Cuisine and Yelp Ratings*

Moohaeng Sohn

April 17, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

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.

TODO: ADD THINGS ABOUT WHAT WE FOUND OUT AND WHY IS IT IMPORTANT

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 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 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

Table 1: Summary Statistics of Star Ratings

Count	Mean	Median	SD	Minimum	Maximum
47231	3.482247	3.5	0.8082371	1	5

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 47,231 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. Notice that 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.

Yelp has removed businesses that had 3 or less reviews posted that were older than 14 days at the date of collection in the dataset (n.d.b). This means restaurants that just opened, or didn't have enough reviews at the time of data collection would not have been included in this dataset.

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.

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

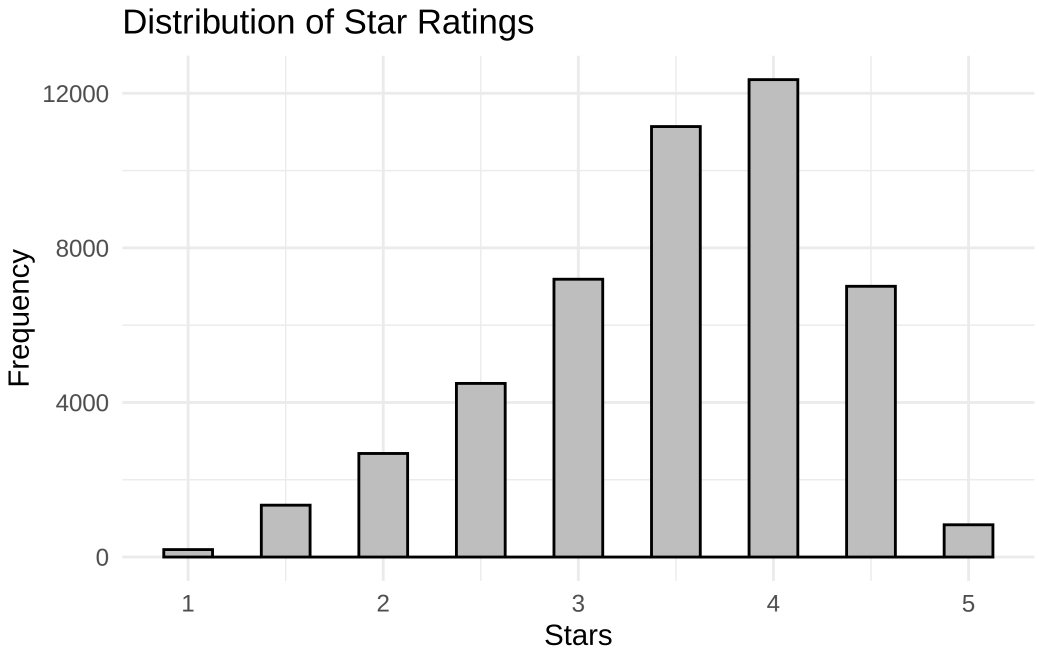

Figure 1: Distribution of Star Ratings

restaurants which didn't include one of the categories: "Restaurants", "Food", "Fast Food" 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.

2.3 Price Range

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

Table 2: Numbers of Restaurants Per Cuisine

Cuisine	Number of Restaurants	Percentage (%)
Other	21750	46.05
American	10845	22.96
Italian	3730	7.90
Mexican	3477	7.36
Chinese	2423	5.13
Japanese	1352	2.86
Mediterranean	738	1.56
Thai	694	1.47
Indian	628	1.33
Vietnamese	619	1.31
Caribbean	343	0.73
French	323	0.68
Middle Eastern	309	0.65

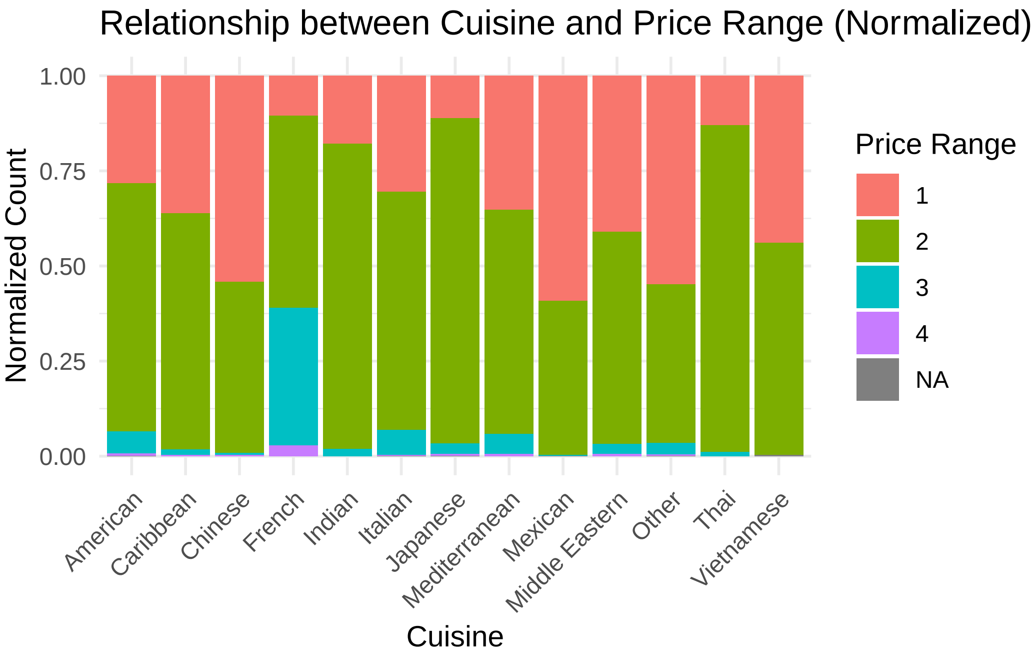

Figure 2: Bills of penguins

$$
\begin{aligned}
y_i|\mu_i, \sigma &\sim \text{Normal}(\mu_i, \sigma) && (1)\\
\mu_i &= \alpha + \beta_i + \gamma_i && (2)\\
\alpha &\sim \text{Normal}(0, 2.5) && (3)\\
\beta &\sim \text{Normal}(0, 2.5) && (4)\\
\gamma &\sim \text{Normal}(0, 2.5) && (5)\\
\sigma &\sim \text{Exponential}(1) && (6)
\end{aligned}
$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

Figure 3: **?(caption)**

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

Checking the convergence of the MCMC
algorithm

Figure 4: **?(caption)**

References

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.

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.

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

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.

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.