Modelling the Yelp Rating of Restaurants: Disparity of Rating Between Restaurant Cuisines and Prices*

Moohaeng Sohn

April 19, 2024

We use the Yelp dataset to explore and model how different attributes of a restaurant play a factor in the review ratings. We found there are some variables which are promising predictors for the rating of a restaurant. Cuisines such as Chinese or Mexican tended to have lower ratings, and lower priced cuisines performed worse. This study can be used to predict whether a restaurant will be successful or not.

Table of contents

*Code and data are available at: https://github.com/alexsohn1126/yelp-analysis

1 Introduction

Thanks to internet technology, we can get hundreds, if not thousands of reviews of restaurants around us. This means we can choose which restaurants we will go to based on those reviews. Therefore, keeping a high review rating is very important for a restaurant's long-term success.

Yelp is a website where people can post reviews about local businesses, containing 287 million reviews (n.d.a). Yelp's reviews are based on the 5-star system. Users can choose between one to five stars to put on their review. These reviews are collected and averaged, which becomes the rating for the establishment. We will use a dataset from Yelp which we will dive into in Section 2.

In this paper, our estimand of interest is how different factors such as the cuisine of the restaurant, or the price of the menu items in the restaurants affect the rating of a restaurant. We will first explore the dataset in Section 2, and discuss the model the relationship between aforementioned variables using logistic regression in Section 3. Then we will look at the results in Section 4, finally discussing about these results in Section 5.

We found out that higher the price of a restaurant, more likely it will have a higher rating, though with a couple of exceptions. And certain cuisines performed significantly worse than other cuisines. We discuss why that may be the case using graphs and tables in Section 5. This study may be important to entrepreneurs whom may be interested in how these reviews may possibly be biased due to the cuisine and the price of a restaurant.

We used the programming language R (R Core Team 2023), along with packages `tidyverse` (Wickham et al. 2019), `rstanarm` (Goodrich et al. 2022), `jsonlite` (Ooms 2014), `arrow` (Richardson et al. 2024), `modelsummary` (Arel-Bundock 2022), `kableExtra` (Zhu 2021), `here` (Müller 2020).

2 Data

The restaurant review dataset we will use is from Yelp. Yelp offers an academic dataset for the public to use, although it is a small subset of their massive database (n.d.b). The

Table 1: Summary Statistics of Star Ratings

Count	Mean	Median	SD	Minimum	Maximum
47213	3.482187	3.5	0.8082359	1	5

dataset is split between multiple JSON files, but we will only focus on businesses JSON. The raw businesses dataset contains 150,346 businesses. These businesses are from metropolitan areas in United States and Canada. Specifically, from metropolitan areas near Montreal, Calgary, Toronto, Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison, and Cleveland (n.d.b).

Another possible source of restaurant reviews could have been from Google reviews of restaurants, but that requires us to use Google Business Profile APIs, which cost money. There is always an option to perform webscraping, but this is a gray area legally, and may be computationally expensive. There are other review websites such as TripAdvisor, but they do not seem to have a dataset open to the public like Yelp does.

The dataset contains multiple variables such as the name of the business, the location of the business, and things such as amenities on site. We have filtered through the raw dataset to only contain restaurants. There were a total of 47,213 restaurants in the final dataset. We will focus on the star rating, categories, and price range of the menu in the restaurant.

2.1 Star Rating

Star rating is the average star rating of a restaurant. One odd thing about the star rating given in the dataset is that it is rounded to the nearest star or half of a star. So all possible values of star ratings are: 1.0, 1.5, 2.0, and so on until 5.0.

Figure 1 shows us that 4.0 is the most common star rating of restaurants. This distribution is left skewed, as we can see the rating gradually increases from 1.0 to 4.0, then quickly drops off from 4.0 to 5.0. We can infer from this that most people consider somewhere around 4-stars an average dining experience.

Table 1 shows the summary statistics of star ratings, and we can see that as we explained, that the minimum is 1.0 and the maximum is 5.0. The standard deviation isn't that big, which is to be expected because most of the restaurants lie between 3.5 to 4.5 rating range.

2.2 Categories (Cuisine)

Categories for a restaurant describes what kind of business it is. These categories are chosen manually by the business owners (2024). Because Yelp is a business review platform, we have decided to filter out non-restaurant businesses from our final dataset. This meant any restaurants which didn't include one of the categories: "Restaurants", "Food", "Fast Food"

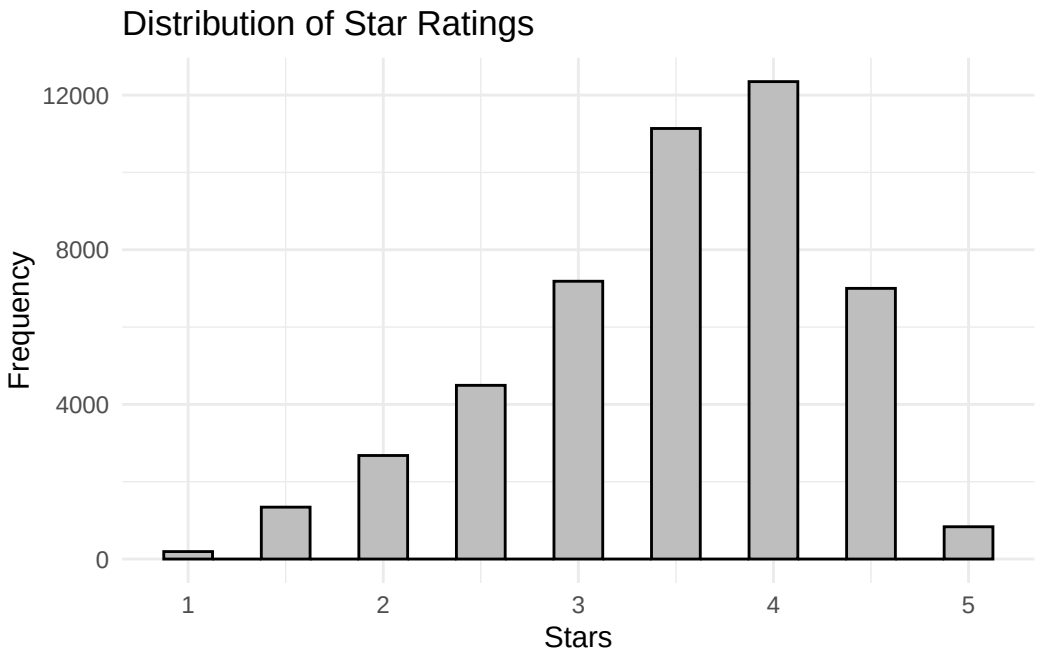

Figure 1: Distribution of Star Ratings

would not be included in the final dataset. After filtering out non-restaurants, we have chosen top 12 cuisines to categorize each restaurant into. The restaurants that did not have these cuisines in their categories were put into "Other" category. This is why we have named this section also cuisine, as we will focus on cuisine categories of restaurants.

Table 2 shows us what types of cuisines we have selected, and the number of restaurants with that cuisine in their category. We can see nearly half of the restaurants are categorized as "Other". The next largest cuisine in our dataset is American. This may be due to the fact that all the restaurants in the dataset are from United States and Canada.

2.3 Price Range

The price range is given as an integer from 1 to 4, inclusive. This is supposed to indicate how expensive the restaurant is per person. 1 being more affordable, and 4 being more expensive. There were restaurants with no price ranges, and those were discarded from the cleaned dataset. On Yelp's website, these price ranges are shown as number of dollar signs equal to the price range. For example, Yelp would show one dollar sign if price range is 1.

Table 3 shows that the over 95% of the restaurants have price range of 1 or 2. About 95.81% of restaurants in the database are in price range 1 or 2. Only 3.73% of the restaurants are in price range 3, and even less, 0.46% of the restaurants are in price range 4.

Table 2: Numbers of Restaurants Per Cuisine

Cuisine	Number of Restaurants	Percentage (%)
Other	21741	46.05
American	10841	22.96
Italian	3729	7.90
Mexican	3476	7.36
Chinese	2423	5.13
Japanese	1351	2.86
Mediterranean	738	1.56
Thai	694	1.47
Indian	628	1.33
Vietnamese	617	1.31
Caribbean	343	0.73
French	323	0.68
Middle Eastern	309	0.65

Table 3: Numbers of Restaurants Per Price Range

Price Range	Number of Restaurants	Percentage (%)
1	20670	43.78
2	24566	52.03
3	1762	3.73
4	215	0.46

2.4 Measurement

For star ratings, the minimum star rating is one and the maximum star rating is five. This is because the minimum star rating that someone can give a restaurant is 1 stars, and the maximum star rating is 5 stars. While the average star rating can have 0.5 star increments, users' reviews can only give ratings in 1 star increments. For example, no 4.5 stars can be given in a review. Each user can rate a business only once, though they are free to change their rating later, one user cannot post multiple reviews about a restaurant. This limits the power each user has to change the rating of a business.

As mentioned above, categories for a restaurant is not given automatically, rather given by the business owners. This means it can be a bit inaccurate when it comes to what exact cuisine that a restaurant serves. Another way the cuisines can be inaccurate is when a restaurant serves foods from multiple cuisines. This makes it complicated to categorize them exactly into each cuisine.

The exact method of how the price range was decided by Yelp is not known. We can make guesses about what kind of information they use to determine the price range, but no clear documentation is available for the price range statistics. We were able to find online discussions about the price range being decided by the check-in feature of Yelp, possibly implying that there are certain price ranges that Yelp uses to display these price ranges (S 2017).

3 Model

We are trying to model the rating y, and as we have discussed in the data section, there are only 9 possible outcomes. From 1.0, to 5.0, in increments of 0.5. Therefore we will use ordinal logistic regression. If we assume that the observed rating y is from a continuous variable y^*, then we can say that:

$$y = \begin{cases} 1.0 & \text{if } y^* \leq c_{1.0|1.5} \\ 1.5 & \text{if } y^* \in (c_{1.0|1.5}, c_{1.5|2.0}) \\ \vdots & \\ 4.5 & \text{if } y^* \in (c_{4.0|4.5}, c_{4.5|5.0}) \\ 5.0 & \text{if } y^* \geq c_{4.5|5.0} \end{cases}$$

Where $c_{n_1|n_2}$ represents the cutoff between n_1 rating and n_2 rating. The above notations are borrowed from Gelman, Hill, and Vehtari (2020) [p.276] and Alexander et al. (2024, 12).

We will use a Bayesian framework to make our model. To do this, we use `rstanarm` (Goodrich et al. 2022) R package. Background details and diagnostics are included in Appendix A.

3.1 Model set-up

We will model the final discrete rating y by a continuous variable y^*. Because we are doing ordinal logistic regression, y^* will have a logistic distribution.

$$y^* = \beta_1 \cdot \text{cuisine}_i + \beta_2 \cdot \text{price}_i$$

We use the R2 prior, because `rstanarm` requires us to do this for this computational process, with the mean set as 0.3 as suggested by Gelman, Hill, and Vehtari (2020, 276). R2 uses the prior beliefs about the location about R^2, which is "the proportion of variance in the outcome attributable to the predictors" Goodrich et al. (2022). We use default priors from `rstanarm` for the cut points ($c_{n_1|n_2}$), which is a dirichlet distribution with concentration of 1.

3.1.1 Model justification

We can expect that if a restaurant has higher prices, then the restaurant will be higher quality. This should be reflected in the ratings by the customers, and those restaurants with higher quality will have higher ratings. However, one could set much higher expectations for more expensive restaurants, effectively cancelling out the higher quality food and service that is expected from a higher priced restaurant. We will keep this in mind when we discuss our results.

We also expect that different cuisines will have different ratings. This dataset is exclusively about restaurants in North America. Therefore we may expect that reviewers are more likely to be from North America, so perhaps they would give a higher rating to a restaurant that serves western cuisine, such as Italian, or American.

4 Results

Our results are summarized in Table 4, and Figure 2.

The rows with numbers and a vertical line between them are the cutpoints. For example, a rating would be considered 1.5 if we had a y^* that is greater than 1.0|1.5 and less than 1.5|2.0.

One thing to note is that `rstanarm` takes the first factor from increasing lexicographic order and sets it as a default. So you won't see American cuisine nor price range of 1 in Table 4 and Figure 2. So every coefficient listed here is how different cuisines/price ranges perform *compared to* American cuisine and price range of 1.

With that in mind, let's explore how different cuisines perform compared to American cuisine. Most cuisines are likely to increase the rating of a restaurant. However, according to our model,

if a restaurant serves Chinese cuisine, it will receive a slightly lower rating than American restaurants. Also, a restaurant serving Italian or Mexican cuisines will receive roughly the same rating, as that of an American restaurant. Rest of the cuisines perform better than American cuisine.

And as for price ranges, we can see compared to restaurants with price range of 1, those with price range 2 gets significantly higher rating. The increase of rating from going from price range of 1 to 2 is larger than that of 2 to 3, and that is larger still than that of 3 to 4. In other words, we have diminishing returns as the price range increases on the rating of the restaurant.

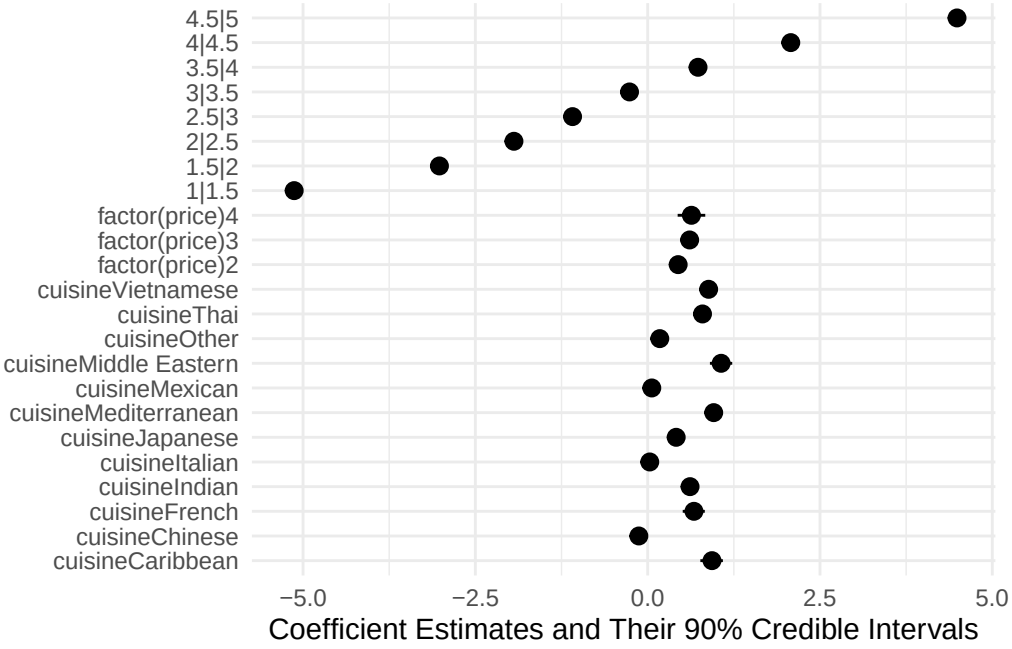

Figure 2: Exploring The effects of cuisines and price ranges on the rating of a Restaurant

5 Discussion

5.1 Rating Disparity Between Cuisines

Certain cuisines had worse effect on restaurant rating than others. Namely, Chinese, American, Italian, and Mexican cuisines had noticeably worse rating than others. While it is plausible to state that restaurants with certain cuisines are simply worse, we will try to explain this backed with research.

Table 4: Table of how different cuisines and price range affects the rating of a restaurant

	Coefficients and Cutpoints
cuisineCaribbean	0.933
cuisineChinese	−0.128
cuisineFrench	0.671
cuisineIndian	0.615
cuisineItalian	0.030
cuisineJapanese	0.413
cuisineMediterranean	0.958
cuisineMexican	0.060
cuisineMiddle Eastern	1.066
cuisineOther	0.176
cuisineThai	0.795
cuisineVietnamese	0.884
factor(price)2	0.442
factor(price)3	0.609
factor(price)4	0.635
1|1.5	−5.129
1.5|2	−3.021
2|2.5	−1.940
2.5|3	−1.090
3|3.5	−0.263
3.5|4	0.731
4|4.5	2.076
4.5|5	4.488
Num.Obs.	47 213
ELPD	−86 205.8
ELPD s.e.	138.5
LOOIC	172 411.6
LOOIC s.e.	277.0
WAIC	172 411.6

As for Chinese cuisine, one possible cause is the stigma around Chinese food. Chinese restaurant syndrome was a list of symptoms supposedly linked by Chinese restaurants' excessive usage of monosodium glutamate (MSG) (Mekkodathil and Sathian 2017). Dr. Robert Ho Man Kwok first coined the term "Chinese Restaurant Syndrome" back in April of 1968, sending a letter to the New England Journal of Medicine (LeMesurier 2017). This stigma against Chinese restaurants could still be remaining today, leading to a lower rating.

American, Italian, and Mexican cuisine's bad performance could be explained by the demographic of the restaurants. All the restaurants in the dataset are in United States and Canada. All 3 aforementioned cuisines have very close ties with these countries. American cuisine is from United States. Italian cuisine is from Italy, but a lot of Italians are living in North America. Around 16 million Italian descendants are living in United States (2023), and 1.5 million in Canada (2022). Mexico shares a border with United states, therefore a ton of cultural influences were made in southern United States by Mexicans. What we propose is that people rate restaurants that serve foods from their own culture harshly. For example, people who have Mexican parents or has been raised with Mexican culture will have higher standards for Mexican food.

We can try attributing this to the fact that different cuisines will have different price ranges. Therefore one can argue that those cuisines with higher proportion of low price restaurants will be more likely to have lower ratings. Figure 3 shows each cuisine's price range proportions. It shows that while Chinese and Mexican restaurants have high proportion of restaurants with price range of 1, but it is not the case for American and Italian cuisine.

Another interesting insight is that these 4 cuisines we have listed as "worse" than others are the top 4 cuisines excluding "Other" category in our dataset (Check Table 2). Perhaps more restaurants there are for certain cuisines, people's expectations are higher, leading to lower rating for "average" experiences in those restaurants. Figure 4 shows the relationship between the number of restaurant and the rating that a cuisine has. We can clearly see there is a downward trend as number of restaurants increase. This could have occurred because people associate rare cuisine as being better. Also, there is a possibility that the owners that would care about their Yelp categories are also more likely to care about the service and the food of the restaurant.

5.2 Price Range and Rating

We also observed that higher price range of a restaurant means the restaurant will higher ratings. We have expected this in the model section, but the diminishing returns of ratings as price increases was not expected. We can plot out the relationship between the price and the rating by cuisine to further illustrate this idea. Figure 5 shows that the lower rating cuisines do have lower average prices. But there are cuisines, such as Caribbean or Vietnamese that has a high rating, yet with lower average prices. The figure shows that for some cuisines, their price range doesn't affect their ratings.

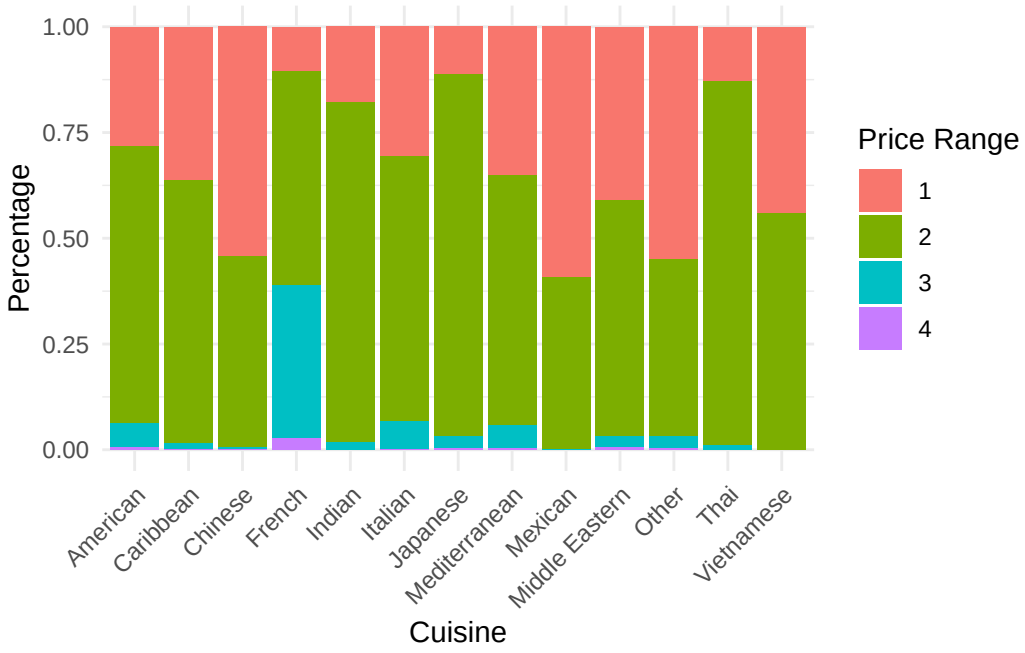

Figure 3: Relationship between Cuisine and Price Range (Normalized)

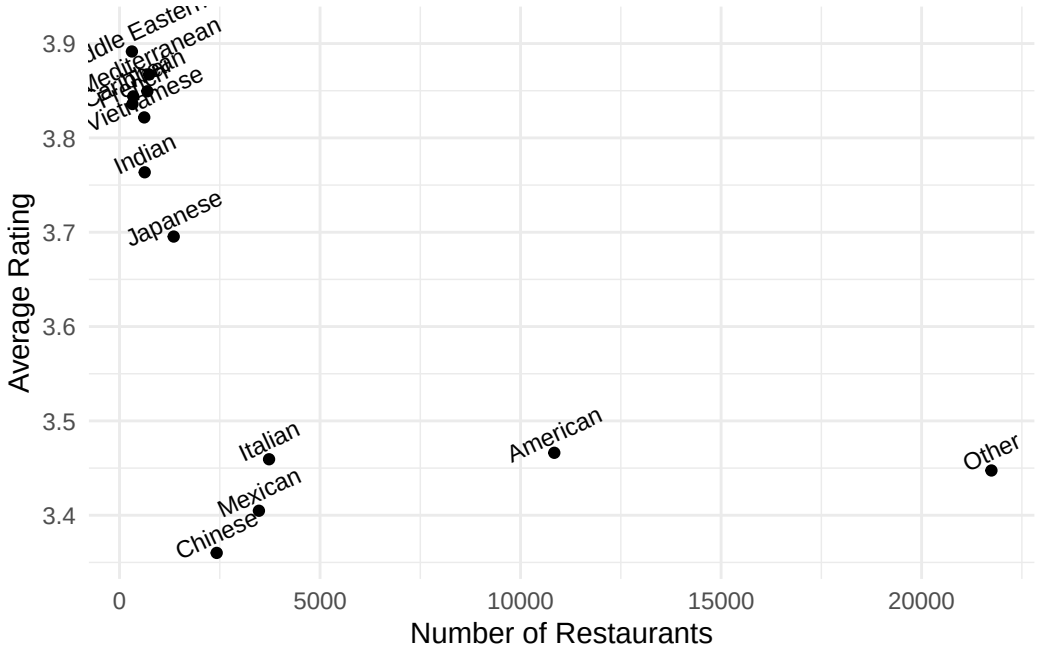

Figure 4: Average Rating vs. Number of Restaurants by Cuisine

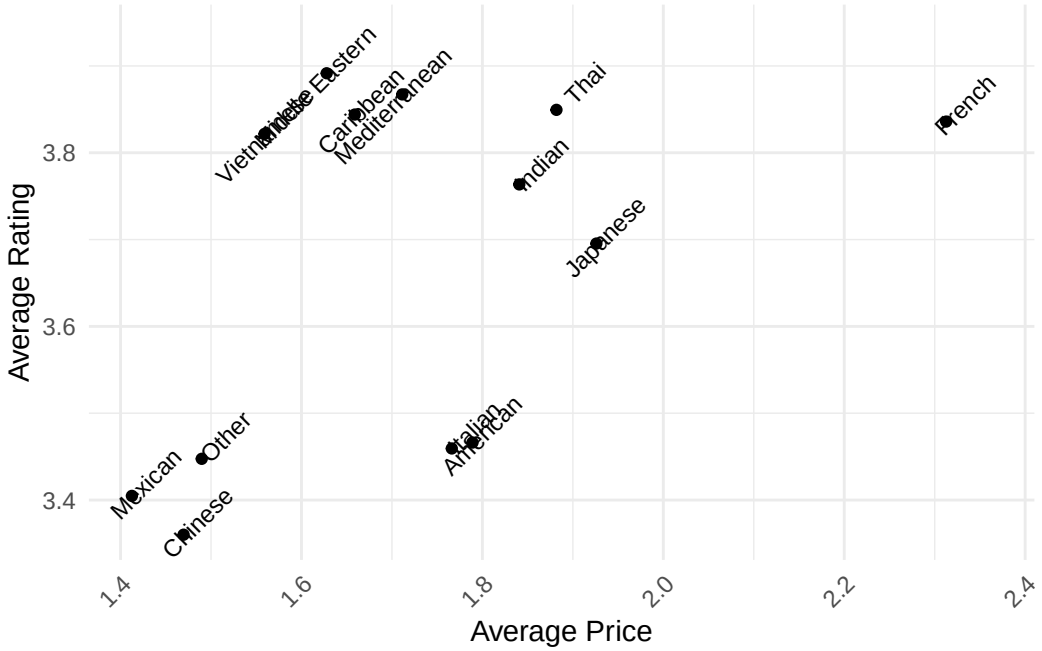

Figure 5: Average Rating vs. Average Price by Cuisine

This could have possibly been caused by simply those with lower prices cannot afford decent service and food, but that still does not explain the low average rating of Italian and American restaurants which have higher costs yet near the same rating as Mexican and Chinese restaurants. Another way this could have been caused is because, again, people rate rarer cuisines higher than common ones. So perhaps in later studies, more research can be done with those rarity in mind.

5.3 Weaknesses and next steps

A weakness that we have observed is that there may not be enough variables in our model to correctly model the rating. We could have added more factors, such as how rare the cuisine is in the dataset, or the number of reviews. We might have enough information to perform post-stratification with the number of ratings for a restaurant, but that would require a reliable, bigger dataset on reviews of restaurants which we do not have access to.

Another limitation could have been in how we clean our data. We selected businesses that had restaurant-related categories in Yelp, but there is likely to be many businesses which forgot to add the restaurant related categories on Yelp, or simply does not care. Therefore, filtering restaurants by categories is unreliable.

The data itself is also a limitation. The data only contains restaurants from big cities, such as Toronto, Vancouver, and Las Vegas. This could have caused a skew in our dataset, over-representing restaurants in bigger cities, and not those in the smaller, suburban communities. Not only that, the ratings of restaurants were given in 0.5 increments which meant it isn't the most accurate rating we could have gotten.

Next steps to take are refining this model to take account of different variables that were missed in this paper and improving the predictability of the model. Perhaps with a newer database, this study shall be repeated too.

Appendix

A Model details

A.1 Posterior predictive check

In Figure 6 we implement a posterior predictive check. This shows that the there is a large disparity between the predictions and the actual data.

In Figure 7 we compare the posterior with the prior. This shows the posteriors are around where the credible intervals are in priors.

A.2 Diagnostics

Figure 8 is a trace plot. It shows the chains cover a constant amount of region. This suggests that a large amount of sample is chosen at around the target space we want to explore.

Figure 9 is a Rhat plot. It shows all of the Rhat values are very close to 1, meaning the chains had enough time to converge.

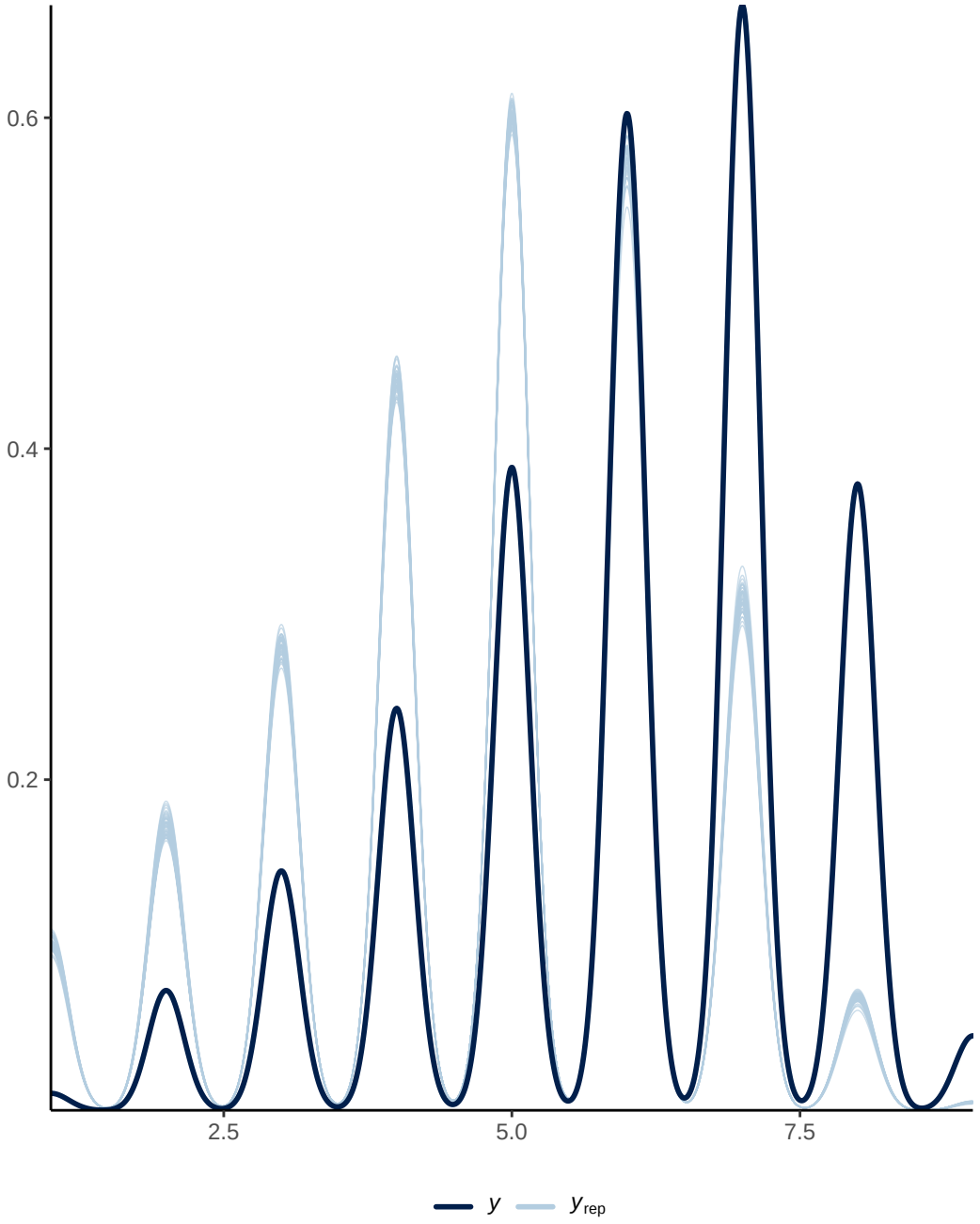

Figure 6: Examining how the model fits, and is affected by the data - Posterior prediction check

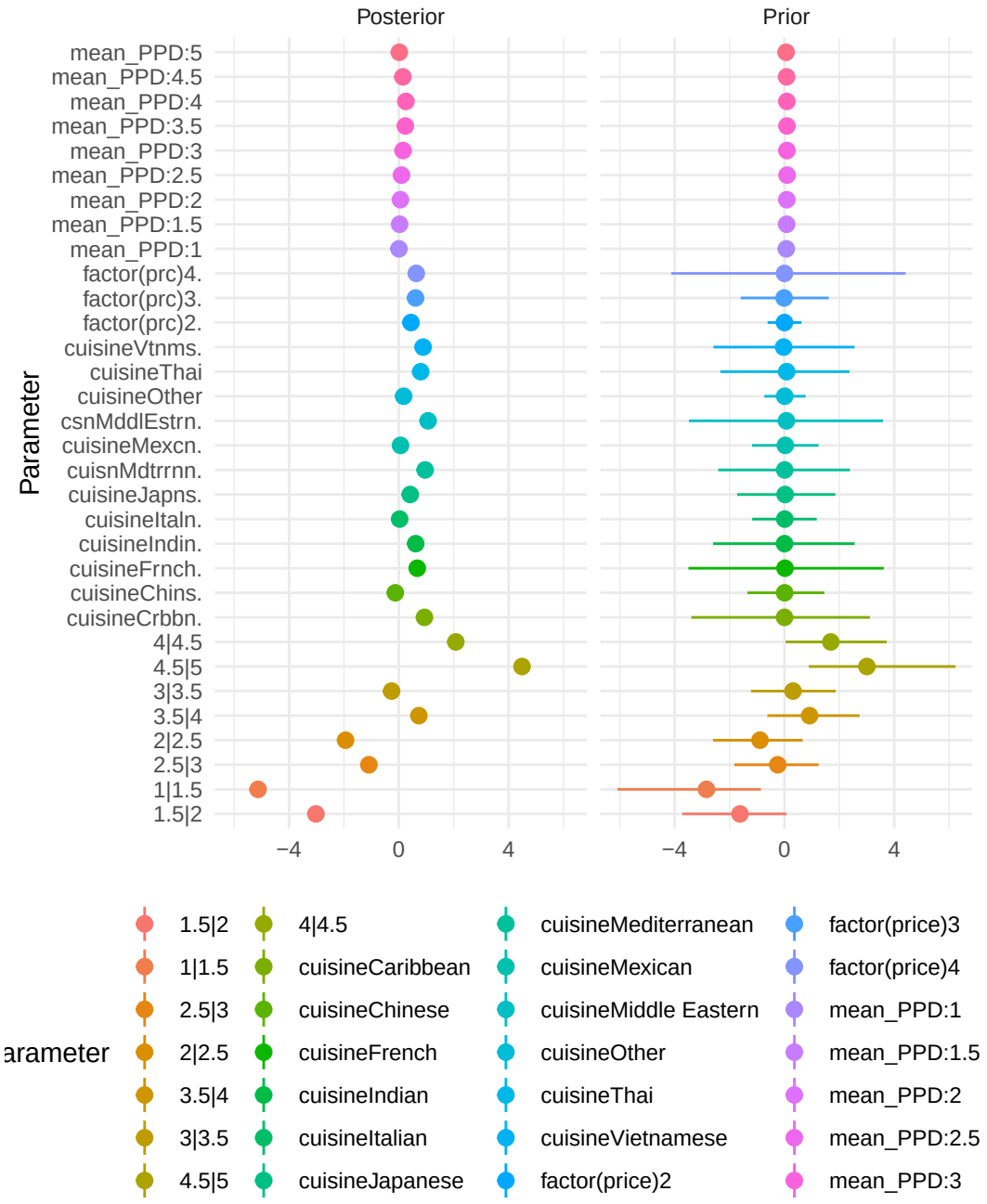

Figure 7: Examining how the model fits, and is affected by the data - Comparing the posterior with the prior

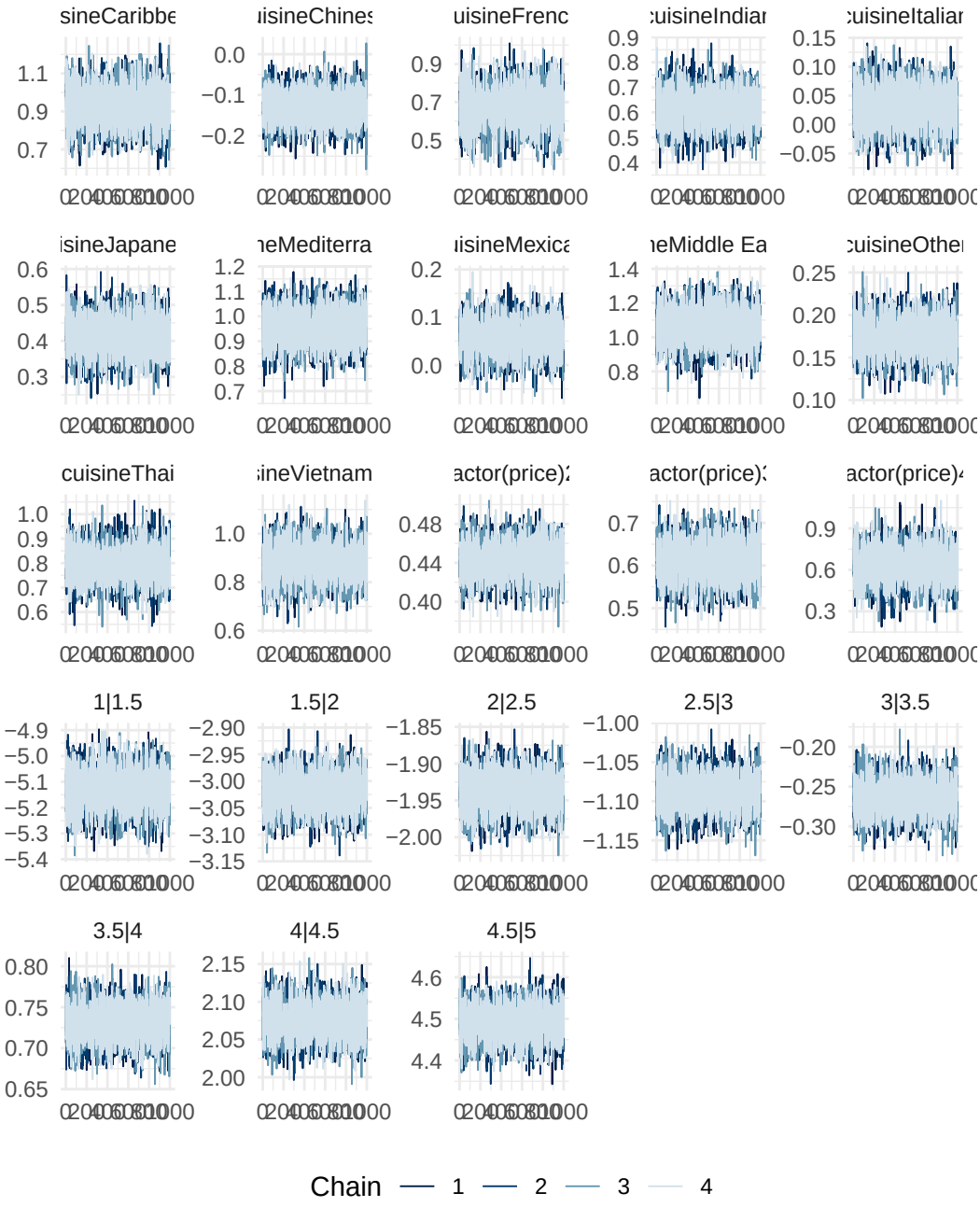

Figure 8: Checking the convergence of the MCMC algorithm - Trace Plot

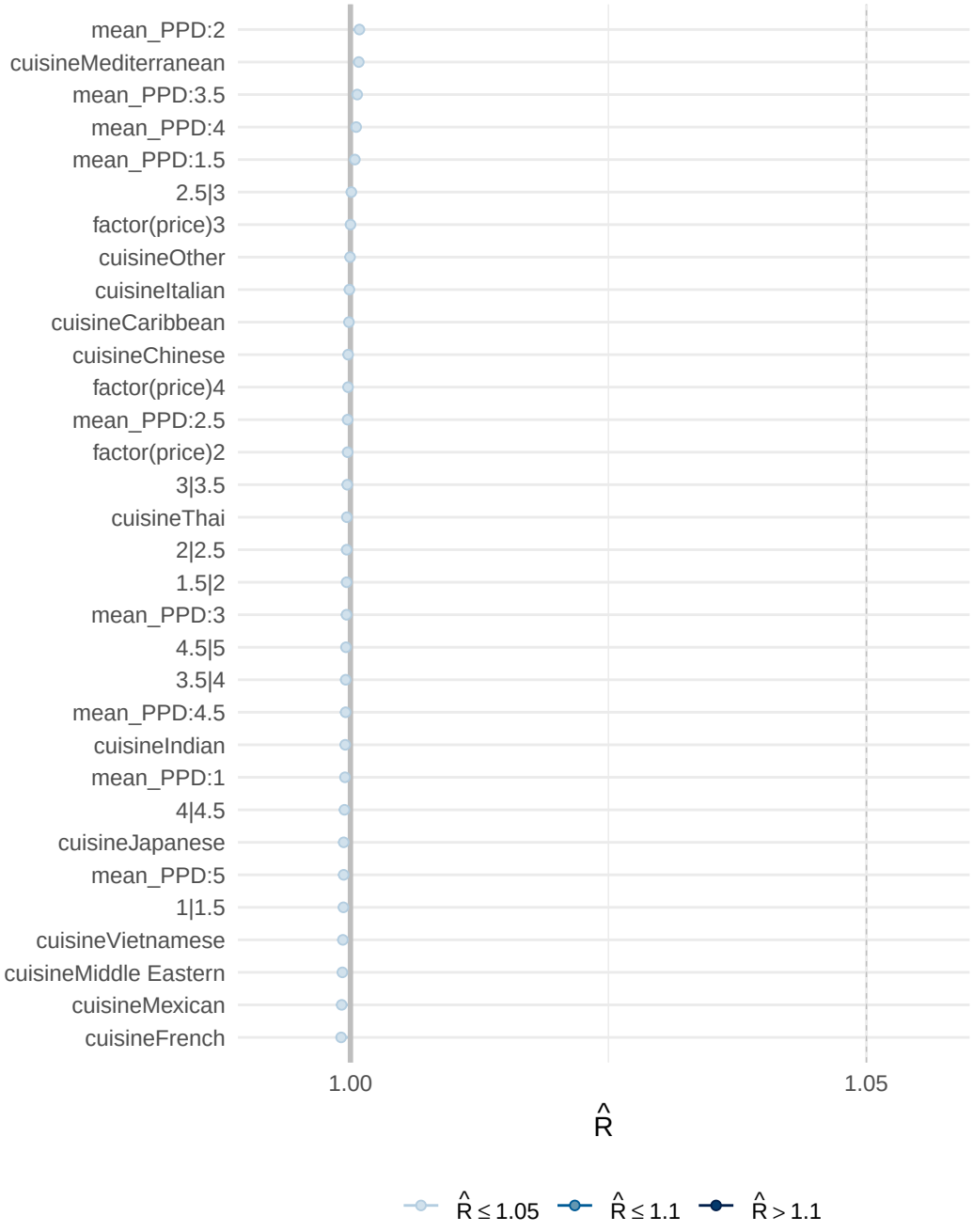

Figure 9: Checking the convergence of the MCMC algorithm - Rhat Plot

References

2022. *Government of Canada, Statistics Canada*. Government of Canada, Statistics Canada. https://doi.org/10.25318/9810033801-eng.

———. 2023. *Census.gov*. United States Census Bureau. https://www.census.gov/newsroom/stories/italian-american-heritage-culture-month.html.

———. 2024. *Yelp*. https://blog.yelp.com/businesses/yelp_category_list/.

———. n.d.a. *Yelp - Company - Fast Facts*. https://www.yelp-press.com/company/fast-facts/default.aspx.

———. n.d.b. *Yelp Dataset FAQ*. https://www.yelp.com/dataset/documentation/faq.

Alexander, Rohan, Lindsay Katz, Callandra Moor, Michaela Drouillard, Michael Wing-Cheung Wong, and Zane Schwartz. 2024. "Evaluating the Decency and Consistency of Data Validation Tests Generated by LLMs." https://github.com/RohanAlexander/evaluating_decency_and_consistency.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press. https://avehtari.github.io/ROS-Examples/.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

LeMesurier, Jennifer L. 2017. "Uptaking Race: Genre, MSG, and Chinese Dinner." *Poroi* 12 (2): 1–23. https://doi.org/10.13008/2151-2957.1253.

Mekkodathil, Ahammed, and Brijesh Sathian. 2017. "Monosodium Glutamate and Chinese Restaurant Syndrome: Separating Facts from Fiction." *Medical Science* 5 (3): 35. https://doi.org/10.29387/ms.2017.5.3.35-36.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Ooms, Jeroen. 2014. "The Jsonlite Package: A Practical and Consistent Mapping Between JSON Data and r Objects." *arXiv:1403.2805 [Stat.CO]*. https://arxiv.org/abs/1403.2805.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

S, Dez. 2017. *Yelp*. https://www.yelp.com/topic/mesa-what-is-latest-definition-and-range-of-the-dollar-sign-on-yelp-for-actual-expense.

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. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.