Analysis on Factors Affecting Apartment Building Scores

Assignment 2 October 25, 2021

Zichen Gong - 1005682469

Introduction

By law, certain requirements are implemented on owners/operators of apartment buildings to make sure these properties still comply with a strict standard. This ensures the welfare of residents. The data set used for this analysis comes from City of Toronto Open Data Portal, published by Municipal Licensing & Standards[4]. Buildings registered in RentSafeTO is strictly enforced by law to go over Apartment Building Standards which runs assessment for apartment buildings with 3 or more storeys, or 10 or more units according to building maintenance standards. Areas under assessment include but are not limited to amenities, common areas, elevators, exterior building, exterior grounds, garbage and recycling management, lighting, mandatory plans (like electrical maintenance plan), mechanical systems, parking facilities and garages, security systems, tenant notification board and overall cleanliness[9]. Each separate branch is graded by score 1 to 5, with 1 being the lowest and 5 being the highest.

The purpose of this analysis is to test potential factors that might affect the score of one property. Since scores are calculated based on categories mentioned above, it is meaningless to also include those as influential factors in our model. That is to say, our model will only incorporate factors that is exogenous of the grading scheme, such as building age.

Since the RentSafeTO building maintenance standard system has already considered enough aspects to measure a good apartment building, we consider the variable SCORE as a demonstration of whether an apartment building is good enough. **This research aims to find out potential factors that have influence on the score of apartment buildings, excluding those that are used to compute the Score measurement.** Ideally consumers who consider purchasing/renting apartment registered with this program can refer back to our model.

Based on what we have from the data set, I set up a multiple linear regression model with score as the response regressed on numerical variables - the year that apartment buildings were built in, the number of storeys of one building, the number of units of one building, and the categorical variable property type - whether one building is owed by private owner, by Toronto Community Housing Corporation (TCHC), or social housing provider.

Data

Data Collection Process

All buildings registered with RentSafeTO have to undergo an assessment enforced by law. Building owners/operators of buildings with 3 or more storeys / 10 or more units have to sign up for this evaluation system. Each building is required to be evaluated at a frequency of at least once every three years[9]. A score is calculated each time and being reported into one data set, which is what we use in this research.

One biggest drawback of this data collection process is this evaluation system only includes apartment buildings that have registered with RentSafeTO. This group of buildings might have certain characteristics in common, and this cannot be discarded during data cleaning process. As what we might see later in this analysis, a large majority of apartment buildings in this data set are private owed buildings[4]. If building type is proved as a significant independent variable to our model, then it is saying that this data set might be biased by itself at the beginning of data collection process.

Data Summary

This data set contains 9028 observations of 40 variables. It contains information about buildings registered with RentSafeTO program. A few variables describe the basic information about these buildings, such as the year that they were built, their location and which type of property they are. Another variable describes scores of each building, and the rest variables are values that are used to calculate the scores. Based on these variables, we have a brief sense about what facilities these buildings have and how these buildings are organized. Higher scores reflect better living environment. Lower scores indicate that these buildings will have more frequent assessments, while buildings that cannot pass this assessments will encounter audits[9].

First, I drop some descriptive variables that are not useful for this analysis, such as building ids, variables describing addresses and interpretation of scores. Then I drop observations with NA values, as these observations are not for model set up.

Since variables type in this data set is character, they cannot be used to building our regression model. As a result, I convert variables made up of numbers into numerical type num, and categorical character variable into factor variables. These new variables created are easier for building our regression model.

There are five **important variables** in total for our regression model:

YEAR_BUILT: This provides the year that the building was built in.

CONFIRMED_UNITS: This provides the number of units in a building.

CONFIRMED_STOREYS: This provides the number of storeys in a building.

PROPERTY_TYPE: This provides information about building types. This will indicate whether one building is owed privately, by TCHC or by social housing provider.

SCORE: This will be the response variable of our regression model. This is the score of one building calculated using sum of scores of each assigned criteria divided by five times the number of unique criteria reviewed.

Here is a **summary of important variables**.

For the variable describing years and buildings were built in (YEAR_BUILT), we see that the oldest building enlisted was built in year 1805, while the newest building was built in year 2021. Both the median and mean value is 1961, and the first quartile is at year 1955, with its third quartile at year 1970, indicating that most buildings were built between year 1955 and 1970 and they cluster relatively symmetric around its mean and median.

For the variable describing the number of units provided by building owners (CONFIRMED_UNITS), the range extends from 10 units - which is just the lower limit of this assessment program - to 719 units. Since

the median value is 48 units, which is much smaller than the mean value which is 87.5 units, we might suspect there might be very large outliers which stretch the mean rightward of the median value. With a first quartile of 24 units and a third quartile of 118 units, together with a larger mean, we expect this distribution be very right-skewed.

For the variable describing the number of storeys (CONFIRMED_STOREYS), the distribution is very similar to the distribution of the variable describing units. The range extends from the lower limit, 3 storeys, to 51 units, with a larger mean of 7.545 storeys compared to its median of 4 storeys, indicating some large outliers. Combining this fact with its first quartile of 3 storeys and third quartile of 10 storeys, this distribution can be very right-skewed since there are at least 25% of all buildings have only 3 storeys, which is the lower limit requirement for this assessment program.

Frequencies of three different types of property (property_type) demonstrate that 7618 apartment buildings are privately owned, which contains 84% of all buildings. The next most type is TCHC buildings, which contains 904 buildings. Social housing is the least frequent occurred building type, which contains 551 buildings.

For the response variable SCORE, the range of building scores extends from 20 to 100. However, with a first quartile of 64, we expect more than 75% of all buildings pass this assessment. A median of 72 and a slightly smaller mean of 71.93 indicate that there are potential small outliers that pull the mean leftward of the median. A first quartile of 64 and a third quartile of 79 show that this is a left-skewed distribution.

A detailed plot grid of all five variables is presented in the appendix at the end of this report.

Plot 1 presents the distribution of variable YEAR_BUILT. This plot fits what we get from numerical summary, besides several outliers around the 1800s, most data points are between 1955 and 1970 and this plot is fairly symmetrical around its mean and median 1961.

Plot 2 presents the variable CONFIRMED_UNTIS, and this is a very right-skewed plot, just like what we get from numerical summary. Most data points cluster below 118 units, but here are some very large outliers locating in the far right.

For plot 3, the distribution of CONFIRMED_STOREY is very similar to plot 2, which makes sense due to empirical experience. Apartment buildings with more units are generally higher and have more storeys. This plot is very right-skewed, with most of its data points locating below 10 storeys. Some very large outliers in the far right indicate some very high apartment buildings. These are the buildings with more units.

Plot 4 demonstrates the frequency of three different property types. It is clear that privately owed housing is dominating among the three types. The number of TCHC housing and the number of social housing is very close to each other, with TCHC housing being slightly more.

Plot 5 presents the distribution of building scores in this assessment program. With a scale of 100, a large majority of buildings have their scores located on the upper half of this scale, indicating that most of these apartment buildings registered with program were able to pass this assessment.

All analysis for this report was programmed using `R version 4.1.1`. Data cleaning function `select()` comes from `dplyr` package[7]. All other functions come from base `R`[8]. The data set comes from `opendatatoronto` package[10].

Methods

Like what we have introduced in previous sections, we have variables that we are trying to investigate whether there are functional relationship between them. It is usually not difficult to discuss whether two variables have correlation between them, but more variables added into one regression model will help to control more potential factors, and it has better explanation power. Although there are multiple different types of function relationships that we can investigate, we usually use linear relationship since it is easier to interpret and easier to understand. In this way, relationships are described with one straight line.

In simple linear regression, we assume that the true relationship between the response variable regressed on the independent variable is $E(Y|X = x) = \beta_0 + \beta_1 x$. With respect to each specific values of X $(x_1, x_2, ..., x_n)$, we are relating the mean value of Y to each value of X. In this model, β_0 and β_1 are unknown parameters representing intercept and slope, respectively. Of course, if we collect data from real population, the mean of Y will not always sit on X values. As a result, we have $y = E(Y|X = x_i) + \epsilon_i = \beta_0 + \beta_1 x_i + \epsilon_i$, in which ϵ_i are fluctuations of y_i values and $E(\epsilon_i|X) = 0$. The relationship between Y and X are linear in parameters β_0 and β_1.

In a multiple linear regression model (which we will be using in this report) with p predictors, we have the relationship between conditional mean and X values as $E(Y|X_1 = x_1, ..., X_p = x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$. And for this time, we add random fluctuations to this model and get $Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon_i$ such that $E(\epsilon_i|X) = 0$. The response Y_i is predicted based on p predictors and each relationship is linear in parameters $\beta_0, \beta_1, ..., \beta_p$.

Although this data set contains many variables, more than half of them are scores for each individual part of this assessments. Since our response variable is Score, it is meaningless if we still use these variables used for calculation as predictor variables of our model. If we also put aside descriptive variables about building locations, variables that might have some association with building score are numerical variables: the year that the building was built, the number of units per building, the number of storeys per building, and the categorical variable: property type.

Since our model output, the response variable Score is a numerical variable, a multiple linear regression model is suitable for this situation in predicting numerical outputs. For each parameter coefficients ($\beta_i s$) on regressors, each coefficient indicates when holding all other regressors constant, one unit increase in this variable will result in β unit(s) change in the response variable. For the intercept coefficient β_0, it explains the situation when all variables' values are equal to zero, the value of response variable Score will be equal to β_0. For independent variables that are categorical, coefficients on these variables represent the effect of each factor within this categorical variable. For a categorical variable with for example, 3 factors, the final model will demonstrate only 2 of them to avoid perfect collinearity. In perfect collinearity situation, if we add all factors together, we will receive a perfect correlation of one. This makes the categorical factors meaningless in this model.

Under this assumption, the notation of our model will be in this form: $SCORE = \hat{\beta}_0 + \hat{\beta}_1 YEAR\_BUILT + \hat{\beta}_2 CONFIRMED\_UNITS + \hat{\beta}_3 CONFIRMED\_STOREYS + \hat{\beta}_4 PROPERTY\_TYPE(SOCIAL\_HOUSING) + \hat{\beta}_5 PROPERTY\_TYPE(TCHC)$.

However, we are not sure whether this is our final model for now. There might be variables that are not statistically significant, and we will remove these variables from our draft model. The most direct answer for what makes a model statistically significant is the t-value. Usually we will set up null hypothesis assuming that parameters on regressors are equal to zero, and we use t-value to test this null hypothesis for our sample data set. T-values measure how far away the distribution of this variable is from the hypothesized mean. If enough values are far enough from our hypothesized mean, usually zero, we will reject this null hypothesis. In this way, the parameter is not equal to zero, which means this variable is statistically significant enough to stay in this model. Usually we look for critical values from the t-value table with respect to the significance level (usually 5%) and the degree of freedom. If the absolute value of t-value is larger than the critical value, this variable is considered as statistically significant. P-value works in a similar way as t-values when sample data has a large sample size and we can assume the distribution as normal distribution.

One value for determining the explanation power of our model is denoted as R^2, or more precisely, adjusted R^2. It is used to determine the proportion of $y_i s$ that can be explained by this model. As denoted by $SSreg = \sum_{i=1}^{n}(\hat{y} - \overline{y})^2$, SSReg explains the amount of $y_i s$ that can be explained by this model. Another notation $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$ explains the total sum of squares. As a result, $R^2 = SSreg/SST$ gives the percentage of variation in $y_i s$ that can be explained by the regression line. However, we do not want a super high R^2 by over-fitting our model, a better evaluation of the explanation power of our model is adjusted R^2, which removes the influence from the amount of regressors in this model.

Since adding more variables into this model means adding controls which were previously conclude by the error term ϵ into this model, otherwise Omitted Variable Bias (OVB) is encountered as our model ignores some influential variables, but rather using error term to conclude them[5]. Adding any additional influential variable into one model will cause other parameters to change. As a result, we will use backward selection to test whether there are certain variables that are not so significant and can be removed from our model.

Results

Table 1

Coefficients	Estimate	Std. Error	t value	p value
$\hat{\beta}_0$	-1.416e+02	1.248e+01	-11.346	< 2e-16 ***
$\hat{\beta}_1$	1.086e-01	6.400e-03	16.964	< 2e-16 ***
$\hat{\beta}_2$	-9.002e-04	2.230e-03	-0.404	0.686
$\hat{\beta}_3$	1.517e-01	3.431e-02	4.421	9.93e-06 ***
$\hat{\beta}_4$	2.760e-01	4.595e-01	0.601	0.548
$\hat{\beta}_5$	-4.500e+00	3.618e-01	-12.436	< 2e-16 ***

Model outputs are included in this table. The second column is estimate values of $\beta_i s$. The third column is $\beta_i s$ standard error values. The fourth column is t values calculated using estimate/standard error since we assume the null of estimate is equal to zero. This is a large data set with 9073 observations after data cleaning, so t critical values are equal to z score values, for which we can use 1.96 as the 5% significance level. The last column is p values, with "***" meaning a p value of smaller than 0.001 which indicates a very statistically significant variable.

With a $\hat{\beta}_0$ of -141.6, this indicates that when all parameters are equal to zero, the assessment score should be -141.6 naturally. This does not make much sense since the lowest score should be 0. This value has a t value with an absolute value of 11.346, which is larger than a critical value of 1.96. The p value is smaller than 0.001, indicating that this constant term is very statistically significant.

With a $\hat{\beta}_1$ of 0.1086, this indicates that when we hold all other variables constant, if one building is built one year later, the score should be 0.1086 higher. Newer buildings are more likely to have higher scores. A t value of 16.964 compared to 1.96 and a p value of smaller than 0.001 all demonstrate that this variable is statistically significant for this model. This is reasonable in reality since newer buildings generally have new facilities, and it is likely that these facilities are up-to-date and easier to use.

With a $\hat{\beta}_2$ of 0.0009002, this means when we hold all other variables constant, if one building has one more unit, the score should rise by 0.0009002. This value is not very significant in reality since the change is so tidy. Besides this extremely small increase of score, we also see that the absolute value of t value is 0.404 compared to 1.96 while the p value (0.686) is very large. This means that this value is not statistically significant either.

$\hat{\beta}_3$ value is 0.1517, meaning that if one building has one more storey, its score should rise by 0.1517 holding all other variables constant. A t value of 4.421 compared to 1.96 and a p value of smaller than 0.001 show that this variable should be statistically significant. This result is reasonable. For buildings with more storeys, they generally have more residents. To avoid chaos, the building management office have to check facilities regularly. Otherwise, more residents are affected.

$\hat{\beta}_4$ estimate is 0.276, which means if this housing is social housing, holding all else constant, the value of this variable is 1 and the score should rise by 0.276. If this building is not social housing, then the value of this variable is 0 and no change is reflected on the response variable. However, since the t value is 0.601 compared to 1.96 and a large p value of 0.548, this variable is not statistically significant.

With a value of -4.5 for $\hat{\beta}_5$, this indicates that if one apartment building is TCHC owed, holding all else constant, this will cause the score to drop by 4.5. If this building is not TCHC owed, then no change is

reflected through the response variable. The absolute value for t value is 12.436 compared to 1.96 and a p value of smaller than 0.001 all tend to show that this variable is statistically significant.

If both the value of social housing variable and TCHC housing variable are equal to zero, then this model stands for when the property is privately owed.

The adjusted R^2 for this draft model is 0.06473, which indicates that for 6.473% of the time this model is able to explain the output fluctuation.

To select the best model based on what we have right now, I implement a backward selection. This selection will remove variables one by one based on their p values. If variables that are not so significant are removed and that do not affect the explanation power of this model, they should be removed. Each time one variable being removed will cause values of remaining parameters to change. There might be variables that are not so significant becoming more statistically significant, or variables that are used to be significant becoming not so statistically significant. Using this selection, we might have a model with fewer variables but is still able to explain similar amount of outputs.

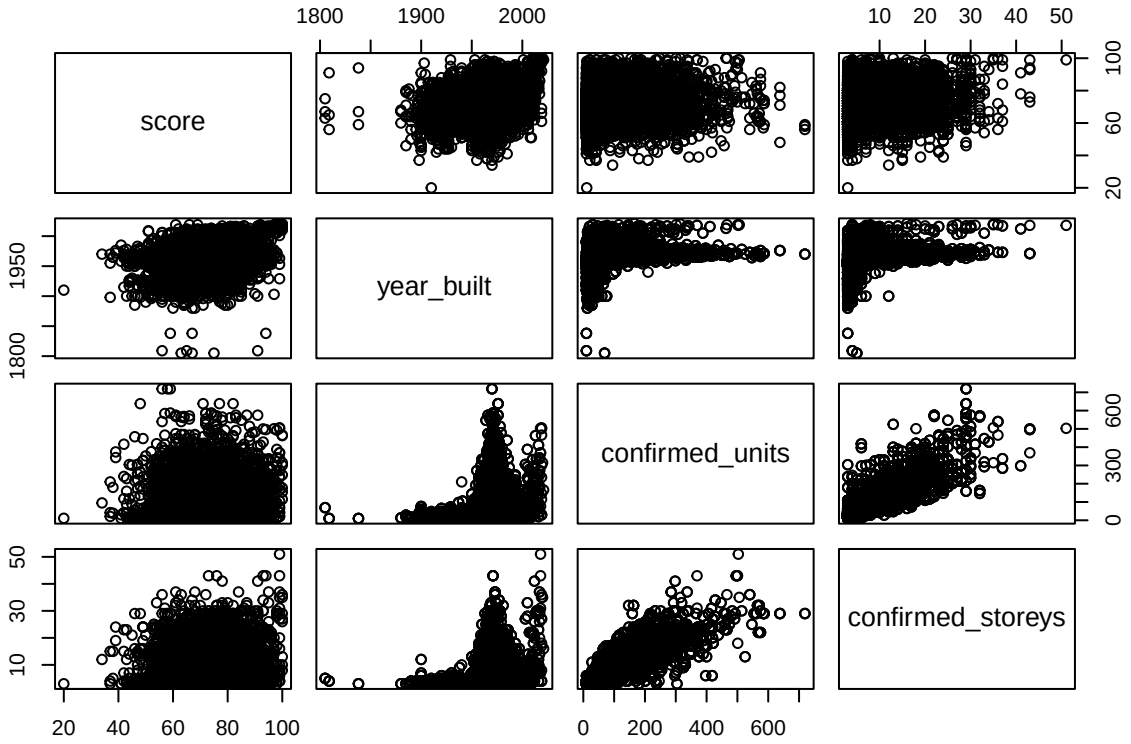
Based on the selection result, we build our final model with the variable CONFIRMED_UNITS removed. Here are the new outputs for parameters:

Table 2

Coefficients	Estimate	Std. Error	t value	p value
$\hat{\beta}_0$	-1.412e+02	1.245e+01	-11.348	< 2e-16 ***
$\hat{\beta}_1$	1.084e-01	6.384e-03	16.978	< 2e-16 ***
$\hat{\beta}_2$	1.400e-01	1.830e-02	7.652	2.19e-14 ***
$\hat{\beta}_3$	2.710e-01	4.593e-01	0.590	0.555
$\hat{\beta}_4$	-4.529e+00	3.546e-01	-12.773	< 2e-16 ***

Our final model will be: $SCORE = -141.2 + 0.1084YEAR\_BUILT + 0.14CONFIRMED\_STOREYS + 0.271PROPERTY\_TYPE(SOCIAL\_HOUSING) - 4.529PROPERTY\_TYPE(TCHC)$

We see that significance levels of remaining variables do not have any drastic change although the variable CONFIRMED_UNITS is removed. The adjusted R^2 actually rises a bit from 0.06473 to 0.06482, indicating that this model has more explanation power than the previous one.



If we take a look at the regression plot matrix above, we see that the linear relationship between score and years buildings built in is the most obvious one among all three numerical variables. Unit numbers does not see to have much linear relationship with score since score values remain relatively constant among buildings with different numbers of units. This seems reasonable according to our final model. Confirmed_storeys has a slightly stronger linear relationship with score compared to unit numbers, but the slope suggests a very inelastic relationship. A large amount increase in storey numbers would only suggest a small amount of increase in score.

All analysis for this report was programmed using `R version 4.1.1`. I used the `lm()` function in base `R` to derive the estimates of a linear regression in this section[8]. Backward selection function `ols_step_backward_p()` is from `olsrr` package[2].

Conclusions

According what we stated at the beginning of this report, we intend to find potential factors that have influence on scores of apartment buildings. Based on what we have from our final model, we know that the year that the building was built in, and number of storeys for each building are the two statistically significant numerical variables, and TCHC owed properties is one categorical factor that is statistically significant. Both years built in and storey numbers have positive linear relationship with score, and property type (TCHC) has negative linear relationship with score.

Based on our final model, we are informed that newer buildings and higher buildings with more storeys generally have higher scores in this assessment program. TCHC owed buildings have lower scores comparatively to other building types on average. This reminds consumers who consider purchasing apartment from buildings registered with RentSafeTO programs should purchase newer buildings and ideally avoid buildings owed by TCHC. If they do not mind, buildings with higher storeys (that is to say these buildings generally have more units and more residents) are better choices.

Weaknesses

The problem with this model comes from this data set. Since this data set is biased at the beginning of its data collecting process, this model only works for buildings registered with RentSafeTO program. We are not sure whether this model has ubiquitous explanation power on other apartment buildings.

In addition, this data set is composed of mainly variables used for calculating the final score, that is to say, most variables are meaningless for our analysis. Only a few of them can be used for model set up, but we know that this model does not suggest strong linear relationship between independent variables and the response variable because of the small adjusted R^2 value. It is not ideal to conclude any causal relationship.

It is understandable that the creator of this data set want their score results to be reproducible, so that all aspects under assessments are included in this data set. However, this makes this data set not very informative for further analysis. This data set is more of a presentation of their assessment process than a data pool for researches.

Next Steps

If we want a more constructive model, we should first try to persuade more building owners to register with this assessment program, so that we have more varieties in data points. To determine whether one apartment building is comparatively advanced, more perspectives should be included in this data set. For example, does the number of faculties working in management office have any impact on score? We simply omitted descriptive variables about building locations at the beginning, but in real life consumers do consider location when they plan to purchase or rent apartment. Can we include variables about the number of convenience stores within 500 meters? the number of gyms? Clinics? These are better variables than simply street names. Once we have more informative variables, ideally numerical variables and meaningful categorical variables, we are able to construct a better model.

Discussion

In this report, we construct a multiple linear regression model with score as the response variable regressed on several independent variables: years that buildings were built in, the number of storeys, and property type (whether this building is owed privately, by TCHC or by social housing). T statistics, p values and adjusted R^2 are the main tools used to determine statistic significance of variables. Backward selection method is used to choose the best model.

Based on what we have right now, our model suggests that years that buildings were built in and the number of storeys have positive linear relationship with score, while the property type TCHC has a negative

linear relationship with score. These three variables are proved to be statistically significant. However, our model overall does not have very strong explanation power, in which only 6.482% output fluctuations can be explained by our model. Ideally, we should include more buildings and more informative variables to our data set. In this way, we can work out a more convincing model. Other researchers and consumers who consider purchasing/renting apartments can refer back to our model then.

Bibliography

1. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: October 12, 2021)

2. Aravind Hebbali (2020). olsrr: Tools for Building OLS Regression Models. R package version 0.5.3. https://CRAN.R-project.org/package=olsrr. (Last Accessed: October 19, 2021)

3. Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot. (Last Accessed: October 19, 2021)

4. City of Toronto Open Data Portal. (2021). *Apartment Building Evaluation* [Data set]. Municipal Licensing & Standards. https://open.toronto.ca/dataset/apartment-building-evaluation/. (Last Accessed: October 17, 2021)

5. Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2021). Introduction to Econometrics with R. *Econometrics-with-r*. Retrieved October 23, 2021, from https://www.econometrics-with-r.org/index.html.

6. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: October 12, 2021)

7. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.7. https://CRAN.R-project.org/package=dplyr. (Last Accessed: October 19, 2021)

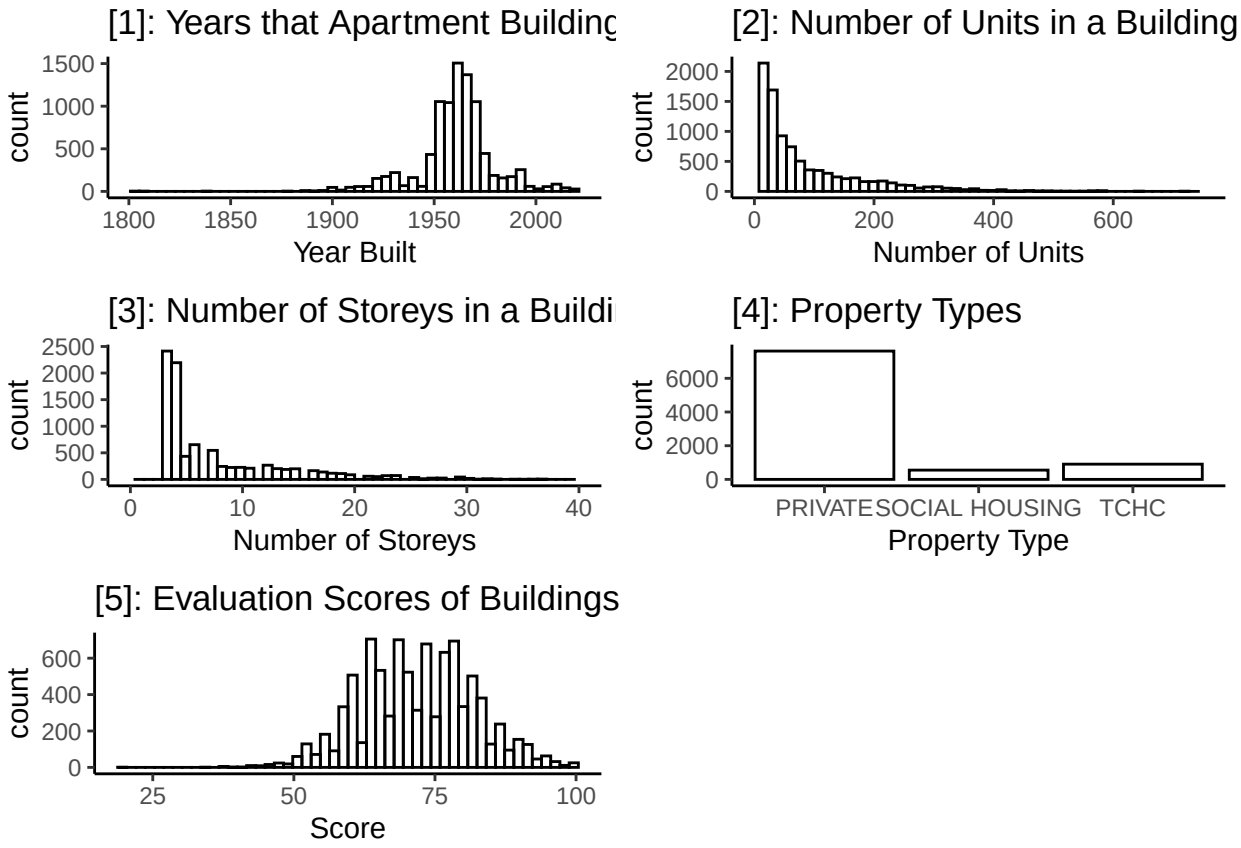
8. Peter Dalgaard. (2008) *Introductory Statistics with R, 2nd edition.*

9. RentSafeTO. (2021, October 19). *RentSafeTO Building Evaluations & Audits*. City of Toronto. https://www.toronto.ca/community-people/housing-shelter/rental-housing-tenant-information/rental-housing-standards/apartment-building-standards/rentsafeto-for-building-owners/rentsafeto-building-evaluations-and-audits/. (Last Accessed: October 19, 2021)

10. Sharla Gelfand (2020). opendatatoronto: Access the City of Toronto Open Data Portal. R package version 0.1.4. https://CRAN.R-project.org/package=opendatatoronto. (Last Accessed: October 19, 2021)

11. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686. (Last Accessed: October 19, 2021)

Appendix



All analysis for this report was programmed using `R version 4.1.1`. I use `ggplot()` function from `tidyverse` package[11] for better graphs. Graph appendix is made from `cowplot` package[3]. This document is created using `R Markdown`[1].