MA678 Midterm Project

Carolyn Wright

December 2, 2021

Abstract

The Yelp Dataset Challenge was put together in order to provide students with the opportunity to conduct analysis or research using Yelp's very large and comprehensive dataset. This data contains information on reviews, businesses, users, tips, and check-ins. Using primarily the businesses and reviews data, I built a multilevel model to better understand some of the predictors of the average proportion of positive reviews a restaurant has from a consumers perspective. Some of these predictors include whether a restaurant is in an area with a tourist attraction, the number of sister restaurants a restaurant has, and the type of cuisine the restaurant serves. The multilevel model uses restaurant type and postal code as random intercepts and the tourism indicator as random slopes effect. The result of this model did not show any major effect due to restaurant type and postal code, however there is an impact of tourism within different postal codes. Although this effect is minor, it does exist.

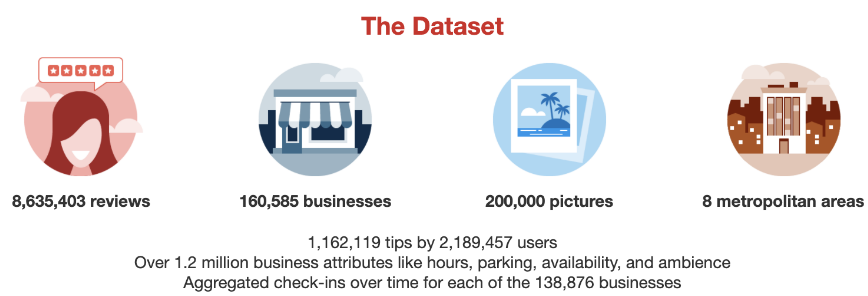


Figure 1: Yelp data breakdown

Introduction

The first thing I look for when exploring a new city is where to eat. Where are the best places to go and what makes them so good? When I first got to Boston, I repeatedly heard that if you want good Italian food you must go to the North End, that you could not possibly go wrong there. So, I went. And since I was told I could not go wrong, I popped into one of the first places I saw. I quickly learned that you can go wrong..very wrong. After being served, quite possibly, the worst chicken parm I have ever had, I began to think about the restaurant ratings in places like this and whether or not they reflect this phenomenon of "you can't go wrong in the North End." Do restaurants in highly trafficked, touristy areas have higher ratings or more positive reviews? Do small mom and pop places rate higher or have more positive reviews than chains? How do Italian restaurants in other areas of the city rate in comparison to those in the North End? Are the best restaurants in the most obvious places? These are the types of the questions I am looking to dig into with this investigation of the Yelp data. I will use a multilevel model to see if these types of factors have any impact on restaurant ranking.

I will focus this investigation on restaurants within Boston, Massachusetts.

Method

The Yelp review data is made up of individual reviews and ratings of restaurants within the US. This data has been pared down to just restaurants within Boston, Massachusetts zip codes within the years of 2015-2020. Additionally, in order to have one observation per restaurant the percent of positive sentiment of the reviews was averaged across all reviews during this time period. Other group level(postal code) variables were added on to the data as supplemental sources of information.

Exploratory Data Analysis

Once the data had been subset down to just Boston, Massachusetts postal codes, I began to explore what relationships might be interesting to look at. During this exploration I primarily used the existing data or external sources to create variables. I was focused on creating variables that would impact ratings from the perspective of a consumer. I quickly realized that ratings or stars have a clear linear relationship with the sentiment of a review, therefore instead of using stars as my outcome I chose to use the average of the positive percentages of reviews for each restaurant.

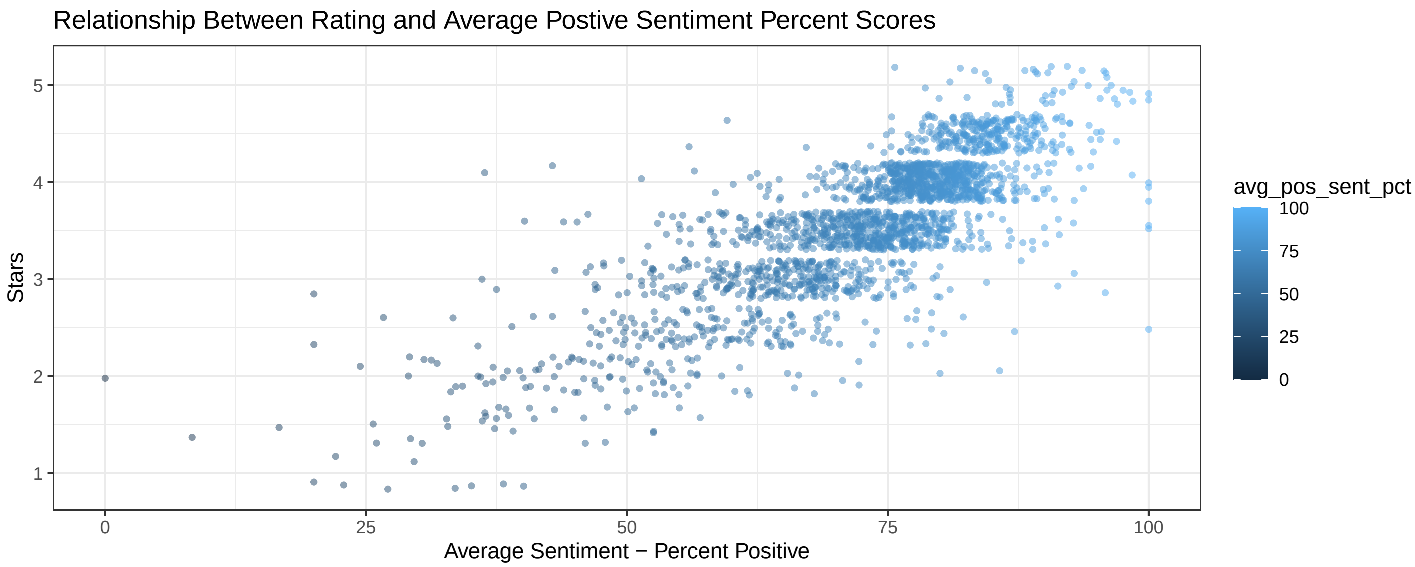


Figure 2: Relationship between stars and Review Sentiment

Additionally, I noticed that there were different frequencies of ratings across postal codes. Furthermore, there was a slight difference in the distribution of ratings between postal codes with at least one tourist attraction and those without(`tourist`).

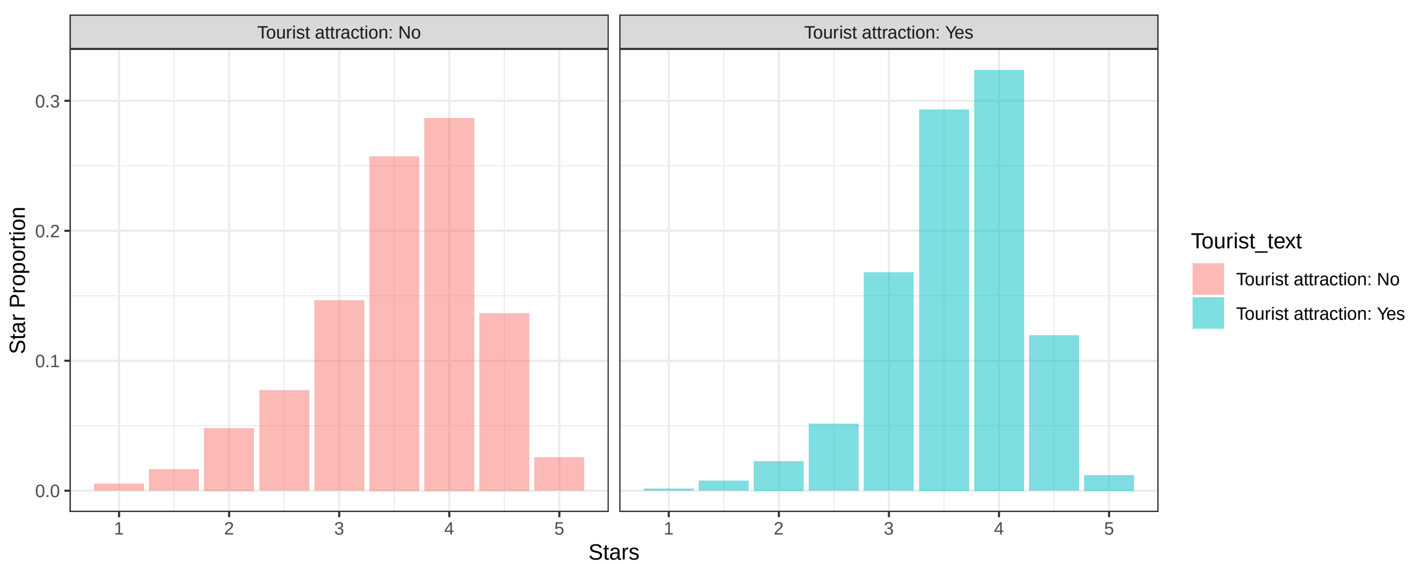


Figure 3: Proportion of Stars by Tourist Indicator

Another interesting relationship that stuck out to me was the relationship between restaurants with at least one other sister restaurant(`relations`) and the ratings. As can be seen in Figure 3, there appears to be a negative relationship between number of relations and the ratings. As number of relations increases, the star rating decreases. These slopes differ slightly between tourist locations and non-tourist locations. Note: The varying colors represent the different postal codes.

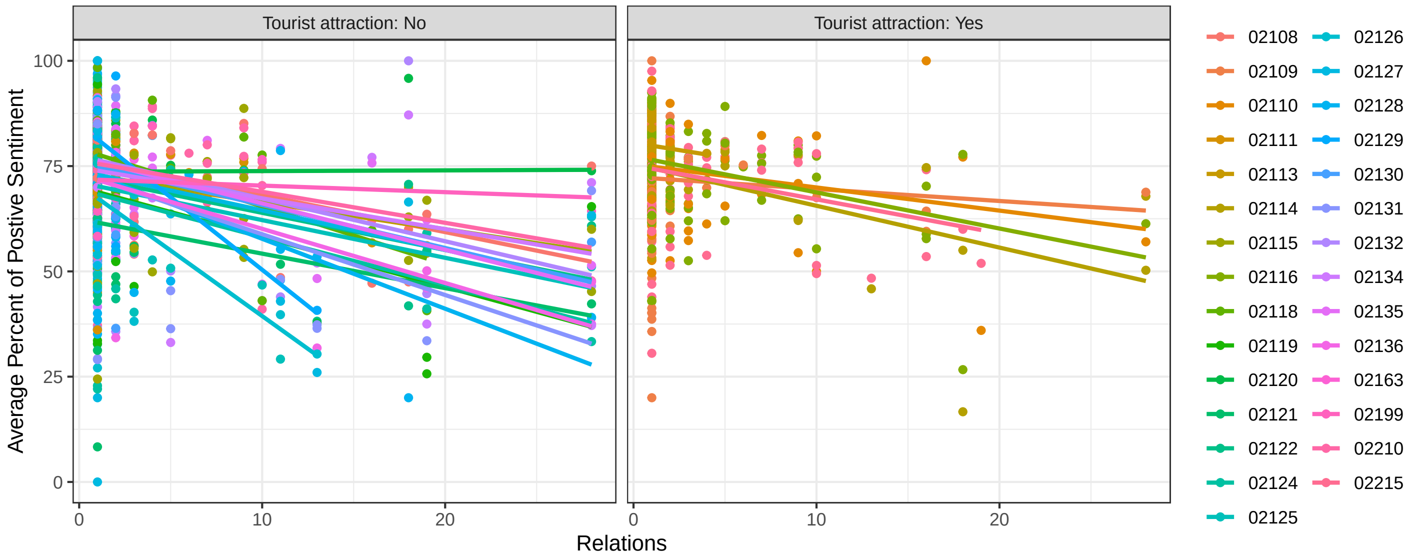


Figure 4: Distribution of Stars by Number of Relations

In addition two those two variables, I added type of restaurant, average number of reviews in a postal code(foot traffic) and population per postal code. More EDA plots can be found in the appendix.

Results

Model Fitting

Below is the multilevel model I used. This model uses postal code as the varying intercept and Tourist as the varying slope.

```
model <- stan_glmer(avg_pos_sent_pct_scaled ~
                    Relations_scaled  + Population_scaled + Tourist +
                    average_num_reviews_scaled  + (1+Tourist|postal_code) +
                    (1+Tourist|Restaurant_type), data = Yelp_data_final)
```

Fixed Effects

Variable	Estimate	s.d.	2.5%	97.5%
Relations_scaled	-0.32	0.02	-0.36	-0.28
Population_scaled	-0.03	0.05	-0.13	0.07
Tourist1	-0.02	0.14	-0.31	0.24
average_num_reviews_scaled	0.15	0.05	0.05	0.26

Random Effects

Varying Intercept

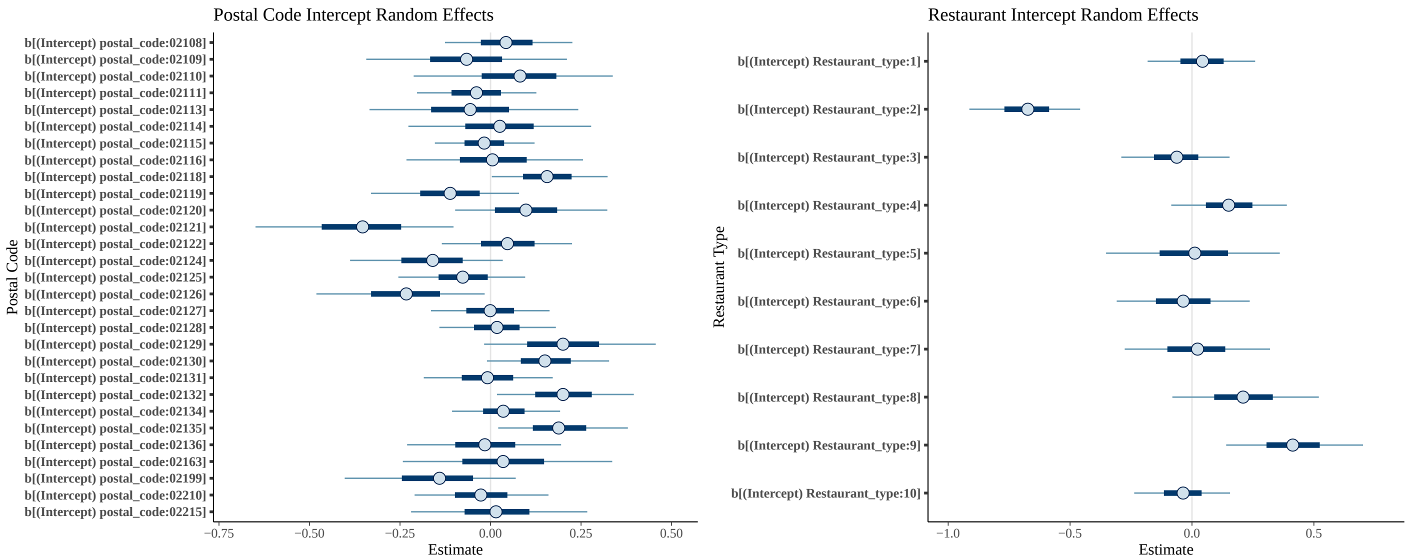


Figure 5: Random Effects Intercepts

Varying Slope

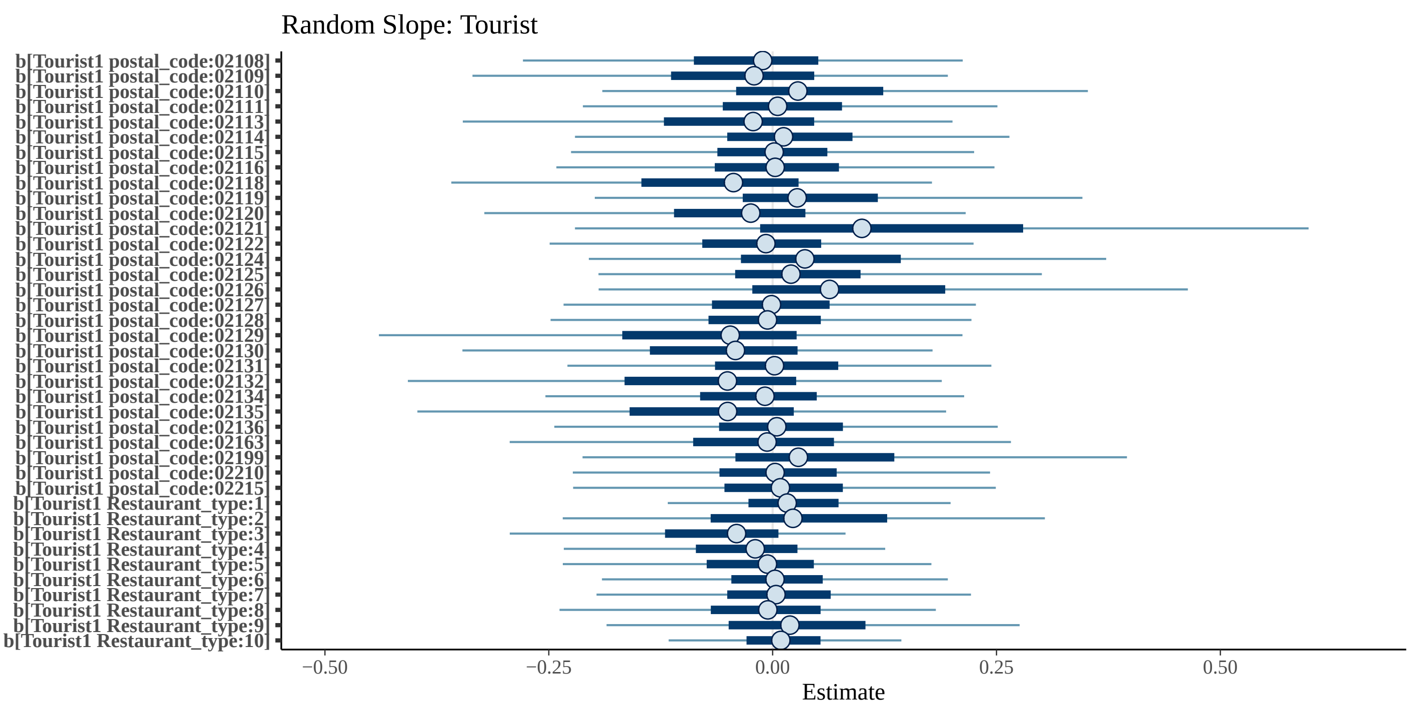


Figure 6: Random Effects Slopes

Discussion

Based on the fixed effects table, it appears that the only fixed effects with any significant impact on the the average positive percent sentiment are relations and the average number of reviews. Although, these two predictors are significant they still only have a very minor effect. However, it is interesting to note that as the number of relations increases the average percent of positive sentiment within reviews for a restaurant will decrease. An example of a restaurant with a lot of relations would be McDonald's, which it is not all too surprising that it may have a lower percent of positive sentiment within it's reviews. It also interesting to note that as the number of reviews increase the average positive percent of positive sentiment in reviews also increases. This could be an indication that people are more inclined to review a restaurant when they have had a positive experience, or that restaurants with positive reviews attract more customers and therefore have a higher number of reviews. It is not possible to say what is causing this relationship, but it is interesting.

Judging by the random effects intercept plots, one can see that roughly 4 of the postal codes (Dorchester, Mattapan, West Roxbury, Brighton) had any true effect on the intercept, and only 2 of the Restaurant types (Chinese and Mediterranean) had any true effect on the intercept. However, across the board there is a minimal effect; all estimates are less than 1 in either direction.

Lastly,the random effects slope plot shows no significant difference in the slopes within an area with a tourist attraction and an area without a tourist attraction. This along with the fixed effects seems to indicate that tourism does not have much of an effect on the percent of positive sentiment within reviews.

This is not all that surprising given what was observed throughout the EDA, however I do believe it was still worth exploring. Given the magnitude and lack of significance of the estimates, this model does not tell us much about how the sentiment of the reviews are impacted by things such as postal code, tourism, or even type of restaurant. Something that may be interesting to explore further would be to replace postal code with neighborhood.For example, 'The North End' neighborhood encompasses 3 different postal codes. However, I am not optimistic that this would make much of a difference. Another interesting thing to potentially explore further would be the relationship between chain restaurants and ratings. Given that `relations` was one of the few variables with significance and a clear connection to the EDA plots, there may be more to explore there. This may shift the question to be from more of a business perspective; seeing what areas respond better to known names such as McDonald's and Dunkin Donuts.

Limitations

Some of the limitations of this model come from the variables themselves. The two variables that lead to the most concern would be `relations` and `average_num_reviews`. `Relations` is created purely based on name and manual research. Although I am fairly confident in identifying the larger chains such as McDonald's and Starbucks, when it come to the smaller local restaurants with one or two sister restaurants this became harder to identify. Additionally, the purpose of using `average_num_reviews` was to account for the foot traffic in the area, however it is possible that this could be swayed by customer experience. For example, if there are bad restaurants in an area there may also be a large number of reviews due reviewers wanting to warn people against going here, therefore this value would not necessarily be solely based on the foot traffic in the area. Looking forward it may be beneficial to attempt to improve the methods for measuring foot traffic and restaurant relations.

Citations/Sources

Citations

- Identifying Boston Zip Codes:
  - https://www.usmapguide.com/massachusetts/boston-zip-code-map/
- Identifying Areas with major Tourist Attractions:
  - https://www.brewsandclues.com/bostons-top-10-must-visit-tourist-destinations/
- Identifying Most Popular 'ethnic' cuisines:
  - https://blogs.voanews.com/all-about-america/2015/05/18/top-10-most-popular-ethnic-cuisines-in-us/
- Model Checking
  - https://www.ssc.wisc.edu/sscc/pubs/MM/MM_DiagInfer.html

Data Sources:

- Yelp Data:
  - https://www.yelp.com/dataset/download
- Population Data:
  - https://www.massachusetts-demographics.com/zip_codes_by_population
- Median Income Data:
  - http://zipatlas.com/us/ma/zip-code-comparison/median-household-income.6.htm

Appendix

Data Cleaning and Processing

Yelp Data

In order to begin exploring the Yelp data there was a bit of processing that needed to occur in order to break it into smaller pieces that could be handled by R. I took the following steps:

1. Extracted the data from json files and converted them into csv files using Python(more specifically the Pandas package within Python).
2. Completed some initial exploratory data analysis within Python to get a sense of what cities/states exist in the data.
3. Created an SQL database to store the CSV files.
4. Explored and subset the business and reviews data down to just Massachusetts Restaurants. Only keeping reviews from 2016 to 2021.
   - Note: I am only interested in looking at information within the past 5 years.
5. Pulled the resulting Massachusetts business and reviews csv into R for more in-depth exploratory data analysis.

Once pulled into R, I was able to take a deeper look into the data and completed the following cleaning/processing steps:

1. Subset down to just Boston, Massachusetts postal codes.
2. Removed grocery stores that had slipped through.
3. Manually cleaned restaurant `names` to exclude odd characters, as well as to make sure restaurants that were apart of chains had the same name spelling.
   - For example: "Flour Bakery & Cafe" was sometimes listed as "Flour Bakery + Caf,àö¬©"
   - Note: This step was very important to defining the `Relations` variable that will be described later.
4. Created variables to potentially be used in the multilevel regression. (See Appendix for detailed descriptions)
5. Aggregated data up to the individual restaurant level.

The resulting dataset has 2177 observations representing each restaurant with a Boston postal code, with 20 variables for potential use.

Supplemental Data

The median income and population information by postal code was manually extracted from websites and entered into csv files. These csv files were then read into R and joined to the overall Yelp data by postal code. There was no manipulation needed for these variables.

Codebook:

Variable names	Definition
avg_pos_sent_pct	The percent of a review that was positive averaged over all reviews(2016-2021) for a business. This was created using sentiment analysis.
pricerange	The price range that the business falls into. This was pulled from the Yelp `attributes` column.
Relations	The number of related/sister restaurants a business has. This was created by counting the number of restaurants with the same name.
alcohol_r	What kind of alcohol is served at a restaurant (None, Beer and Wine, Full bar).
Restaurant_type	1 - Italian, 2 - Chinese, 3- Mexican, 4 - Japanese, 5 - Greek, 6 - Thai, 7 - Spanish, 8 - Indian, 9 - Mediterranean, 10 - Other
italian	Whether a restaurant indicated that they serve Italian food.
chinese	Whether a restaurant indicated that they serve Chinese food.
mexican	Whether a restaurant indicated that they serve Mexican food.
japanese	Whether a restaurant indicated that they serve Japanese food.
greek	Whether a restaurant indicated that they serve Greek food.
thai	Whether a restaurant indicated that they serve Thai food.
spanish	Whether a restaurant indicated that they serve Spanish food.
indian	Whether a restaurant indicated that they serve Indian food.
mediterranean	Whether a restaurant indicated that they serve Mediterranean food.

Variable names	Definition
average_num_reviews_scaled	The average number of reviews within a postal code over the last 5 years (2016-2021).
Population	The population by postal code.
median_income	The median income by postal code
postal_code	The postal code indicator.
Tourist	Whether or not restaurant exists in a postal code that has a major tourist attraction.

Model Checking

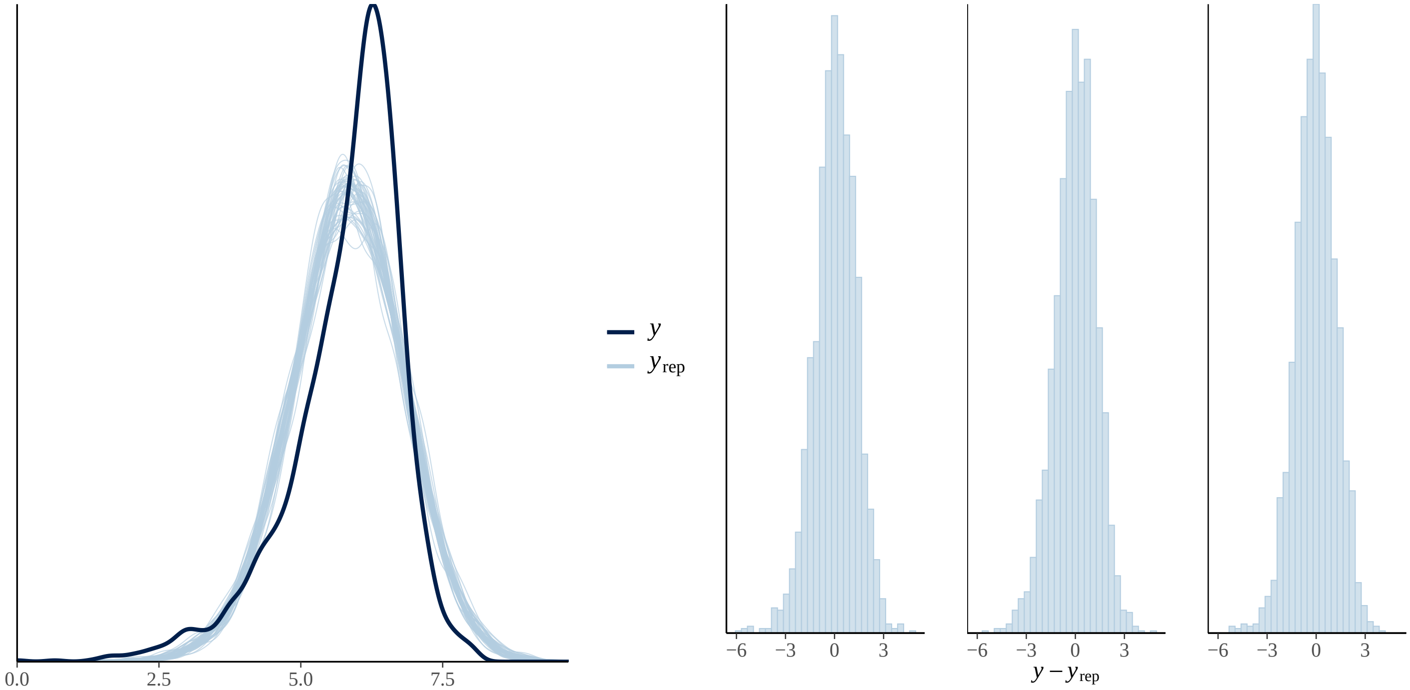


Figure 7: PP Check

Sentiment Analysis

Word Clouds

Sentiment Plots

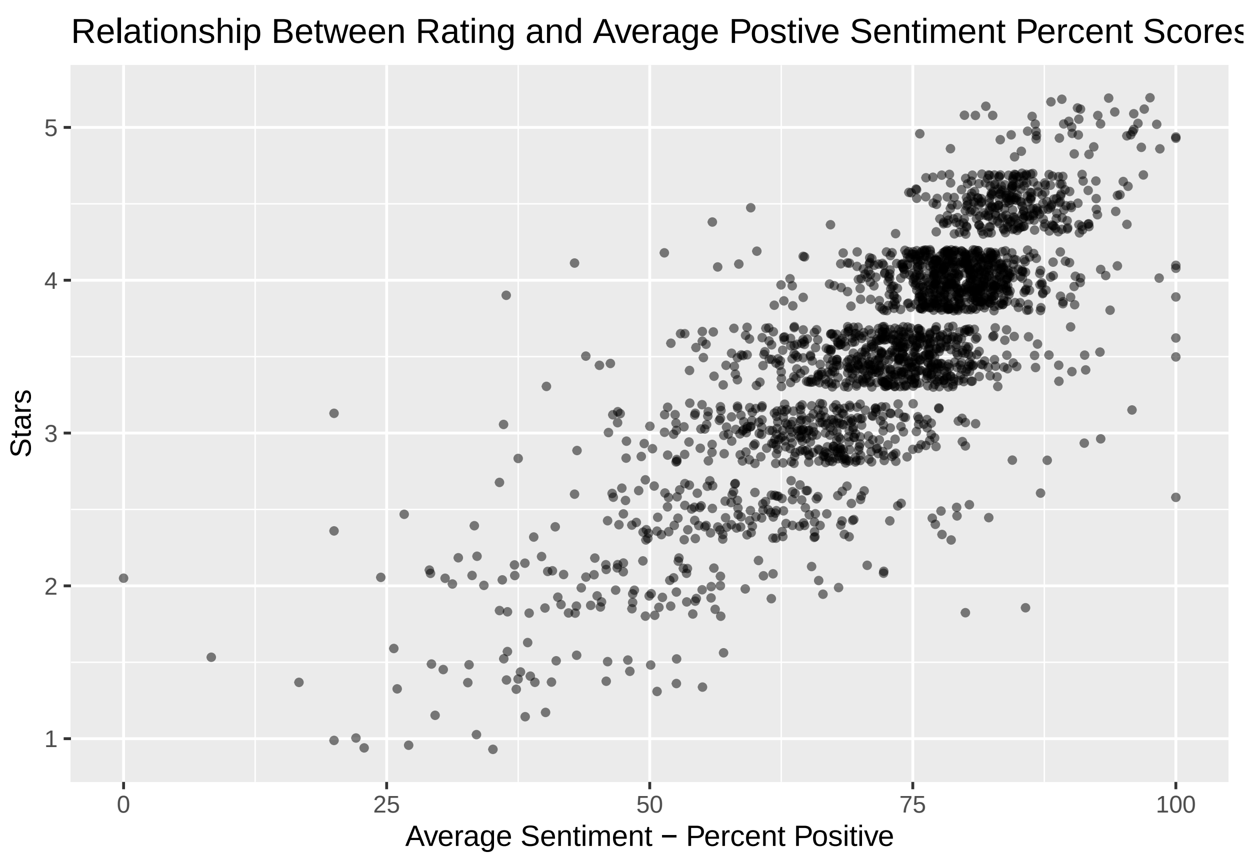


Figure 8: One Star Reviews

Figure 9: Two Star Reviews

Figure 10: Three Star Reviews

Figure 11: Four Star Reviews

Figure 12: Five Star Reviews

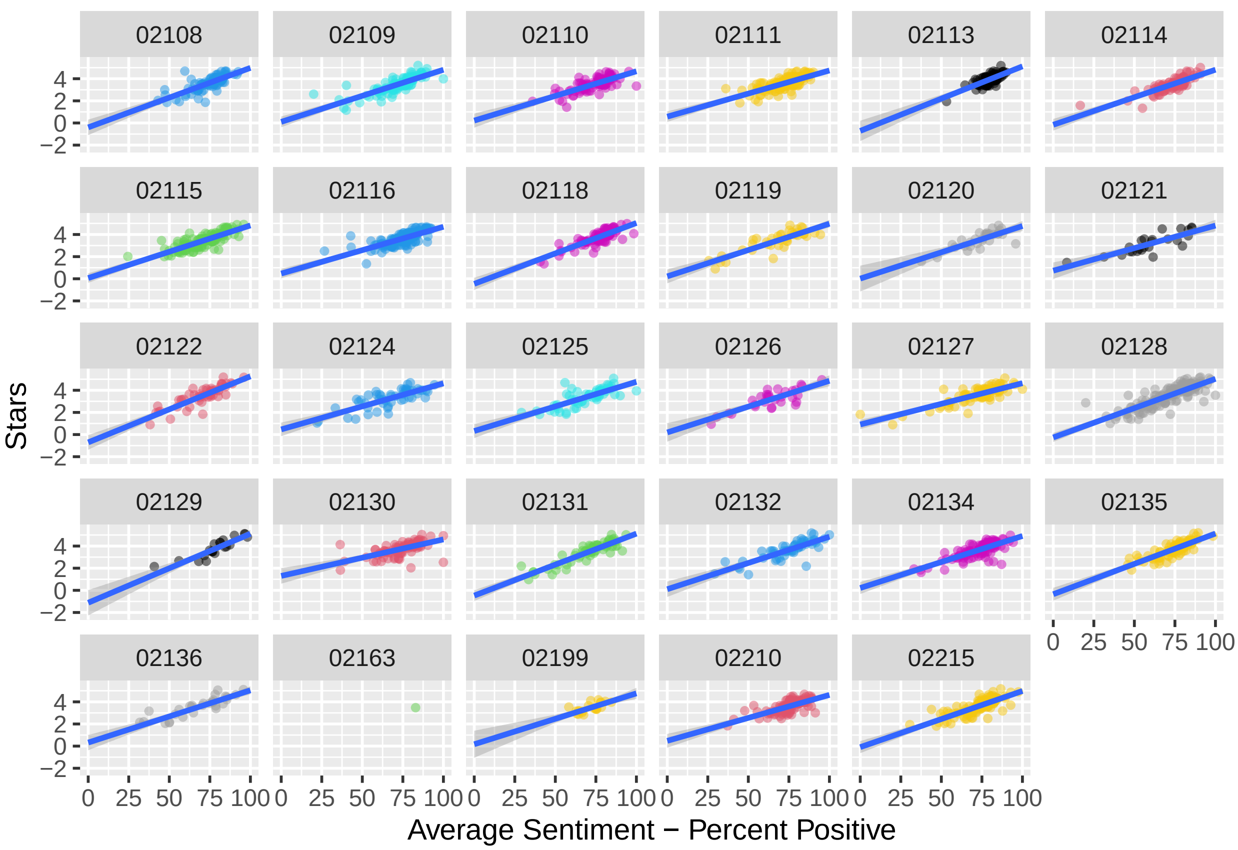


Figure 13: Relationship Between Rating and Average Postive Sentiment Percent Scores

Plots

Distribution of Stars by Postal Code

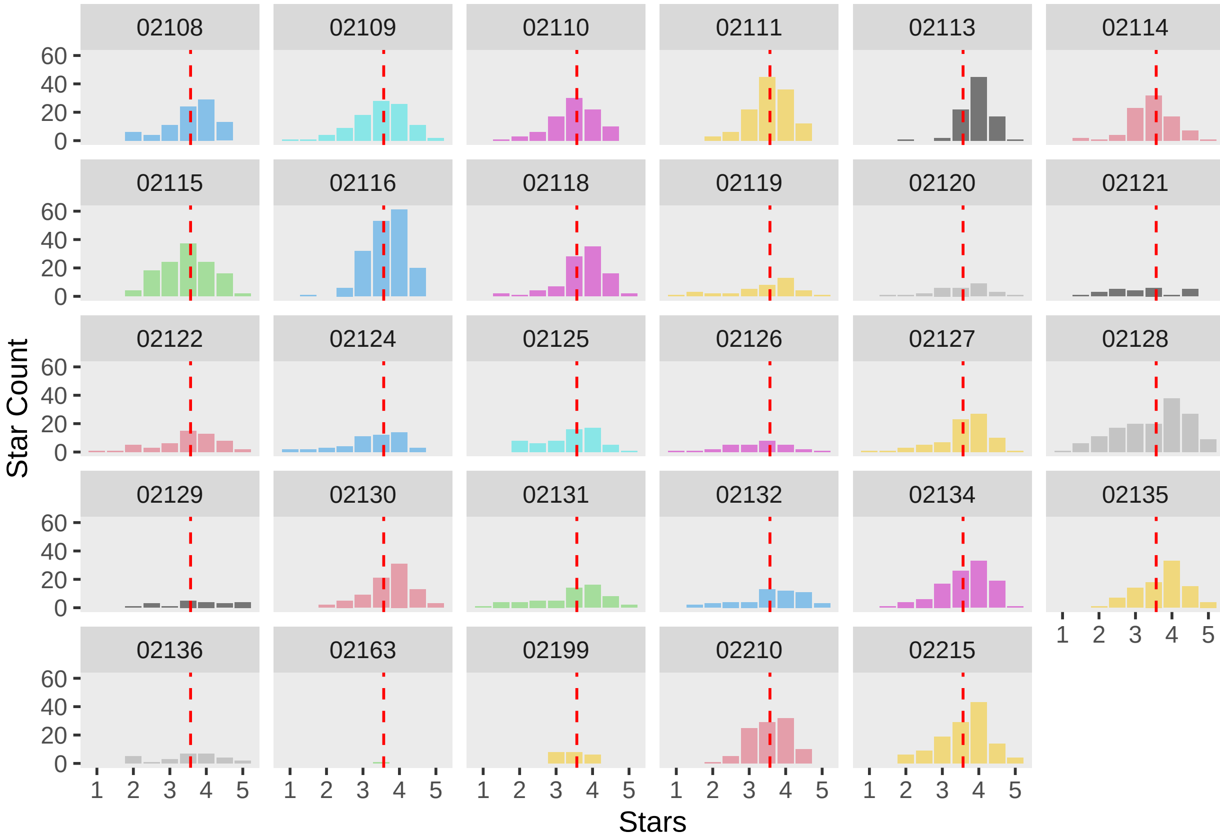
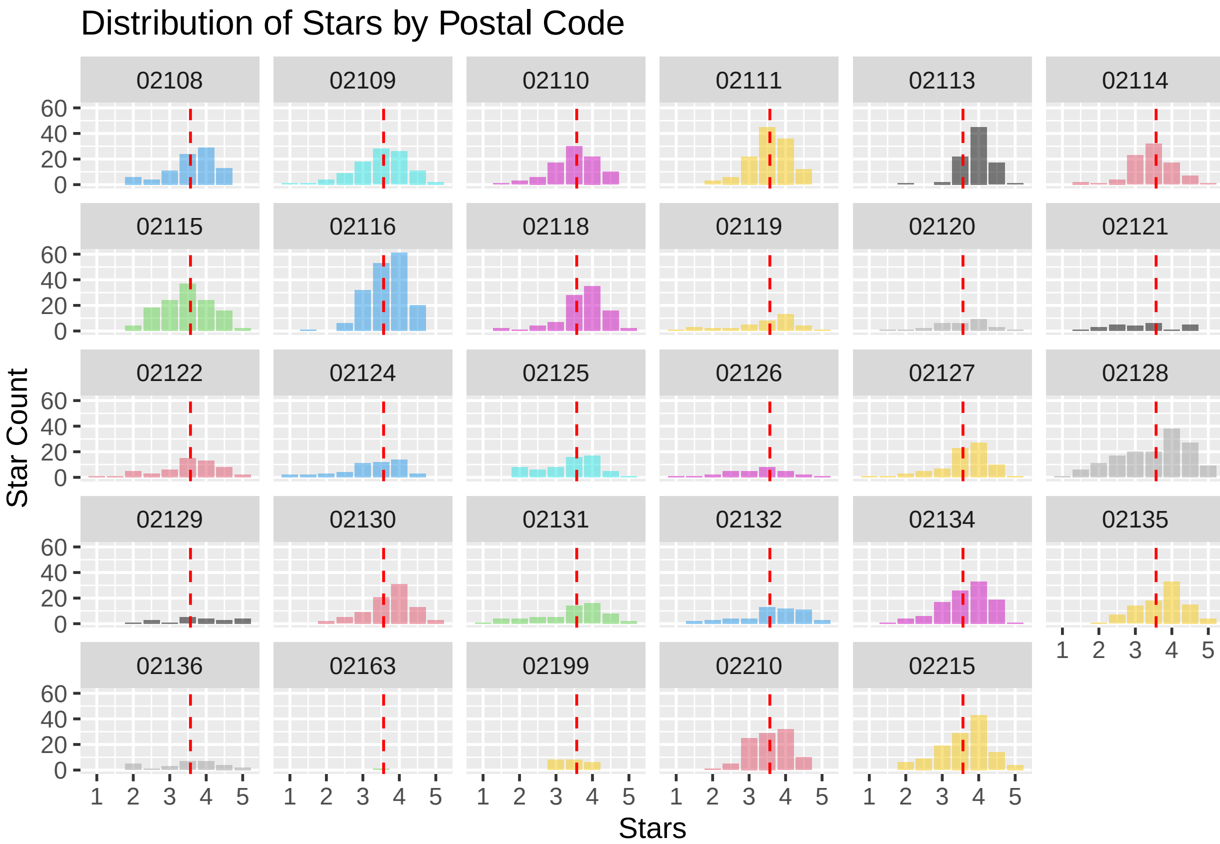
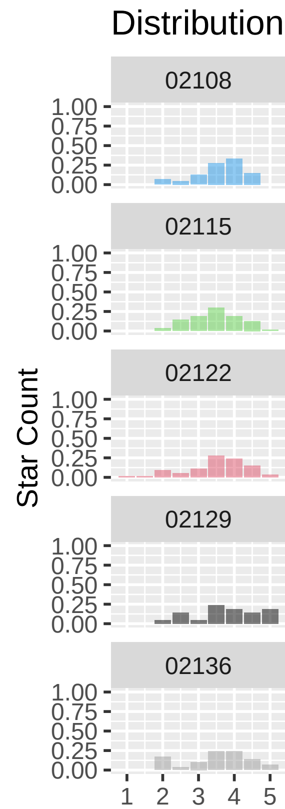
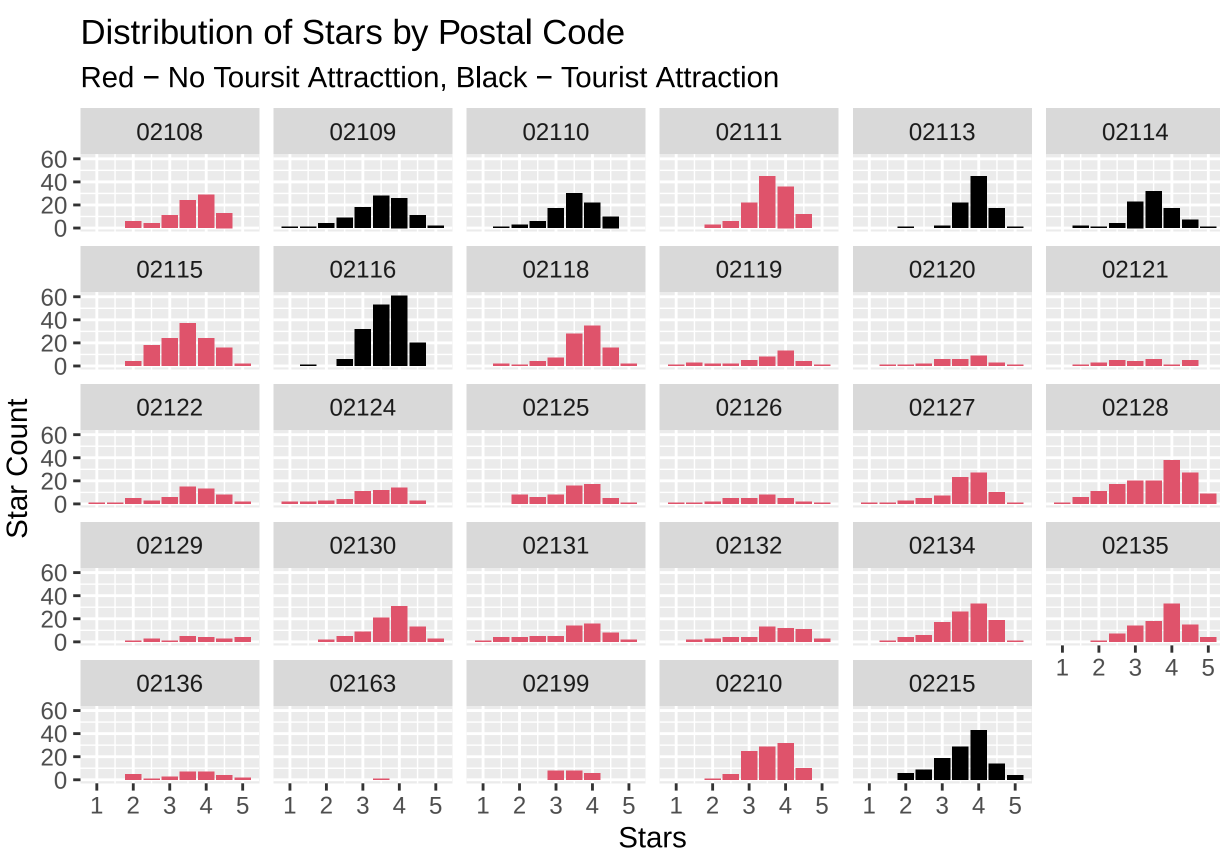


Figure 14: Distribution of Stars by Postal Code

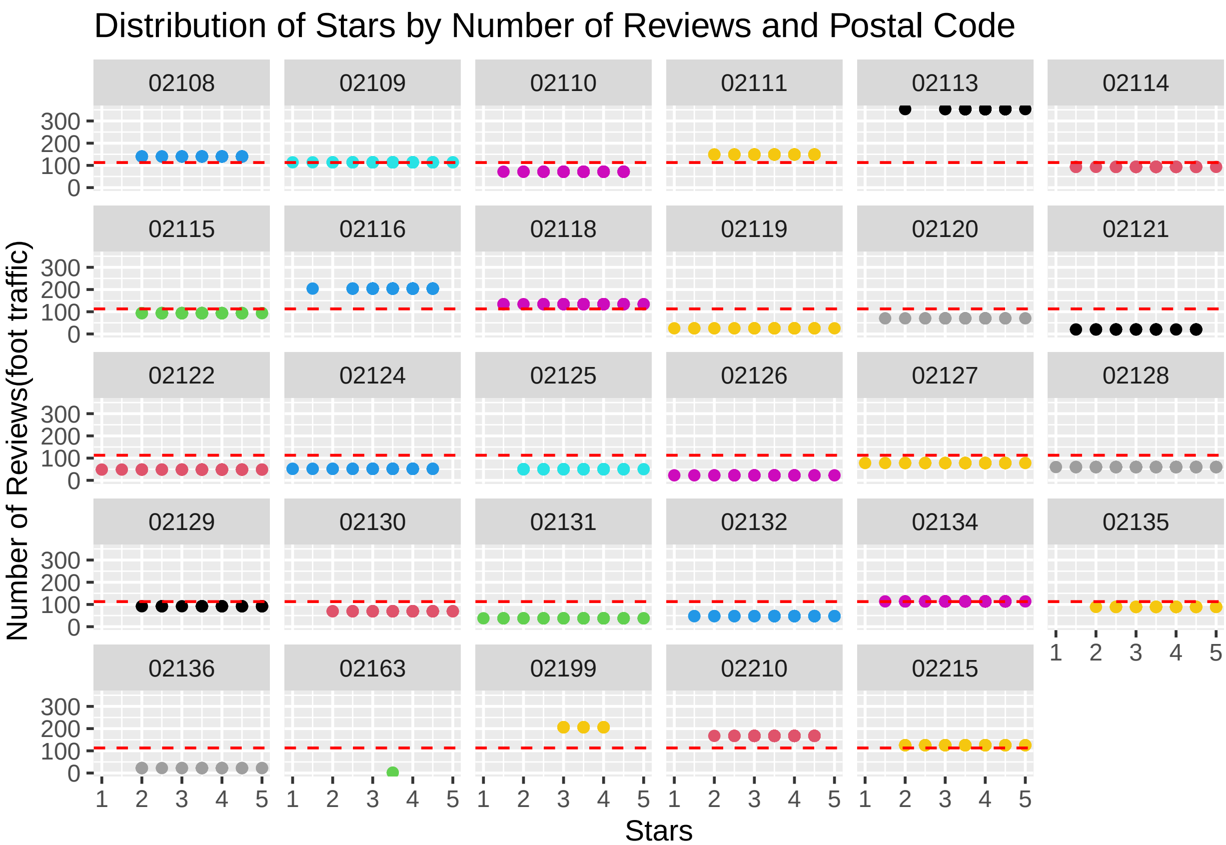




Tourist Plots

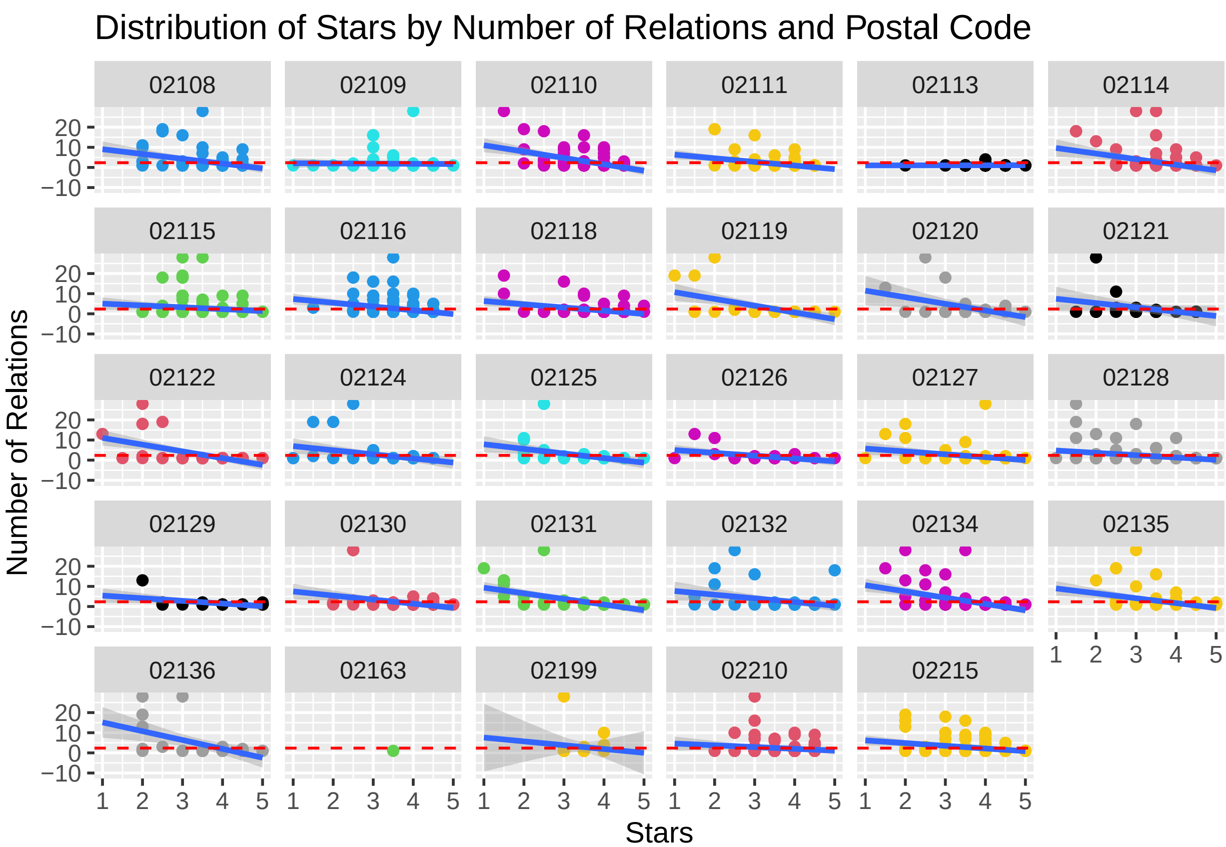


Foot Traffic Plots

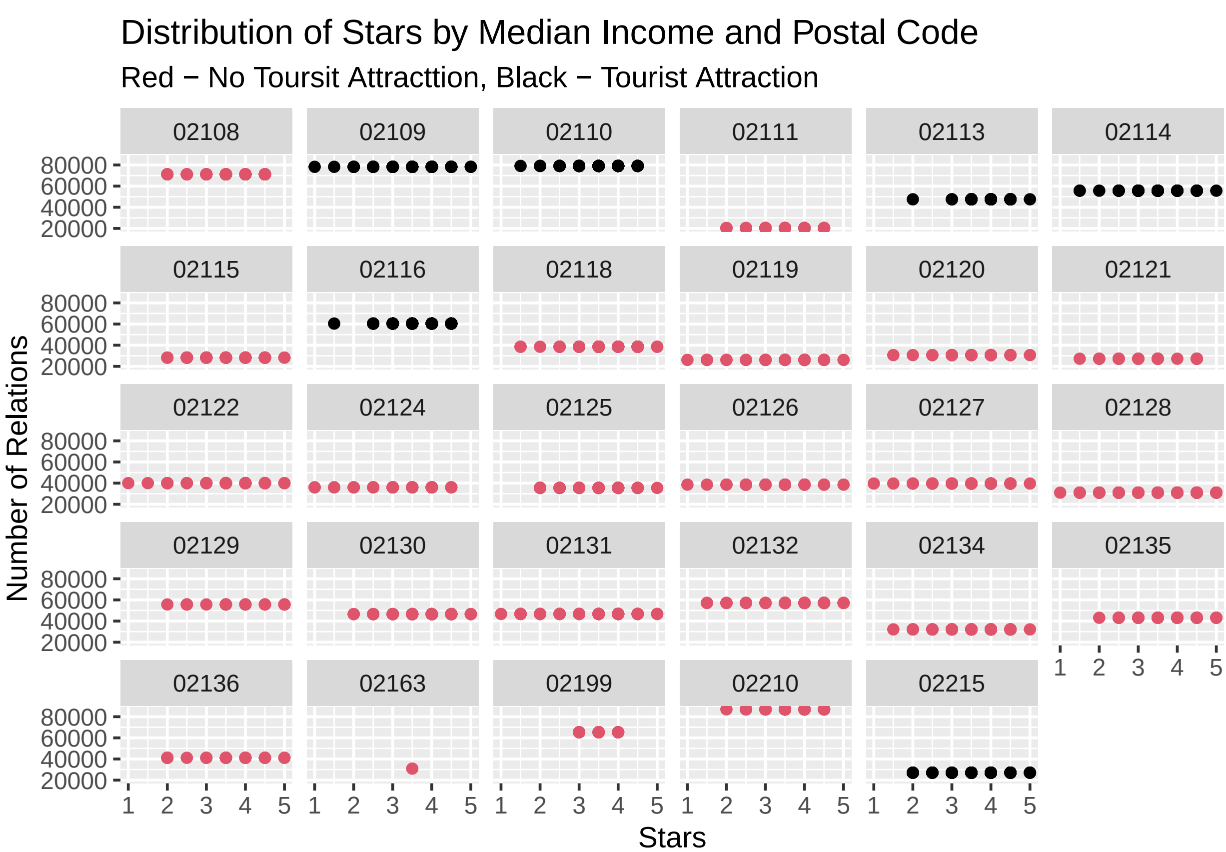


Relations Plots

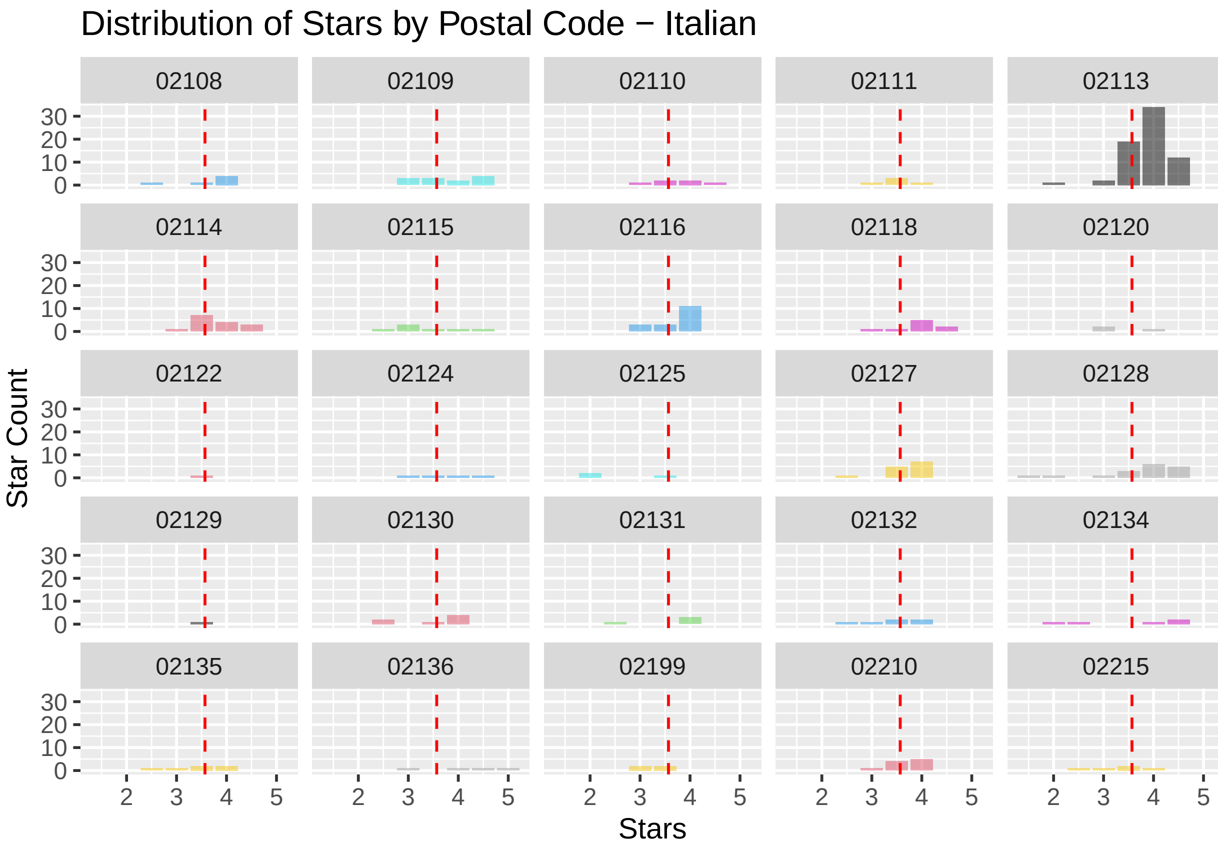
```
## `geom_smooth()` using formula 'y ~ x'
```

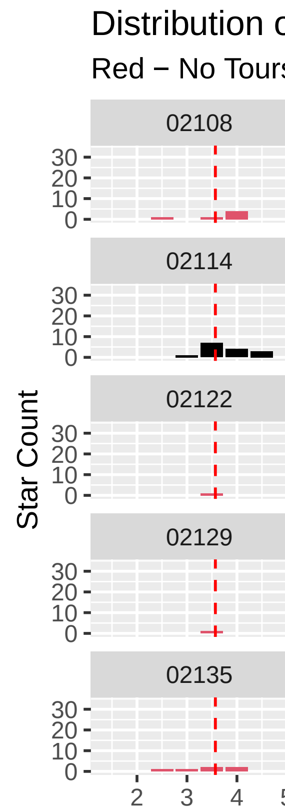


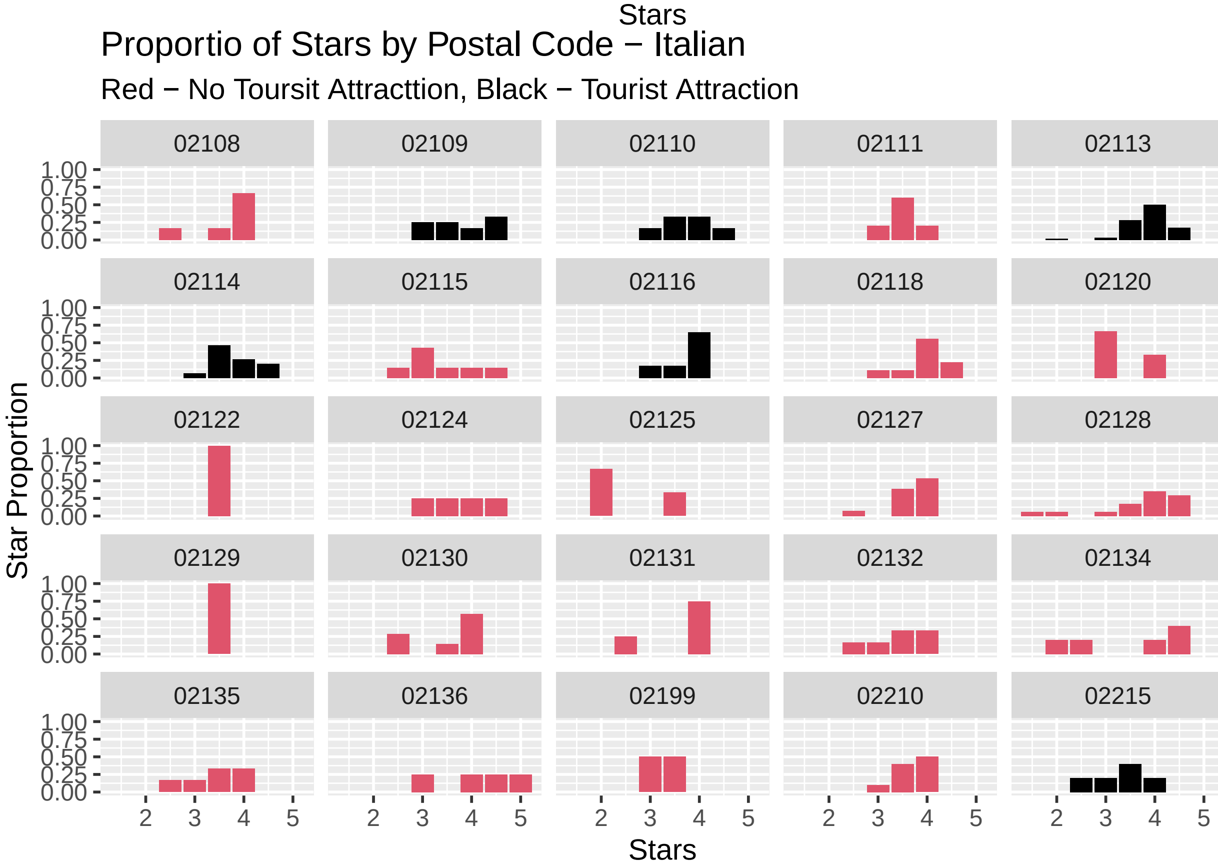
Median Income Plots

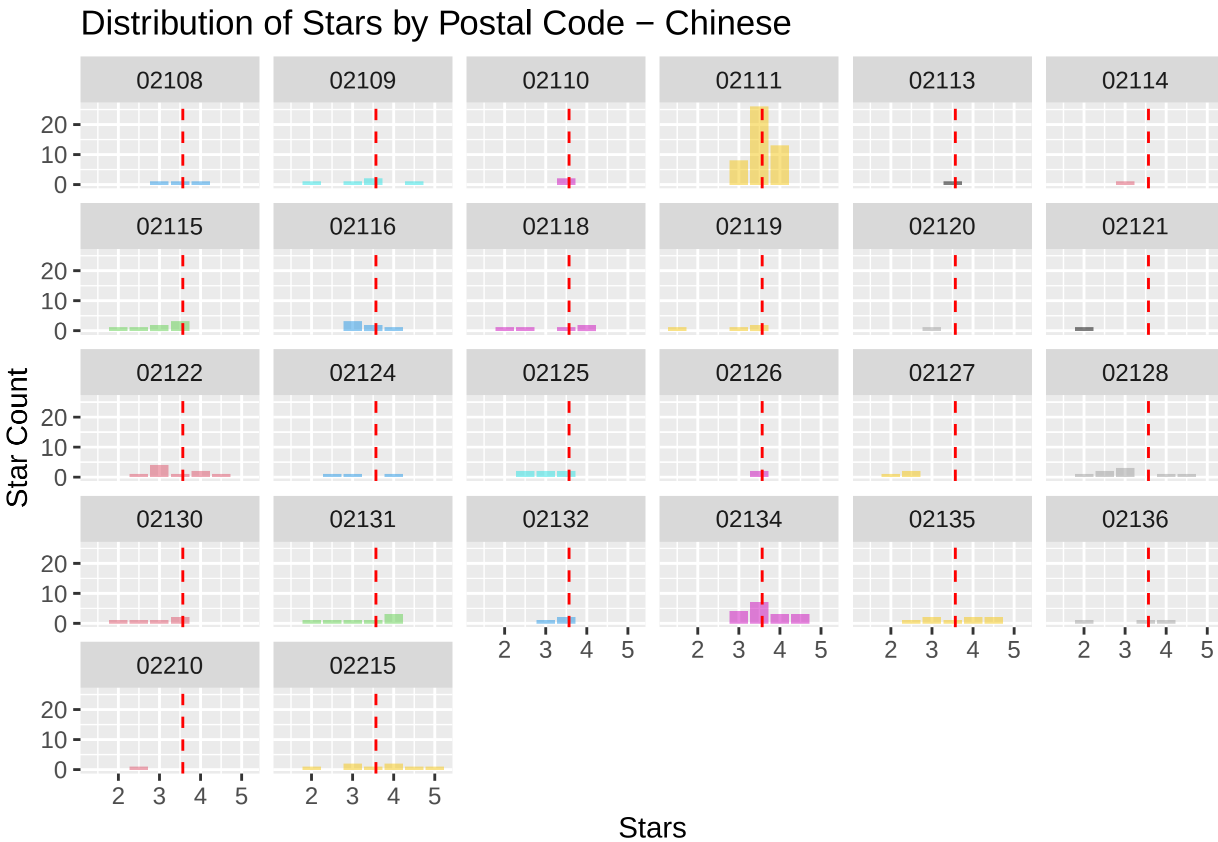


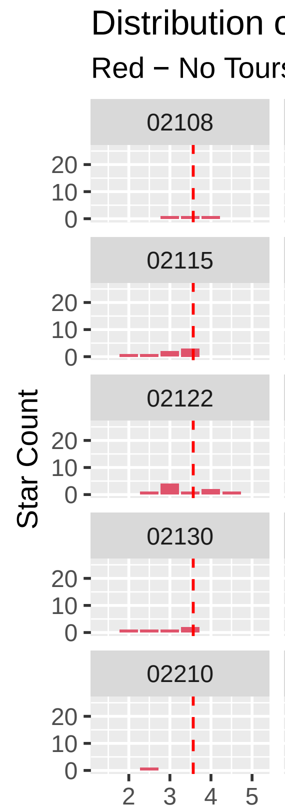
Restaurant Type Plots

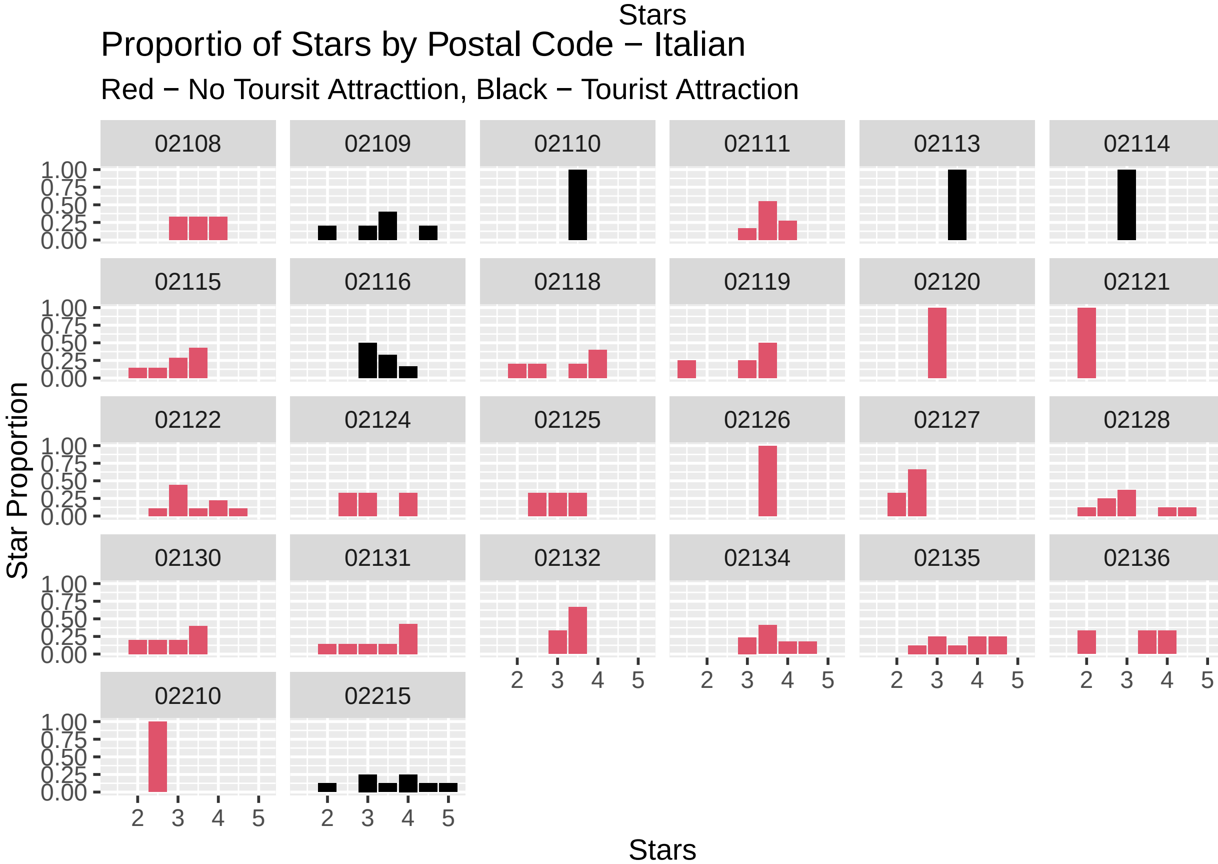







Distribution of Stars by Postal Code − Chinese
02108
02109
02110
02111
02113
02114
02115
02116
02118
02119
02120
02121
02122
02124
02125
02126
02127
02128
02130
02131
02132
02134
02135
02136
02210
02215
Star Count
Stars


Red − No Tours
02108
02115
02122
02130
02210
Star Count


Proportio of Stars by Postal Code − Italian
Red − No Toursit Attracttion, Black − Tourist Attraction
02108
02109
02110
02111
02113
02114
02115
02116
02118
02119
02120
02121
02122
02124
02125
02126
02127
02128
02130
02131
02132
02134
02135
02136
02210
02215
Star Proportion
Stars