Analysing the Causality Between Neighbourhood Groups and Airbnb Price in New York City Using Propensity Score Regression

Ruolan Zhang

12/22/2020

Abstract

Tourism is one of the most essential components of New York City economy, it portrays the modern metropolis image of the city and boosts the rapid development of local Airbnb business. After conducting propensity score matching, the dataset " New York City Airbnb Open Data" builds a multiple linear regression model for causal inference and estimates that the price of airbnb in boroughs with high popularity are 56 usd more expensive than those in non-popular boroughs. Also, other factors such as room type, availability days in one year, number of reviews and minimum night stays are significantly associated with Airbnb price. Understanding those elements that contribute to the change of price is necessary for both investors and guests, so that it can promote economic growth in the long run.

Keywords

Accomodations, Airbnb, Price, New York City, Neighbourhood Groups, Propensity Score Matching, Multiple Linear Regression, Causal Inference.

Introduction

Airbnb is a popular online short term rental platform that provides gigantic worldwide housing resources for tourists to rent. Compared to hotels, Airbnb accommodations are privately owned by each host. People who have unused spaces at home are able to post the information on the website, and rent out these spaces for travellers. Additionally, it's diverse price levels for one night allow tourists to choose their suitable accommodations according to their demands. Since the establishment of Airbnb in 2009, it offers over 800,000 listings of accommodations in 34,000 cities across approximately 90 countries.

Using the dataset " New York City Airbnb Open Data" from Kaggle, this study mainly analyzes the causal link between locations of airbnb and their prices in New York City, accompanied with some other potential factors. NYC is divided into 5 boroughs, which are Manhattan, Brooklyn, the Bronx, Queens and Staten Island. Airbnb prices differ by their neighbourhood groups. Tourism is strongly tied with the local economy. In 2018, NYC welcomed 65.1 million visitors from all around the world, including both leisure and business travels. In this way, this study provides predictions of airbnb prices, it enables tourists who would like to travel to NYC to make decisions on the borough they want to stay according to their travelling budgets and purpose of visiting. Another point is, from the perspective of landlords, the report provides reference for them to make rational assessment on their accommodations before pricing. Consequently, it benefits the overall Airbnb industry in New York City, promoting tourism development at the same time.

In the methodology section, Propensity score matching(PSM) approach is applied to examine whether there is a causal link between geographic location of Airbnb and it's price. There do exist some other potential variables such as room_type, minimum_nights, number_of_reviews that could distribute to the price. As a consequence, using PSM effectively eliminates interventions of these covariates and mimics the process of randomization to reduce the selection biases. After finishing matching treatment and controlled observations, analysis will start by building a multiple linear regression model to predict the price of airbnb within neighbourhood groups . Results of propensity score analysis will be presented in the Results section with several tables and plots.

Methedology

Data

This dataset, "New York City Airbnb Open Data" shows Airbnb listings and metrics in NYC, NY, USA in 2019, which can be downloaded from Kaggle. Data collection process uses a secondary data collection method. This dataset had already been released on a non-commercial website called "Inside Airbnb". Thanks for Dgomonov to get the acknowledgement of using the dataset from the website. "Inside Airbnb " uses the way of web crawling to gather the raw data that is publicly available on the official Airbnb website, and compiles these information together to form the final dataset. Each accommodation on Airbnb will be automatically recorded. Having a glimpse on the dataset, each observation stands for one private accommodation with its relevant information. In total, there are 48,896 observations with 16 variables listed below (Table 1).

Table 1: Variable Description

Number	Variables	Type	Description
1	id	int	listing id
2	name	character	name of the listing
3	host_id	int	host ID
4	host_name	character	name of the host
5	neighbourhood_group	character	location
6	neighbourhood	character	area
7	latitude	double	latitude coordinates
8	longitude	double	longitude coordinates
9	room_type	character	listing space type
10	price	int	price in dollars
11	minimum_nights	int	amount of nights minimum
12	number_of_reviews	int	total number of reviews
13	last_review	character	latest review date
14	reviews_per_month	double	number of reviews per month
15	calculated_host_listings_count	int	amount of listing per host
16	availability_365	int	number of days when listing is available for booking

Among these 16 variables, ten of them are numerical variables, while the rest six are categorical. The main focus of the study is to examine the causal inference between the location of Airbnb and its price. At the same time, the correlation between other potential factors and the outcome of interest should also be concerned. In the data cleaning process, 7 of 16 variables , "neighbourhood_group", "neighbourhood", "room_type", "price", "minimum_nights","number_of_reviews" and "availability_365", are kept. Specifically, "neighbourhood_group" and "price" are key variables for this study. Removing observations with "Unknown" category from "neighbourhood_group" and observations with 0 in "price" make the whole dataset more operable.

Although variables such as "last_review" and "review_per_month" are similar to "number_of_reviews", both of them have approximately 21% of missing values, which are not qualified as good variables. Therefore, these two variables are not selected. Similarly, "latitude" and "longitude" are not included, they just provide the exact coordinates for each accommodation, which are duplicate. In addition, there is a variable called "neighbourhood", it is kept in the data cleaning process but will not appear in the model section. Under this variable, there exist 221 categories, and each of them represents a specific area. Considering the aim of the study, not only because it is better to use borough as the smallest unit of a region in the model, but also this categorical variable has too many levels. But in the discussion section, it can be used to support further analysis.

By intuition, accommodations with larger private spaces are more expensive. Thus Airbnb price should be closely connected with its room_type, depending on whether it is an entire home, or private room , or a shared room. "minimum_nights" stands for the number of minimum night stays. Most listings allow tenants to rent for at least 1 day, there do exist some hosts that only expect long rent. Therefore, the length of the rental period could affect its price per day. Not all listings are available for the whole year, whether the accommodation has more available days could also be a potential factor for price in common sense. It's worth noting that the first two key variables are similar to each other, and both of them are included. The most significant variable, "neighbourhood_group", has 5 categories, and each stands for one borough in NYC. When conducting propensity score matching, the treatment group and the control group will be classified according to Airbnb's location. Whereas "neighbourhood" subdivides boroughs into decades of small areas based on their communities. Among these 7 variables, "price" is the outcome of interest, and the rest 6 will be used to predict the price. Specifically, whether there exists a causal link between "neighbourhood_group" and "price" will be analyzed by performing a propensity score regression.

Originally, "neighbourhood_group" is a variable with 5 categories. In order to be prepared for propensity score matching, a new boolean variable is created for the treatment group called "popular_borough" which includes Manhattan and Brooklyn. Based on the statistical data, Time Square in Manhattan is listed as the second place of " The world's 50 most visited tourist attractions", which has around 39 million visitors annually, followed by Central Park that ranks the third. Thus there is no doubt that Manhattan attracts most tourists every year and it is a well-deserved metropolitan area in NYC. Brooklyn is famous for its Brooklyn Bridge Park all around the world. As reported on Brooklyn Daily Eagle, there will be millions of travellers visiting there in summer, which the annual count of visitors is remarkable. Consequently, from the perspective of tourists, these two boroughs attract them more compared to other 3 boroughs. For all observations, lists that are located within these 2 boroughs return TRUE under this new variable, and vice versa.

Model

All models are generated using R. This report will use a multiple linear regression to predict how 6 variables affect the airbnb price. In addition to infer a causality between the key variable, "neighbourhood_group", and the outcome of interest, "price", a propensity score matching must be applied to ignore influences of covariates.

Starting the propensity score matching process, popular_borough is the treatment, and price is the outcome of interest. Essentially, propensity score matching mimics the process of randomized controlled trial in an observational study, in case that the treatment group is not randomly assigned. The propensity score matching will be for popular_borough propensity, so first construct a logistic regression on the treatment to examine the log odds that each observation is located in popular boroughs:

Mathematical Equation of Logistic Regression Model:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Interpretation of x_k $(k = 0, 1, ..., 4)$:

x_1 is a numerical variable standing for the number of minimum nights stays.

x_2 is a numerical variable standing for the number of reviews in total.

x_3 is a numerical variable standing for the number of days available in one year.

x_4 is a dummy variable representing 3 categories of room type.

Observing p-values, "number_of_review" is insignificant since it has a p-value equals 0.0525, which is slightly higher than 5%. But it should not be removed from the model. All rest variables are significant. The overall adequacy of the model is good.

Reduce the dataset by removing observations that do not find a similar propensity score which is calculated from logistic regression. Then, use the reduced dataset to perform a multiple linear regression to predict price, the set of independent variables is composed of all predictors in logistic regression and also the binary variable "popular_borough".

Mathematical Equation of Multiple Linear Regression Model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$$

Interpretation of x_k $(k = 0, 1, ..., 5)$:

x_1 is a numerical variable standing for the number of minimum nights stays.

x_2 is a numerical variable standing for the number of reviews in total.

x_3 is a numerical variable standing for the number of days available in one year.

x_4 is a dummy variable containing 3 categories of room types.

x_5 is a dummy variable to indicate whether or not this accommodation is located in popular boroughs.

All variables have p-values smaller than 5%, which are significant. Therefore, the variables for this model are adequate. In this way, the effect of location will be clearly displayed by directly comparing the price of airbnb, and there is no need to consider whether the price level is caused by other covariates instead of "popular_borough".

Results

Result of Data

The subsection Results of Data includes Figure 1.1 and 1.2. Two plots are generated purely based on analyzing the well-cleaned dataset which has 48,884 observations. Main prices of Airbnb in each borough are summarized in Figure 1.1.As figure 1.2 combines two boroughs , "Manhattan" and "Brooklyn", into a large category "popular_borough", the rest 3 are classified as "not popular_borough".

Figure 1.1: Mean Price of Airbnb in Different Neighbourhood Groups

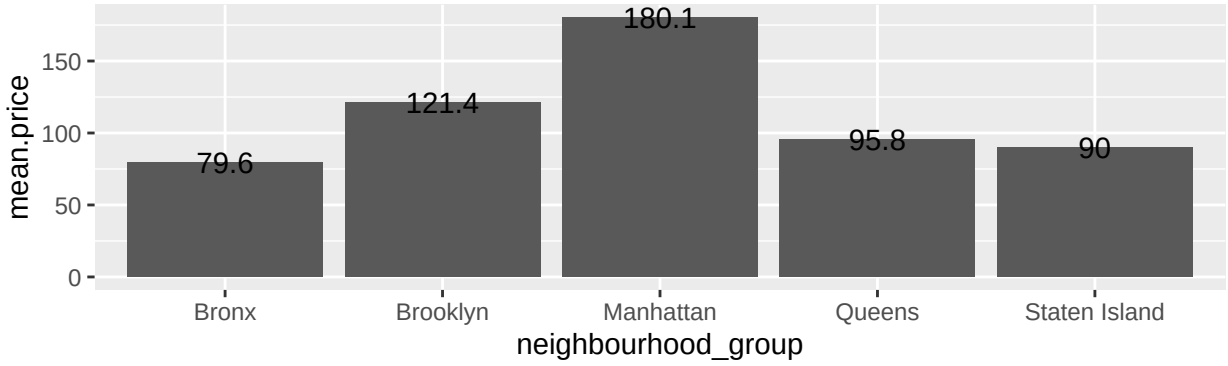
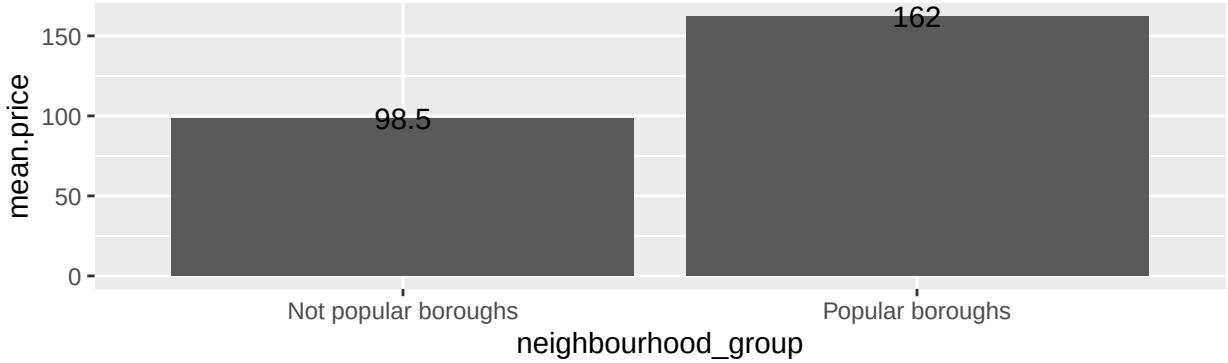


Figure 1.2: Mean Price of Airbnb in non−popular/popular boroughs



Interpretation of Figure 1.1 and 1.2

Figure 1.1 is a bar plot that compares the mean price of Airbnb in five boroughs. Since the number of airbnb in each borough are different, the better way to compare their prices is to calculate their averages. As indicated from barplot 1.1, significantly, Airbnb in Manhattan has the highest mean price among the 5 boroughs, around 180.1 US dollars. Followed by that in Brooklyn, 121.4 dollars. In comparison, it seems that tourists who visit Bronx borough will pay the least money for booking Airbnb, only 79.6 USD on average. In order to better compare the relationship between region and Airbnb mean price, two boroughs that attract the most tourists annually are combined into a category called "popular_borough". All the observations that have the neighbourhood_group of either Manhattan or Brooklyn will be filtered together, counting their mean prices, so does the other category. As can be seen from Figure 1.2, Airbnb's mean price in popular_borough is almost 1.7 times of the price of Airbnb in not popular boroughs. The conclusion can be drawn that popular boroughs do have higher Airbnb prices.The result of data is summarized here, but waited to be confirmed whether there exist causality between the variable and the outcome of interest. Result of the model can help to explain it.

Result of Logistic Regression Model

Interpretation of β_k $(k = 0, 1, ..., 5)$:

β_0 in this logistic model represents the intersection when minimum_night, number_of_reviews, and availability_365 all equals to 0, and categorical variable Room_type is at baseline of Entire Home. The odds of accommodation to be in popular boroughs is equals to exp(2.363)

β_1 is the coefficient of x_1, which means as minimum_nights increases by 1, log odds of being in popular boroughs increases by 0.01836.

Table 2: Logistic Regression Results

Variables	Coefficients	P_value
(intercept)	2.363e+00	< 2e-16 ***
minimum_nights	1.836e-02	< 2e-16 ***
number_of_reviews	-5.294e-04	0.0525
availability_365	-2.632e-03	< 2e-16 ***
room_typePrivate Room	-6.480e-01	< 2e-16 ***
room_typeShared Room	-7.803e-01	< 2e-16 ***

β_2 is the coefficient of x_2, which means as the number_of_reviews increases by 1, log odds of being in popular boroughs decreases by 5.294e-04.

β_3 is the coefficient of x_3, which means as the availability_365 increases by 1, log odds of being in popular boroughs decreases by 2.632e-03.

β_4 is the coefficient of dummy variable x_4, which means as the room_type changes from Entire home to Private room, log odds of being in popular boroughs decreases by 0.6480.

β_5 is the coefficient of dummy variable x_5, which means as the room_type changes from Entire home to Shared room, log odds of being in popular boroughs decreases by 0.7803.

By observing p-values from the result table 2, all variables are significant except "number_of_reviews" that has a relatively large p-value, 0.0525. This variable should still be retained in the logistic model although it is insignificant, it is slightly higher than the significant interval chosen in the model.

Result of Multiple Linear Regression Model

Table 3: Multiple Linear Regression Results

Variables	Coefficients	Standard_Err	P_value
(intercept)	165.97127	4.72979	< 2e-16 ***
minimum_nights	0.30894	0.15266	0.043 *
number_of_reviews	-0.26669	0.03285	5.08e-16 ***
availability_365	0.13701	0.01584	< 2e-16 ***
room_typePrivate Room	-128.85767	4.57700	< 2e-16 ***
room_typeShared Room	-165.61881	8.80826	< 2e-16 ***
popular_borough1	56.44211	4.34164	< 2e-16 ***

Interpretation of β_k $(k = 0, 1, ..., 6)$:

β_0 represents the intersection when minimum_night, number_of_reviews, and availability_365 all equals to 0, and categorical variable Room_type is at baseline of Entire Home.The Airbnb which is located in popular boroughs will have a price around 166.0 USD.

β_1 is the coefficient of x_1, which means when minimum_nights increases by one night, the price of Airbnb will increase by 0.31USD.

β_2 is the coefficient of x_2, which means when number_of_reviews increases by one, the price of Airbnb will decrease by 0.27USD.

β_3 is the coefficient of x_3, which means when availability_365 increases by one, the price of Airbnb will increase by 0.14USD.

β_4 is the coefficient for dummy variable x_4, which means when the baseline for the room_type changes from Entire home to Private Room, the price of Airbnb will decrease by 128.86USD.

β_5 is the coefficient for dummy variable x_5, which means when the baseline for the room_type changes from Entire home to Shared Room, the price of Airbnb will decrease by 165.62USD.

β_6 is the coefficient for dummy variable x_6, which means when the baseline for the room_type changes from not popular_borough to popular_borough, the price of Airbnb will increase by 56.44USD.

From tabel 3, all variables in the multiple linear regression are significant including the treatment variable, "popular_borough". Their p-values are all within 5% significant interval. However, minimum_nights is not that significant compared to others.

The first logistic function ensures the comparability between treatment and control group, which minimizes the selection bias when distributing observations into two groups. Uses the produced odds to calculate the probability that each observation's location is in either Manhattan or Brooklyn. Each two observations that have similar propensity scores will be matched together as a pair. Both observations will have almost the same covariates except for the "popular_borough" variables, which means, after distributing them into treatment and control groups, their independent X variables, "minimum_nights", "availiability_365" and "number_of_reviews" preduce all similar responses. In total, there are 3,7833 observations left after removing all miss values. But the whole dataset is reduced, for observations that cannot find a good match among the dataset, they will be removed. There are 1,1460 treated observations, so it expects 2,2920 observations left.

Using the reduced dataset to conduct another multiple linear regression on the price of Airbnb efficiently reduces the influence of other covariates. In this way, the effect of neighbourhood_group will be clearly displayed by directly comparing the price of airbnb, and there is no need to consider whether the change of price level is caused by other covariates instead of "popular_borough". Result indicates that there exists a causal link between the neighbourhood_group and the price of Airbnb.

Discussion

Summary

The general idea of previous sections is to prove the existence of causality between the neighbourhood groups and these boroughs' Airbnb prices in New York City. By mimicking the process of observational study, propensity score matching is applied to reduce the effect of other potential confounding variables associated with price. As expected, building multiple linear regression models after a series of data processing, the result indicated that the causal link truly existed.

Conclusion

" New York City Airbnb Open Data" is a high-quality dataset from Kaggle which updates once a year. Using data by 2019, the report tells a story about the relationship between potential factors and Airbnb price. In common sense, the price of accommodation is highly associated with local popularity, whereas causality is not that easy to build by intuition, unless the randomized controlled trial is performed. The remarkable result will be closely discussed is whether Airbnb is located in a popular borough directly causing the difference in Airbnb price.

Propensity score regression estimates that the average price of Airbnb in popular boroughs such as Manhattan or Brooklyn, will be approximately 56.44 US dollars more expensive than those in other boroughs per day, when keeping all other factors the same. Especially the average price of Airbnb in Manhattan is incredibly higher than other boroughs. For new visitors who are aimed to relax themselves, they would prefer either visiting popular attractions such as Central Park and Empire State Building, or experiencing a new lifestyle in metropolis and enjoying the local entertainments. It is better to live in a convenient place that everywhere they want to visit is easy to reach. Therefore Airbnb in those places will definitely have higher prices. Another

reason is, originally the price of land in densely inhabited districts will be more expensive than those in the countryside, which is calculated in hosts' sunk cost.

Neighbourhood groups in causal relation is truly the key component among all variables. However, some other independent variables that are associated with price in multiple linear regression are non negligible as well.

Accommodations that have more available days tend to have higher prices. These accommodations usually have shorter time lags between different bookings, so their hosts need to pay for cleaners to clean the room more often, which potentially increases the daily cost. This will be reflected on the increase of Airbnb price. In addition, travellers will have more options to choose their desired dates.

Besides, Airbnb with different room types vary on their price differently. By test statistics, the entire home is the most welcomed room type among those three, also the most expensive one. Keeping all other factors the same, a private room is incredibly 129 dollars cheaper than the entire home, and a shared room, which is the least people's choice, is about 166 dollars cheaper than the entire one. A reasonable explanation is the entire home accommodates more people than other two types. For family travels with more than 2 people, an entire room is their best choice. Additionally, humans prefer to own their own space instead of sharing with others. Therefore, accommodations with larger private spaces will cost more money.

Another factor that is associated with price is total number of reviews. Tourists from all over the world may not be able to afford the sky-high hotel price, Airbnb is a good choice for them. So people who choose Airbnb in NYC are mostly limited in their travelling budgets, and they would prefer cheaper accommodation. Consequently, Airbnb with more reviews are likely to have favorable prices, which is less expensive.

Applying all these variables, the prediction for Airbnb price would give a general guidance for visitors from both domestic places or international countries about how the price levels of Airbnb would differ by variables, especially the location, in New York City. Helping to choose suitable accommodation to live based on their expected budget is one of the benefits. From the perspective of Airbnb hosts, before posting their room information on the official website, this study works as a reference to help them evaluate their room values, and calculate a proper price in the market.

Weakness and Next Step

The multiple linear regression model is composed of 5 variables, which can only represent parts of the big world. All variables that are associated with Airbnb prices in NYC are uncountable, and there must be something else important that this study doesn't include. One of the concerns in data processing is, "popular" can never be objectively defined. This study defines boroughs with significantly more tourist amount as popular boroughs, which there exists bias when converting categorical variables with multi-levels into binary variables. Considering the data collecting process, observations are collected by secondary data collection method instead of directly downloading from the official website. There may exist small errors when web crawling.

In the future, other difficult models can be applied instead of the linear one, the outcome of interest should not result in a certain value, instead, it can be considered to follow a specific distribution. Another point is, in order to make the study with strong timeliness, the result should be re-analyzed once a year after the update of the dataset. Last but not the least, Airbnb business is now competing with the hotel industry in the local area, but the study does not consider any competitive effect from the hotel industry. It is better to collect data from both industries, build separate models and compare them together.

Reference

"The Boroughs." Encyclopædia Britannica, Encyclopædia Britannica, Inc., www.britannica.com/place/New-York-City/The-boroughs.

Dgomonov. "New York City Airbnb Open Data." Kaggle, 12 Aug. 2019, www.kaggle.com/dgomonov/new-york-city-airbnb-open-data.

Bram, Jason. Tourism and New York City's Economy . 1995, www.newyorkfed.org/medialibrary/media/research/current_iss 7.pdf.

"Inside Airbnb. Adding Data to the Debate." Inside Airbnb, insideairbnb.com/about.html.

"Data Collection Methods - Research-Methodology." Research, research-methodology.net/research-methods/data-collection/.

"The World's 50 Most Visited Tourist Attractions." Love Home Swap, www.lovehomeswap.com/blog/latest-news/the-50-most-visited-tourist-attractions-in-the-world.

"Travel and Tourism Trend Report." Indd.adobe.com, indd.adobe.com/view/e91e777a-c68b-4db1-a609-58664a52cffd.

Travel and Tourism Trends, NYC & Company, 2019, indd.adobe.com/view/fcc4cd9f-7386-4b52-a39b-c401266a137f.

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

T. Lumley (2020) "survey: analysis of complex survey samples". R package version 4.0.

Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.28.

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

Andrew Gelman and Yu-Sung Su (2020). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.11-2. https://CRAN.R-project.org/package=arm

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra

David Robinson and Alex Hayes (2020). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.6. https://CRAN.R-project.org/package=broom

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham and Evan Miller (2020). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files. R package version 2.3.1. https://CRAN.R-project.org/package=haven

Appendix

Code and data supporting this anaylysis is available at: https://github.com/zhangruolanlan/STA304-final-project