Explore the causal inference between the population projection and safety of neighborhoods, with respect to the crime rate of different crimes, violent crime percentage, property crime percentage in Toronto

Jianheng Chen - 1005680746

December 17, 2021

Abstract

In the **Introduction** section, we are going to cover some background of crime rates under the global pandemic. Description of data, research question, and hypothesis will also be in this section.The data cleaning process and important data, variable we want to study the existence of casual relationship showcased in the **Data** section. Moreover, in the **Method** section we will introduce the process of propensity score matching to match the data with respect to the treatment group that we can reduce bias from confounding variables. In **Result** section will reveal the result of the propensity score matching to explain the casual inference of safety of the neighborhood and predicted population projection. Finally, we will conclude the key results and analyze the drawbacks of our model and the possible progress we can make in **Conclusion**.

Keywords

Global pandemic, Crime rate, Neighborhood, Population loss, Safety, Confounding variable, Causal Inference, Propensity Score Matching, Hypothesis test

Introduction

Background

It is hard to imagine it has been over two years since the arise of COVID-19. The global pandemic brings irreversible changes to our lives, from the way we connect with other people to the way we live. However, the pandemic in fact brings some advantages, for example, the reduction of many types of crimes in the world. According to Lederer's study of the US's crime cases (2020), 19 out of 20 policy officers from Chicago reported that a significant decrease in the number of criminal incidents from March to April 2020. Even though the report also indicated that the number of domestic violence had a trend to increase, the overall crime rate had decreased by 10% in Chicago. After more and more restrictions were taken to contain the spread of the virus the crime rate had a foreseeable trend of decrease globally, concluded Leverer (2020).

Moreover, street crime, a main type of crime including robbery, stealing without assaulting the victim, has been reported to have a huge decrease in number as there were fewer people on the streets (Edwards, 2020). Even Karachi, the city called "Asia's most crime-ridden city", has no report of auto theft in the entire March.

With the overall crime rate decreasing, are the cities being safer compared to before the pandemic? The answer was not that simple. Logez (2021) indicated that the US has seen the biggest increase in the murder rate in decades in 2020 and it was 15% higher than 2019. The rising homicide rate and shooting by firearm rate had become an unavoidable social problem. A 6-year-old was killed in a drive-by shooting, which shocked the mayor of DC (Logez, 2021). What is more, Logez (2021) considers that it was not because of year-to-year fluctuations in crime but the stress under lockdowns, the disappointments of policy and abuse of gun usage. Many people were forced to leave their homes lived for decades.

Even though the case was not severe as the US, Canada suffered from the same problem of emerging of a higher murder rate (Logoz, 2021).

Dataset and problem

The data I collected is from City Toronto (2021) about the neighborhood crime rate from 2014 to 2020 created by the Toronto Police service. The information collected are the predicted population projection in different neighborhoods in 2020, the number of different types of crime including assault, auto theft, break and enter, robbery, theft over, homicide, shooting, and firearm discharges from 2014 to 2020, the rate of different types of crime per 100,000 population from 2014 to 2020.

I choose this dataset as it is generally believed that the residents are always concerned with the safety of the neighborhood. People's opinion of a community mostly relies on the rate of violent crime in the community (Office Of Policy Development And Research, 2016). What is more, the people may choose to move to other communities if the number of severe crimes significantly increased in the neighborhood they lived.

Based on the study of violent crime and murder and non-negligent manslaughter rate in the US from 1995 to 2014, violent crime wreaks a terrible impact on the neighborhoods' development, people's health and leads to vicious circles of decay (Office Of Policy Development And Research, 2016). Although with distinct preferences, people from different ethnic groups will tend to exit neighborhoods with higher rates of violent crime. Hipp (2011) made a further conclusion that people with housing choices consider the safety of neighborhood as their top priority. Moreover, the population loss because of the neighborhoods with a high rate of violent crime will push the neighborhood to a worse scenario. Economic drop in the neighborhoods was not the cost of the population loss, coercive sexual environments could emerge and the number of sexual harassments, molestation, and exploitation would rocket (Hipp, 2011). The Government would be the final payer to these consequences and quite a lot of investments and policy supports would be needed to revive the economy and environment of the neighborhoods.

Research question

Noticing the great costs from population loss of neighborhoods, to conclude, my research question will be **How does the 2019 crime rate of different crimes, violent crime percentage, property crime percentage affect safety level towards the 2020 population projection in different neighborhoods in Toronto?**.

This analysis plays important role in checking the population change with the safety level of communities under the effect of the global pandemic of COVID-19. We may be able to predict the level of safety and incoming population change in the neighborhood by our model if there is causal inference. Referring to the background this is a possible way to support the government to see the trend of population loss based on the crime rate, which could save a lot of financial expenditure.

Hypothesis

With the literature review above we notice the rate of violent crime, including homicide, assault, robbery, and shooting (Statistic Canada, 2021), will greatly decrease the population projection in a community as they are considered to be a sign of an unsafe environment. According to the Office Of Policy Development And Research (2016) property crime such as auto theft, theft over (also called pickpocket) has an unignorable effect on the perception of safe of a community. These factors affect the level of safety in a neighborhood. For the hypothesis part, we assume whether a neighborhood is safe has a causal inference with the population projection of the neighborhood.

Data

Data Collection Process

The dataset is created by the Toronto Police Service and the data collection was started in 2014, published May 29, 2018, and most recently get updated on October 20, 2021(Open data dataset, 2021). The data is titled "Neighborhood Crime Rates (Boundary File)" discussing the crime data by neighborhoods. To be specific, the Toronto Police Service used the geographical division on the map licensed under Open Data Commons Open Database License (ODbL) to classify each neighborhood (Leaflet, 2018). By using the record of where the crime occurred, the Toronto Police Service was able to match each crime to a specific neighborhood to create this dataset. Moreover, the 2020 population estimated for different neighborhoods was achieved from a non-governmental organization named Environics Analytics (Open data dataset, 2021). A clarification was made that only long-term residents were recorded and temporary populations such as the commuters and business patrons were not included. The aim of creating the population projection data is to compare the number of crimes in different geographic areas with different population sizes. Furthermore, the crime was calculated by the **crime count per 100,000 population** under the standard of Statistic Canada (Open data dataset, 2021). The crime rate provides a more precise perspective of the crime by taking into account the change in security in the region compared with the prime count (Open data dataset, 2021).

Drawbacks

One foreseeable drawback of the data collection process is the precise categorization of where the crime occurs. There is no criterion on how to categorize a crime that happened at the junction of two neighborhoods or a crime that crossed a few neighborhoods. How to record the case can be ambiguous and not objective, which biases the dataset we have. Unfortunately, we cannot remove this bias in the data cleaning process so we would mark it as a limitation.

Data cleaning and Summary

As described in the **Dataset and problem** section the information collected is the id and names of the neighborhoods, the population projection of these neighborhoods in 2020, the number of cases of assaults, auto thefts, break&enter, robbery, theft over, homicide, and shooting from 2014 to 2020, the rate of these crimes per 100,000 population and the geometry locations. In conclusion, the dataset has 140 observations with 104 variables.

The next step we will do is to clean the data. Firstly, the dataset has several columns indicating the id and name of the neighborhoods, which are irrelevant to the research question we want to discuss, so they will firstly be removed. Secondly, recall that we want to study how these crime rates affect the safety of the community towards the prediction of 2020 population projection, the data lacks timeliness is not useful in predicting. Under the background of the global pandemic of COVID-19, the data before 2019 was not valid enough. Similarly, there is few data in 2020 to complete our study so we will choose the data from 2019.

Moreover, in the **Data Collection Process** we know the crime rates are the best predictors for change in security in the region so we will make no modification to the crime rates in 2019. Next, as we also want to study how the percentage of violent crime and property crime affect the level of safety. We need to create two new variables to represent them. Fortunately, we can do the job by dividing the number of violent or offenses against the property by the total number of crimes, based on their definition in **Hypothesis**. Then, we are also interested in the fatal crimes proportion in the environment. Among the crimes we study, homicide and shooting can be treated as the most serious violations as they are fatal (Statistic Canada, 2016). Based on Becker's study of within neighborhood disadvantages, collective efficiency, and homicide rate (2019), a high homicide rate led to disadvantages including extremely violent environment and poverty, which influenced the collective efficiency of neighborhood. By here, we will categorize the neighborhoods with the percentage of life-threatening crime lower than the mean as relatively safe neighborhoods. We will then remove the crime counts as we no longer need them.

Finally, we check the information by rows to identify if there are any outliers or blanks. Fortunately, there are no outliers and blanks so we do not need to remove anything.

Important variables

After cleaning the data and creating the new variables we need in the research questions, there are overall 11 variables in the new dataset. Reminded what we want to study in the research question, the crime rates variables are homogeneous in describing the crime counts per 100,000 population. By here, the population projection, violent and property crime percentage, and safety of the neighborhood are the main four variables of our interest and concern that we will do a showcase here.

Population projection The variable indicates the predicted population in different neighborhoods in Toronto in 2020.

Percentage of violent crime This variable measures the percentage of homicide, assault, robbery, and shooting over the total crime counts.

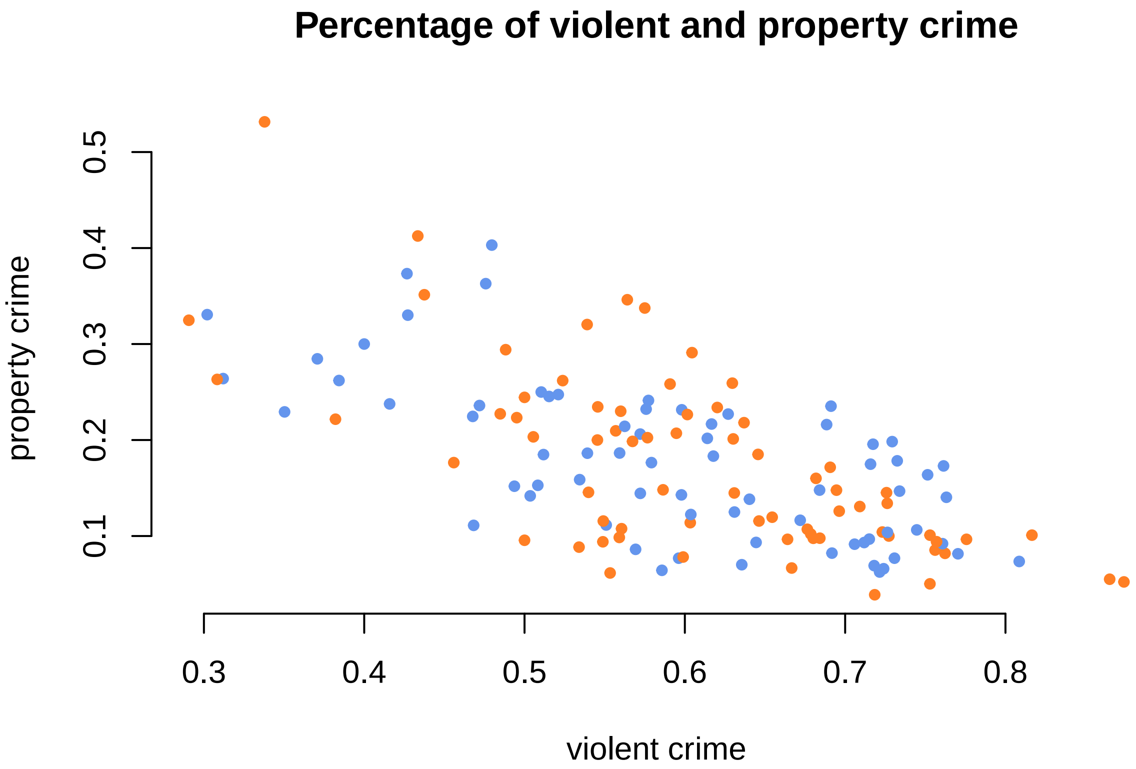
Percentage of property crime This variable measures the percentage of auto theft and theft over, over the total crime counts.

Safety This is a variable describing if the neighborhood is relatively safe by measuring the percentage of lethal crimes.

Table 1: Distribution of the percentage of violent and non-violent crime

	vars	n	mean	sd	min	max	range	se
percent_of_pro	1	140	0.1750612	0.0885287	0.0388350	0.5314618	0.4926268	0.0074820
percent_of_violent	2	140	0.6016887	0.1213952	0.2905983	0.8739130	0.5833148	0.0102598

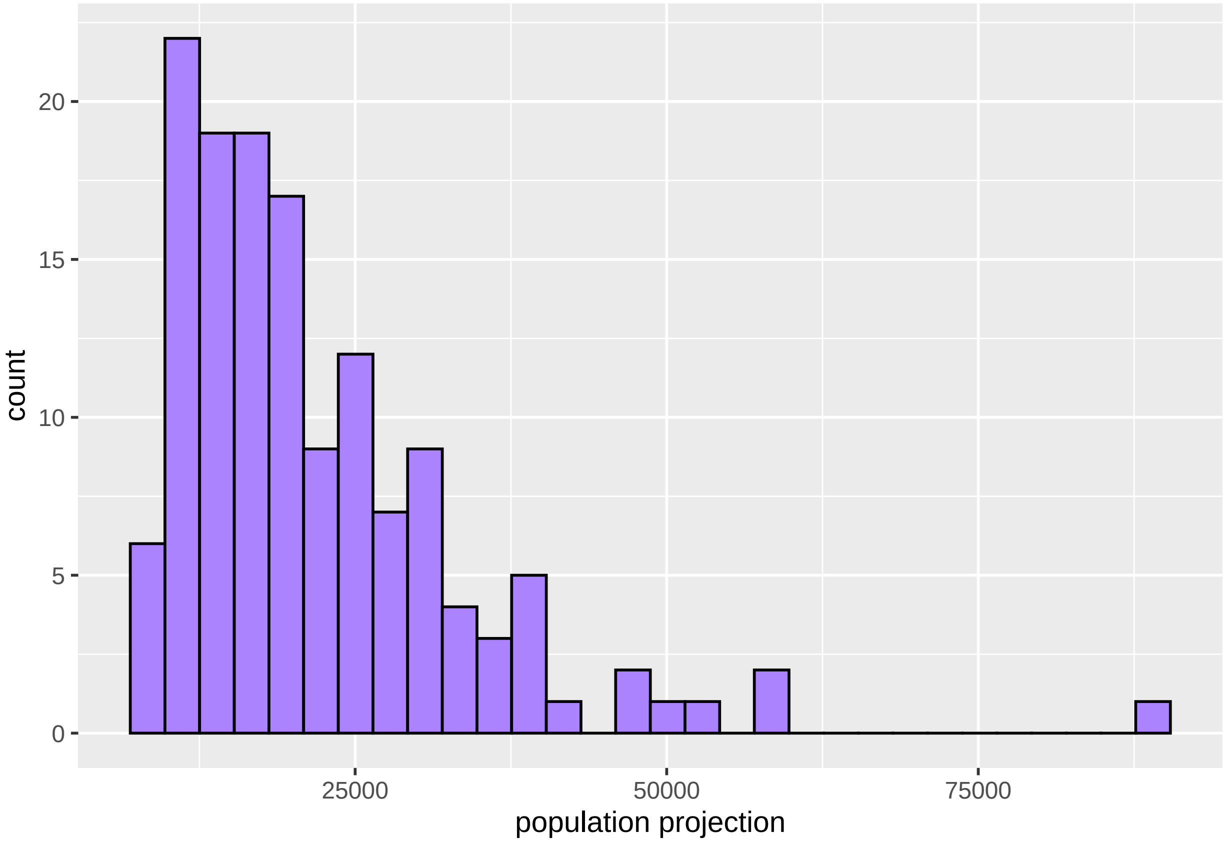
Graph 1: Percentage of violent and property crime in all crimes.



According to **table1**, in the 140 observations of neighborhoods, we surprisingly see that on average there are over 60% of cases of crime are violent crimes, with a standard deviation of 0.121. The percentage of property crimes takes a small proportion of crimes that they occupy 17.5% of total crime counts with 0.089 standard deviations. In fact, this result complies with what we studied in the **background**.

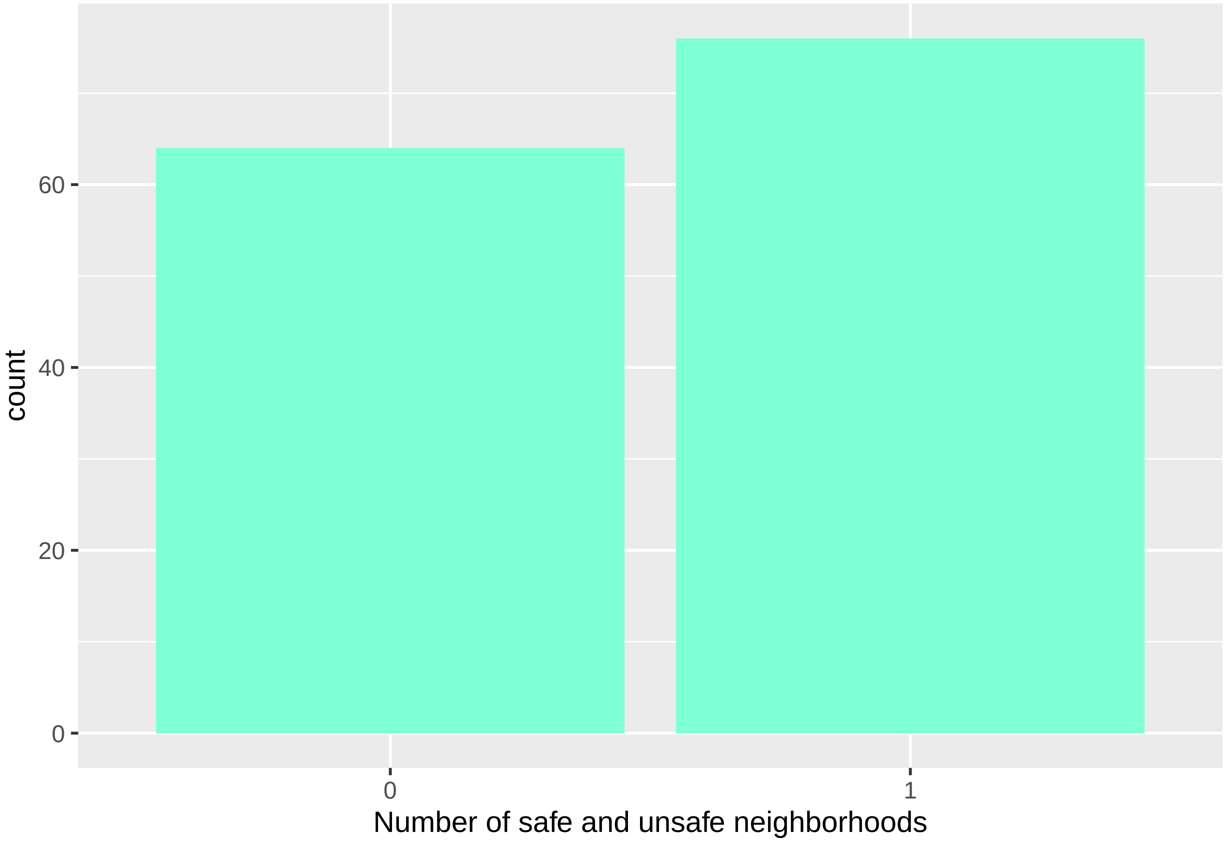
Referring to **Graph1**, the blue points stand for violent crime and the chocolate points stand for property crime. Based on the 140 neighborhoods, we can see the points representing the percentage of violent crime are quite concentrated around 0.6. There are seldom any observations of a relatively high property crime rate. We can relate these back to the **table1**. For the distribution graph please see **Appendix**.

`Graph 2: Distribution of population projection in 2020`



As reported by **Graph2**, a significant right skewness of the population projection was shown. Merely few neighborhoods would have a population greater than 87,500 while most of them will have only 13,000 population.

`Graph 3: Number of safe and unsafe neighborhoods`



According to **Graph3**, we will see there are 64 neighborhoods are classified as safe and 76 are classified as unsafe. We will then perform a methodology named propensity score matching to help us to reduce the bias due to confounding variables, which are the variables associated with each other and the effect of one factor on an outcome can be influenced by another (Resnik, 2019). We will have **safety** as our treatment and **population projection** as our outcome of interest as we want to study their causal inference. An explicit explanation will be made in the next **Method** section.

All analysis for this report was programmed using `R version 4.1.1` with `tidyverse 1.3.1` package (Wickham et al., 2021) and `pysch 2.1.9` package

Methods

Methodology

For the construction of the model we first need to figure out what methodology is needed for our model. Recall from **Important variable** section we showed there are not equal observations of unsafe and safe neighborhoods. We can use the methodology of propensity score matching to get a new model with an equal number of safe and unsafe communities when other predictors are controlled.

Generally speaking, propensity score matching is a technique to estimate the effect of a treatment by accounting for the covariates that predict receiving the treatment.(Abadie, Alberto, Imbens&Guido, 2006). Before using the methodology, noted we need to check the assumptions: conditional on observational data, untreated units can be compared to the treated (Rosenbaum, Paul, Rubin&Donald, 1983). In our case, the data we use is from the observations of crimes reported by the Toronto Police Service and the treatment variable is safety (categorical), which means the comparison is possible. We can conclude that we are not violating the assumptions of propensity score matching.

Next, we will cover the process of propensity score matching. As mentioned above, safety will be our treatment and the response variable population projection will be our outcome of interest. The propensity score matching will be for the safety propensity. We will then run a logistic regression model where the treatment is the response variable based on other covariates as we think these variables help to explain it. A fitted value will be created based on the propensity score. What is more, our forecast based on the propensity score will be added to the dataset and we will use it to create matches. To be specific, for every neighborhood that was treated(classified as safe), we want the untreated neighborhood (classified as unsafe) to be considered as similar to the treated (based on propensity score) as possible. We will do this by finding the closest treated neighborhoods to the untreated neighborhoods (also called nearest neighbor matching). Then, evaluation for quality of matching will be processed, we should verify if covariates are balanced across treatment and comparison groups in the matched observations. Finally, we will get a reduced dataset with an equal number of treated and untreated observations.

Moreover, we can now run a Frequentist linear regression on our new dataset, where population projection will be our response variable, the crime rate of different crimes, violent crime percentage, property crime percentage, and safety will be our predictors. The introduction of the model is covered as supplementary methods in the **Appendix** section.

Mathematical model

Understanding that the propensity score matching helps us in reducing the confounding bias but not completely remove them. Here we will construct a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + ... + \beta_{10} x_{10} + \epsilon$$

. With what we discussed in the **Important variable**, we will go through the parameter of interest.

The first parameter β_0 here represents the intercept of the regression line, which is the value of population classified as unsafe and all other predictors unchanged. The parameter $\beta_1, \beta_2, ..., \beta_7$ are quite homogeneous. They in order represent how the population projection count will change with respect to one unit change of the assault rate, auto theft rate, break&enter rate, robbery rate, theft over rate, homicide rate, and shooting rate on average. Next, β_8 plays the role of what changes will be in the population projection with identification of safety (printed as 1 if safe, 0 if unsafe). Moreover, β_9 and β_{10} stand for the response of changing 1 percent in the percentage of violent crime and property crime. Recall that we want the propensity score matching to reduce the influence of confounding variables, by applying this model to a hypothesis test we can find if there is a significant relationship between the independent variables to check how well we did in the propensity score matching.

Hypothesis test

Keep in mind that we will discuss the relationship of population projection with these independent variables. To be specific, we will use a hypothesis test here. It is appropriate that we want to test whether to reject a hypothesis for population and the p-value can do the job (Rumsey, 2021). To be specific, our null hypothesis is $H_0 : \beta = 0$ which means no significant relationship exists in the predictor and response variable, and the alternative hypothesis is $H_A : \beta \neq 0$ which means there is a significant relationship exist. In general, we will choose a 5% level of confidence, and when the p-value is smaller than 0.05 we are sufficient enough to reject H_0.

Results

After the propensity score matching, we will get a new dataset with 128 observations. Based on our methods we come to a summary of our multilevel linear regression, as shown in **Table2**.

Table 2: Results of the multiple linear regression model

$Parameter$	$Estimate$	$Std.Error$	$t - value$	Pr>$\|t\|$
$\hat{\beta}_0$	39771.114	23071.278	1.724	0.0874
$\hat{\beta}_1$	9.556	6.769	1.412	0.1607
$\hat{\beta}_2$	5.325	13.165	0.404	0.6866
$\hat{\beta}_3$	-19.295	14.449	-1.335	0.1844
$\hat{\beta}_4$	-18.364	21.373	-0.859	0.3920
$\hat{\beta}_5$	66.046	42.468	1.555	0.1226
$\hat{\beta}_6$	-165.489	254.648	-0.650	0.5170
$\hat{\beta}_7$	-81.907	66.195	-1.237	0.2184
$\hat{\beta}_8$	1791.062	3627.127	0.494	0.6224
$\hat{\beta}_9$	-273.31	392.60	-0.696	0.4877
$\hat{\beta}_{10}$	-256.13	305.33	-0.839	0.4033

To be noticed, as long as we are comparing the p-value with 0.05 to check the significance of the variables with respect to the response variable. However, in this case, we get no significant predictor. We will then run a simple linear regression between the outcome and treatment alone to see if it is because of confounding variables.

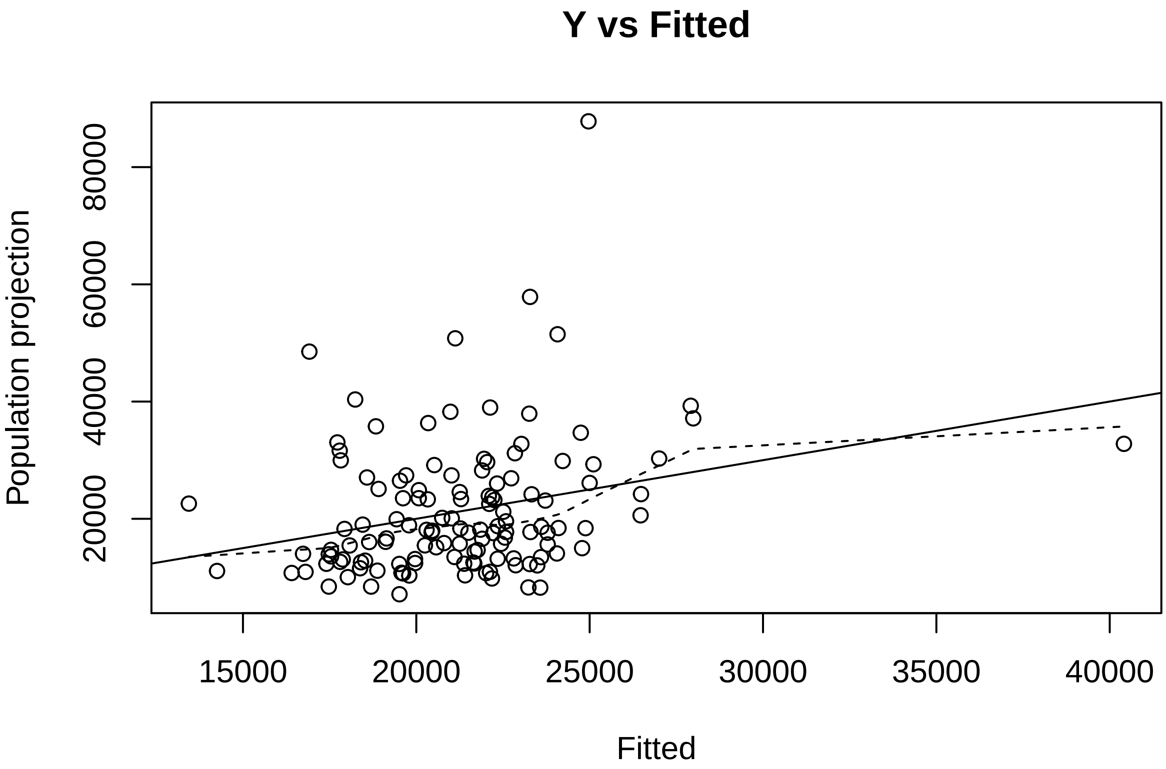
Table 3: Results of the simple linear regression model

$Parameter$	$Estimate$	$Std.Error$	$t - value$	Pr>$\|t\|$
$\hat{\beta}_0$	20405	2004	10.181	<2e-16
$\hat{\beta}_8$	1765	2314	0.763	0.447

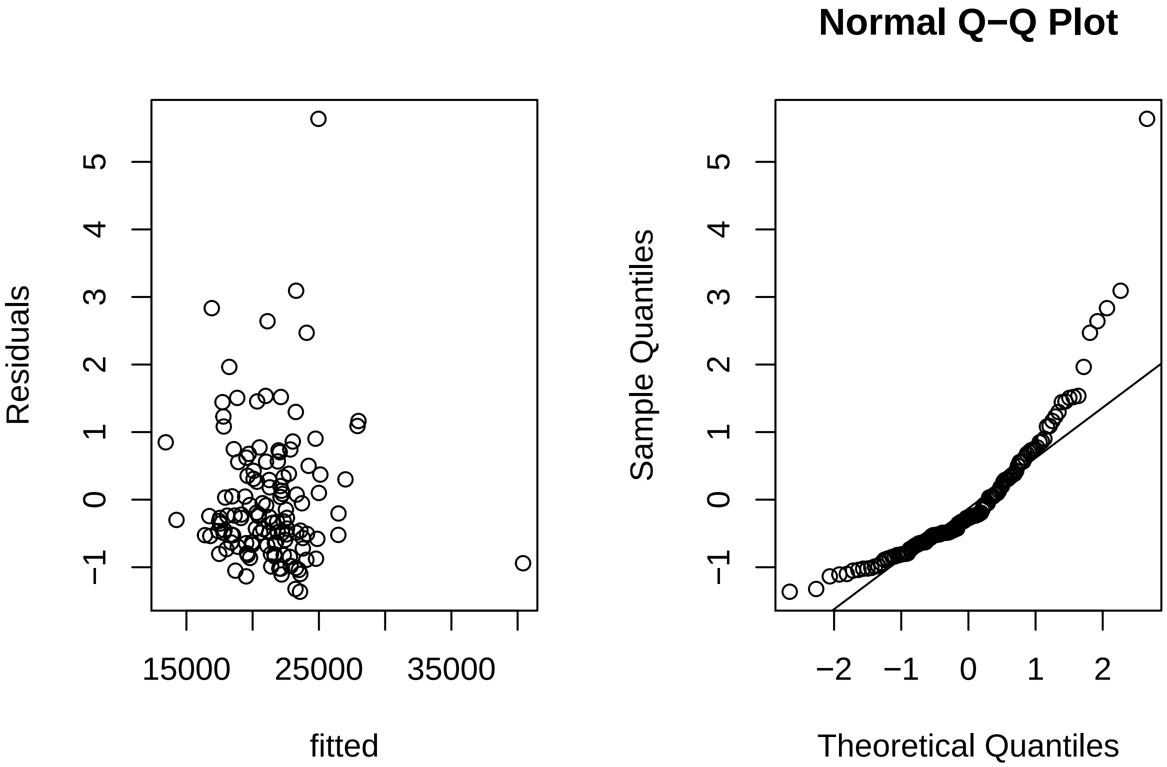
In this case, we will have a significant intercept, but the treatment we select, safety, still can not be concluded as affecting the population projection. To be noticed, in both of these models we have R^2 smaller than 0.1.

According to Remsey (2021), the reasons of significance can be the minor effect of the independent variable on the response variable, the variance of the response data too large and too few samples. We can see the estimate of $\hat{\beta}_8$, which indicates how the population projection will change if the neighborhood is classified as safe, can merely affect the population projection. Then, the standard error in the treatment is larger than the mean, indicating an abnormal distribution of data. We will then use residual plots to check the possible reasons for this result.

`Graph 4: Response variable and fitted multiple regression`



`Graph 5: Residual plots for our final model`



In **graph 4** we apply the regression line of our full model on the outcome. We can see a severe variation appears in our response variable where most observations are around 20,000 but the others are relatively random. For the residual vs fitted value plot, the residuals do not follow a linear relationship and the residuals are not spread randomly around the horizontal line of 0. What is more, the normal Q-Q plot tells us if the residuals are distributed normally with a mean 0 and constant variance. However, in this Q-Q plot, there are lots of points on the lower and upper tail not falling on the line. Here we will say the result of no casual inference exists in whether the neighborhood is safe and the predicted population projection is reasonable, as we find out that we are not building a reasonable linear model. In conclusion, we see that treatment with only the effect of the other variables in our study is too weak to affect the outcome and there are too big variances in the response variable. The assumption of normality in the model is violated that the linear regression goodness of fit is not powerful enough.

Conclusions

In conclusion, this report applied the method of propensity score matching and hypothesis test to predict the causal inference of treatment and outcome. A hypothesis we had was whether a neighborhood is safe with respect to the crime rate of different crimes, violent crime percentage, property crime percentage has a causal inference with the population projection of the neighborhood. We choose safety as our treatment and population projection as our outcome. The main methodology, propensity score matching, was used to match the unsafe and safe neighborhoods by accounting for the covariates that affect the treatment. A reduced dataset was created and we used a hypothesis test to detect the significance of the treatment and how it affect the outcome. However, the treatment was shown not statistically significant in predicting the outcome under a 95% confidence level as a result. We then found out that some assumptions of our model were violated and our final model was not good enough to fit the data, which could be the reason for this null result.

Weaknesses

According to our whole process of study, one weakness of our data is the notable gap in our outcome data. The variance of the population projection is too high that constructing a proper model is a difficult task (but ethically we can not remove them as outliers). What is more, all of the data in the original dataset was numerical that the confounding variable issue can be huge. We are performing a propensity score matching technique based on our hypothesis threshold which can be inaccurate. The threshold we set to distinguish the safety is the mean of fatal crime percentage, which can be an incorrect standard.

When the result shows that no casual relationship exists in the treatment and outcome, we also notice that the linear regression model we choose was not a good model to fit the data. It is likely that the overlaps of the covariates affect the modeling of our data.

Next Steps

Even though we derive a null result in our study of this dataset. The research question is still meaningful and interesting. Understanding that whether a neighborhood is safe can significantly determine the population loss in a neighborhood (Hipp, 2011), to study their causal inference we can try to define safety with more covariates besides crime data. According to Becker (2019) the formation of a neighborhood, including the residents' income, age and sex also help to identify the security of the community. It is not impossible to find another dataset to better develop our research. With the assistance of these extra variables, we can better estimate the safety of the neighborhood. What is more, we can try a different threshold on determining if a region is safe with the support of more academic reviews. In the propensity score matching process, we can match the safe and unsafe neighborhoods with a fairer propensity score, which is helpful in efficiently reducing the bias of confounding variables.

Discussion

In the study of the Office Of Policy Development And Research (2016), it discusses the intertwined characteristics relevant to a high crime rate, including poverty, segregation, inequality, collective efficacy, job access, residential instability, foreclosures, vacancy rates, and land use. These characteristics can be both the cause and result of a hazardous environment. The cost of a violent environment is also discussed in that the average rates of obesity and stress levels are comparatively high in the dangerous regions. People refuse to go outside or choose to move as a result.

Indeed, government programs can help to prevent crimes by helping youth and other people with financial support and medical care. A cognitive-behavioral therapy session was proceeded in Chicago to help youth to slow down their thinking and consider whether their automatic thoughts fit the situation (Office Of Policy Development And Research, 2016). It was expected that the violent crime rate can be reduced by 45 to 50 percent and the graduation rate would have a 19 percent increase. It was possible the population loss problem be solved with the increased number of new-born and higher population quality.

Finally, I hope my study provides some insights into the importance of neighborhood safe towards the population loss problem.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: January 15, 2021)

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

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: January 15, 2021)

4. Lederer, E. M. (2020). Crime rates plummet around the world as the coronavirus keeps people inside. *Time. Available online at https://time. com/5819507/crime-drop-coronavirus.*

5. Edwards, R. (2020). Crime and the Coronavirus: What You Need to Know. *SafeWise:[site]. May, 20.*

6. Lopez, G. (2021, July 21). *Murders are up. crime is not. what's going on?* Vox. Retrieved December 15, 2021, from https://www.vox.com/22578430/murder-crime-2020-2021-covid-19-pandemic

7. Open data dataset. City of Toronto Open Data Portal. (2021, May 6). Retrieved December 3, 2021, from https://open.toronto.ca/dataset/neighbourhood-crime-rates/.

8. Statistic Canada. (2021, July 27). *Police-reported hate crime, by most serious violation*, Canada (selected police services). Open Government Portal. Retrieved December 15, 2021, from https://open.canada.ca/data/en/dataset/7c4e7d38-bbe5-447c-85ec-b714f3a06a4e#rate

9. Office Of Policy Development And Research. (2016). *Neighborhoods and violent crime: HUD USER.* Neighborhoods and Violent Crime | HUD USER. Retrieved December 15, 2021, from https://www.huduser.gov/portal/periodicals/em/summer16/highlight2.html

10. Hipp, J. R. (2011). Violent crime, mobility decisions, and neighborhood racial/ethnic transition. Social Problems, 58(3), 410-432.

11. Leaflet. (2018, May 29). OpenStreetMap. Retrieved December 15, 2021, from https://www.openstreetmap.org/copyright

12. Becker, J. H. (2019). Within-neighborhood dynamics: disadvantage, collective efficacy, and homicide rates in Chicago. *Social problems*, 66(3), 428-447.

13. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4 (43), 1686.

14. Revelle, W. (2021). *psych: Procedures for Psychological*, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.1.9, https://CRAN.R-project.org/package=psych

15. Abadie, Alberto; Imbens, Guido W. (2006). "Large Sample Properties of Matching Estimators for Average Treatment Effects". *Econometrica.* 74 (1): 235–267. CiteSeerX 10.1.1.559.6313. doi: 10.1111/j.1468-0262.2006.00655.x.

16. Resnik, R. (2019). *Confounding Variable.* Confounding Variable - an overview | ScienceDirect Topics. Retrieved December 16, 2021, from https://www.sciencedirect.com/topics/nursing-and-health-professions/confounding-variable

17. Rosenbaum, Paul R.; Rubin, Donald B. (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika.* 70 (1): 41–55. doi:10.1093/biomet/70.1.41.

18. Mondal, S. (2020, December 2). *Linear vs logistic regression: Linear and logistic regression.* Analytics Vidhya. Retrieved October 21, 2021, from https://www.analyticsvidhya.com/blog/2020/12/beginners take-how-logistic-regression-is-related-to-linear-regression/.

19. Causevic, S. (2020, August 30). **Frequentist vs. bayesian approaches in Machine Learning. Medium. Retrieved October 21, 2021, from https://towardsdatascience.com/frequentist-vs-bayesian-approaches in-machine-learning-86ece21e820e.

20. Rumsey, D. J. (2021, July 13). *How to determine a P-value when testing a null hypothesis.* dummies. Retrieved October 21, 2021, from https://www.dummies.com/education/math/statistics/how-todetermine-a-p-value-when-testing-a-null-hypothesis/.

21. Song, F.; Parekh, S.; Hooper, L.; Loke, Y. K.; Ryder, J.; Sutton, A. J.; Hing, C.; Kwok, C. S.; Pang, C.; Harvey, I. (2010). "Dissemination and publication of research findings: An updated review of related biases". *Health Technology Assessment.* 14 (8): iii, iix–xi, iix–193. doi:10.3310/hta14080. PMID 20181324.

Appendix

A1: Ethics Statement

Reproducibility Reproducibility is carefully considered in this paper. The source of the data, any literature review that helps to derive the important/interesting variables are properly cited in APA format. The data cleaning process is explicitly described to get a reduced dataset with the interesting variables.

P-hacking In our case of study we get a null result. Instead of changing the data or modifying the variables to create a significant result, we discuss how a null result happens and the appropriate process to refine it. We are avoiding p-hacking in this study.

A2: Materials

Glimpse of data

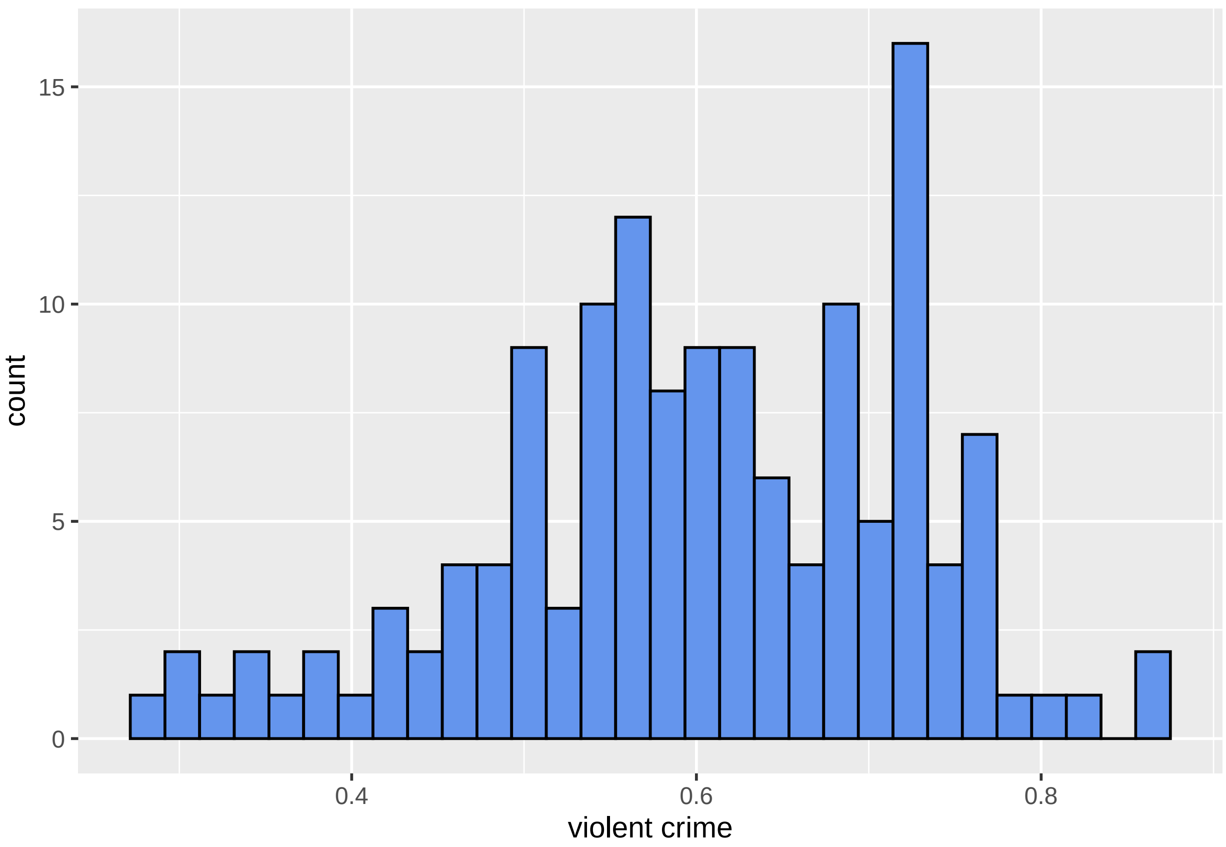
As there are too many variables (104) in the original dataset, I will instead showcase the data after cleaning (with only main variables).

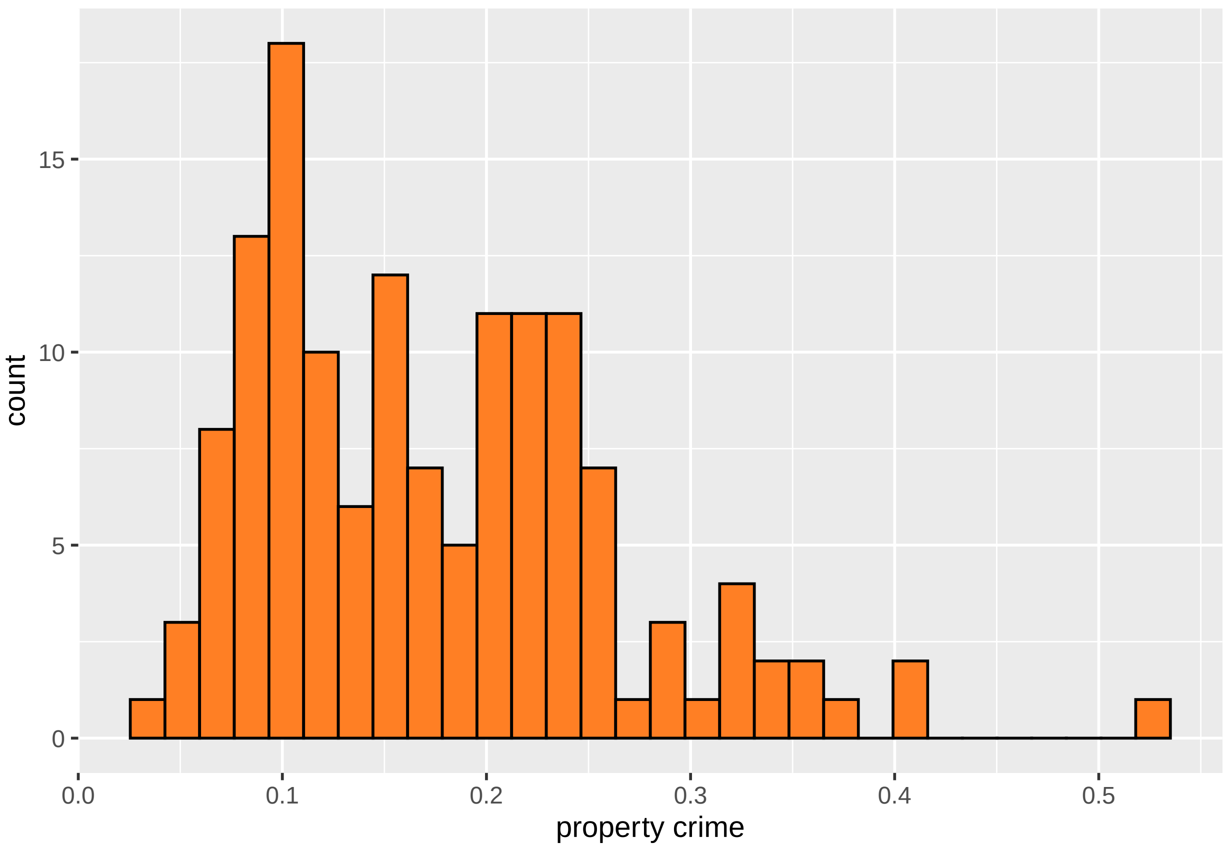
```
## Rows: 140
## Columns: 11
## $ F2020_Population_Projection <int> 14083, 30277, 18146, 17560, 27410, 29970, ~
## $ Assault_Rate2019            <dbl> 253.8071, 1279.8610, 393.5901, 1237.4050, ~
## $ AutoTheft_Rate2019          <dbl> 43.50979, 492.51180, 168.68150, 377.11390,~
## $ BreakAndEnter_Rate2019      <dbl> 203.04570, 371.89670, 219.28590, 530.31640~
## $ Robbery_Rate2019            <dbl> 29.006530, 281.435300, 67.472590, 247.4810~
## $ TheftOver_Rate2019          <dbl> 43.509790, 103.863000, 56.227160, 176.7721~
## $ Homicide_Rate2019           <dbl> 0.000000, 0.000000, 0.000000, 5.892405, 0.~
## $ Shootings_Rate2019          <dbl> 0.000000, 20.102520, 11.245430, 100.170900~
## $ safety                      <int> 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, ~
## $ percent_of_pro              <dbl> 0.15189873, 0.23390276, 0.24539877, 0.2070~
## $ percent_of_violent          <dbl> 0.4936709, 0.6202365, 0.5153374, 0.5947137~
```

Supplementary

Supplementary plots

Graph: Distribution of percentage of violent and property crime.


15
10
5
0
count
0.4
0.6
0.8
violent crime



Supplementary methods

Logistic regression A regression to clarify elements to a set into two groups based on the independent variable (Mondal, 2020).

Linear regression A regression focus on the numerical relationship of a dependent variable and independent variable (Mondal, 2020).

Frequentist and Bayesian model In the Frequentist model we assume the data is sampled from some distribution while in the Bayesian model we assume the parameter is also a variable and follow some distribution(Causevic, 2020). In the question we study, we have a numerical response variable and the parameters are measuring how one unit change of predictors will affect the response variable. The parameters are not variables that we will build a Frequentist linear model in conclusion.