PUBLIC HEALTH 252D FINAL REPORT

June 21, 2019

Asem Berkalieva
Philippe Boileau
Edie Espejo
Naomi Wilcox
University of California, Berkeley
Division of Biostatistics and Epidemiology

Contents

1 Specify the Scientific Question

Over the last two decades, foodie culture has taken the western world by storm. Cooking shows, once only considered suitable programming for daytime television, are now watched by millions and their stars, celebrity chefs, are venerated by the masses. From Gordon Ramsay's fiery reality shows to the late Anthony Bourdain's culinary expeditions, food culture has a niche for everyone. Americans now flock to the latest and greatest eateries to taste new experiences, whether they be in the concrete jungle of New York City or the rural expanses of the Midwest. Although the new-found importance of food has benefited society in many ways, such as increased conscientiousness of farming practices, it has done nothing but increase the competitiveness of the already cutthroat restaurant industry. In fact, median life-span of an American restaurant is only 4.5 years, a statistic which decreases to 3.75 years for smaller restaurants with five employees or less [1].

Given the vast number of restaurants and the relatively limited amount of food that one can consume in a day, customers have turned to the internet to help choose where they will eat their next meal. Websites such as Yelp have been created to quickly and easily convey relevant information about eateries: contact information, location, price, type of cuisine, as well as a five star review system that summarizes the experiences of other diners. As one would expect, restaurants with the highest reviews tend to be the most successful. Or are they?

The goal of this research project is to assess the importance of favorable Yelp reviews on restaurant closure. More precisely, we wish to see whether having a rating above or equal to 3.5 stars in 2017 affects the probability of restaurant closure by 2019 when compared to having a rating below 3.5 stars, controlling for age, type, number of reviews, and whether the restaurant is part of a chain. To answer this question, Yelp review data from the Las Vegas metropolitan area was collected. More information on this data is provided in section 4.

2 Specify a Causal Model

We represent our knowledge of the data generating distribution using the following structural causal model (SCM):

$$X = \{W, A, Y\}$$

where W represents age, type, number of reviews and chain status, A represents whether the restaurant received a rating above or equal to 3.5 stars in 2017, and Y represents the closure status of each restaurant in 2019. The random input from the deterministic system is encapsulated in U, which is defined as:

$$U = \{U_W, U_A, U_Y\}$$

The structural equations of this model are as follows:

$$\begin{aligned}
W_{age} &= f_{age}(W_{type}, W_{chain}, U_{W_{age}}) \\
W_{type} &= f_{type}(W_{chain}, U_{W_{type}}) \\
W_{reviews} &= f_{reviews}(W_{age}, W_{type}, W_{chain}, U_{W_{reviews}}) \\
W_{chain} &= f_{chain}(U_{W_{chain}}) \\
A &= f_A(W, U_A) \\
Y &= f_Y(W, A, U_Y)
\end{aligned}$$

As we can see from the SCM, a number of exclusion restrictions are made among the covariates. First, we assume that a restaurant's chain status influences its type, age and number of reviews since, in many cases, the corporate branch of the restaurant will select the menu of the restaurant, and thus it's type. Additionally, the success and popularity of a chain are likely to influence the number of reviews that a restaurant receives; more popular chains are likely to attract crowds. We also believe that chain status influences age, since older, successful chains are likely to have older franchises.

Next, we assume that the type of food a restaurant serves will influence its age and number of reviews. This assumption is grounded in the belief that the popularity of certain cuisines fluctuate over time, influencing the age of the restaurant and the number of customers, and hence the number of reviews.

Finally, we assume that the age of a restaurant will influence the number of reviews that a restaurant receives. We expect that, all other variables being equal, older restaurants will have served more customers than their younger counterparts, and therefore will have received more Yelp reviews.

No independence assumptions are made. The economic and social phenomenon that govern the success and failures of restaurants are practically impossible to disentangle, especially in an simple model such as ours. For this reason, we assume there to be a dependence between the exogenous variables of our model. Figure 1 illustrates directed acyclic graph defined by the SCM and it's assumptions.

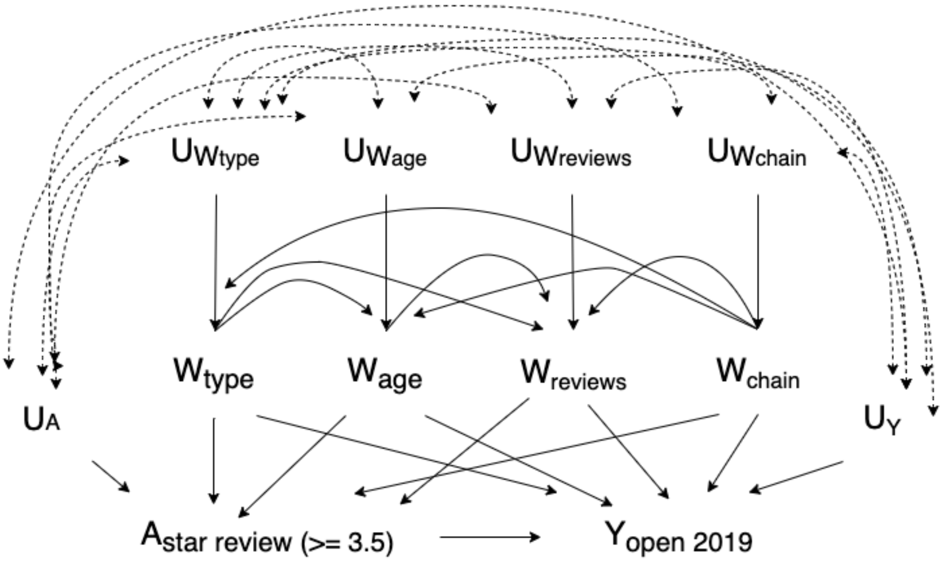

Figure 1: The directed acyclic graph representing the SCM.

3 Translation of Scientific Question into Causal Question

The causal question assesses the effect of average Yelp review on two-year survival in Las Vegas restaurants. Using a threshold of 3.5 stars as the intervention variable, the counterfactual outcomes of interest are defined as:

Y_1: Restaurant survival at year 2 having received an average Yelp rating above or equal 3.5 stars

Y_0: Restaurant survival at year 2 having received an average Yelp rating below 3.5 stars

Thus, the target causal parameter is the difference in counterfactual probability of 2 year survival had all restaurants received an average Yelp rating above or equal to 3.5

stars and the counterfactual probability of 2 year survival had all restaurants received an average Yelp rating below 3.5 stars:

$$\Psi^F = E_{U,X}[Y_1] - E_{U,X}[Y_0] = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1)$$

where Y_a denotes the counterfactual outcome (survival) if the restaurant had average rating $A = a$.

4 Data Preparation

Our causal dataset includes business covariates/attributes from the 2017 Yelp Challenge (YC) datasets and from the 2019 YC dataset. The 2019 Yelp dataset downloaded via the current YC webpage contained 192,609 records, and its 2017 counterpart downloaded via a web archive contained 174,567 records. These data were originally in JSON format and converted into CSV. After filtering the 2017 dataset to only include Yelp businesses that were in Las Vegas, NV, the two datasets were joined by Yelp business ID to create a subset of only Las Vegas, NV businesses.

Our dataset was further cleaned to target our population of interest, Las Vegas restaurants. We initially chose to keep the business in our dataset that matched either "Restaurant" and "Food" in their business categories. However, after visual inspection, many businesses were not restaurants, but rather resorts or convenience stores. To further clean our dataset, we filtered out more categories including "Specialty Food", "Grocery Stores", and "Car Wash".

Two variables were added to the original dataset. Restaurant age was estimated by counting number of days since the date of the first review post. Additionally, restaurant category was converted into a binary variable. Any restaurant that self-categorized as "American", "Burgers", and/or "Steak" in the original "Type" variable was designated as "American", while all others were labeled "Other".

Our final Las Vegas dataset has 4,239 rows and 7 relevant columns. Our dataset is available on our **Github site** (https://palautatan.github.io/yelp-for-causal/).

5 Specify the Observed Data and its Link to the Causal Model

We assume the observed data $O = (W_{age}, W_{type}, W_{reviews}, W_{chain}, A, Y)$ were generated by sampling 3,644 i.i.d. times from a data generating system compatible with the causal model, $\mathscr{M}^{\mathscr{F}}$. This provides a link between the causal model and the observed data. The distribution of the exogenous variables, U, and the structural equations, F, identify the distribution of the endogenous variables, X, and thus the distribution of the observed data, O. The statistical model is non-parametric because we have not placed any restrictions on it. Information on the outcome, exposure, and covariate distributions used in this model is presented in Table 1 (below) and densities of the covariates are presented in the appendix (Figureshttps://palautatan.github.io/yelp-for-causal/ 1-7A).

Table 1. Characteristics of 3,644 Las Vegas restaurants reviewed on Yelp by survival status in 2019.

Variable n (%)	Closed in 2019 248 (7)	Open in 2019 3396 (93)	p-value
Number of stars			0.00
< 3.5	58 (4.3)	1298 (95.7)	
≥ 3.5	190 (8.3)	2098 (91.7)	
Days open			0.00
≤ 2606	158 (8.7)	1664 (91.3)	
> 2606	90 (4.9)	1732 (95.1)	
Number of reviews			0.55
≤ 65	129 (7.1)	1695 (92.9)	
> 65	119 (6.5)	1701 (93.5)	
American restaurant			0.19
No	187 (7.2)	2423 (92.8)	
Yes	61 (5.9)	973 (94.1)	
Chain restaurant			0.00
No	209 (8.9)	2151 (91.1)	
Yes	39 (3.0)	1245 (97.0)	

Values are N (%).
Fisher's exact test was used for categorical variables.
Median values were selected as cut-off points.

6 Identify

To see if the positivity assumption was violated, we first cross-tabulated the data to check for zero cells, and found none, however, some covariate combinations had less variation (i.e. are more rare) than others (Table 1A – Appendix). We then made histograms of the weights created for IPTW by intervention status (see Figure 2 below). Some observations had very high weights, indicating that they were rare to observe. We accounted for this by stabilizing the weights in the IPTW estimation phase.

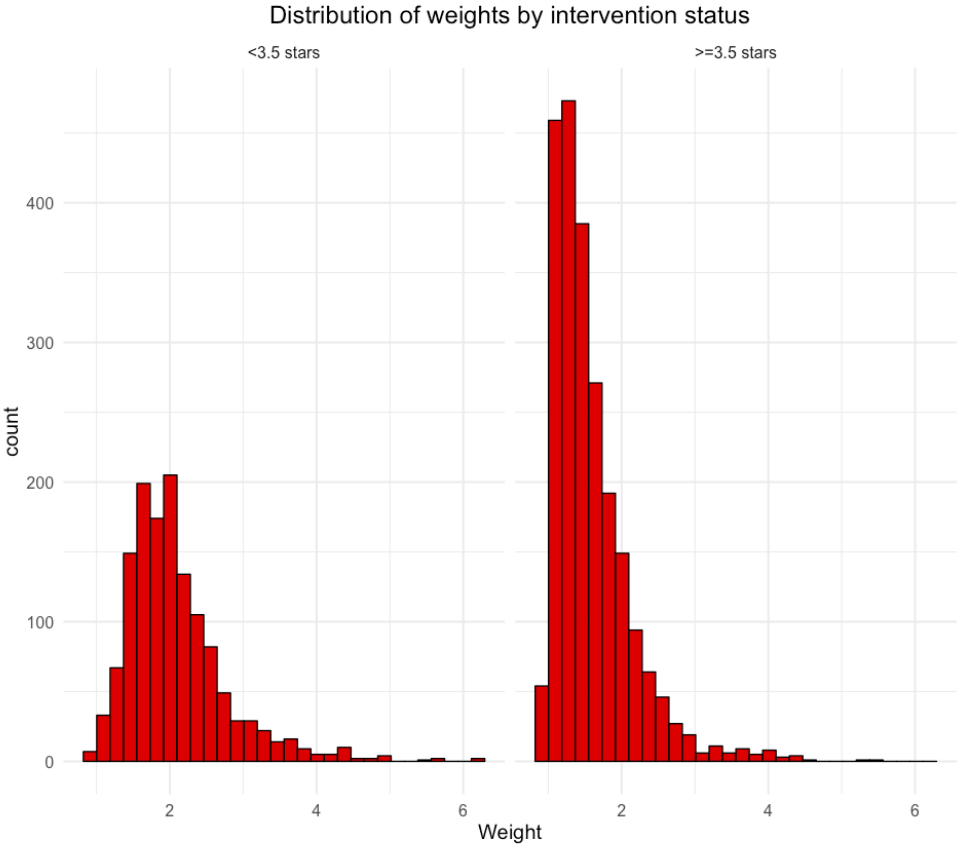

Figure 2

Although we believe that the SCM meets the positivity assumption

$$\forall w \in W \ni P_0(W = w) > 0, \; min_{a \in A} P_0(A = a|W = w) > 0$$

it does not satisfy the backdoor criterion (Figure 1). We believe that the counterfactual survival of the restaurants two years after baseline are dependent on the their average Yelp rating given the baseline covariates. To circumvent these limitations and proceed with our analysis, a number of assumptions must be made to identify our target causal parameter as a parameter of the observed data distribution.

To fulfill the backdoor criterion, we must assume that at least two pairs of exogenous variables are independent. Although these assumptions are required to insure the identifiability of our target causal parameter, they are implausible since it is impossible to disentangle the relationship of the outcome's, treatment's and covariates' exogenous variables due to the complexity of the system. For example, U_A, U_W and U_Y may share exogenous variables like restaurant location, cleanliness and price range. For the same reason, it is unclear which additional data or modification to the data collection process could remedy this situation. Thus, we make the additional assumptions that the exogenous variables of the SCM are independent from one-another (i.e. $U_A \perp\!\!\!\perp U_Y$, $U_A \perp\!\!\!\perp U_W$ and $U_Y \perp\!\!\!\perp U_W$).

Under these assumptions, the modified SCM, denoted $\mathcal{M}^{\mathcal{F}*}$, satisfies the backdoor criterion, conditioned on the covariates. Therefore, the mean effect of an average Yelp review score equal to or above 3.5 stars on the two-year survival of restaurants is identified via the G-computation formula. However, since our analysis is based on a working SCM, its results must be interpreted with caution.

7 Commit to a Statistical Model and Estimand

We define the statistical model to be the working SCM, $\mathcal{M}^{\mathcal{F}*}$, described in the previous section: the original SCM augmented by the additional assumption of independence among the exogenous variables.

The statistical estimand is defined as follows:

$$\begin{aligned}\Psi(P_0) &= E_W[E_0[Y|A=1,W] - E_0[Y|A=0,W]] \\ &= E_W[Pr_0[Y=1|A=1,W] - Pr_0[Y=1|A=0,W]]\end{aligned}$$

where $\Psi : \mathcal{M}^{\mathcal{F}*} \rightarrow I\!R$. If the statistical model reflected reality, the estimand could be interpreted as the average effect of having an average Yelp review score above or equal to 3.5 stars on the probability of two-year restaurant survival in the Las Vegas metropolitan area.

8 Estimate

SuperLearner was used to implement the G-computation, IPTW, and TMLE estimators. Six models were used in estimating TMLE (generalized linear models, penalized generalized linear models, classification trees, pruned classification trees, intercept model, and Random Forests). The empirical risk of each estimator was evaluated through cross-validation. The results for each of the three methods are reported in Table 2. Statistical uncertainty was also quantified for the TMLE estimate by utilizing the sample variance of the estimated influence curve. In particular, our point estimate from TMLE was -0.027 with a 95 % confidence interval of [-0.045, -0.008]. The p-value was 0.004, signifying statistically significant results (Table 3).

We also implemented the non-parametric bootstrap for variance estimation of the three classes of estimators. The data was bootstrapped 500 times and sampling distributions of all three estimators were generated to assess estimator behavior (Figure 3). All three distributions follow a fairly normal, symmetric distribution. Both normal-based and quantile-based 95 % confidence intervals were generated based on these bootstraps (Figure 4). The normal-based and quantile-based confidence intervals are in agreement with one another, with none containing the null value of 0.

Table 2: Estimations Results

Method	Value
Simple Substitution	-0.024
IPTW	-0.011
S-IPTW	-0.028
TMLE	-0.027

Table 3: TMLE Inference Results

	Result
TMLE Point Estimate	-0.027
Asymptotic Variance	0.316
95 confidence interval	[-0.045, -0.008]
p-value	0.004

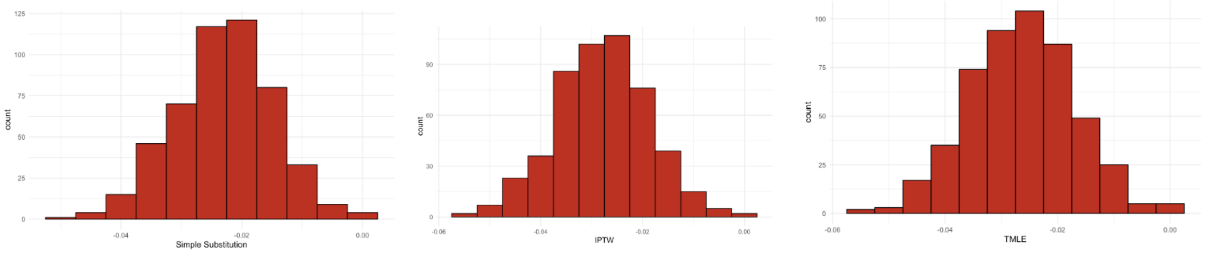

Figure 3: Sampling distribution of the three classes of estimators

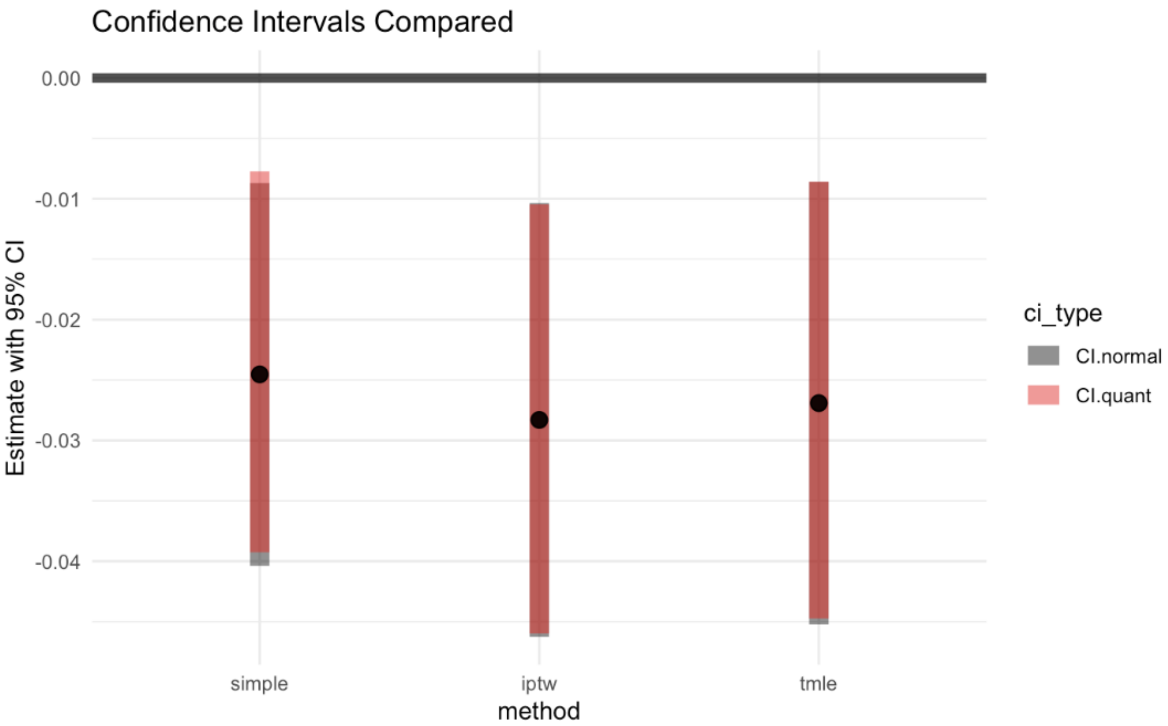

Figure 4: Confidence interval coverage of the three classes of estimators

Table 4: Bootstrapped 95 % Confidence Intervals

Method	Normal-based CI	Quantile-based CI
Simple Substitution	[-0.0404, -0.0087]	[-0.0393, -0.0077]
S-IPTW	[-0.0462, -0.0104]	[-0.0460, -0.0105]
TMLE	[-0.0452, -0.0086]	[-0.0447, -0.0086]

9 Interpret

After controlling for our baseline covariates, the marginal difference in 2-year-survival probability from 2017 to 2019 between restaurants above or at 3.5 stars and those below 3.5 stars is 0.027. The average effect of having a Yelp review score above or equal to 3.5 stars on the probability of two-year restaurant survival in Las Vegas is about -0.027 based on TMLE methods, and according to the associated confidence intervals, this value is statistically significant.

Under the assumptions that our causal model and convenience assumptions are true, then the marginal probability difference can be interpreted such that restaurants with at least 3.5 stars have a 2.7% lower chance of 2-year-survival than do restaurants with less than 3.5 stars. However, the convenience assumptions we made were too lax for us to conclude with this causal interpretation.

While we remain surprised by our negative result implying Las Vegas restaurants on Yelp that have reviews of above or equal to 3.5 stars have a lower survival probability than do lower rated restaurants, we have reason to believe that this result was made possible by (1) our assumptions being extreme, (2) that Las Vegas has a different relationship between Yelp stars and closure probability than what we'd expect, or (3) because we failed to account for time-dependent confounding. We treated our Yelp review data as cross-sectional though Yelp review averages take into account all-time reviews. The 2017 Yelp review per business we used in this analysis was actually a cumulative mean.

10 Limitations and Future Work

We acknowledge the limitations of this study. Restricting the model to four binary covariates caused a loss of information. Additionally, in order to properly apply the causal framework, the working model required the use of extreme convenience assumptions that are not accurate in the real world. Finally, we acknowledge that reviews may not be representative of restaurant quality, and anticipate some other measure may exist to more accurately measure quality.

In the future, we hope to extend the covariates to be continuous and to include spatial aspects, such as accounting for neighborhood of each restaurant. We would also like to consider the temporality of this model; reviews were averaged from restaurant opening to 2017, and time dependent covariates were ignored. Finally, more research should be done on the system that governs restaurant closure in order to make less assumptions and work with more covariates.

11 Team Contributions

Asem Berkalieva: Sections 3, 5, 8, 10, data preparation and bootstrap based inference
Philippe Boileau: Sections 1, 2, 6, 7 and estimation of target parameter using `SuperLearner`
Edie Espejo: Sections 4, 9; data download, preprocessing, and filtering; initial data visualization; estimation of target parameters; bootstrap code
Naomi Wilcox: Sections 5, 6, analysis of practical positivity assumption violations (Table 1A), Table 1

References

[1] Tian Luo and Philip B. Stark. Only the Bad Die Young: Restaurant Mortality in the Western US. *arXiv e-prints*, page arXiv:1410.8603, Oct 2014.

Appendix

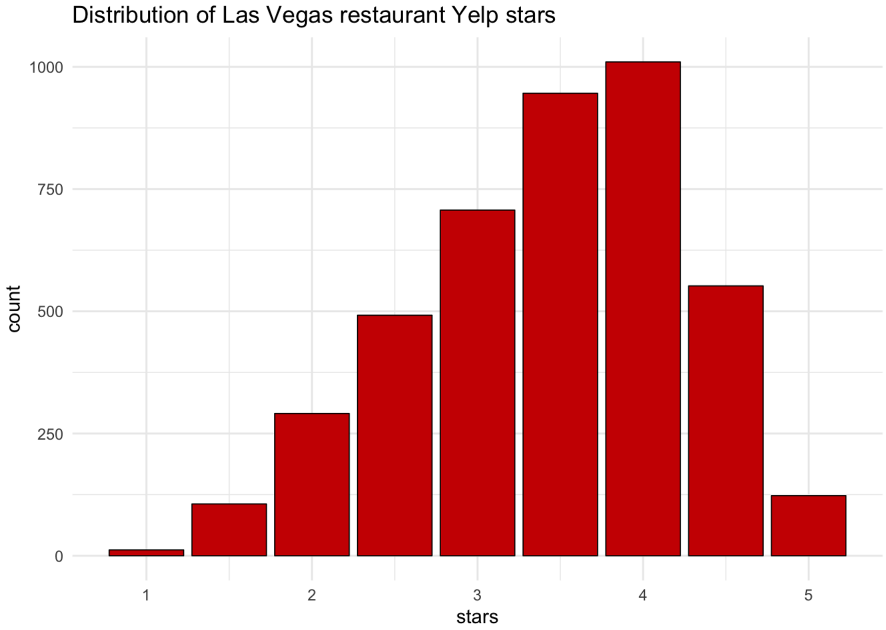

Figure 1A: Distribution of 2017 Yelp star reviews of Las Vegas restaurants.

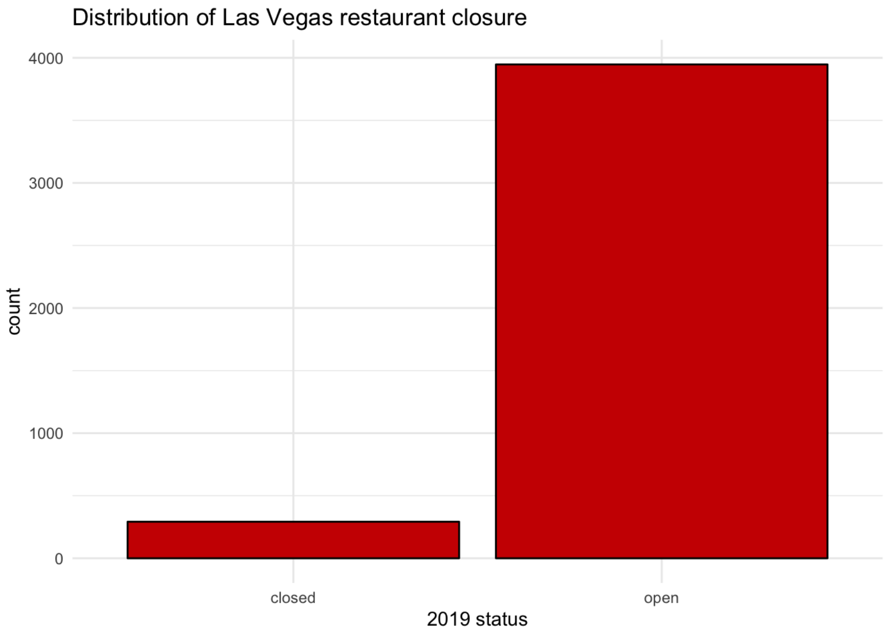

Figure 2A: Las Vegas restaurant closure in 2019. About 4% of restaurants closed.

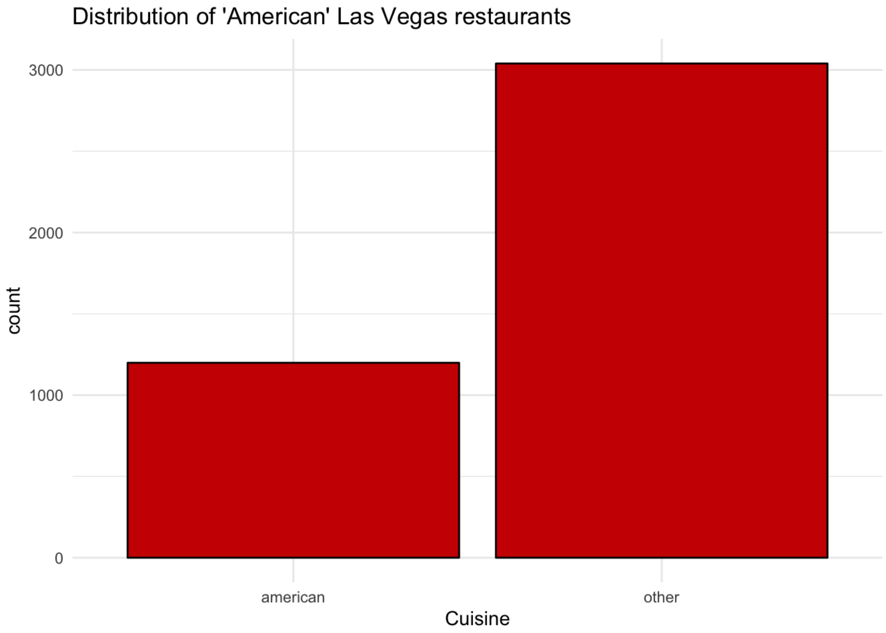

Figure 3A: Distribution of "(US) American restaurant" or not. Clearly, there are more restaurants from other origins than (US) American.

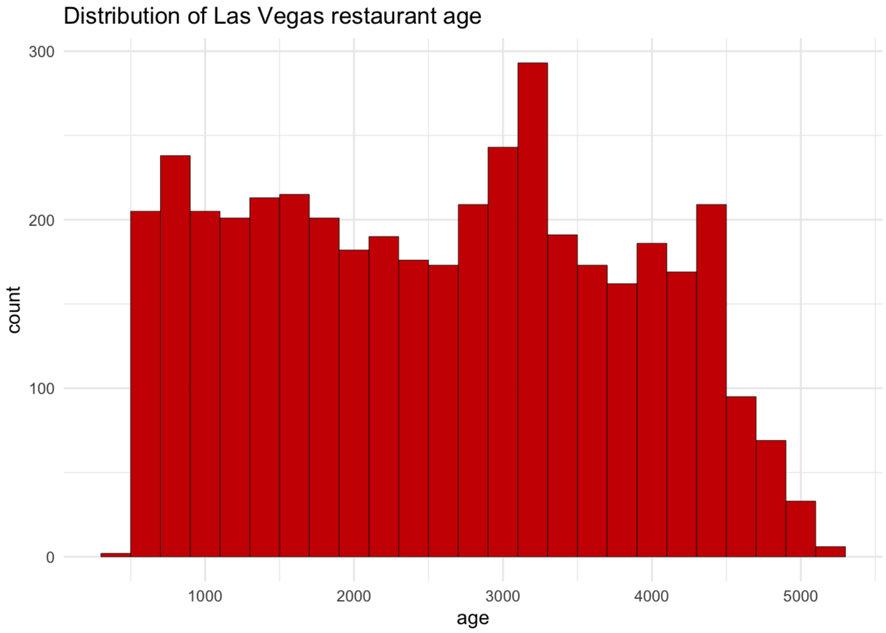

Figure 4A: Age distribution in days of Las Vegas restaurants. The median number of days was 2,606.

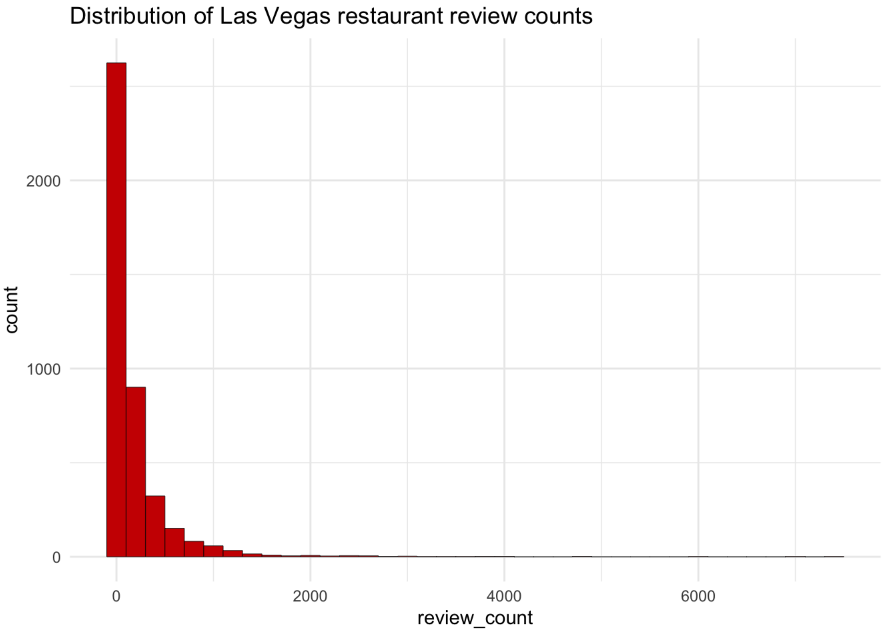

Figure 5A: Review count distribution of Las Vegas restaurants.

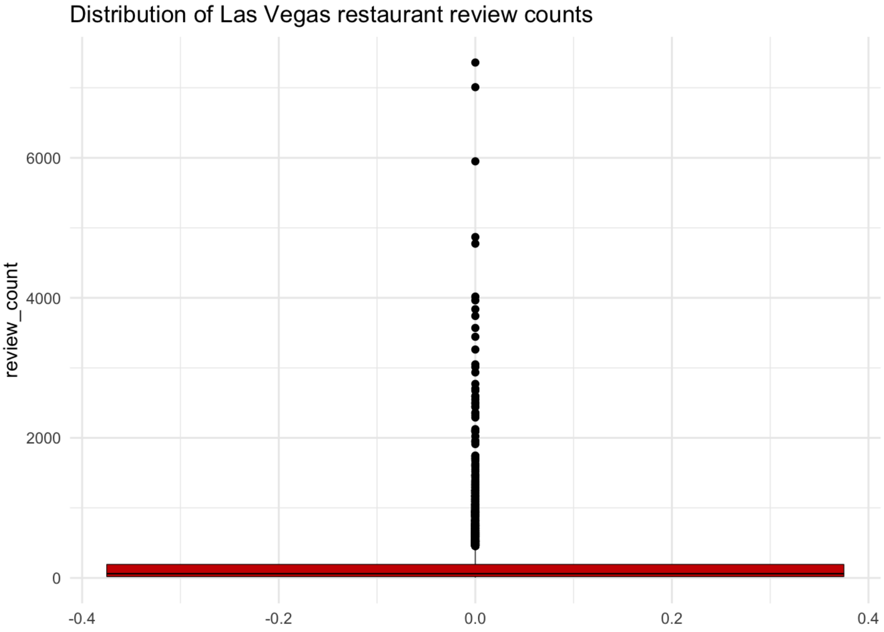

Figure 6A: There were plenty of outliers in the review count distribution. Since we binarized review counts, these did not affect our analyses much.

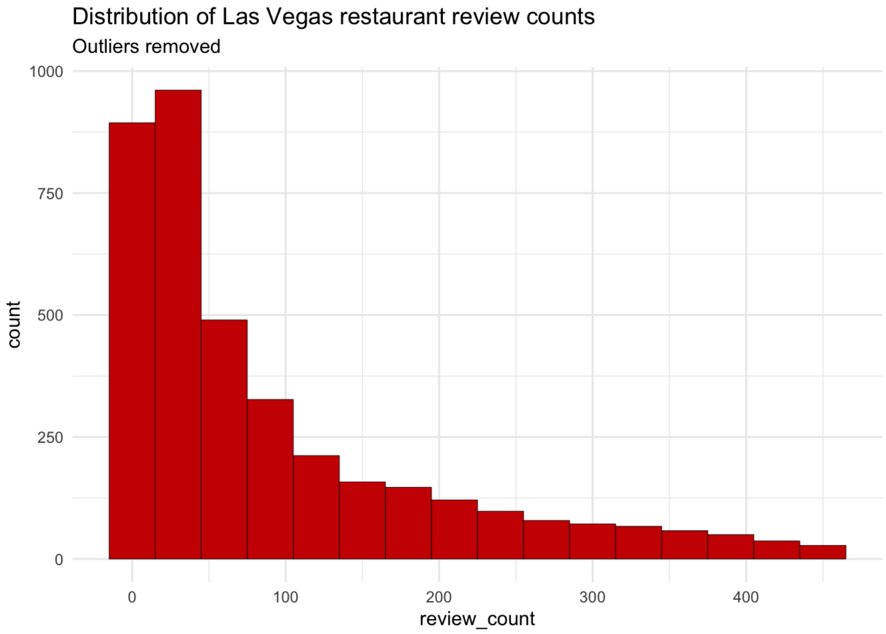

Figure 7A: This plot shows a plot of review counts without outliers just to provide a "zoomed-in" look of the review count histogram.

Table 1A. Cross tabulations of W (covariate) combinations for 3,644 Las Vegas restaurants reviewed on Yelp by intervention status.

agebinary = FALSE, reviewbinary = 0, chain = 0

American	< 3.5 stars	≥ 3.5 stars
No	142	30
Yes	126	108

agebinary = TRUE, reviewbinary = 0, chain = 0

American	< 3.5 stars	≥ 3.5 stars
No	203	25
Yes	234	78

agebinary = FALSE, reviewbinary = 1, chain = 0

American	< 3.5 stars	≥ 3.5 stars
No	46	29
Yes	42	55

agebinary = TRUE, reviewbinary = 1, chain = 0

American	< 3.5 stars	≥ 3.5 stars
No	55	14
Yes	123	46

agebinary = FALSE, reviewbinary = 0, chain = 1

American	< 3.5 stars	≥ 3.5 stars
No	358	439
Yes	114	383

agebinary = TRUE, reviewbinary = 0, chain = 1

American	< 3.5 stars	≥ 3.5 stars
No	112	44
Yes	82	132

agebinary = FALSE, reviewbinary = 1, chain = 1

American	< 3.5 stars	≥ 3.5 stars
No	96	167
Yes	46	179

agebinary = TRUE, reviewbinary = 1, chain = 1

American	< 3.5 stars	≥ 3.5 stars
No	26	36
Yes	19	55