The Occupancy of Shelters in Toronto - focusing on room-based shelters during the COVID-19 pandemic

Assignment 2

Erica Zhou - 1005687678

Introduction

COVID-19 as a global pandemic has greatly impacted people's daily life. For people who are experiencing poverty and homelessness, the pandemic is more harmful and deadly. Thus, in this project, we will focus on the shelter system in Toronto, which provides protection and support for people facing difficulties and homelessness. We are interested in the occupancy rate of the shelters (room-based shelters) and especially its connection with the scale of the shelters. According to the report from the City of Toronto, there has always been pressure on shelter beds (for bed-based shelters) since 2018[1], and thus the occupancy rate of beds is always high. Thus, in this project, we focus on another type of shelter, the room-based shelters, that is, in these shelters, the rooms will not be shared between households. It is more friendly to family programs. Especially during the pandemic, more families and groups are facing financial challenges and seeking supports. Thus, we also want to explore if being included in a "COVID-19 response" program makes the shelter more highly demanded in 2021. Therefore, this project specifically focuses on the COVID-19 period and does not apply to any other time. The topic is not as popular as the one of the bed-based shelters, but it is necessary to explore this topic for this special period to support the system and prepare for any future change.

Hypotheses

Before the analysis, I hypothesize that there will be a positive relationship between the scale of a shelter and the occupancy rate. My opinion is that a larger shelter may be better known and hold a higher reputation so that it may be more popular and has a higher occupancy rate. I also predict that being included in a "COVID-19 response" will increase the occupancy rate of the shelter in this period of time.

Terminology

According to the topic, all the "shelters" mentioned in the analysis are "room-based shelters", that is, the shelters providing separate rooms for households. The "occupancy rate" also refers to the occupancy rate of the room-based shelters. EDA refers to Exploratory Data Analysis[4], which is a method that summarizes the main characteristics of the data with visualization.

The **Data** section in this project gives an overview of the data with some interesting characteristics and visualization. There will be introductions of the important variables as well as numerical summaries such as the center and spread of the data. In the **Method** section, there will be a description of the linear model construction, and the results will be presented in the **Result** section. There will be a conclusion about the research in **Conclusion** at the end of the project. And finally, there will be some massages for further studies.

Data

Data Collection Process

The data **Daily shelter overnight occupancy** is collected from Open Data Toronto[2]. One can download the csv/json/xml file directly from the portal and reproduce the research. An alternative is to copy and

paste the code also provided in the same link directly to the R markdown. The dataset gives a daily list of active overnight shelters in the Shelter Management Information System (SMIS) database. The data provides daily updated information about the shelters and the overnight service programs like the program's operator, location, classification, occupancy and capacity, and so forth. As a result, the individuals that are not covered by such programs are not counted towards the data. This may cause bias and inaccuracy of the research result. In addition, the data does not include the information on the overall quality of the shelters, therefore the results may be not accurate in practice. Plus, this research focuses on the COVID-19 period only and cannot be applied to other situations.

Data Summary

The data **Daily shelter overnight occupancy** provides a daily list of active overnight shelters in the SMIS database. The data is updated every day from 2021-01-01 to 2021-10-14. It contains information about the overnight programs, the shelters, the capacities, and the occupancy, and so on. It includes all the basic information related to our research question.

I kept only ORGANIZATION_NAME, SHELTER_GROUP, CAPACITY_TYPE, CAPACITY_ACTUAL_ROOM, SERVICE_USER_COUNT, OCCUPIED_ROOMS, OCCUPANCY_RATE_ROOMS and PROGRAM_AREA, which are probably more helpful for this research and further studies on room-based shelters. Then I removed all the observations collected as a bed-based shelter to keep only the room-based shelters in the dataset. The values of PROGRAM_AREA are converted into 1 if it is "COVID-19 Response" and 0 otherwise. Plus, there are no missing values in this new dataset.

Important variables for this research:

ORGANIZATION_NAME: It refers to the name of the organization providing the overnight service.[2]

SHELTER_GROUP: It refers to the shelter group to which the program belongs in the SMIS database.[2]

CAPACITY_ACTUAL_ROOM: It refers to The number of rooms showing as available for occupancy in the Shelter Management Information System.[2]

OCCUPANCY_RATE_ROOMS: It refers to the proportion of actual room capacity that is occupied.[2]

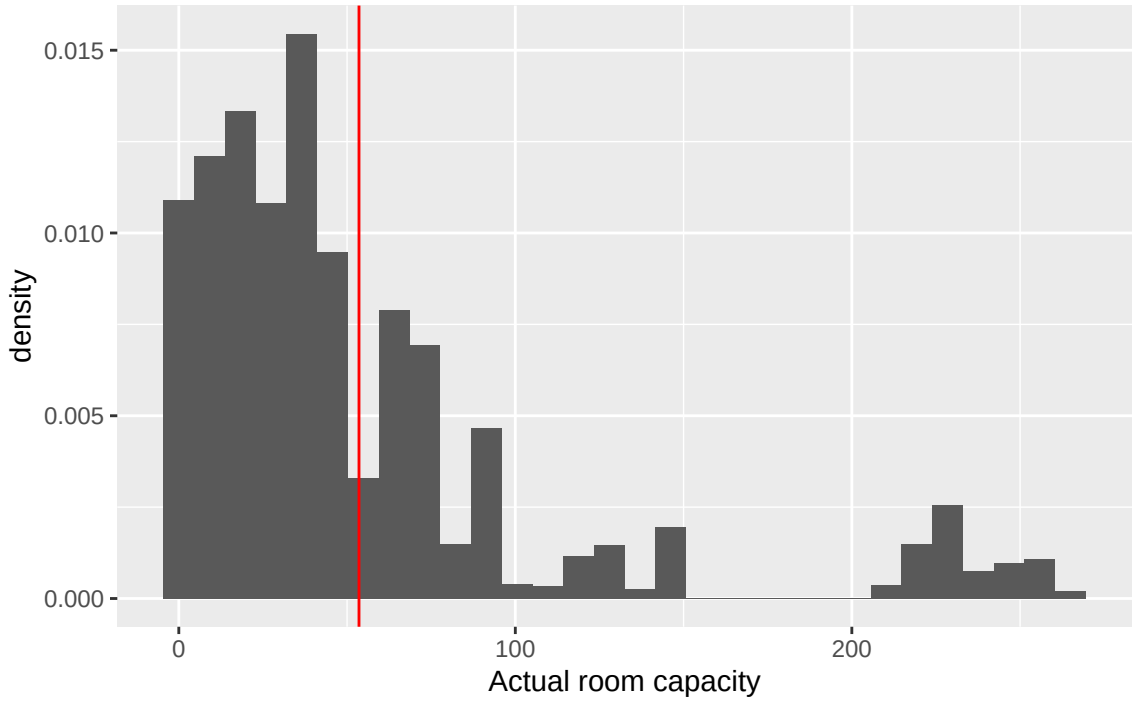
PROGRAM_AREA: It refers to the program in which the shelter is covered, including Base Shelter and Overnight Services System, Winter Response, Temporary Refugee Response and COVID-19 Response. It is converted into a binary variable that equals 1 if the program area is COVID-19 Response and 0 otherwise.[2]

Variable	Mean	Standard Deviation
CAPACITY_ACTUAL_ROOM	53.546	58.298
OCCUPANCY_RATE_ROOMS	93.057	16.391
PROGRAM_AREA	0.786	0.41

From the numerical summaries, we notice that the actual room capacity is 53.546 on average, that is, there are about 53 available rooms in the shelter. However, it has a high spread represented by the high standard deviation, which means that there are fewer observations located close to the mean. In practice, it indicates lots of small shelters and big shelters rather than average-size shelters. On the other hand, the occupancy rate of the rooms achieves an average of 93.057%, which indicates a high level of occupancy on average. It has a lower spread compared to the actual room capacity, which means that the mean is in a better way representative of the sample data. The mean of PROGRAM_AREA tells us that about 78.6 of the listed shelters are covered in a COVID-19 program. Both the numerical variables are highly spread and the categorical variable has a lower spread.

```
df1 %>% ggplot(aes(x = CAPACITY_ACTUAL_ROOM, y = ..density..)) + geom_histogram(bins = 30) +
  xlab("Actual room capacity") +
  ggtitle("Fig.1") +
  geom_vline(xintercept = round(mean(df1$CAPACITY_ACTUAL_ROOM), 3), col = "red")
```

Fig.1




```
df1 %>% ggplot(aes(x = OCCUPANCY_RATE_ROOMS, y = ..density..)) + geom_histogram(bins = 30) +
  xlab("Occupancy rate of rooms")+
  ggtitle("Fig.2") +
  geom_vline(xintercept = round(mean(df1$OCCUPANCY_RATE_ROOMS), 3), col = "red")
```

Fig.2

Fig.1 is a histogram of the actual room capacity. The red line is the sample mean of the data. We see an obvious left skewness[4], which means that most observations have a lower capacity than the average. According to the graph, most shelters have fewer than 50 rooms. On the other side, we can see some outliers on the right side above 200. It means that there is a small portion of large-scaled shelters that have more than 200 rooms. The outcome is reasonable as it is costly to construct and organize a room-type shelter in Toronto. Larger-scaled shelters generally need more land, budgets, and are more time-consuming. However, they are necessary for the shelter system of the city to avoid an increasing amount of family homelessness especially during some challenging periods such as the COVID-19 pandemic. Statistically, the skewness and the outliers will cause the violation of normality[8] and create biased estimates.

Fig.2 is a histogram of the occupancy rate of the room-based shelters. The red line is the sample mean of the data. We see that most rooms have a full occupancy, which causes a great right skewness. However, we also notice that there are shelters with extremely low occupancy rates close to 0. Overall, the occupancy rate is high, which means the room-based shelters are highly demanded and sufficiently used but equivalently it indicates the pressure of the shelter system of Toronto in 2021. Statistically, the linear regression model will violate the assumption of normality because of the skewness. It will lead us to biased estimates[4].

Fig.3

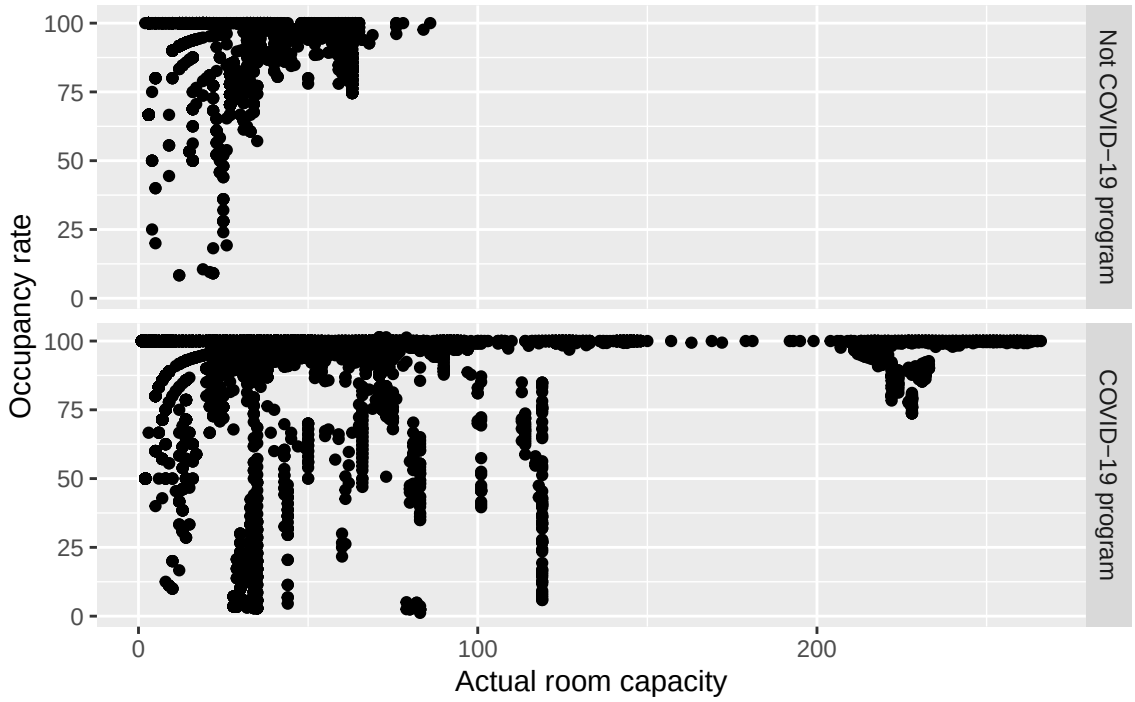


Fig.3 is a scatter plot showing trends between the actual room capacity and the occupancy rate by program area. We noticed that the occupancy rate is overall increasing when the room capacity raises. Therefore, we can predict a positive relationship between the occupancy rate and the room capacity. However, we notice that a big portion of the shelters have a 100% occupancy rate, and it is also reflected in Fig.2. A difference between the two groups is that the shelters covered by a COVID-19 program are more likely to have a greater capacity, and these large-scales shelters usually have a higher occupancy rate. We may predict that a larger-scaled shelter will have a higher occupancy rate. On the other hand, we may guess that COVID-19 has given big pressure on the shelter system and shelter rooms are more highly demanded in response to COVID-19. According to the visualization, we may expect our linear regression model to have a high response variance as it did not satisfy the assumption of constant variance[8] and will not explain the population behavior well. It may also violate the linearity[8] as we see a great increasing pattern on the left side of the plot.

All analysis for this report was programmed using `R version 4.1.1`.

Methods

In this section, we will explore the linear regression between the occupancy rate and the scale of the shelter assumed that the data satisfies all assumptions[8] of a linear regression. We will be using a multiple linear regression model[8]. In statistics, a multiple linear regression model is a technique that uses more than one predictor (independent variables) to predict the response (dependent variable). With this model, we can better predict the dependent variable as a function of the independent variables. Graphically, we are creating a non-vertical straight line with a slope, an intercept, and some errors that represent the relationship. We look at the p-value to see if our results are significant. The p-value is the probability of obtaining test results at least as extreme as the results observed, under the assumption that the null hypothesis is correct[4]. If the p-values are smaller than the significant level 0.05[4], then we say that our results are significant. On the other hand, we want our R^2[4] to be big. R^2 measures the proportion of the variation in the response that is explained by the predictors. Finally, we want to interpret the results appropriately in practice to check their rationality.

Regression model:

The multiple linear regression model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Where β_0 represents the intercept of the regression line. In our case, y refers to the occupancy rate, x_1 refers to the actual capacity of the shelter and x_2 is a **binary variable** that equals 1 when the shelter is covered by a COVID-19 program and 0 otherwise. Therefore, β_1 represents the average change in the occupancy rate with one unit increase in the capacity of the shelter, and β_2 represents the extra effect if the shelter is categorized as "included by a COVID-19 program" comparing to the ones not covered. To think of the model equation differently, if the shelter is included by a COVID-19 program then the intercept of the regression would be $\beta_0 + \beta_2$, otherwise, the intercept would be only β_0. As mentioned before, we **assumed that our linear model satisfies all four assumptions**[8].

Results

Our estimated linear relationship is $y = 90.923 + 0.019x_1 + 1.442x_2$, where x_1 refers to the actual capacity of the shelter, and x_2 represents "being covered by a COVID-19 program". y is the occupancy rate of the shelter. Recall that y and x_1 are continuous variables, and x_2 is a categorical variable.

term	estimate	std.error	statistic	p.value
(Intercept)	90.9231322	0.2985121	304.587739	0.00e+00
CAPACITY_ACTUAL_ROOM	0.0186696	0.0023523	7.936900	0.00e+00
PROGRAM_AREATRUE	1.4424007	0.3346149	4.310629	1.64e-05
[4]				
R^2: 0.00672				

The intercept would be 90.923 if the shelter is not in a COVID-19 program. Otherwise, it would be 92.365. It indicates the occupancy rate of a shelter covered by a COVID-19 program to be generally 1.422% higher than **the ones not covered** regardless of the capacity (see the functions below). The value of β_1 tells us that when the capacity of the shelter increases by 1 unit, the occupancy rate will gain a 1.9% improvement. It turns out to be a subtle positive connection between the occupancy rate and the capacity. On the other hand, the positive value of β_2 represents a positive effect on the occupancy rate when being covered by a COVID-19 program. We may believe that COVID-19 does contribute to the demand for shelters.

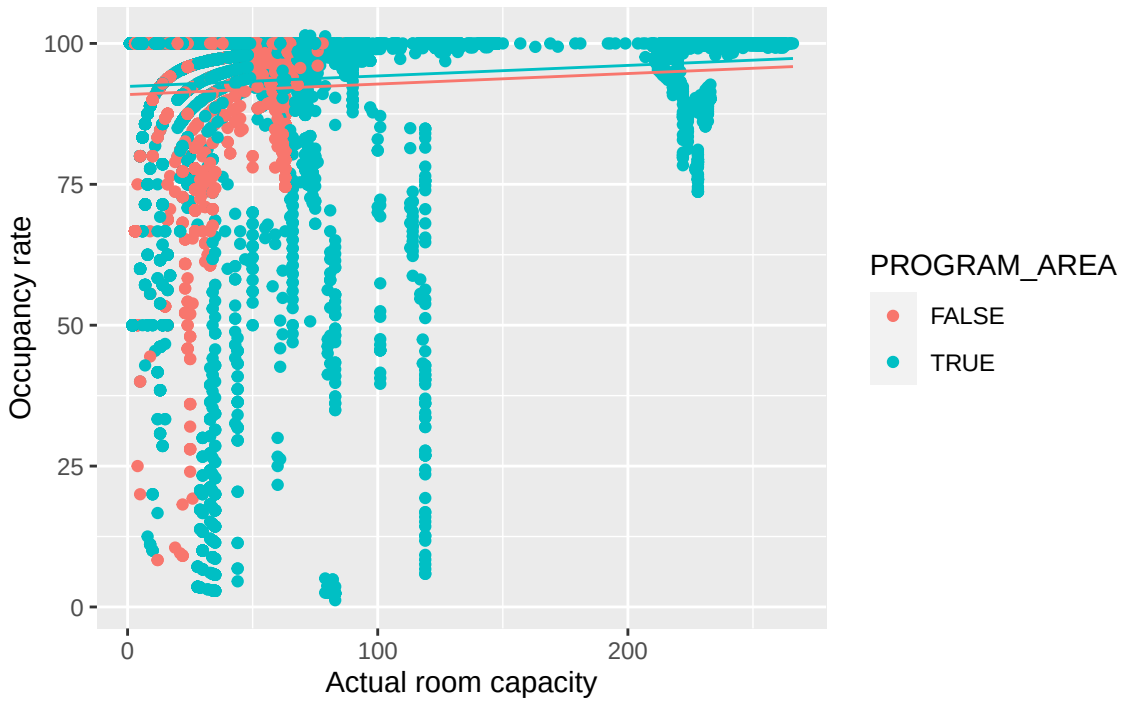
$$Covered\ by\ a\ COVID-19\ program: y = 92.365 + 0.019x_1$$

$$Not\ covered\ by\ a\ COVID-19\ program: y = 90.923 + 0.019x_1$$

The null hypothesis[4] refers that there are no relationships between the response and predictors. The p-values here are all extremely small and smaller than the significance level of 0.05, so we may think that our results

are significant. However, the coefficient of determination[8] is 0.00672, which means that only 0.672% of the variation in the response is explained appropriately by the model. The result is reasonable in practice because there are lots of factors not included in this research such as the management and the construction of the shelters. Some factors probably have a much stronger impact on the occupancy of a shelter. Although the results are significant, we need to add better predictors to the model in the future.

Fig.4



In Fig.4, **PROGRAM_AREA = TRUE** means that the shelter is covered by a COVID-19 program. It shows the patterns between the capacity and the occupancy rate. The two straight lines represent the linear relationship for the covered ones and the uncovered ones correspondingly. It confirms that the shelter covered by a COVID-19 program generally has a higher occupancy rate. However, there are great variances on the left side of the graph. It turns out that the linear model doesn't appropriately explain the population behavior. We may predict a fractional function to better fit the relationship in the plot.

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 frequentist logistic regression in this section [9].

Conclusions

According to the linear model, there **is a slightly positive linear relationship** between the occupancy rate and the scale of the shelter. The capacity only causes a small amount of increase in the occupancy rate. It makes sense in practice because most homeless people or ones suffering from poverty would not pay as much attention to the scale or the reputation of a living place as mentioned in **Hypotheses**. Thus, the government may not need to concern about the scale when predicting the occupancy rate of an existing or future shelter. On the other hand, being covered by a COVID-19 program **has some positive impact** overall on the occupancy of the shelters. In other words, COVID-19 causes some extra demand for the shelters and thus increases the occupancy rate. It confirms the hypotheses in **Hypotheses**. It turns out that COVID-19 has brought pressure to the shelter system especially to the room-based shelters for families and groups. Before COVID-19, these families and groups didn't require shelters because they had jobs and incomes. However, many of them either lost their jobs or had a reduction in income due to the recession of the economy during the pandemic. In the future, the government could pay more attention to the room-based shelter system to

support some families and groups during hard times. More supportive programs could be added to the system if possible. At the same time, reducing the pressure of the overall shelter system would be an important and emergent topic.

Weaknesses

The weakness of the model is the dissatisfaction of normality, linearity, or constant variance[4] of linear regression. Thus, the outcomes are not reliable enough. Since it violates linearity, there is model misspecification[8], which can result in biased estimation of the coefficients. The dissatisfaction of constant variance will cause the variances to be too large, and thus our estimates have less precision and the estimates of variance are not representative of the truth. As it violates normality, our p-values are not trustable. This indicates that the significance we have shown in the **Result** section may not be trustable.

To solve these problems, we have to select more useful predictors to make the model more efficient and complete. Some useful predictors might be clients' satisfaction, room qualities, food & drinks support, and so forth. In this research, a single dataset contains limited variables, and thus, it is necessary to find varieties of datasets that can provide more useful and appropriate variables and check their satisfaction with the assumptions.

Next Steps

The next step is to fix the problems mentioned above by collecting more useful and appropriate data. For future studies, researchers can consider the bed-based shelters that are more widely used by the homeless. With more possible datasets, researchers can explore the quality of the programs and the services, and the satisfaction and health level of the clients of the shelters. Then there can be a final report that combines both room-based and bed-based shelters to provide a complete overview of the occupancy of the Toronto shelter system during the pandemic.

Discussion

In conclusion, this project constructs a linear model to explore the connection between the occupancy rate and the scale(capacity) and the COVID-19 programs. It is a small step in the topic of the shelter system and COVID-19 that still needs improvement. By doing more research and analysis in the future, we will finally be able to relieve the pressure of the shelter system, improve the overall quality of shelters and gradually reduce the homelessness in the city during the pandemic or even generally.

Bibliography

1. Mary-Anne Bedard. (2018) *2018 Issue Briefing: Pressures on Toronto's Shelter, Housing and Homelessness System.* City of Toronto. https://www.toronto.ca/city-government/council/2018-council-issue-notes/pressures-on-torontos-shelter-housing-and-homelessness-system/. (Last Accessed: October 23, 2021)

2. City of Toronto. (2021, Oct 20) *Daily Shelter & Overnight Service Occupancy & Capacity.* City of Toronto. https://open.toronto.ca/dataset/daily-shelter-overnight-service-occupancy-capacity/. (Last Accessed: October 20, 2021)

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

4. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.

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

6. Allaire, J.J., et. el. *Table.* RStudio. https://rmarkdown.rstudio.com/lesson-7.html. (Last Accessed: October 23, 2021)

7. Aschwanden, Christie (2015-11-24). *Not Even Scientists Can Easily Explain P-values.* FiveThirtyEight. https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/. (Last Accessed: October 23, 2021)

8. Sheather, Simon (2008) *A Modern Approach to Regression with R.* Springer Texts in Statistics.

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