS LOCKERS, PROPERTY TYPE.

For Model3 in table 4, the stepwise selection method is used to calculate the minimum AIC value on the basis of Model2, so as to select the most preferred model. The value of AIC calculated by Model3 is

Table 3: Model2

term	estimate	std.error	statistic	p.value
(Intercept)	-85.0228768	35.1736511	-2.417232	0.0161774
CONFIRMED_STOREYS	0.0402956	0.0273569	1.472961	0.1417090
YEAR_BUILT	0.0475921	0.0179907	2.645368	0.0085484
SECURITY	2.4364948	0.2967869	8.209578	0.0000000
ELEVATORS	2.1086002	0.3189756	6.610538	0.0000000
PARKING_AREA	1.4775993	0.2328424	6.345921	0.0000000
STAIRWELLS	2.9071102	0.2713602	10.713104	0.0000000
ENTRANCE_DOORS_WINDOWS	2.6609175	0.2990326	8.898421	0.0000000
ENTRANCE_LOBBY	1.6450917	0.3553291	4.629769	0.0000053
GARBAGE_CHUTE_ROOMS	2.0208201	0.2959934	6.827247	0.0000000
STORAGE_AREAS_LOCKERS	2.3657824	0.3129845	7.558784	0.0000000
WARD	0.0521936	0.0285516	1.828044	0.0684403
PROPERTY_TYPESOCIAL HOUSING	-1.1147807	0.6898499	-1.615976	0.1070494
PROPERTY_TYPETCHC	-0.9217524	0.5904320	-1.561149	0.1194413

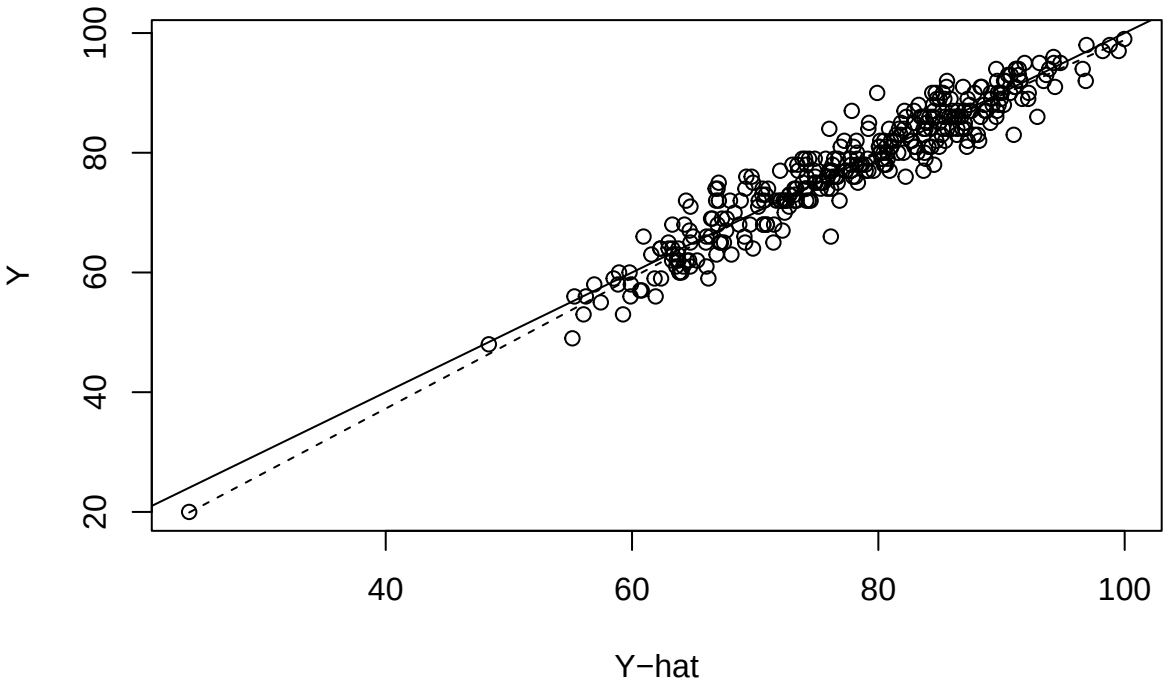
Table 4: Model3

term	estimate	std.error	statistic	p.value
(Intercept)	-85.0228768	35.1736511	-2.417232	0.0161774
CONFIRMED_STOREYS	0.0402956	0.0273569	1.472961	0.1417090
YEAR_BUILT	0.0475921	0.0179907	2.645368	0.0085484
SECURITY	2.4364948	0.2967869	8.209578	0.0000000
ELEVATORS	2.1086002	0.3189756	6.610538	0.0000000
PARKING_AREA	1.4775993	0.2328424	6.345921	0.0000000
STAIRWELLS	2.9071102	0.2713602	10.713104	0.0000000
ENTRANCE_DOORS_WINDOWS	2.6609175	0.2990326	8.898421	0.0000000
ENTRANCE_LOBBY	1.6450917	0.3553291	4.629769	0.0000053
GARBAGE_CHUTE_ROOMS	2.0208201	0.2959934	6.827247	0.0000000
STORAGE_AREAS_LOCKERS	2.3657824	0.3129845	7.558784	0.0000000
WARD	0.0521936	0.0285516	1.828044	0.0684403
PROPERTY_TYPESOCIAL HOUSING	-1.1147807	0.6898499	-1.615976	0.1070494
PROPERTY_TYPETCHC	-0.9217524	0.5904320	-1.561149	0.1194413

857.07. And no matter how we increase or decrease the value of variables AIC continues to increase, which means that we can stop and get the most appropriate model. This includes the variables of CONFIRMED STOREYS, YEAR BUILT, SECURITY, ELEVATORS, PARKING AREA, STAIRWELLS, ENTRANCE DOORS WINDOWS, ENTRANCE LOBBY, GARBAGE CHUTE ROOMS, STORAGE AREAS, LOCKERS, PROPERTY_TYPE.

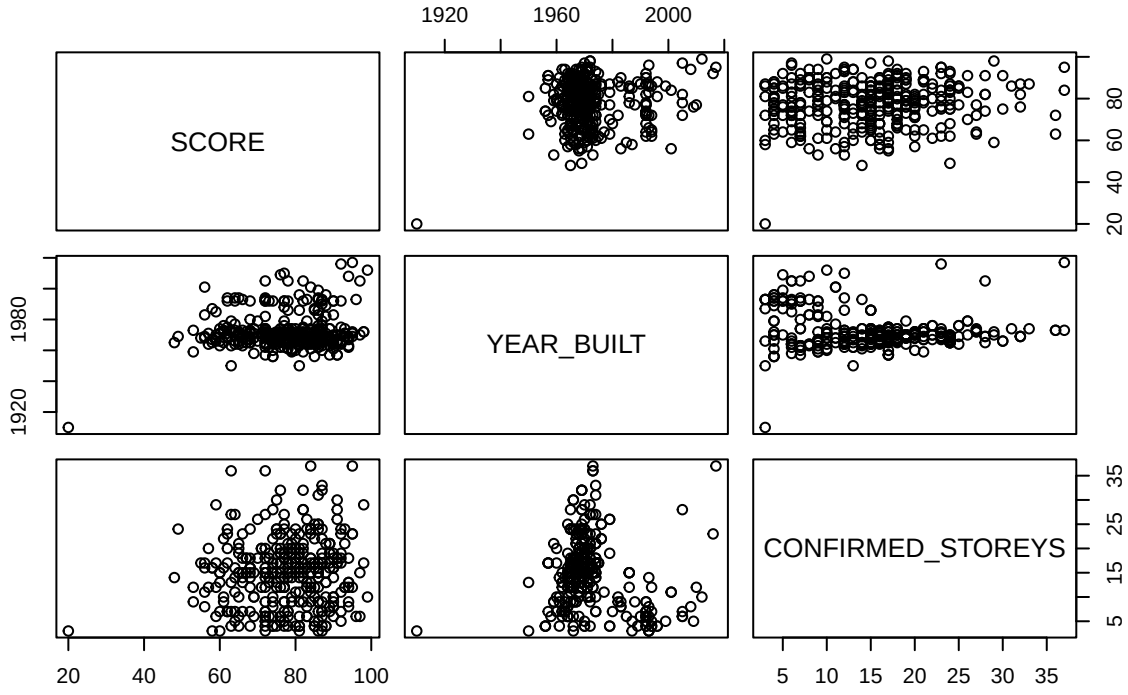
4.3.2 Model checking

Figure 7: Y versus Y−hat

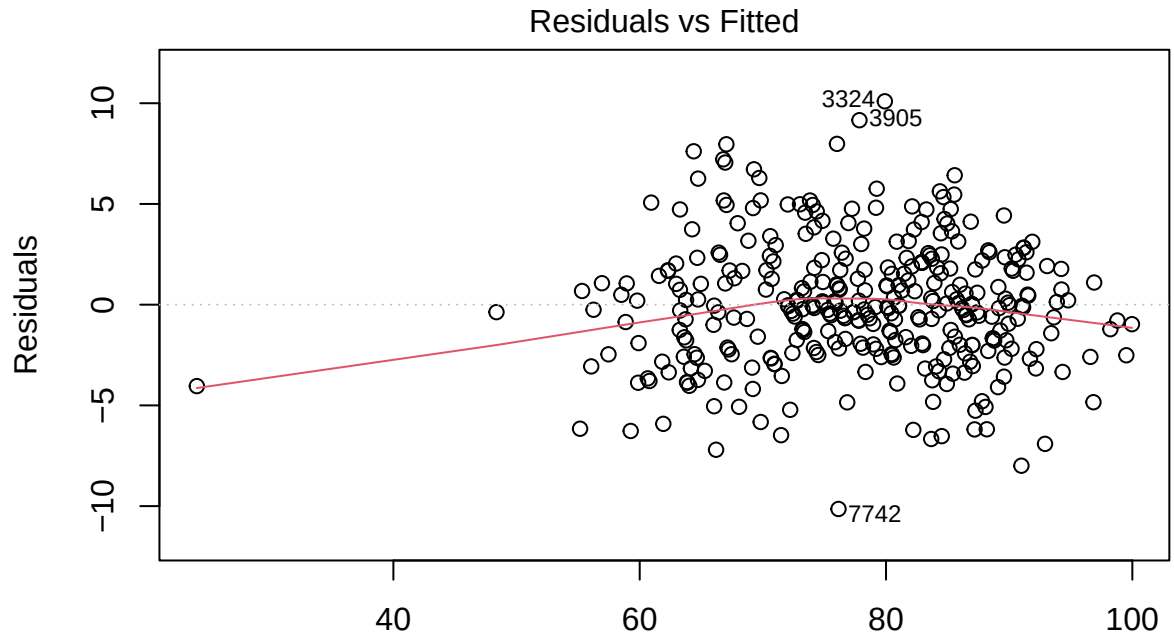


We can start model checking. Firstly, we need to check if Model 3 meets conditions 1 and 2 to prove the model's linearity by Figure 7. So on the left, we can see that the graph is about response versus fitted values, all the points are basically on or close to a straight line, which means the model satify the condition 1. And then start cheking the condition2.

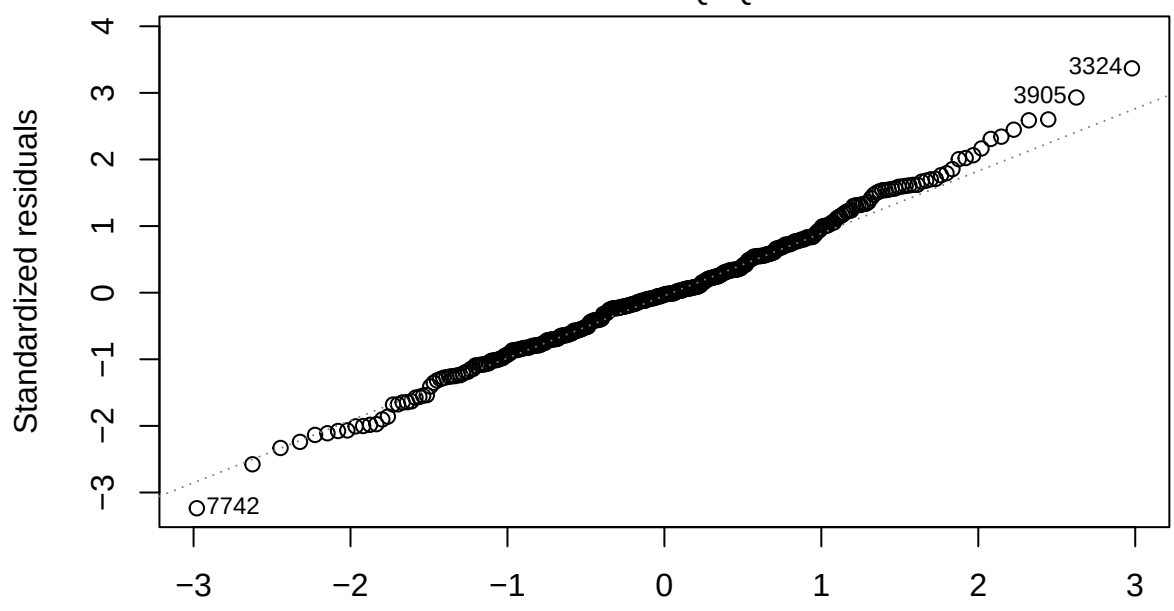
Figure8:Checking condition 2 by scatter plots of numerical variables



Selecting 3 numerical variables of score, year built and confirmed storeys from model 3 to make scatter plots in Figure 8. And then we can see that the points on the figure are almost linear. There are no problematic patterns which means the figure holds condition2.



SCORE ~ CONFIRMED_STOREYS + YEAR_BUILT + SECURITY + ELEVATORS + PAR



SCORE ~ CONFIRMED_STOREYS + YEAR_BUILT + SECURITY + ELEVATORS + PAR

Based on the residual plot, there is no systematic pattern, cluster pattern, or fanning pattern. And all the points are near or on the straight line and there are very few points of deviation in the QQ plot. This means model 3 can satisfy the 4 assumptions

Check if there are any problematic observations in model 3.There are 40 outlier points that exist whose standard residuals are not between -2 and 2. And there are 10 leverage points in model 3, which fitted values are bigger than the value in training set. However, there is no influential point in model 3.

Using partial F test to select the preferred model. Creating a new model which includes some but not all variables in model 3 randomly. Then use the ANOVA formula to compare the 2 models. P-value is lower than 0.05, then choose model 3.

Create a Model 4 that has the same variables as Model 3, and then repeat the Model checking the steps from model 3. Only this time we are using data from the test set. This is done to check how well the model fits the test data set. First, check whether model 4 satisfies conditions 1 and 2. It can be seen from the figure that the point is near or on the straight line, and there is no pattern in the figure, indicating that the final model satisfies linearity. The four hypotheses and normal distribution satisfy conditions 1 and 2. Then after examining the problematic observations in model 4, the situation is the same as in the model 3, meaning that model 4 has the same limitations as the model3. Then model 3 can be the final model. Specific graphs are presented in the Appendix which are shown in figure 11, figure 12, residual plot and QQ plot.

5 Discussion

In the result section above, the scores of external facilities and internal facilities are analyzed and how they affected the evaluation of apartments. The model is also established for the related variables. The linear regression model is used to show what and how factors influence the evaluation of apartments. As for external facilities, we selected the construction year, parking area, and property type of apartments for research in this paper. Then the internal facilities are discussed and analyzed about the floors number and the number of units of the apartments, and then the situation of security and elevators evaluation. Through these variables, the relationship between scores and these variables is studied and analyzed through the Linear Regression Model.

5.1 Exterior Facilities

In figure1 in the result section, we can see the data about the building year of the apartments. In fact, the year of apartment construction is still very important in many people's minds, because one of the important factors for people to choose to buy or rent a house is the year built. In figure1, we can see that the years of construction accounted for a large part between 1965 and 1970, and the renovation of apartments was another important factor in this process. Because the degree of renovation will further affect the value of the Apartments (Chew (2018)). however, it was not mentioned in the data set, and we hope to learn this aspect in our further study. Then it is about the evaluation of the parking lot of the apartment. Since the current definition of the parking lot can maximize the utilization of space, and in order to facilitate people or improve the overall rate of return, it is not hard to find that developers are paying more and more attention to the planning and construction of parking area (Son and Prestwich (2022)). Therefore, we can also see that the evaluation of the parking area is relatively high in figure2. As for the third factor about property type, it can be seen from figure3 in the result part that apartments of private type occupy the largest proportion in this society. However, the proportion of social housing and TCHC type is relatively low. Housing is fundamental to human beings, then society and the government should step up their intervention in housing and try their best to provide everyone with a place to live. Therefore, the improvement of social housing is indispensable.

5.2 Interior Facilities

As for the number of floors and units in the apartment, we can see the points in Figure 4. People give low scores to those with few floors and few units, those with few units but many floors, or those with many units but few floors. As opposed to high scores, which are concentrated within 300 units and 30 stories, the scores are good and concentrated. Nowadays, people choosing an apartment will concern about the number of floors and units. Then comes the second and most important element of the apartment's interior, security. Security has become a very important factor for people to choose an apartment because, with the development of society, people's awareness of security is constantly improving. People also begin to regard security as one of the most important factors that will affect the value of an apartment (Goslett (2012)). In Figure 5, we can also see that the security factor of private property type apartments is very high, and developers have also continuously improved the security factor of apartments from various aspects. Finally, this article describes the installations in the Apartment as an important factor in evaluating the interior facilities of an apartment. Elevators are also an indispensable part of apartments, whether old or new because adding elevators to apartments can improve the social problems caused by the aging population and increase the use of old apartments (Chen et al. (2020)). As can be seen from the data in Figure 6, the ratings of concealed provisions were relatively high, indicating that many apartments were suitable for the use and safety of concealed provisions.

5.3 Apply Regression model

Regression models were used to help us find and explore linear relationships between the response of score and the predictors that affect evaluation. From the partial F test, we can see the coefficients of the Model 3 and Model 5. The detailed steps for building the best model are given in the Result section. The limitation of this model is about after testing the validity of Model4 with the test set, we found that the difference in variable coefficients between Model3 and Model4 was caused by some contrast with actual observations and leverage points and outliers, which led to the failure of the final model to fit the data well. However, there are no influential points in the final model. For the integrity and accuracy of the whole analysis process, we can not delete these actual points.

5.4 Limitation and Future

There are also some limitations in the whole process. For example, there are a large number of incomplete observations, which may lead to certain limitations in the analysis of the whole data. However, in order to make a more accurate analysis, So it is necessary to filter out these missing values. At the same time, when analyzing the topic of apartment evaluation, there are certain limitations in the data category. For example, when I analyzed the year of apartment establishment, we could redefine the score by the renovation potential of the apartment at the same time. I hope I can learn more about the factors that can really affect the apartment evaluation and the comprehensive relationship between them in future studies.

Apendix

Figure 11: Y versus Y−hat for testing dataset

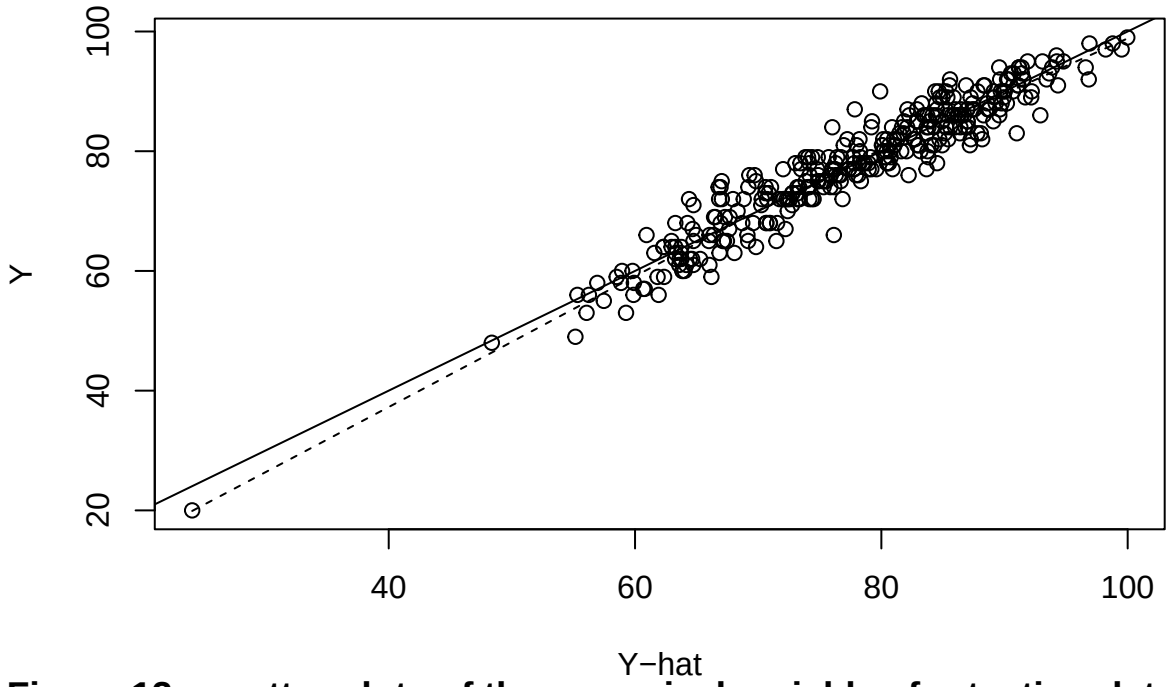
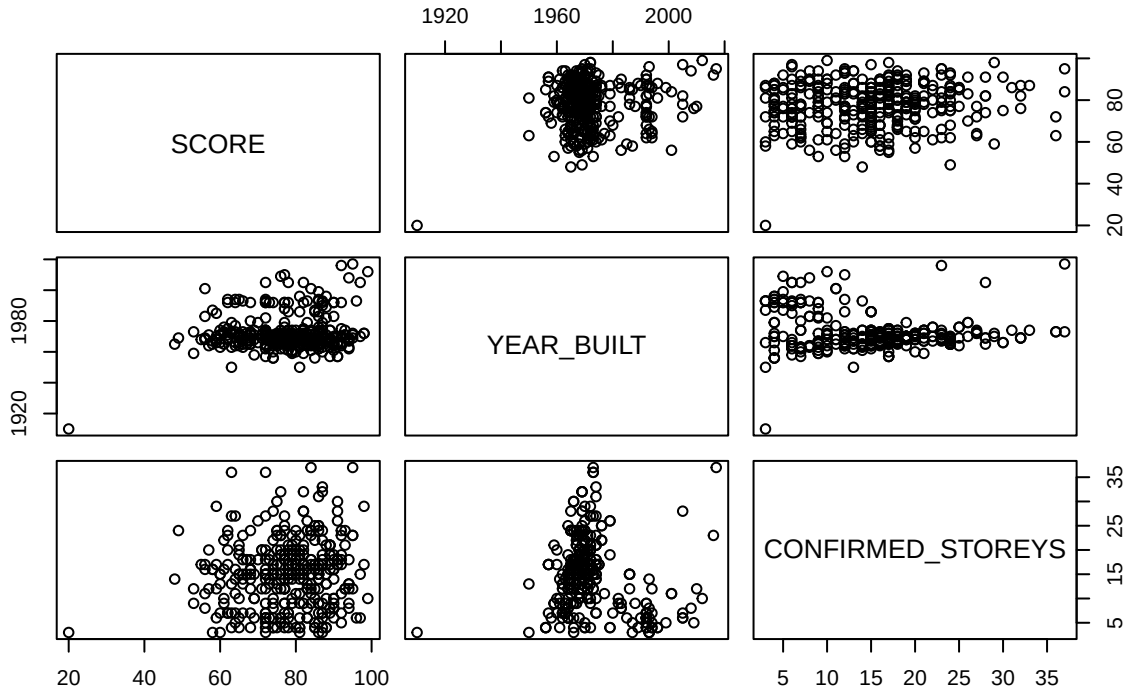
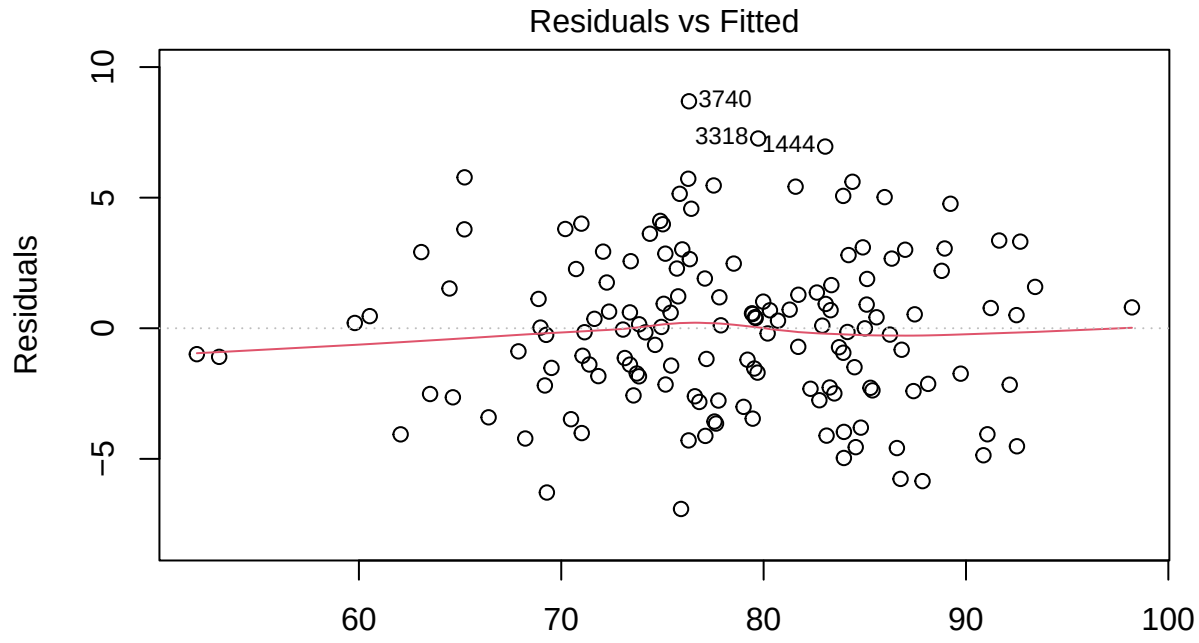
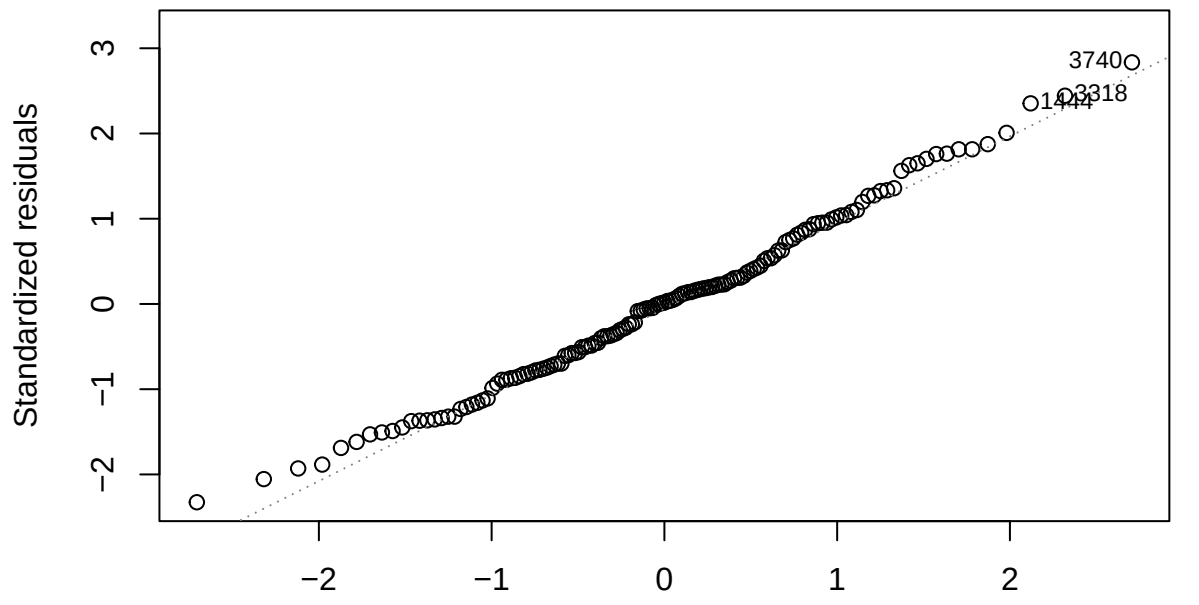


Figure 12 : scatter plots of the numerical variables for testing dataset





SCORE ~ CONFIRMED_STOREYS + YEAR_BUILT + SECURITY + ELEVATORS + PAR



SCORE ~ CONFIRMED_STOREYS + YEAR_BUILT + SECURITY + ELEVATORS + PAR

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 aims to ensure that owners and operators of apartment buildings with three or more storeys or 10 or more units comply with building maintenance standards in Toronto. Data documents are provided on the Open Toronto website, through which the evaluation of apartments can be analyzed and studied.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The data set was created by a Bylaw Enforcement Officer from Municipal Licensing & the Standards.
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 creation was funded by the Canadian government.
4. *Any other comments?*
    - None.

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 instance represent the evaluation of apartment building.The types are ratings of apartments facilities, year built, information about the apartments, scores,property types of the apartments building.
2. *How many instances are there in total (of each type, if appropriate)?*
    - There are 40 instances in the data set.
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).*
    - This data contain most of possible instances. Because in this data set there are many uncompleted observations and there are some which will really influence the evaluation of the apartments not shown in the data set, for example the old apartments need to be evaluated by the renovation potential levels.
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.*
    - In the raw data, the instance contains 40 variables.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - None.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

- Yes ,there is a big amount of missing value from the data set. Because the respondents didn't give feedback or the apartment buildings didn't receive the question lists.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
   - There are no relationships between individual instances.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - There are no recommended data splits.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - There is a big amount of uncompleted obeservations which will influence the analysis of the report, but there are no errors in the data set.
10. *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 self-contianed.
11. *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.*
    - There is no confidential data, and the dataset is public data.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - The Dataset's data does not contain any disturbing information or data.
13. *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.*
    - The data set classifies different factors according to the internal and external facilities of different apartment buildings. For example, the internal facilities can include elevators, number of floors, number of units, etc., while the external facilities are the building year of the apartment, parking area, or property type.
14. *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.*
    - We can not identify individuals in any way.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - The dataset does not contain data that could be considered sensitive in any way.
16. *Any other comments?*
    - TBD

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.*

    - Each year, the city compiles a list of all buildings that need to be assessed that year. An enforcement officer is appointed for each building assessment and building owners/operators are notified when their buildings need to be assessed. The assessment shall be conducted in the presence of the owner or designated person to provide law enforcement officers with access to locked common areas and/or facilities. Law enforcement officers will take notes or photos and upload them to the city's mobile investigative app.

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?*

    - By arranging a Bylaw Enforcement Officer for inspection, and by uploading notes and photos to the City's mobile Investigation application

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 data set was evaluated for apartment buildings throughout Toronto so it wasn't sampled.

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)?*

    - This is a Bylaw enforcement program that ensures owners and operators of Apartment Buildings. All apartments in Toronto need to be evaluated and checked with the help of building owners/operators.

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.*

    - Data is updated every year. The dataset used in this paper was updated on April 22, 2022.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - The owner or designated person is present to provide law enforcement officers with access to locked common areas and/or facilities. During a building evaluation, one bylaw enforcement officer will complete an inspection of the apartment building, and another officer will record notes and photos of their inspection, which will be uploaded to the city's mobile survey app. After inspection, the total building assessment score is calculated.

7. *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 collected through Open Toronto's website

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - There is no need for individuals to collect data.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

- It does not involve any information that requires the collection of personal data.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - It does not involve any information that requires the collection of personal data.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - The analysis of the potential impact of the dataset and its use on data subjects has not been conducted.
12. *Any other comments?*
    - None.

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.*
    - Data cleaning was performed on this dataset, uncompleted observations were deleted in order not to affect the subsequent data analysis. So these missing values need to be deleted.
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, a new dictionary was created to store the cleaned data and keep raw data.
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 Software can satisfied the data cleaning, at https://www.R-project.org/.
4. *Any other comments?*
    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - This dataset has not been used in this course so far, but it has been used in another course to check the running status of model. However, the varibales mentioned and the research direction are different.
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.*
    - At this moment, I didn't find any other papers or other databases use this dataset.
3. *What (other) tasks could the dataset be used for?*
    - We can use this dataset to test the validity of different models.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- Because the data set contains a large number of missing values, it will lead to data integrity, which still needs to be adjusted with raw data according to different needs in future research and analysis.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - Sorry, I still learning in this question, and I think if we would like to have the deeper exploring about how the factors influence the evaluation of apartment buildings.
6. *Any other comments?*
   - None.

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.*
   - Details of the latest building assessment must be posted on the tenant noticeboard. They must also be shared with all potential tenants.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - The dataset will be distributed using Github.
3. *When will the dataset be distributed?*
   - The dataset will be distributed in April 2022.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - The dataset will be released under the Open Government License-Toronto.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - There are no restrictions.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - None.
7. *Any other comments?*
   - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Qiu Wantong
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - This dataset will be uploaded to Github and can be contacted through Github.
3. *Is there an erratum? If so, please provide a link or other access point.*

- There is no any erratum.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - The dataset will not be updated when it is uploaded to GitHub.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - This data set is public data and will not be restricted.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - Raw data will be uploaded to GitHub together and will not be deleted.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
   - None.
8. *Any other comments?*
   - None.

References

Chen, Chunyan, Zhiqi Gong, Hui Yang, and Taimin Zhao. 2020. "Analysis of Adding Elevator to Multi-Storey Residential Buildings in Xining Based on Cost Benefit Analysis," 3. https://doi.org/10.1088/1755-1315/495/1/012053.

Chew, Thomas. 2018. "4 FACTORS THAT WILL AFFECT YOUR FLAT's RESALE VALUE." *Redbrick*.

Goslett, Adrian. 2012. "How Does Security Affect Your Property's Value?" *Private Property*.

McAfee, Ann. 2017. "Housing and Housing Policy." *The Canadian Encyclopedia*.

Ovation. 2015. "10 Benefits of Living in an Apartment," 1.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Son, Petty, and Prestwich. 2022. "How Important Is a Property's Parking Space?"

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.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.