STA304-Paper 1

Xueru Ma

02/02/2022

Abstract

The objective of this experiment is to get the evaluation of current building characteristics to optimize the designing for buildings in the future. The optimization is important for construction companies since appropriate features of the building will attract more customers and promote the selling of the apartments. This report explores dataset Apartment Building Registration Data on the website 'Toronto Open Data Portal'. It will mainly focus on variables 'YEAR_BUILT','CONFIRMED_UNITS', 'HEATING_TYPE', and 'PROPERTY_TYPE', mutating a new variable 'AGE'.and tried to find tenant preferences for building characteristics and form a feedback for construction companies. Finally we find people tend to live in the building with age around 60, they prefer to live in the building with hot water. Moreover,there's negative relationship between 'AGE' and 'CONFIRMED_UNITS',the new buildings would mostly have more units inside

```
citation("knitr")
```

```
##
## To cite the 'knitr' package in publications use:
##
##   Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report
##   Generation in R. R package version 1.37.
##
##   Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition.
##   Chapman and Hall/CRC. ISBN 978-1498716963
##
##   Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible
##   Research in R. In Victoria Stodden, Friedrich Leisch and Roger D.
##   Peng, editors, Implementing Reproducible Computational Research.
##   Chapman and Hall/CRC. ISBN 978-1466561595
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.
```

Data Source

This report explores dataset on building information for the buildings that are registered in the Apartment Building Standard (ABS) program. The ABS program is part of RentSafeTO which is a multi-residential inspection program that establishes standards on how building owners operate their building(s) and communicate with their tenants, in addition to establishing a schedule for continuous building evaluation, audit and enforcement of these standards.This information was collected from building owners/managers during the initial registration process, and is updated on an annual basis as part of the registration renewal process. We obtained the building information dataset in csv format from the City of Toronto Open Data Portal, and analyzing that dataset with R packages opendatatoronto as well as dplyr. The data was lastly refreshed on Feb 6, 2022.

Methodology and Data Collection

The dataset contains all building information reported to the Apartment Building Standard (ABS).ABS collects this information from the building owner/manager during the initial registration process. Since the observations of this dataset include all rental apartment buildings in the GTA area with 3 or more storeys and 10 or more units, this dataset provides a very comprehensive overview of Toronto's building rental market. The information reflected in this dataset is also unique to Toronto. This data has been collected from a very early stage and is updated annually, so it is a strong reference for Toronto's current building evaluation, audits, and tenant requirements. According to our common sense, the older a building is, the less unit amount it should have. So we expect that 'CONFIRMED_UNITS' and 'AGE' will have a negative relationship. After our observation and investigation of the data, we found that the data meets our expectation and it basically matches with the state it should present theoretically. As this data is updated annually, the rental market has been affected by covid-19 in the last two years, and people may prefer to work from home rather than coming to the office. People will have less need living in buildings located near downtown and the number of buildings with less units would be chosen. A possible problem with this data is that since some of the observations in the data were obtained too far back in time, previous rental trends may have an impact on current construction demand. Our dataset collects basic information about the buildings, but does not emphasize the different locations of the buildings. Since our analysis was performed ignoring the building locations, there are some potential bias in our dataset. Also, due to the large amount of information collected, the dataset contains some extreme values. These extreme values may have a relatively large impact on our overall analysis. Our population is all the rental apartment buildings in the GTA area with 3 or more storeys and 10 or more units. Our sample represents the building information collectd by ABS program here. There are 3488 observations and 70 variables in our sample data. The frame illustrates the way we find our sample in our population. Since this information from the building owner/manager during the initial registration process, it is mandatory for building owner or building manager to provide, not through questionnaires, etc. Therefore, there is no non-response situation, and the information obtained by the ABS organization is also more realistic and reliable.

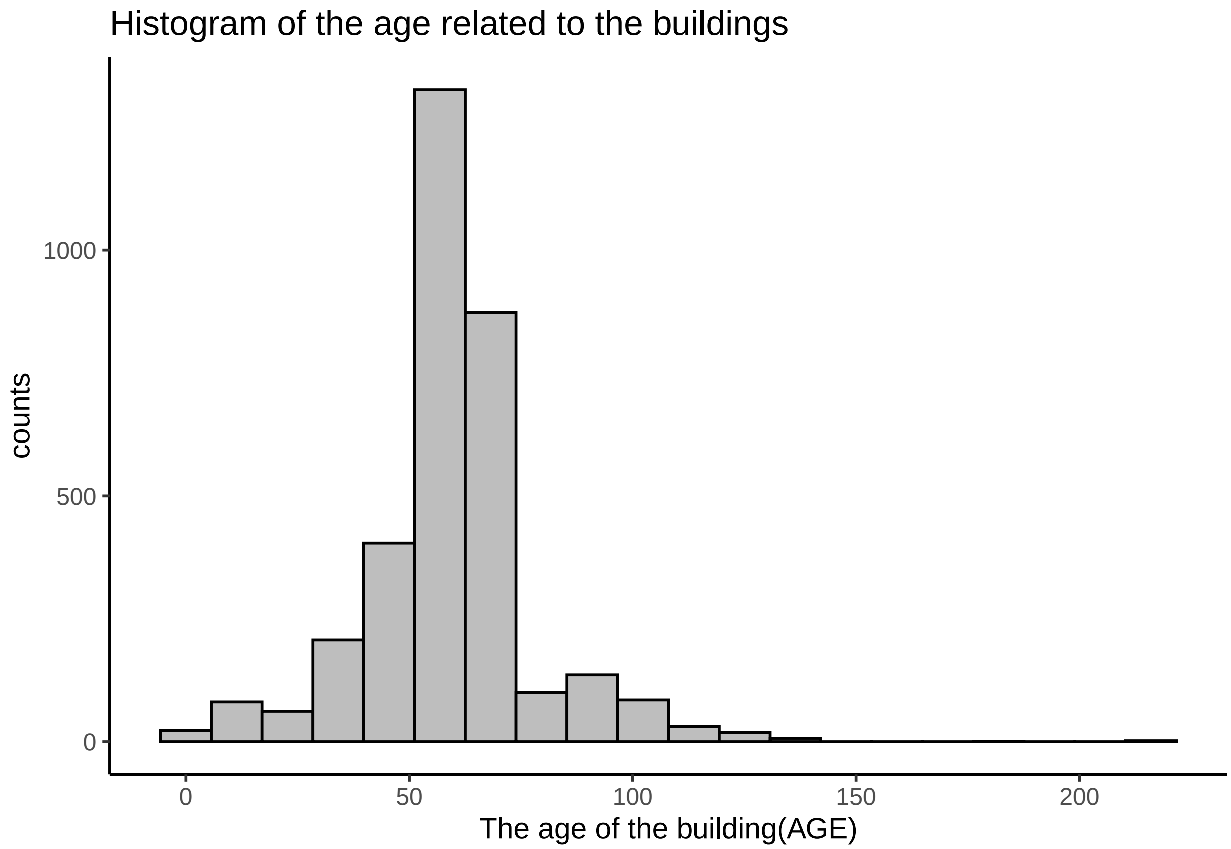
Data Characteristic (A description of the important variables)

After observing our original dataset, we find there are 3488 observations and 70 variables. However, I find there are missing values under some variables. Moreover, there are too many variables and I would like to focus on exploring several variables I am interested in. Here I did the data cleaning process. First, within the r package, I use the function mutate to create a new variable called age, which is obtained by subtracting the year the building was built from our current year. The variable 'AGE' is a numerical variable. It represents the age of the building. I found there are five relatively important variables in the dataset, They are AGE, CONFIRMED_UNITS, HEATING_TYPE ,NO_OF_ACCESSIBLE_PARKING_SPACES, PROPERTY_TYPE. The 'CONFIRMED_UNITS' variable describes the total amount of units in a building.It's a quantitative variable as well. The variable called 'HEATING_TYPE' shows the way that the building is heated. The 'HEATING_TYPE' variable is a categorical variable. As a numerical variable, NO_OF_ACCESSIBLE_PARKING_SPACES shows the number of parking space in the building. PROPERTY_TYPE illustrates whether a building is owned privately, through TCHC or social housing. Next, I use the function filter to remove the observations with missing value under variables AGE, CONFIRMED_UNITS, HEATING_TYPE, NO_OF_ACCESSIBLE_PARKING_SPACES, PROPERTY_TYPE. Later, I found that there are some extreme data in variable CONFIRMED_UNITS, which may greatly affect our next data analysis, so here I still applying the variable filter, only intercepting the cases of confirmed_units less than or equal to 4000 in our later analysis. Finally, I use the function select, choosing the five variables: AGE, CONFIRMED_UNITS, HEATING_TYPE, NO_OF_ACCESSIBLE_PARKING_SPACES, PROPERTY_TYPE to conduct the main analysis. Especially, I am looking forward to exploring the relationship between the age of the building and the number of units in that building in terms of different heating type, because I find that the new building in downtown seems to have more units.

Data

Data summary

Here is the histogrm of the interested variable: The AGE of the buildings.



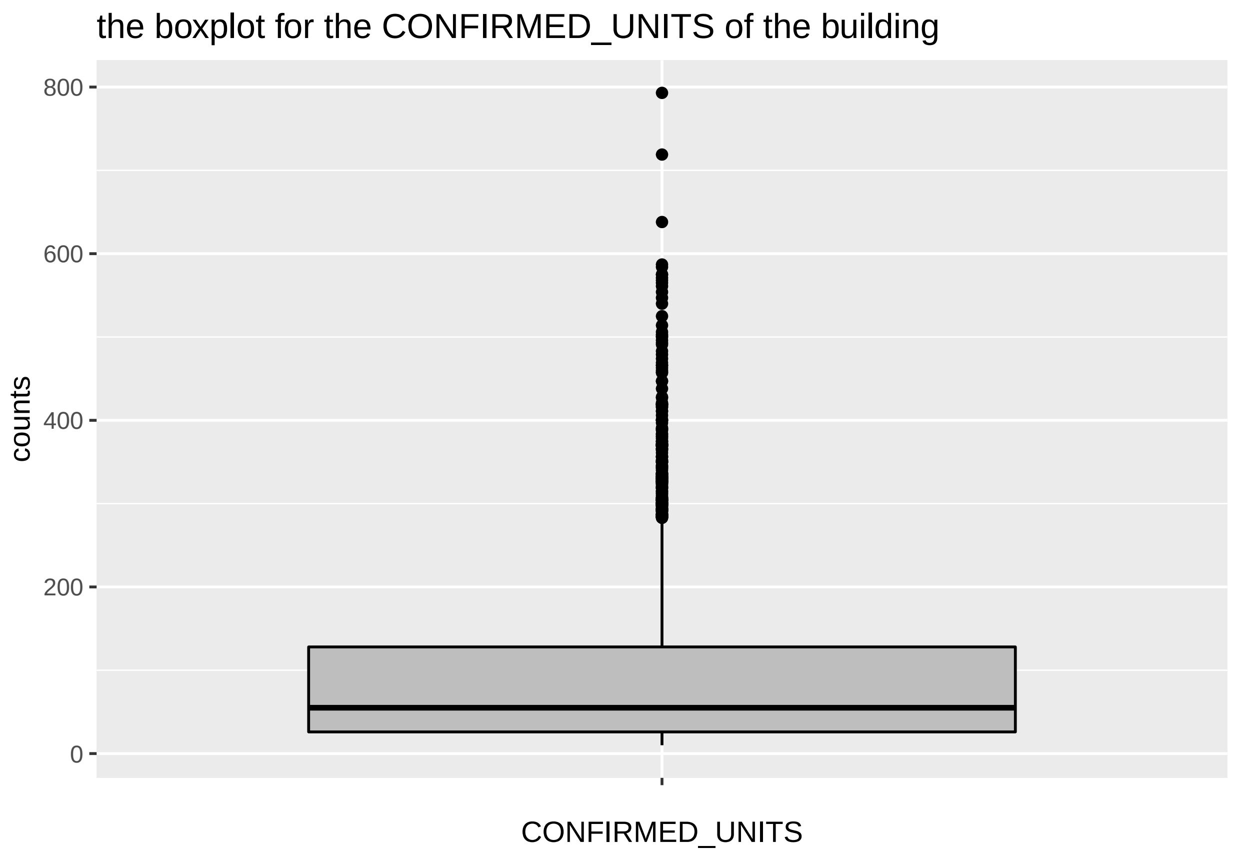
Here we create the histogram of AGE.We observed that the center of the distribution of buildings' AGE is around 60. People seemed to be less interested in living in buildings that were too old or too new. The age of the building spreads from 1 to 217, but most age of the building distributes between 30 to 90. Beside that, the age of the building is in right-skewed distribution and it is a uni-mode graph. We can observe more details related to the age of the buildings in the summary below.

Table 1: The summary for distribution of AGE

min	Q1	median	Q3	max	IQR	mean	sd	Small_Outliers	Large_Outliers
1	52	60	67	217	15	59.5922	19.22135	197	266

Based on the information provided in the numerical summary part, I find that the mean and median of the age of the building are so similar, they are all around 60. Since the interquartile range is from 52 to 67, it is relatively small and it supports our opinion above that people seemed to be less interested in living in buildings that were too old or too new. There are 197 small outliers and 266 large outliers in our model. Otherwise, we find the sample standard deviation for our model is 19.22 so there is relatively high uncertainty in our model.

Here is the boxplot for another interested variable:"CONFIRMED_UNITS"



It can be observed that most of the CONFIRMED_UNITs are located below 200. The median of the confirmed unit is around 55. We find the Q1 is about 30 and Q3 is around 130. It represents that most people tend to live in the building with not that large scale. As it is shown in the graph, the boxplot is right-skewed. We also tried to create a summary below in order to see the exact distribution data for the CONFIRMED_UNITS of the building@R.

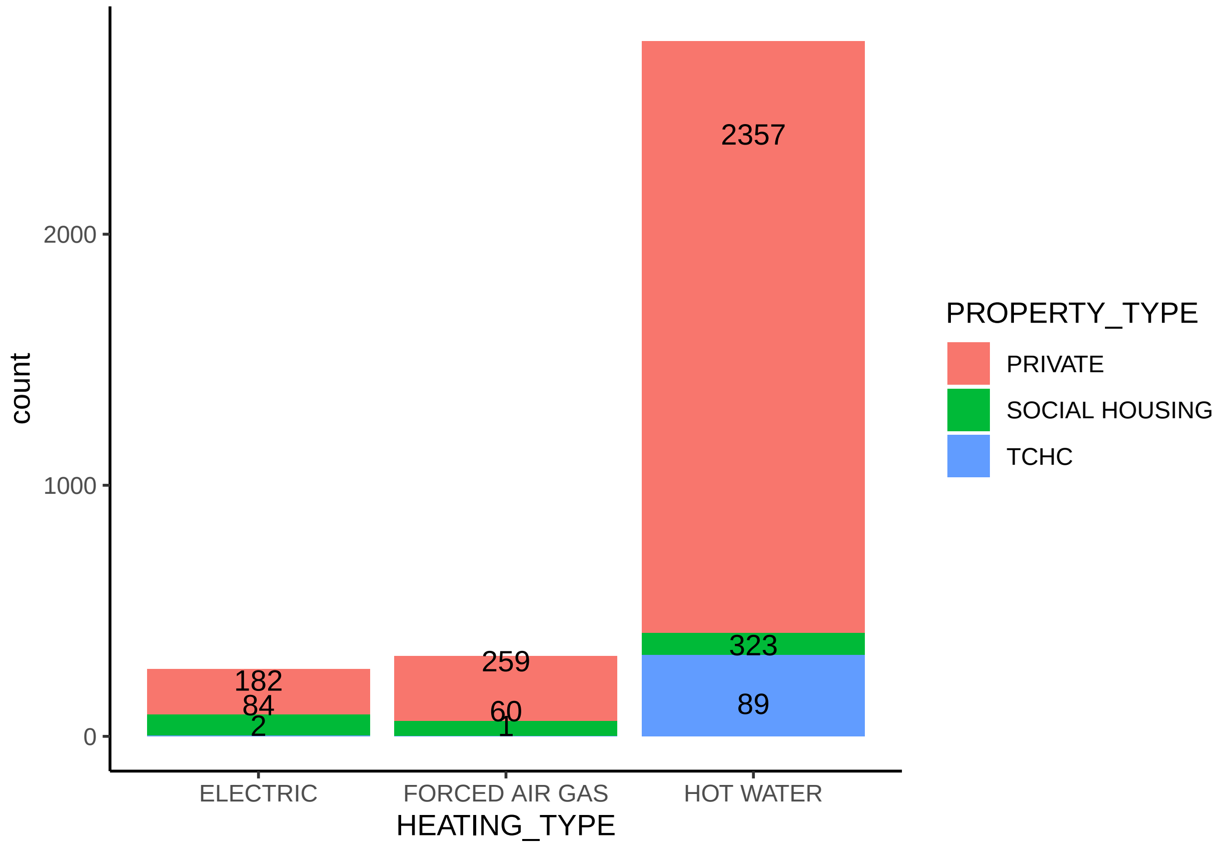
Table 2: The summary for distribution of CONFIRMED_UNITS

min	Q1	median	Q3	max	IQR	mean	sd	Small_Outliers	Large_Outliers
10	26	55	128	793	102	92.7751	97.48057	0	198

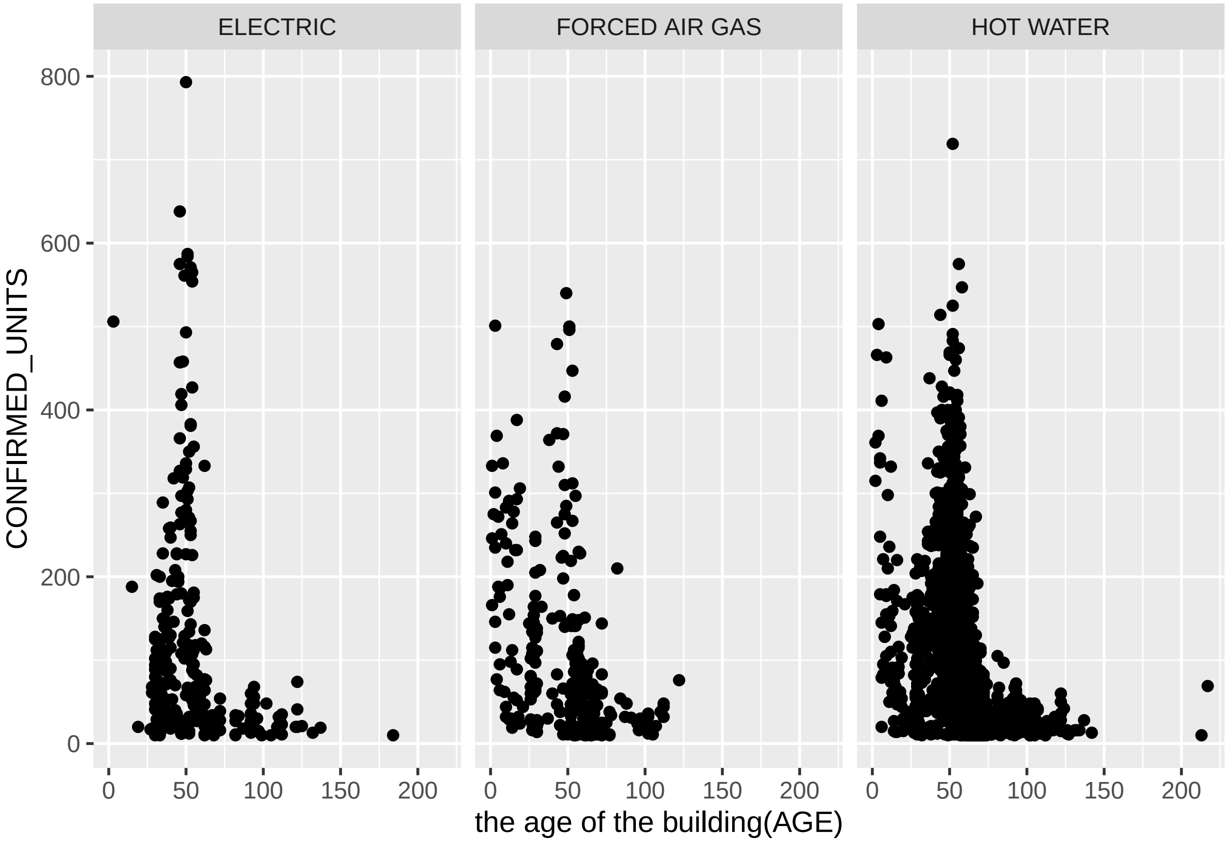
I got the numerical summaries of CONFIRMED_UNITS here. The minimum number of units in the building is 10 and the maximum number of units in the building is 793. We can get that the amount of the units in the building distributes between 10 to 793. Moreover, we found that Q1 is 26 and Q3 is 128.00, so the interquartile range is 102(128-26=102). Beside that,the table also shows there are nearly 92.78 units in each building on average. The median for the number of units is 54. The standard deviation of the CONFIRMED_UNITS is 97.48057 thus there are huge uncertanty in our model. We observed there are no small outliers but 198 large outliers in our data in terms of CONFIRMED_UNITS.People generally prefer to live in buildings with fewer units, but there are also a significant number of people who prefer to live in residential buildings with a large capacity.

More Plots

Here is the bar plot for the HEATING_TYPE and we can also observe different amounts of PROPERTY_TYPE in each HEATING_TYPE



In this plot, we can observed that most buildings choose hot water as their heating type, it's even more than the the amount that the other two types adding up together. For the HEATING_TYPE of the building, we find the amount of ELECTRIC is very closed to that of FORCED_AIR_GAS. Most buildings are private. SOCIAL HOUSING is also a common property type among all buildings. Only 92 buildings are with TCHC property type.



Here we created the three scatter plot for AGE(on the x-axis) and CONFIRMED_UNITS(on the y-axis) in terms of different types of heating approach. For the buildings whose heating type is ELECTRIC,we could observe that there's a negative relationship between the AGE and the CONFIRMED_UNITS.The scatter plot we create here is in the nonlinear form.There's a moderate relationship between the age of the building and the number of units. In terms of the buildings whose heating type is FORCED AIR GAS,we see that the relationship between the AGE and the CONFIRMED_UNITS is negative. The scatter plot for that is in the linear form. There's still a moderate relationship between the age of the building and the number of units. There's a negative relationship between the AGE and the CONFIRMED_UNITS for the buildings whose heating type is HOT WATER.The scatter plot we create here is in the nonlinear form.The relationship between the age of the building and the number of units is still moderate. Since there are the most points in the third scatter plot, we know that the amount of buildings that apply HOT WATER as the heating approach is more than that of ELECTRIC and AIR GAS. By comparing to the graph, we could also find that only the scatter plot for FORCED_AIR_GAS is in the linear form. All of the relationship between the age of the building and the number of units in different heating types is negative.There are the most observations in the HOT_WATER type which means tenants would prefer to live in the buildings with hot water.

This project is explored by using tidyverse package @MXR. This project is explored by using opendatatoronto package @M. This project is explored by using dplyr package @X.

Reference