TITLE*

SUBTITLE

Adrian Wong, Yingying Zhou, Xinyi Xu, Yang Wu

26 February 2021

Abstract

ABSTRACT

Contents

*Code and data are available at: https://github.com/yangg1224/groupproject-.git

1 Introduction

2 Data

In this report, we use the R statistical programming language (R Core Team 2020). To be specific, we use the "tidyverse" package to process data (Wickham et al. 2019), "kableExtra" package to generate tables (Zhu 2020), ADD ANY PACKAGE HERE. The survey conducted by Petit Poll collects the following data:

- Geographical information (FSA, Region)
- Type of restaurant (Fast food, fast casual, casual dining, premium casual, family style, fine dining)
- Number of years opened so far
- Services provided (delivery, take-out)
- Number of employees and salary of employee
- Cost to run the restaurant (Fixed, variable)
- Revenue per month
- Potential COVID-19 cases

Appendix C shows screen captures of the online version of the survey.In addition, we consulted several open datasets about the health inspection from each region. (CITATION)

The data section illustrates the intervention and data gathering methodology. Also, we describe detailed information on our dataset, with several data visualizations such as tables and graphs.

2.1 Intervention

In this experiment, we conducted a randomized controlled trial and randomly assigned subjects to two groups: the treatment group and the control group. Within this experiment, the treatment group receives an intervention and the control group is not being treated.

In our case, the intervention is having restaurants open for indoor dining. Our experimental subjects are restaurants in the GTA that are currently in operation. We assumed all restaurants in Ontario are closed for indoor dining at the moment. During the shut-down, all restaurants in the control group were forbidden to offer dine-in and patio services according to the COVID-19 lockdown policy set out by the Ontario government ("Salle de Presse de l'ontario," n.d.). However, take-out and delivery services were acceptable. Our intervention involves randomly selecting restaurants to reopen for indoor dining in the GTA. The restaurants in the treatment and control group were picked through stratified sampling. This experiment was conducted by Petit Poll and was authorized by the Ontario Department of Public Health.

We randomized the control and treatment groups through stratified sampling. In other words, we divided all restaurants in Ontario into smaller strata. Each strata was grouped based on region. After stratification, we took random samples from each region, in proportion to its share of the population within the total population (Nickolas 2020). The distribution of all strata shows as following (see more information in Appendix A):

- Toronto - 29.5%
- Durham - 12.8%
- York - 21.9%
- Peel - 24.6%
- Halton - 11%

We then use a random number generator to randomly select a certain number of restaurants from a list of restaurants that was ordered alphanumerically. This list was procured by health inspection records that

indicate all of the open restaurants within each region. Procedurally, we set the minimum number and the maximum number based on the number of restaurants within the given list. If the randomly generated number was 59, we chose the 59th restaurant in the list. We repeated the process until the treatment group and the control group were established with the corresponding number of restaurants.

To ensure the separation of treatment and non-treatment groups, we relied on stratified random sampling based on regions. After stratification, the number of restaurants in the treatment and control groups were the same. This ensures that the treatment is the only source of potential differences in outcomes between the two groups and not based on convenience of access for citizens within each region (Nickolas 2020). Moreover, ad-hoc analysis showed that the proportion of restaurant types were the same between control and treatment groups. This reduces the likelihood that citizens would migrate between groups for particular restaurant types.

The experiment ran for two and a half months from December 1, 2020 to February 15th, 2021. We distributed a consent form to 11,325 restaurants within the GTA on December 1st, 2020, and received 3,454 responses to participate in our survey by December 15th, 2020. We randomly sampled an equal number of respondents into the treatment group and control group. On December 16th, the intervention was announced for the selected 1,637 restaurants in the treatment group. We reserved half a month for the restaurant to prepare for reopening. The treatment group reopened the restaurant from January 1st to January 31st, 2021. We considered one month as the minimum effective period for a validity reopening. During the intervention, all restaurants in the treatment group were allowed to offer dine-in, patio services, delivery service, and takeout services. The survey itself was conducted for a half month from February 1st, 2021 to February 15th, 2021. Finally, we collected 3,274 responses in total from both the treatment group and control group by February 15th, 2021.

2.2 Data Gathering Method

The population included all restaurants in the GTA, excluding the ones that were completely closed. The sampling frame was all restaurants listed in the health inspection reporting program of all five regions. Currently, GTA has approximately 25,351 restaurants, including 7500 in Toronto, 3260 in Durham, 5553 in York, 6235 in Peel, and 2803 in Halton. We use a stratified sampling method to randomly select restaurants in each region. We sent consent forms to 11,325 restaurants, and received 3,454 responses, which reflects a 30.5% response rate. The final sample was 3,274 restaurants that responded to the paper survey. We arrived at 3000 as a sample size to ensure enough statistical power in our sample.

We use stratified random sampling to obtain a sample that best represents the entire population. Unlike simple random sampling, which randomly selects data from the entire population, stratified random sampling takes each stratum in direct proportion to the population in each region compared to the total population in GTA. Stratified random sampling reflects the population more accurately than simple random sampling. With simple random sampling, it is not guaranteed that any particular subgroup or type of restaurant is chosen (Murphy 2020). In contrast, SRS ensures each subgroup within the population has proper representation within the sample. In other words, it captures key population characteristics in the sample. Stratification gives a smaller error in estimation and greater precision than the simple random sampling method. The greater the differences between the strata, the greater the gain in precision(Hayes, n.d.).

The collection instrument for this survey was paper questionnaires sent by mail. We first sent consent forms to obtain permission within the early stages of the experiment to randomly selected restaurants in a list of all restaurants per region. We did not use electronic questionnaires or surveys sent by email in order to prevent a non-response bias from the few restaurants that do not have a commercial email address or who don't check their email frequently. Additionally, we avoided choosing in-person interviews in order to reduce close contact and lessen the chance of virus transmission.

The total estimated cost was approximately $20,958. To be specific, the average cost to print a page for black and white is around 5 cents ("Printing Costs: How to Accurately Calculate Your Printing Cost Per Page," n.d.). According to Stamps, the cost of a one-ounce First Class Mail stamp is $0.55 for one way ("How Much Is a Stamp?" n.d.). Each envelope cost 15 cents ("USPS® rate Change Effective January 26,

2020," n.d.). We sent 11,000 informed consents first, and then sent 3454 paper surveys after. When we mail the consent form and survey, we include a prepaid envelope with a stamp in each single package to encourage survey response. Therefore, we spent $15,950 on consent forms and $5,008 on paper surveys. See more details in Appendix B.

We took three steps to deal with the non-responses. First of all, we tested the survey before sending them out. We kept the survey short. A one-minute survey normally has a higher response rate than a 15 minutes survey (Stephanie 2020). Besides, we sent reminders to the participants through mail or phone, if we did not receive responses in the first week. It ensured the surveys were sent to the right address. In addition, we would offer incentives in exchange for completing the survey.

During the early stage of designing the survey, we determined that it was unnecessary to collect the personal information of the restaurant owner. Also, the survey did not ask the restaurant's name and address but only the first three digits of postal code (FSA code), which encodes region-level information. The survey was conducted anonymously; therefore, the risk of unauthorized collection, use, and disclosure of personal information was kept to a minimum.

We did not use an electronic questionnaire sent by email. An email survey will require participants to send their responses back to us by personal email. Meanwhile, physical questionnaires can be sent back anonymously, without the sender's name and return address.

In the questionnaire, most questions are multiple-choice, and we were not asking any open-end questions about the personal experience. In particular, when it comes to the cost structure of restaurant operations, which is deemed as sensitive questions by business owners, only the direction of change was asked instead of a specific amount. Thus, it would be impossible to reverse engineer the financial details and to infer private attributes about restaurants such as store names and locations using this dataset.

When conducting the survey, we strictly follow the terms of reference for the safe collection, retention, use, disclosure, and disposal of personal information, in accordance with the Act. During the survey, we reviewed the term of reference periodically. At the final stage of the experiment, we checked all conditions and made sure they were fully complied with.

Participation in the survey was considered on a voluntary basis. Thus, we sent out the informed consent in order to provide participants as much information as possible. The consent should include the following:

- The main purpose of the research.
- The name of the institution that conducts the research.
- How the information will be disclosed.
- How much time the survey will take.
- How the respondent will be informed about the final result.

We have also considered how the data will be stored for future use. All data will be replaced with code and stored separately from the survey response. Petit Poll will be responsible for ensuring the confidentiality, integrity, and accessibility of the dataset under supervision of the Ontario government.

2.3 Descriptive Analysis

After discussing data gathering method, we sampled data in R (R Core Team 2020). We totally have **3274** observations, and 14 of following features according to the questionnaires.

- `type` : Categorical identifier ["Treated" or "Control"] for each observation
- `Q1` : First three digits of the postcode
- `Q2` : Categorical identifier for distinguishing the type of restaurants
- `Q3` : Region name in GTA
- `Q4` : Describe whether the restaurant is a franchise ("Franchise" or "No")

Table 1: First 6 rows Raw data

type	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13
Control	M5W	Toronto	Family Style	Franchise	11	Yes	No	1-10	20.11	No	No change	No change	44140
Control	L7C	Peel	Fine Dining	No	10	Yes	No	1-10	23.31	No	No change	No change	42217
Control	L7C	Peel	Family Style	No	2	Yes	Yes	1-10	16.74	No	No change	Decrease	37507
Control	L6A	York	Fast Casual	No	2	Yes	Yes	1-10	19.21	No	No change	No change	41194
Control	L6H	Halton	Premium Casual	No	1	Yes	No	1-10	15.22	No	No change	No change	56615
Control	L4Z	Peel	Fast Food	No	3	Yes	No	1-10	15.60	No	No change	No change	51303

- `Q5` : The length of the operation years for each restaurant
- `Q6` : Describe whether the restaurant offer takeout service ("Yes" or "No")
- `Q7` : Describe whether the restaurant offer delivery service ("Yes" or "No")
- `Q8` : Number of employees in the restaurant (category type)
- `Q9` : Average employee hourly rate (CAD)
- `Q10` : Describe whether the restaurant has been a site of a potential COVID case ("Yes" or "No")
- `Q11` : Describe the restaurant's fixed costs change situation
- `Q12` : Describe the restaurant's flexible costs change situation
- `Q13` : The restaurant's past month revenue (CAD)

The first six rows of raw data is shown in the Table1. (Table 1)

2.3.1 EDA

Taking a deep look at all the features from survey questionnaire, we learned some demographic features about the restaurants in GTA:

- From figure1(Figure 1) and figure2(Figure 2), we noticed that more restaurants are located in Toronto (around 500) and Peel (around 400). The number of restaurants in Hilton is similar to the number in Durham. Meanwhile, Casual dining takes the lead in the restaurant type in GTA, with around 26%. Then it comes to Family style restaurant, accounting for 20%. Fast food restaurant, fine dining restaurant and Premium casual restaurant almost equally make up 10%. There is no big difference between treated group and control group in terms of restaurant number and type distributions.

- In terms of employees salary, the average hourly rate before and after intervention is both around 17 CAD. In contrast with two boxplots, we can see there is a slight increase in the treated group. The reason behind might because the employee take more risks to go for work, accordingly they will receive higher salary. (Figure 3)

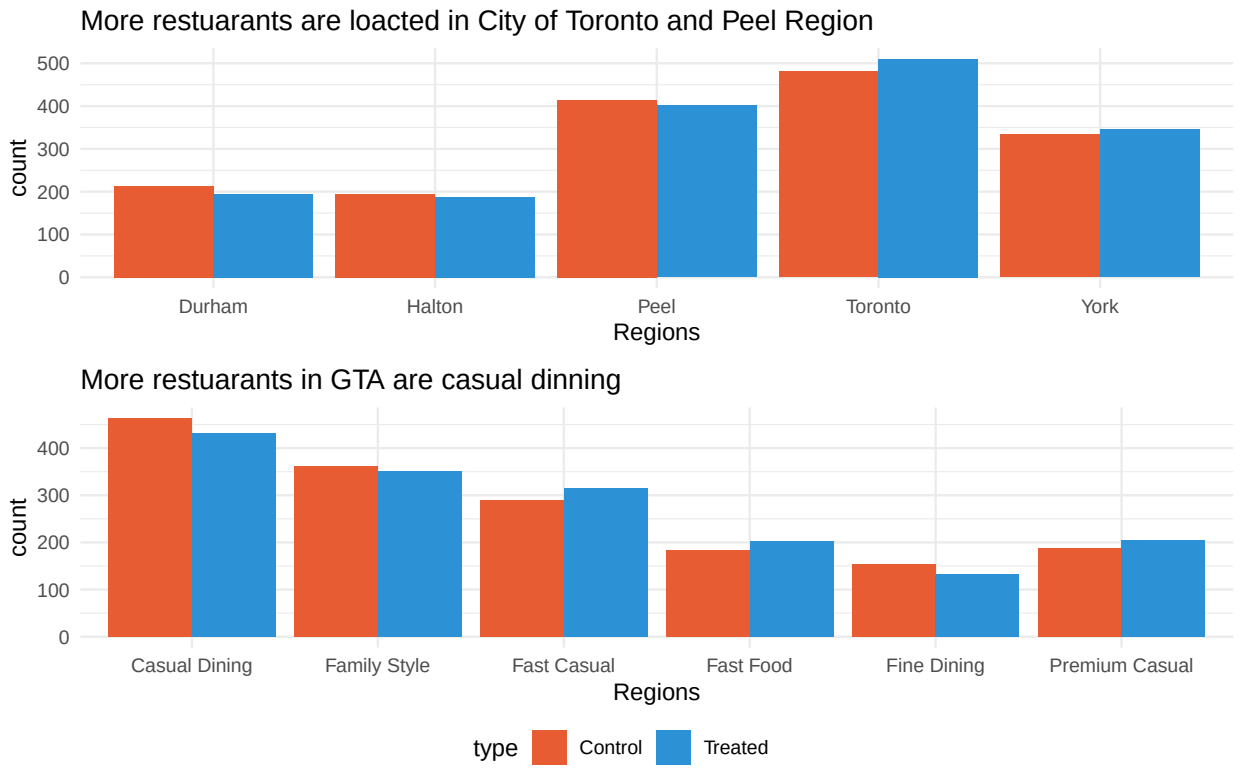


Figure 1: Restaurant numbers and types

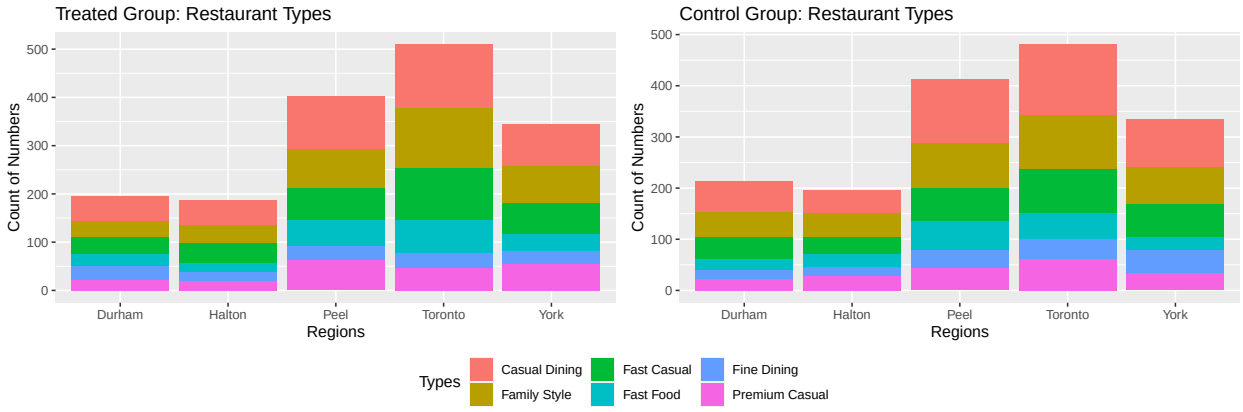


Figure 2: Restaurant types

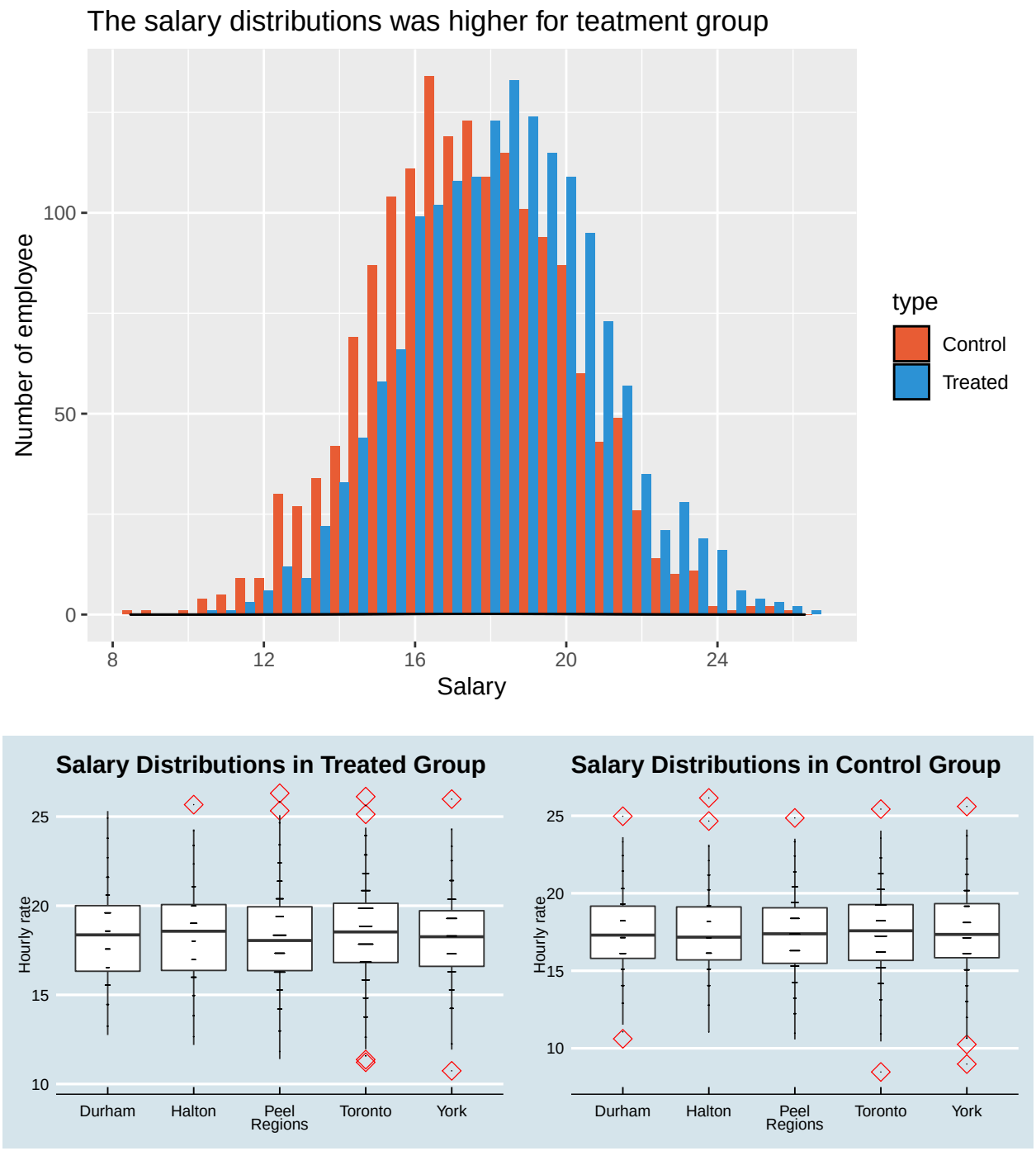
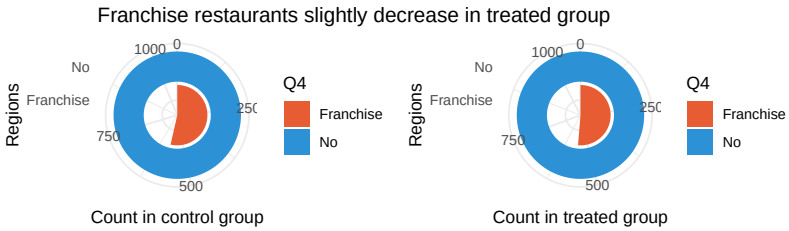


Figure 3: Employee salary distribution

- The attributes of the restaurant determine its management mode, so whether the restaurant is a franchise is quite important factor. From the pie charts, we found the portion of franchise rate in treated group is slightly lower than control group. We assume in treated group, the branch franchise restaurants should follow the rules by the head office. Considering the potential cost of COVID issues, chain restaurants will face greater risks, which is why they are less likely to be in the treated group. (Figure **??**)



* The polar chart illustrates the em-

Table 2: T Test on the Restaurant's revenue

mean_of_Treated	mean_of_Control	p.value	conf.low	conf.high	method	alternative
63143.44	49358.97	0	11638.18	15930.76	Welch Two Sample t-test	two.sided

ployee numbers distribution, as can be seen in figure below. (Figure 4) Because of COVID rule, no restaurant is allowed to open for large group dine in. So in the control group, there is 0 restaurant which has more than 30 employees. Most of restaurant has 10 to 20 employees.

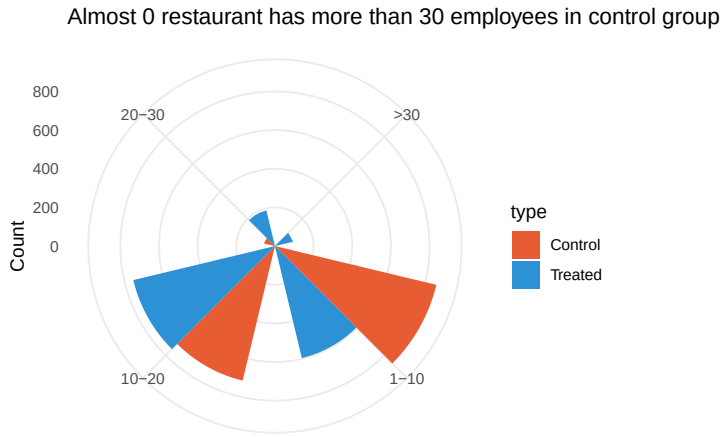


Figure 4: Employee numbers distribution

2.3.2 T-Test

The T Test is used to compare the sample mean of our Treated group and Control group. The goal is to determine whether the intervention has an effective effect on the treated group. Our hypothesis is the intervention will have positive impact towards the restaurant's revenue. (Kim 2015) The T test results is represented in the Table2(Table 2). The package **Broom**(Robinson, Hayes, and Couch 2021) is used to clean the t test results and convert it into the dataframe. The p value we get is $<$ 2.2e-16, as the p value would indicate a significant result, meaning that the actual p value is even smaller than 2.2e-16 (a typical threshold is 0.05, anything smaller counts as statistically significant).(Kim 2015) So we can interpret hypothesis not rejected which means the intervention has a significant effect on treated group.

2.3.3 Correlation matrix

Correlation matrix shows internal relationships between x variables and y variable. (Figure 5) Intensity is indicated by the color(from red to blue). No significant coeffiency is barred with symbol "x." More detailed analysis will be conducted in finding part.

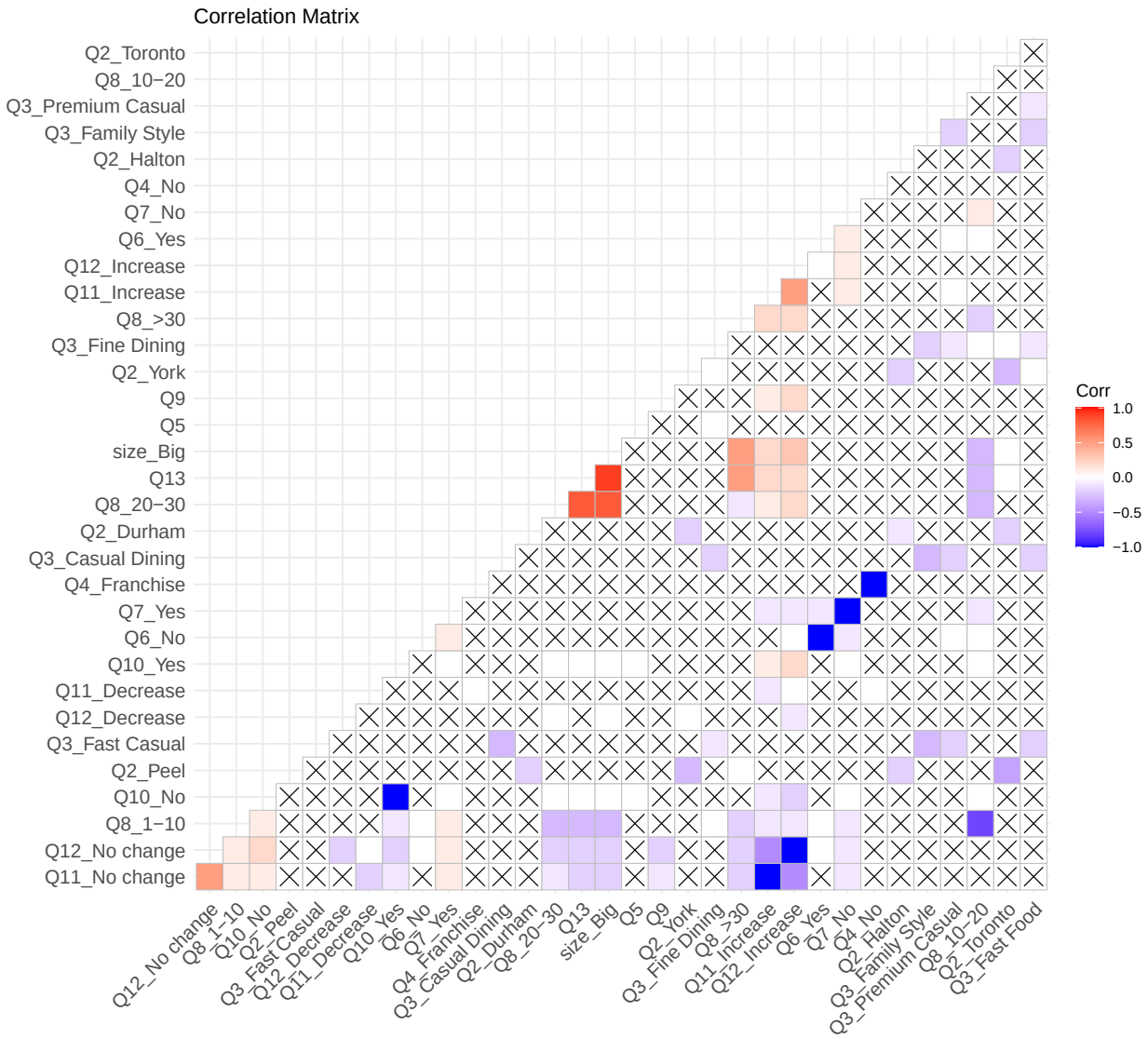


Figure 5: Correlation matrix

3 Discussion

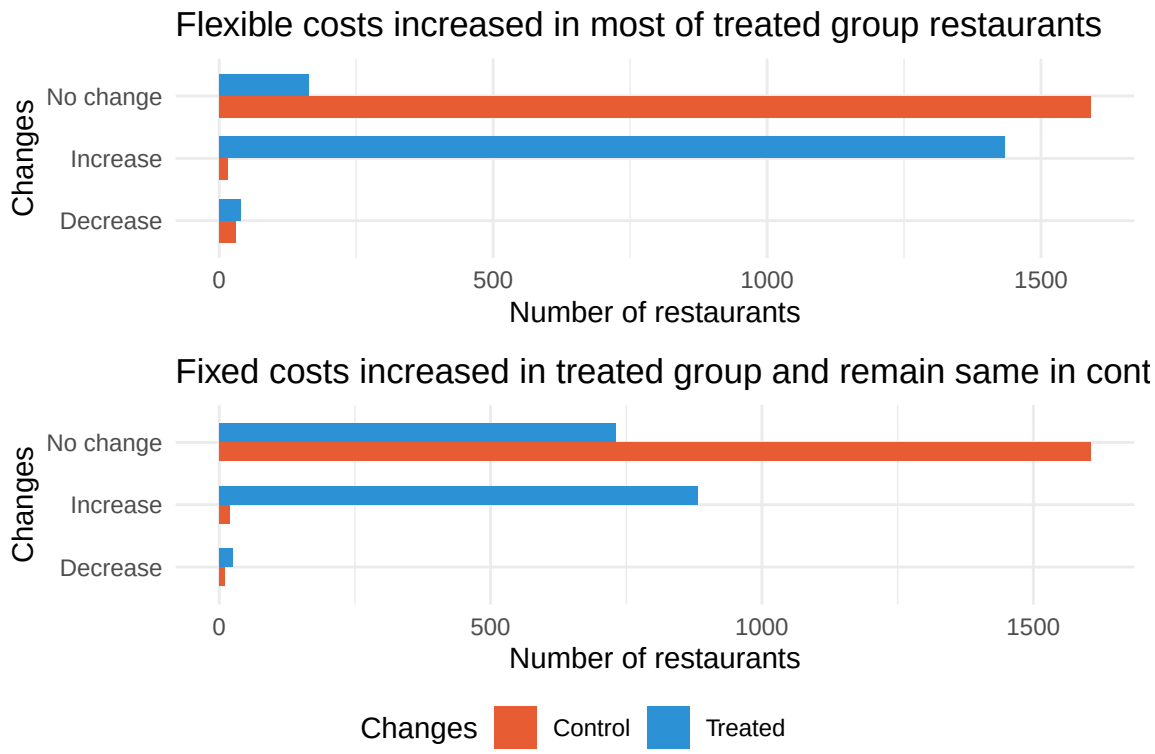
3.1 Overview

3.2 Findings

3.2.1 Finding ONE

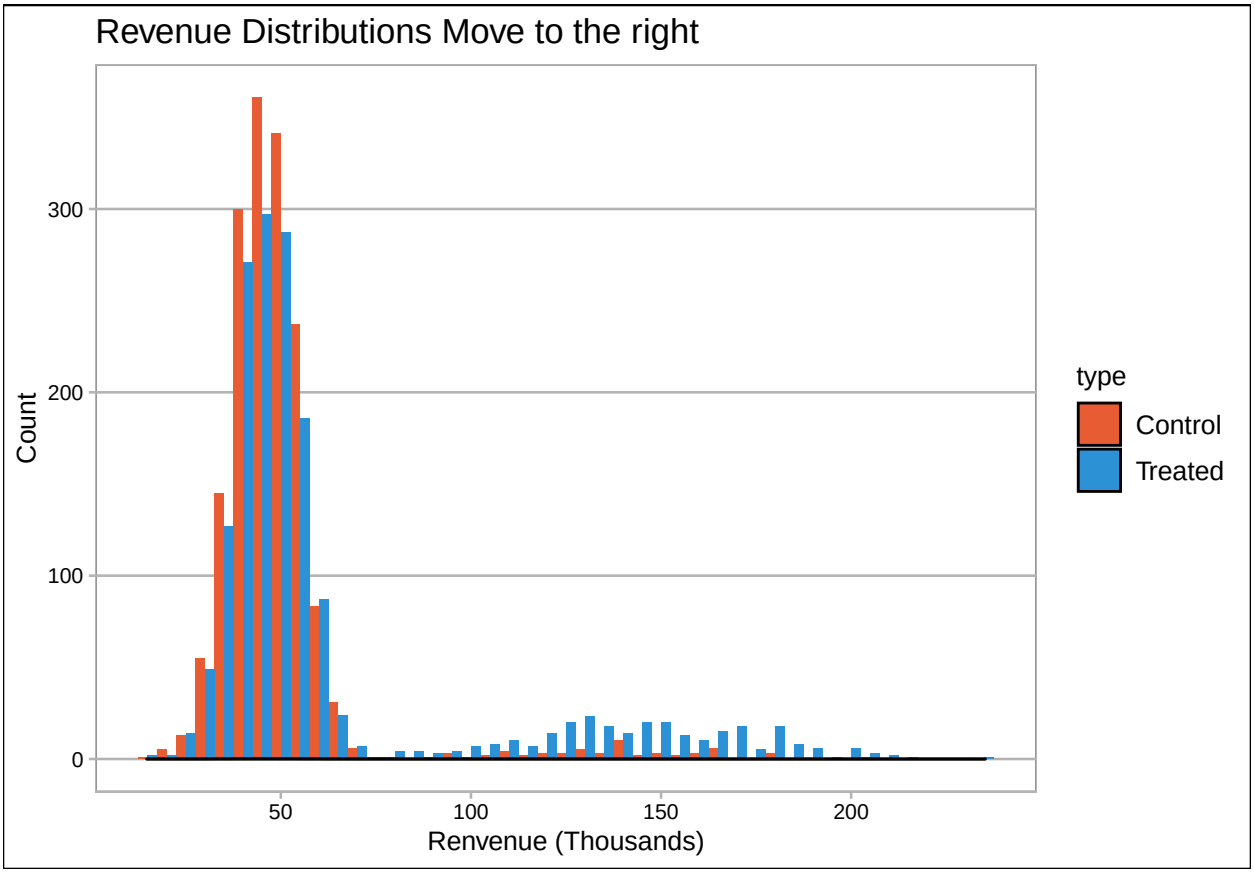
3.2.2 Finding TWO

Invention effect on Flex and fiexed cost



###

Finding THREE Invention effect on Revenue distributions



Limitation

3.3 Future Directions

Table 3: Detailed information for stratification

Region	Number of Restuarants	Proportion(%)	Sample Selected
Toronto	7500	29.58	48430
Durham	3260	12.86	21051
York	5553	21.90	35858
Peel	6235	24.59	40262
Halton	2803	11.06	18100
Total	25351	100.00	1637

Table 4: Estimated Cost

Components	Cost per unit	Total cost for each component
Printing Cost	0.05	738.95
Envelope Cost	0.15	4433.70
Stamp Cost	0.55	16256.90

4 Appendix

4.1 Appendix A

4.2 Appendix B

4.3 Appendix C: Screenshot of the survey

References

Hayes, Adam. n.d. "Reading into Stratified Random Sampling." https://www.investopedia.com/terms/stratified_random_sampling.asp#:~:text=Advantages%20of%20Stratified%20Random%20Sampling,proportional%20to%20the%20overall%20population.

"How Much Is a Stamp?" n.d. *Stamps.com - How Much Is a Stamp?* https://www.stamps.com/usps/how-much-is-a-stamp/.

Kim, Tae Kyun. 2015. "T Test as a Parametric Statistic." *Korean Journal of Anesthesiology* 68 (6): 540.

Murphy, Chris B. 2020. "Pros and Cons of Stratified Random Sampling." *Investopedia*. Investopedia. https://www.investopedia.com/ask/answers/041615/what-are-advantages-and-disadvantages-stratified-random-sampling.asp.

Nickolas, Steven. 2020. "How Stratified Random Sampling Works." *Investopedia*. Investopedia. https://www.investopedia.com/ask/answers/032615/what-are-some-examples-stratified-random-sampling.asp.

"Printing Costs: How to Accurately Calculate Your Printing Cost Per Page." n.d. https://www.tonerbuzz.com/blog/printing-costs-how-to-accurately-calculate-your-printing-cost-per-page/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

"Salle de Presse de l'ontario." n.d. *Ontario Newsroom*. https://news.ontario.ca/en/release/58790/ontario-continues-to-support-restaurants-during-covid-19-pandemic.

Stephanie. 2020. "Non Response Bias: Definition, Examples." *Statistics How To*. https://www.statisticshowto.com/non-response-bias/.

"USPS® rate Change Effective January 26, 2020." n.d. *Home*. https://www.fp-usa.com/usps-rate-change-effective-january-26-2020.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Zhu, Hao. 2020. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.