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 ratings from a consumers perspective.

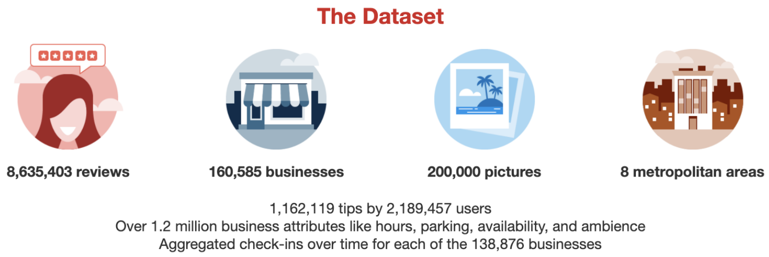

The Dataset
8,635,403 reviews
160,585 businesses
200,000 pictures
8 metropolitan areas
1,162,119 tips by 2,189,457 users
Over 1.2 million business attributes like hours, parking, availability, and ambience
Aggregated check-ins over time for each of the 138,876 businesses

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? Do small mom and pop places rate higher 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

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.

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 first 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`).

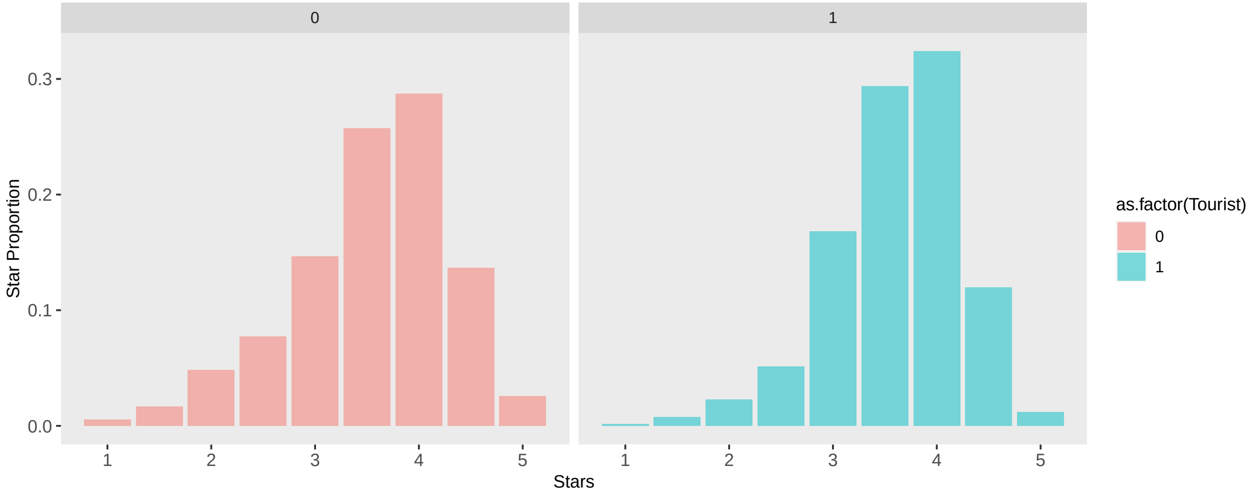


Figure 2: Proportion of Stars by Tourist Indicator

Another interesting relationship that stuck out to me was the relationship between restaurants with at least one other sister restaurant and the ratings(`relations`). 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.

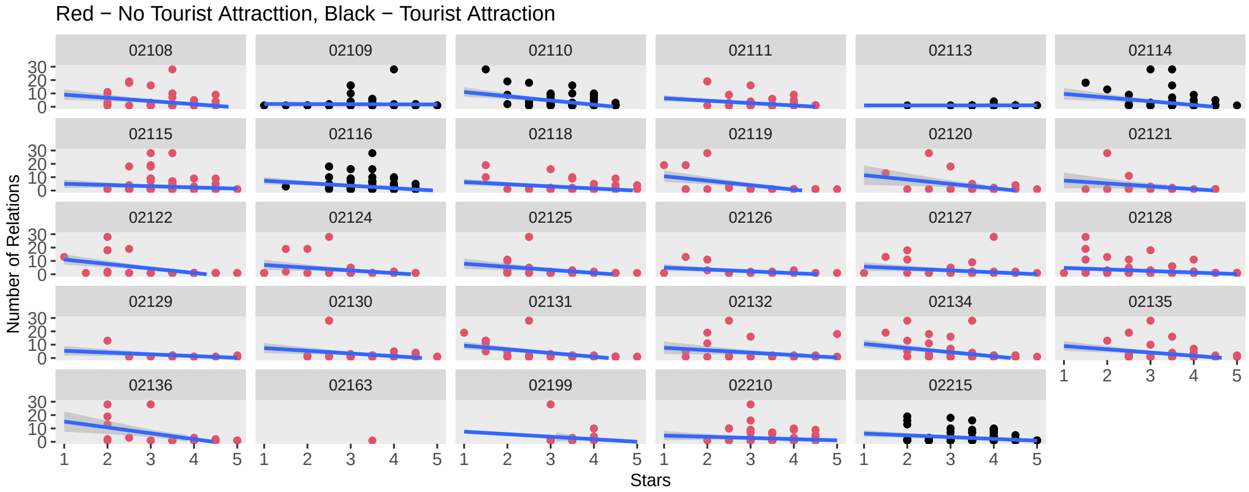


Figure 3: Distribution of Stars by Number of Relations and Postal Code

In addition two those two variables, I added type of restaurant, average number of reviews in a postal code(foot traffic), average positive sentiment review score per business, 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.

```r
model <- lmer(bus_stars ~ avg_pos_sent_pct_scaled  + Relations_scaled  +
                 italian + chinese + mexican + japanese + greek + thai +
                 spanish + indian + mediterranean + Population_scaled +
                 Tourist + average_num_reviews_scaled  +
                 (1+Tourist|postal_code), data = Yelp_data_final)
```

Fixed Effects

	Estimate	S.E.	t val.	d.f.	P-Value
avg_pos_sent_pct_scaled	0.57	0.01	57.12	2145.18	0.00
Relations_scaled	-0.05	0.01	-5.23	2158.42	0.00
italian2	0.05	0.03	1.67	777.85	0.10
chinese2	-0.07	0.04	-1.81	1557.55	0.07
mexican2	0.03	0.04	0.89	2138.51	0.37
japanese2	0.00	0.04	0.10	2157.02	0.92
greek2	-0.11	0.08	-1.32	2161.66	0.19
thai2	-0.10	0.06	-1.64	2159.17	0.10
spanish2	0.02	0.08	0.24	2155.57	0.81
indian2	0.07	0.08	0.80	2157.36	0.43
mediterranean2	-0.06	0.05	-1.14	2159.29	0.26
Population_scaled	0.04	0.01	2.64	15.48	0.02
Tourist1	-0.03	0.03	-1.16	9.47	0.28
average_num_reviews_scaled	0.02	0.01	1.42	8.93	0.19

Random Effects

Postal Code	Intercept	Tourist
02108	-0.004	0.001
02109	0.006	-0.001
02110	-0.007	0.002
02111	0.010	-0.003
02113	0.018	-0.004
02114	-0.014	0.003
02115	-0.028	0.008
02116	-0.018	0.004
02118	-0.011	0.003
02119	0.020	-0.005

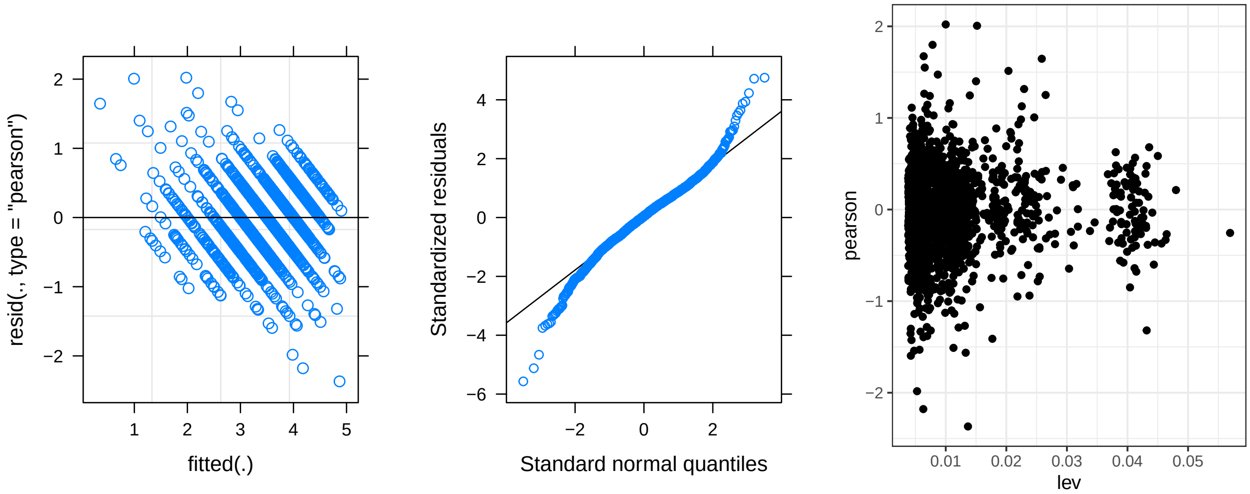


Figure 4: Residual, QQ and Plot

Model Checking

From the plots in Figure 4, the model seems to be doing okay. There is no clear pattern or curve present in the residuals. Additionally, the QQ plot appears to follow a normal distribution. Lastly, the leverage plot does not show any one point having a major influence on the model.

Discussion

Conclusion

Looking at the fixed effects not all of the predictors used are significant at an alpha level of .05, and the majority have very small estimates. Additionally, based on the random effects table it appears that postal code and tourist actually have a very minimal effect. 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 ratings 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 dig into there.

Limitations

Some of the limitations of this model come from the variables themselves. The three 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. Lastly, there is some concern with the use of `avg_pos_sent_pct`. This variable is highly correlated with the outcome(`bus_stars`), and although this does help to predict the stars accurately it does not necessarily assist in answering the question at hand. In a sense, given that `bus_stars` and `avg_pos_sent_pct` are so highly correlated, `avg_pos_sent_pct` could serve as a proxy for `bus_stars`. Looking forward it may be beneficial to attempt a model without `avg_pos_sent_pct`, as well as 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

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).
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.
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.

Sentiment Analysis

Word Clouds

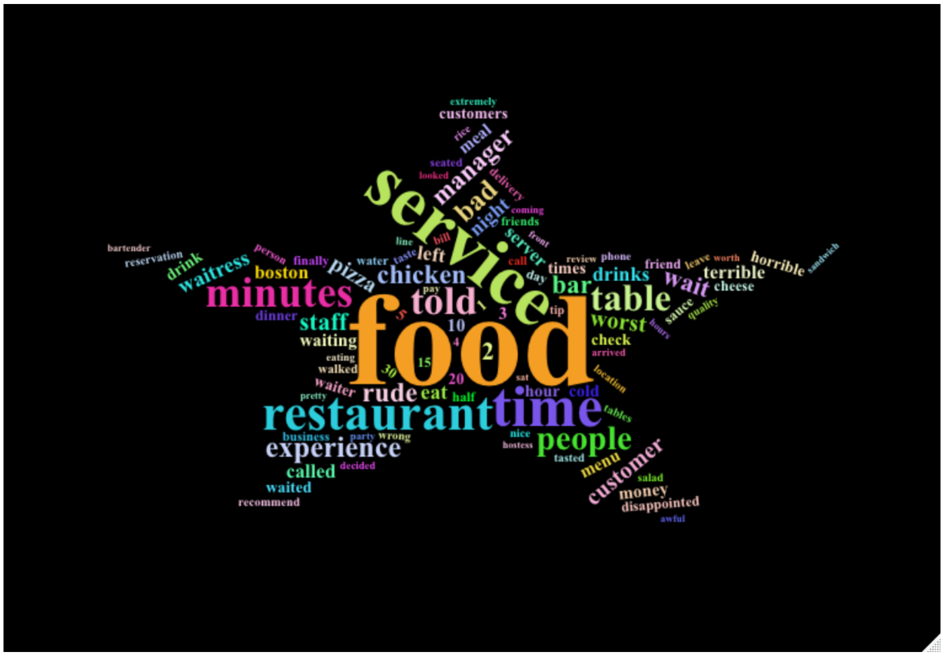


Figure 5: One Star Reviews

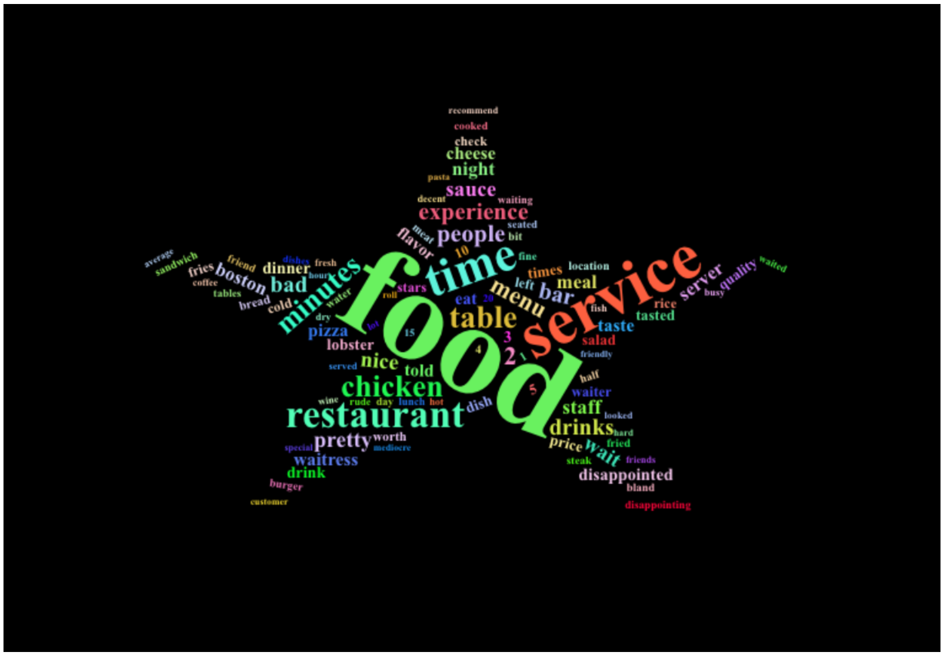


Figure 6: Two Star Reviews

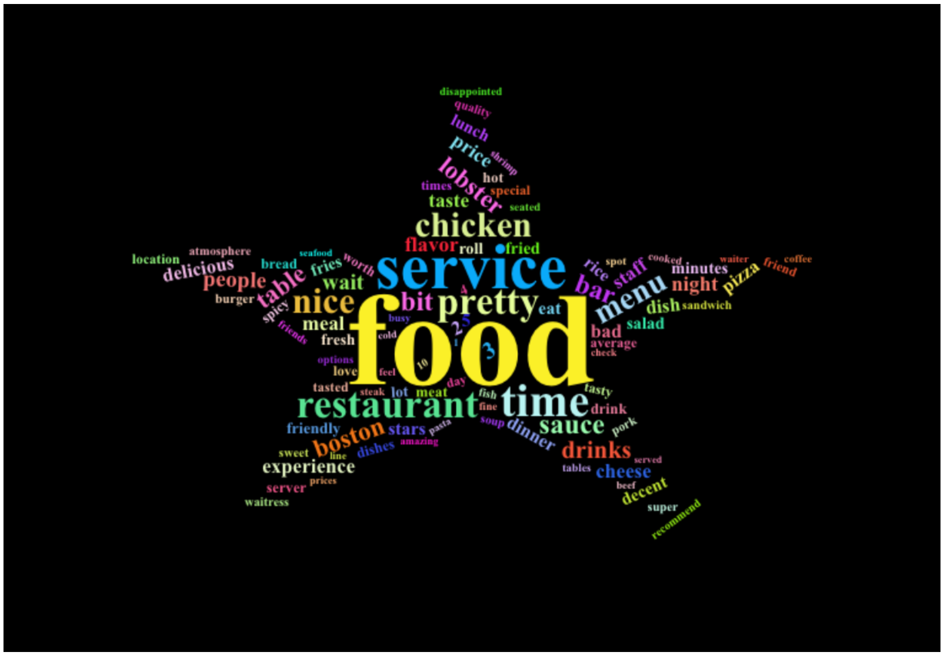


Figure 7: Three Star Reviews

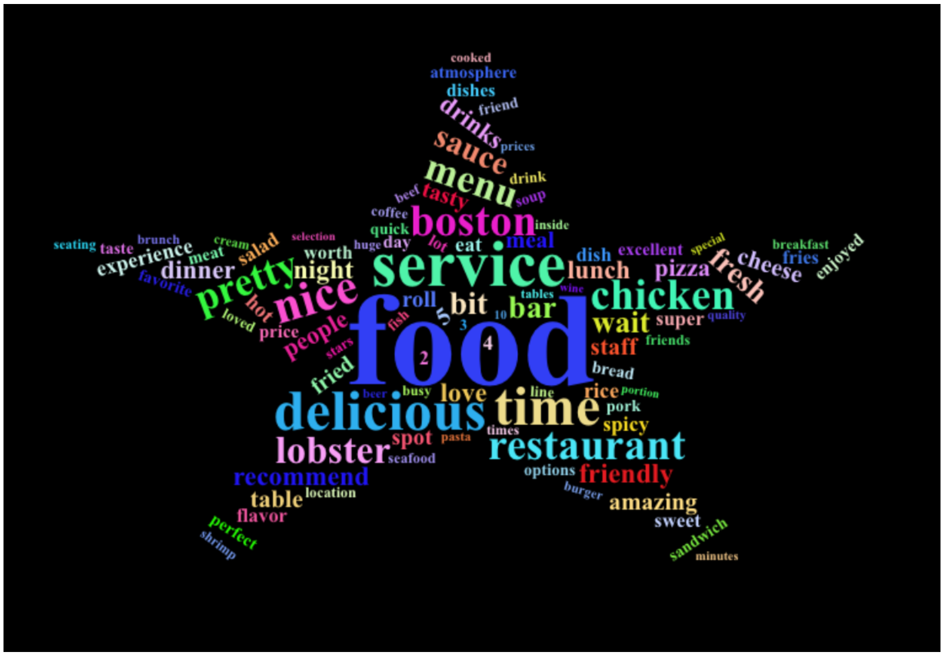


Figure 8: Four Star Reviews

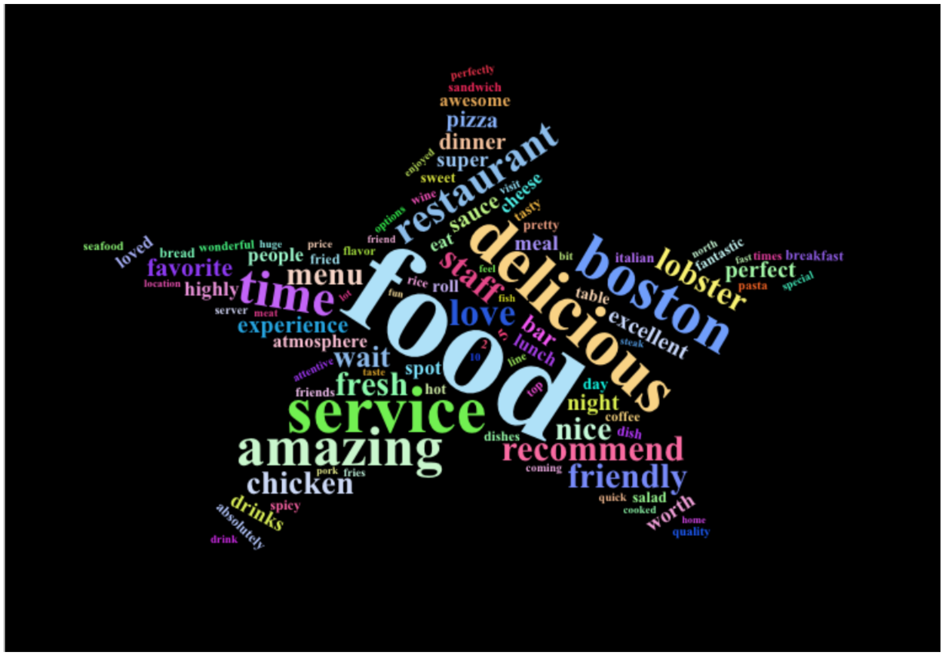
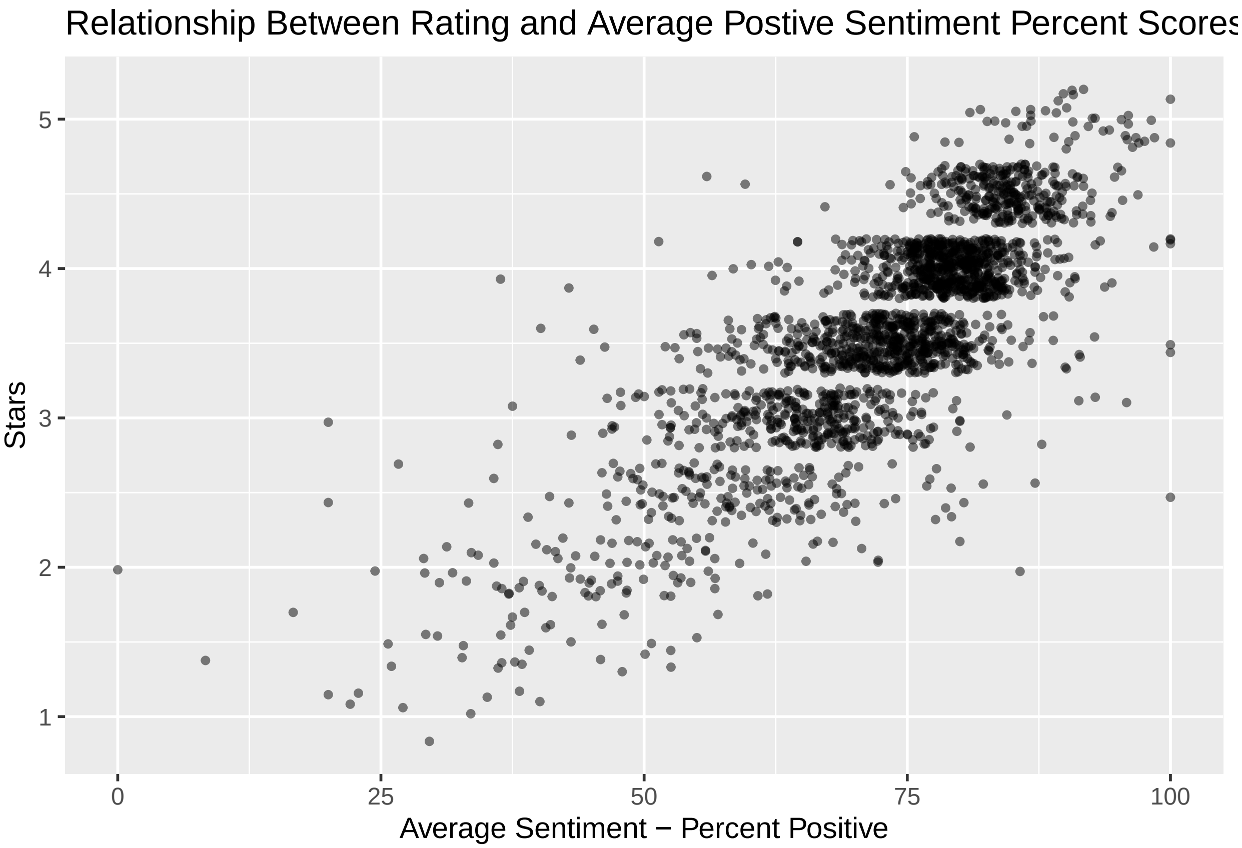


Figure 9: Five Star Reviews

Sentiment Plots



Plots

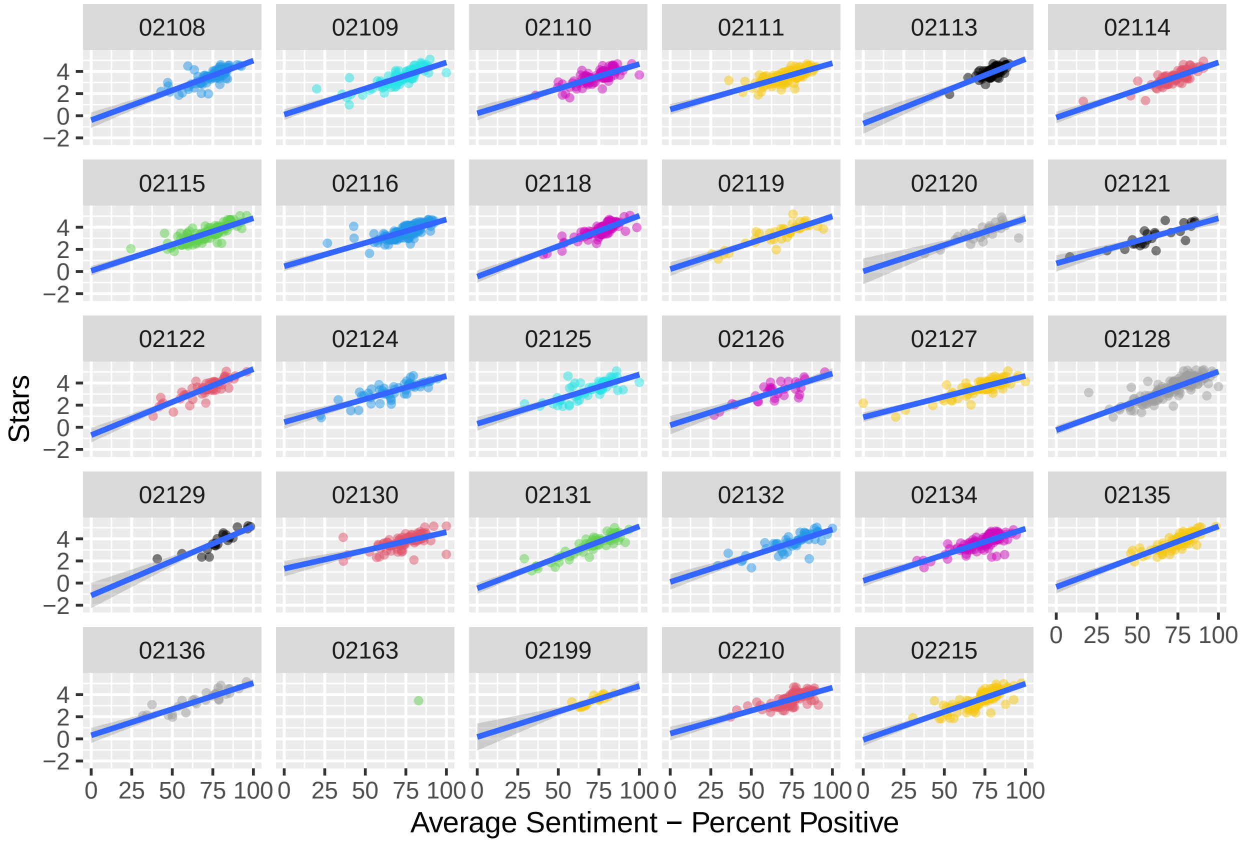


Figure 10: Relationship Between Rating and Average Postive Sentiment Percent Scores

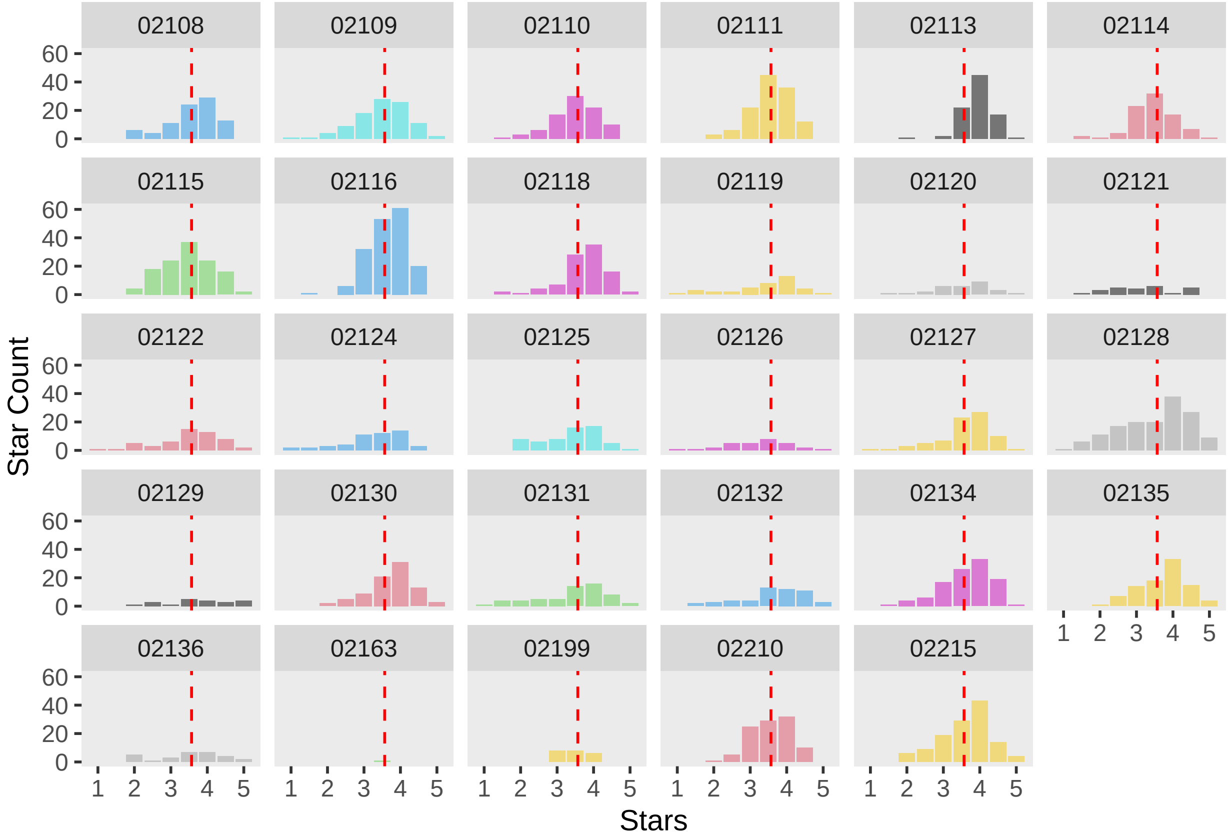
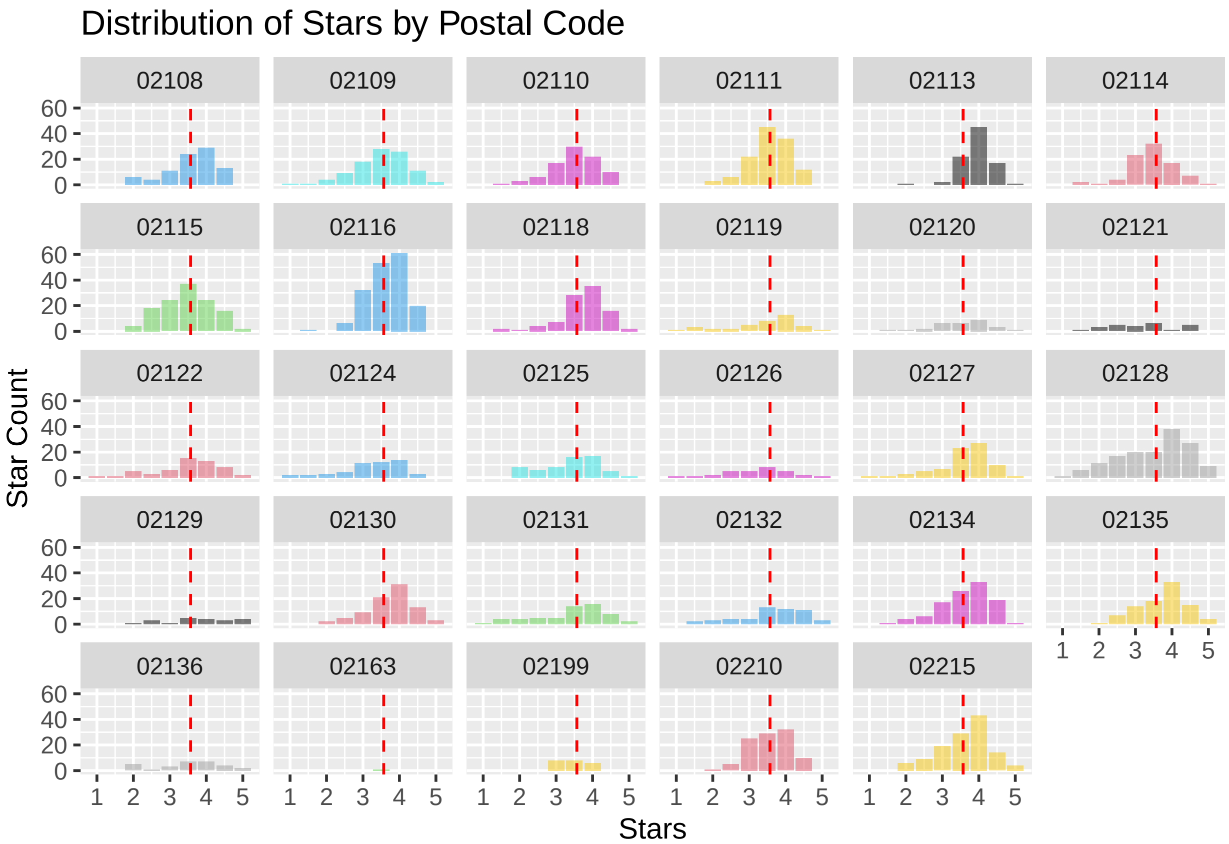
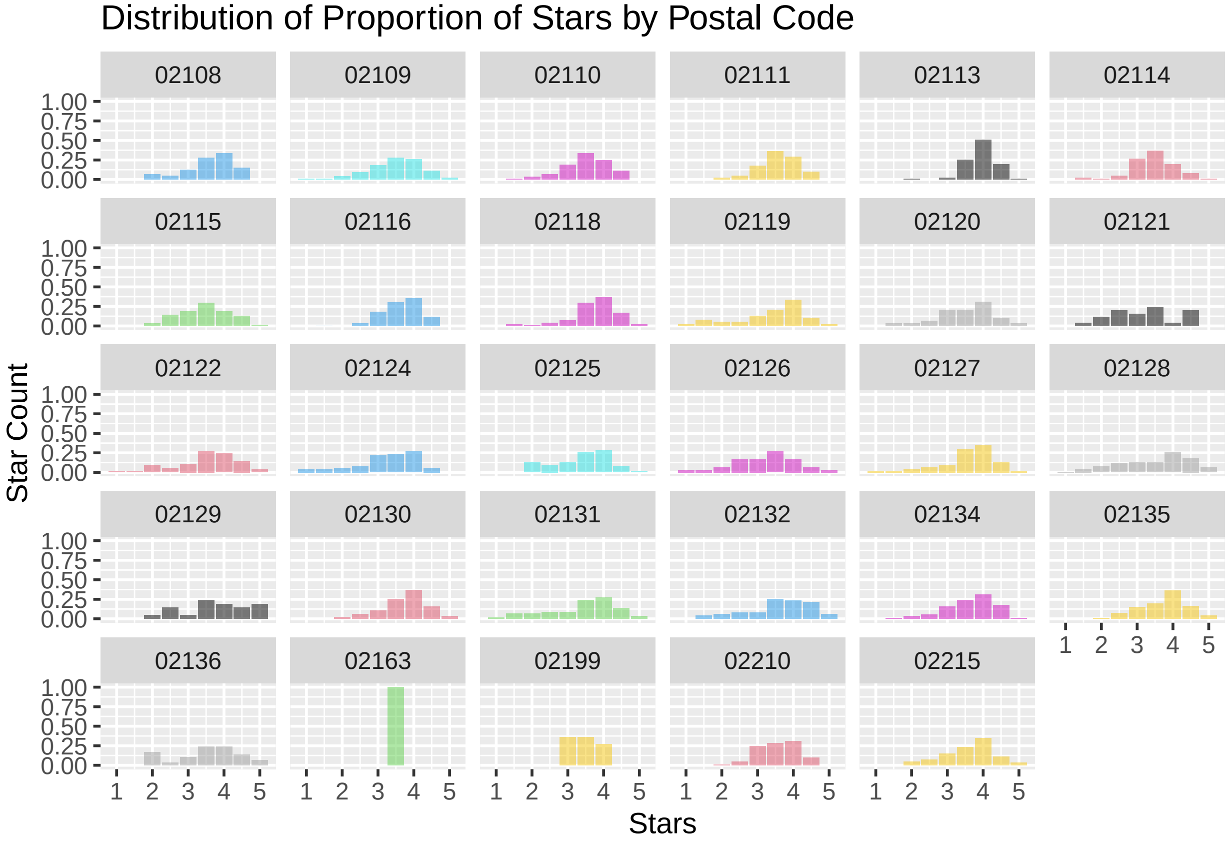


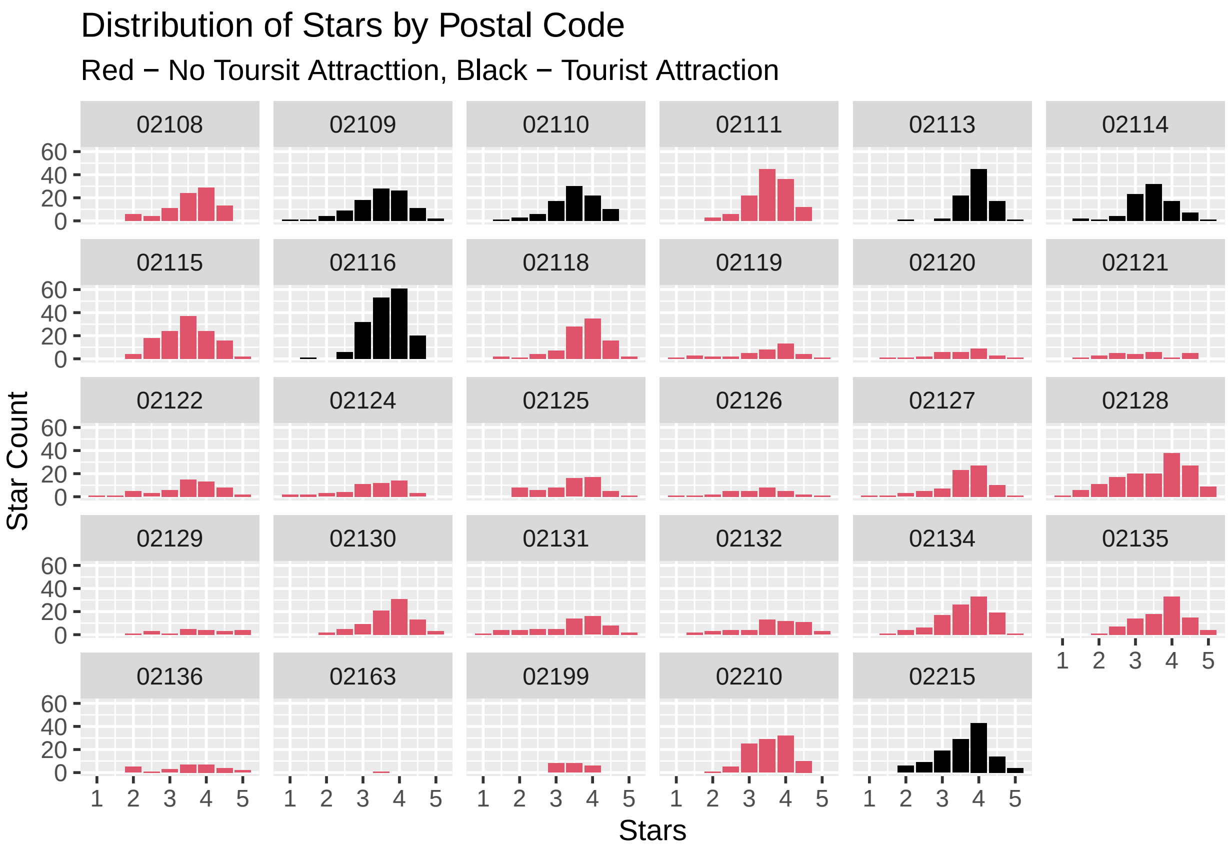
Figure 11: Distribution of Stars by Postal Code

Distribution of Stars by Postal Code

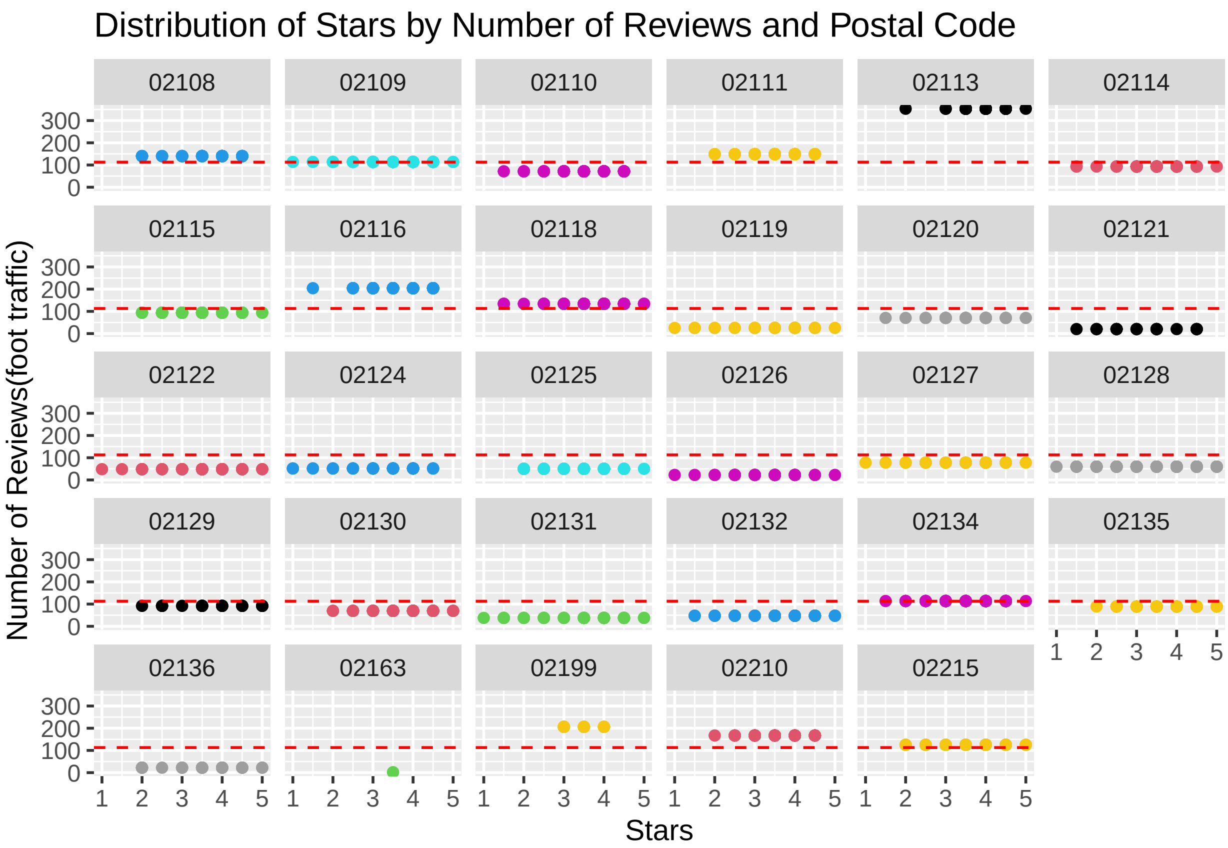




Tourist Plots

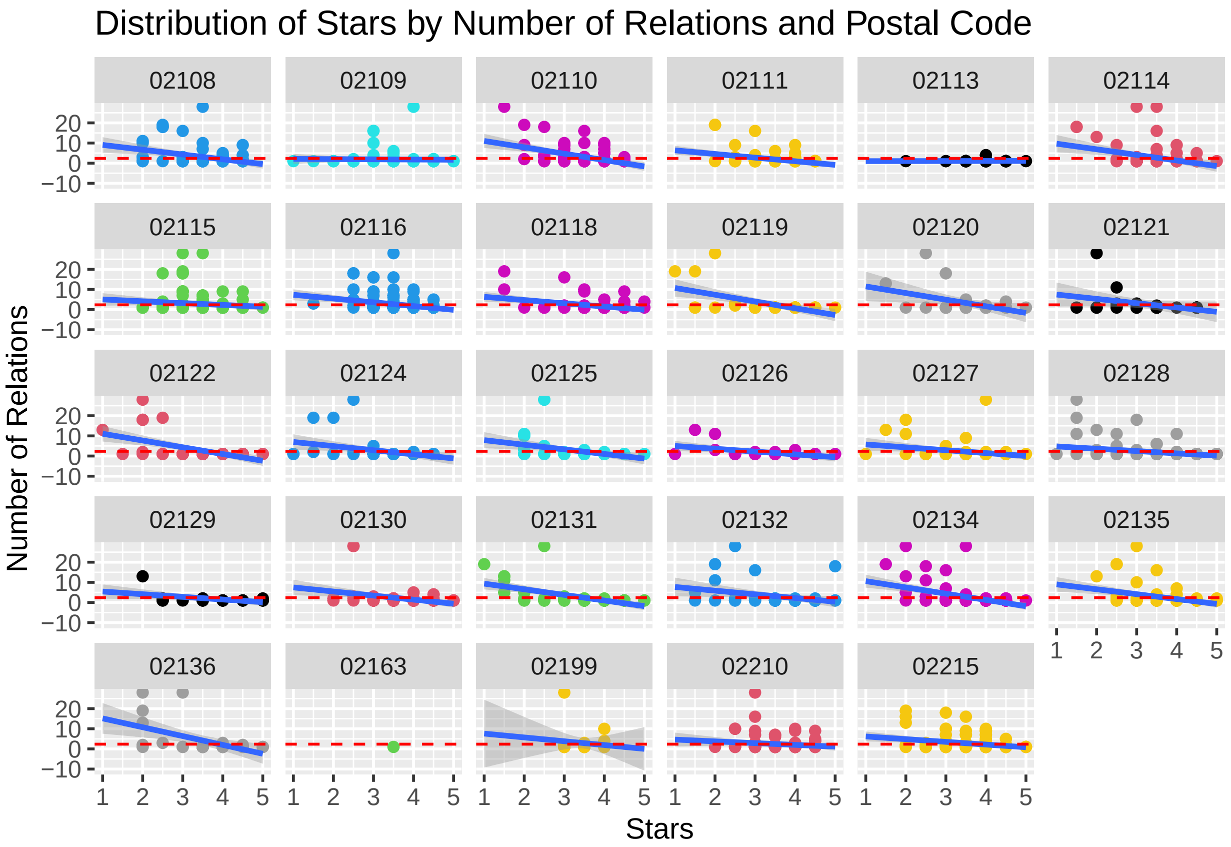


Foot Traffic Plots

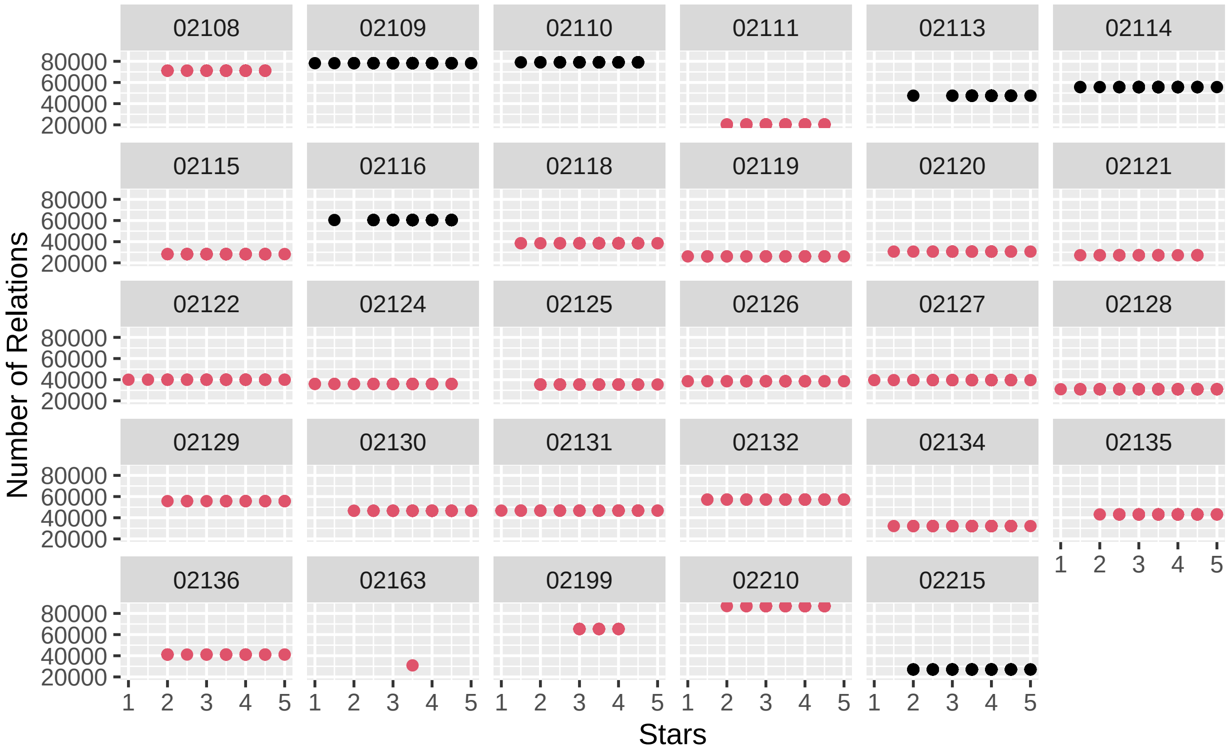


Relations Plots

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


Distribution of Stars by Number of Relations and Postal Code
02108
02109
02110
02111
02113
02114
02115
02116
02118
02119
02120
02121
02122
02124
02125
02126
02127
02128
02129
02130
02131
02132
02134
02135
02136
02163
02199
02210
02215
Number of Relations
20
10
0
−10
Stars
1
2
3
4
5

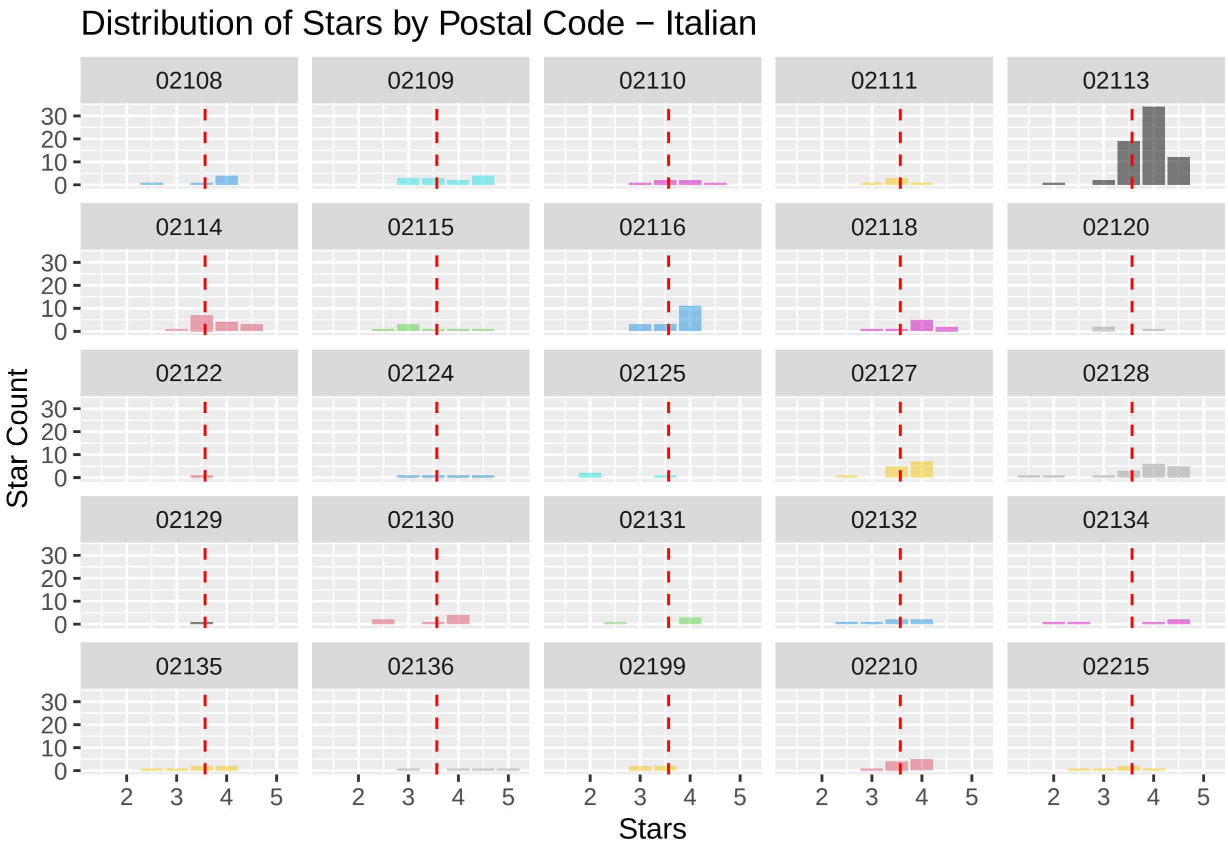
Median Income Plots

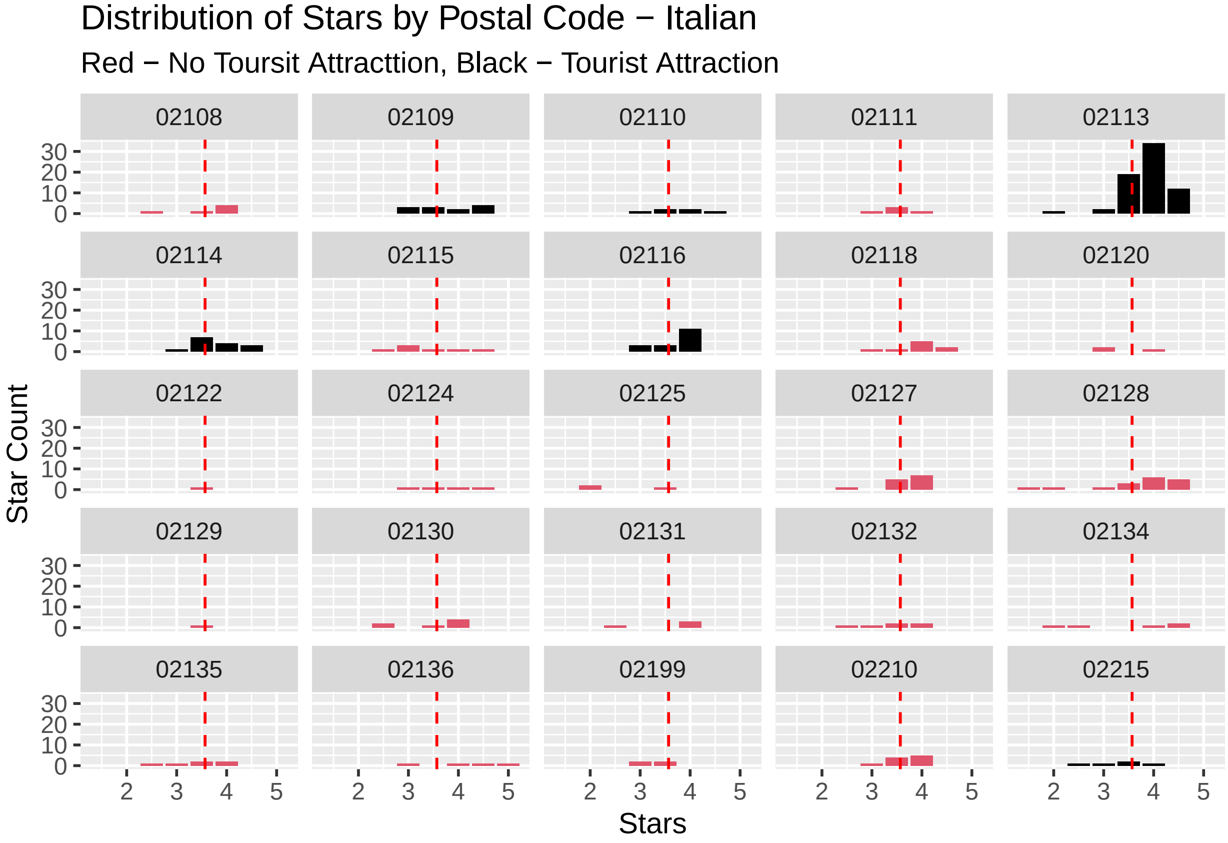


Distribution of Stars by Median Income and Postal Code

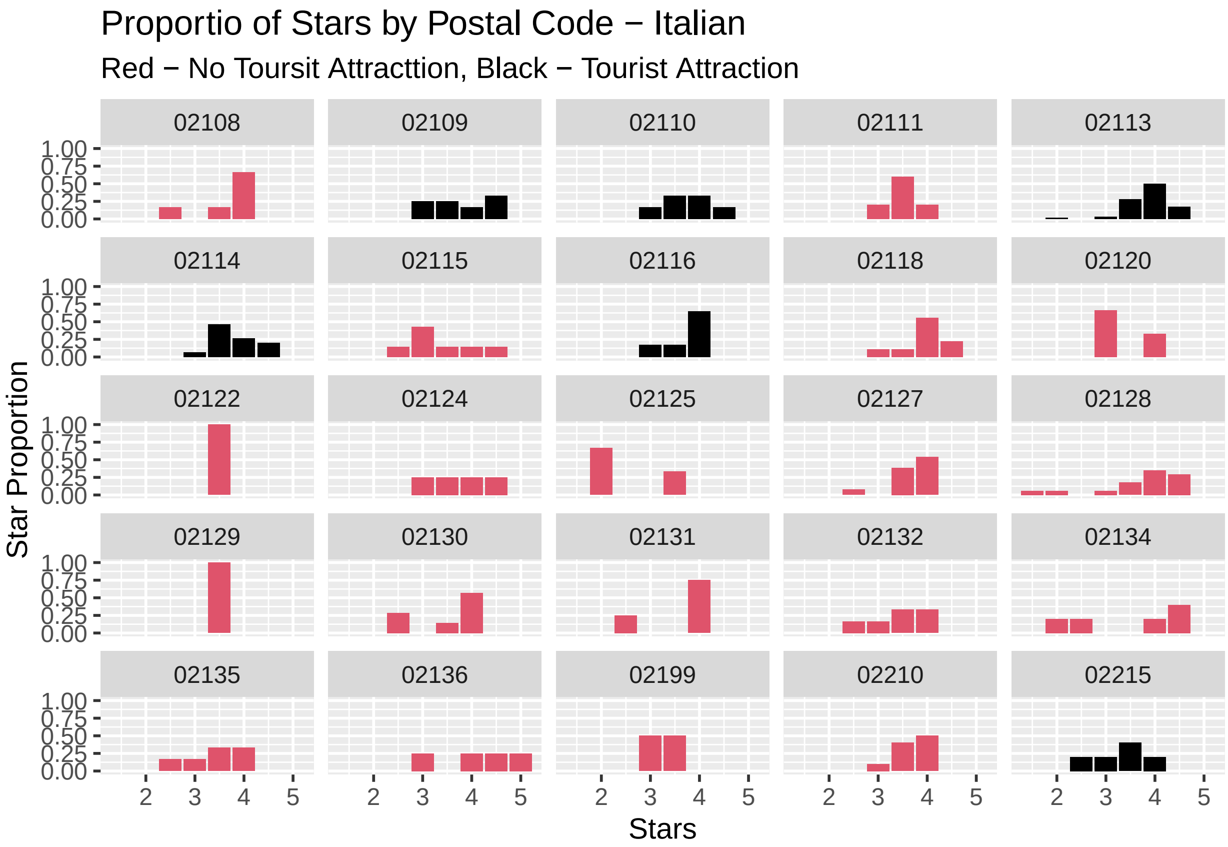
Red − No Toursit Attracttion, Black − Tourist Attraction

Restaurant Type Plots

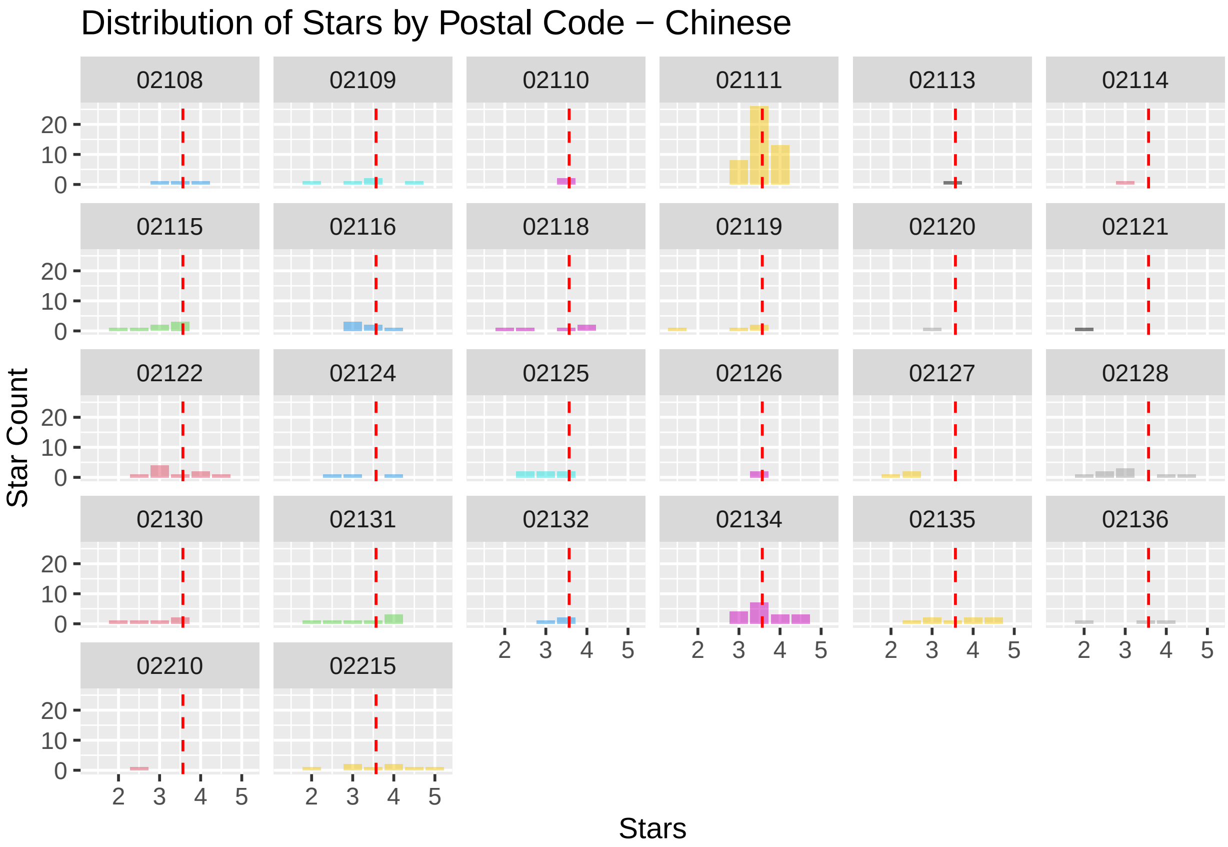

Distribution of Stars by Postal Code − Italian

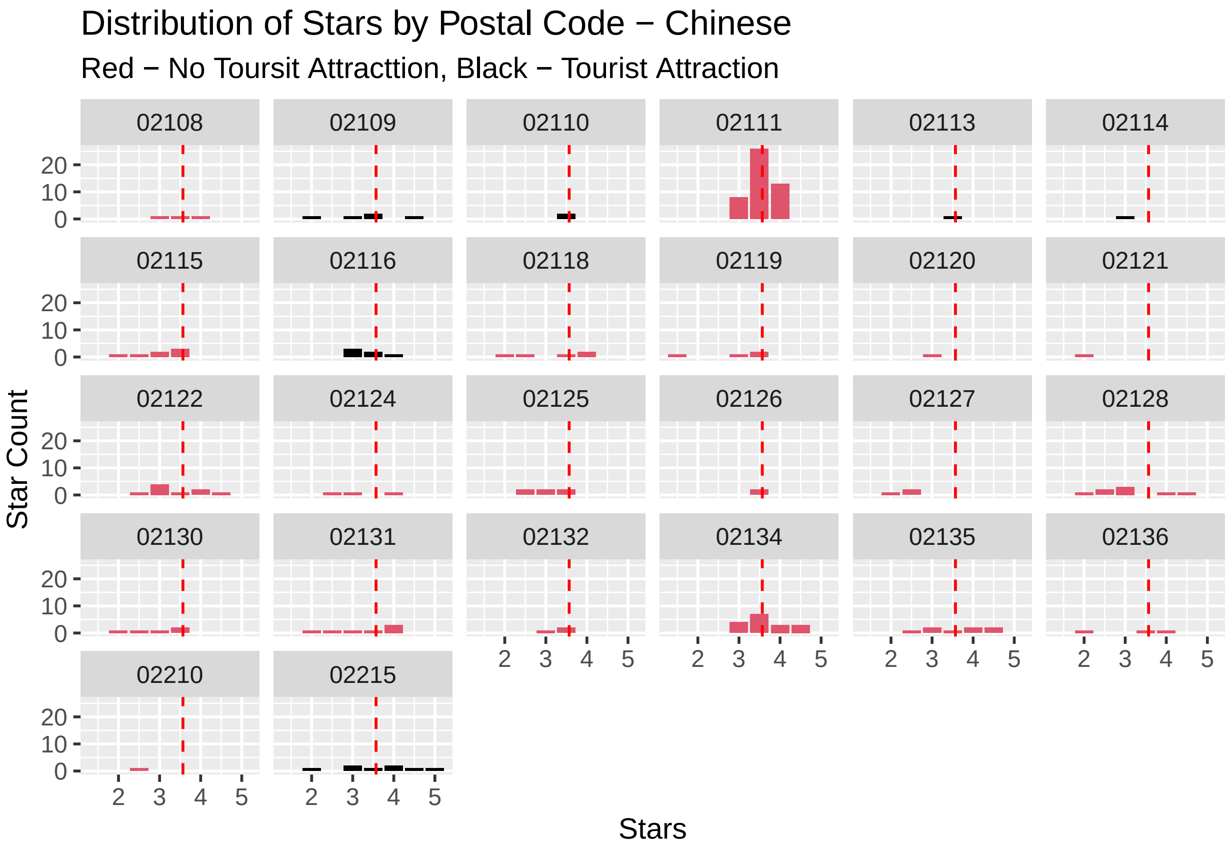

Distribution of Stars by Postal Code − Italian

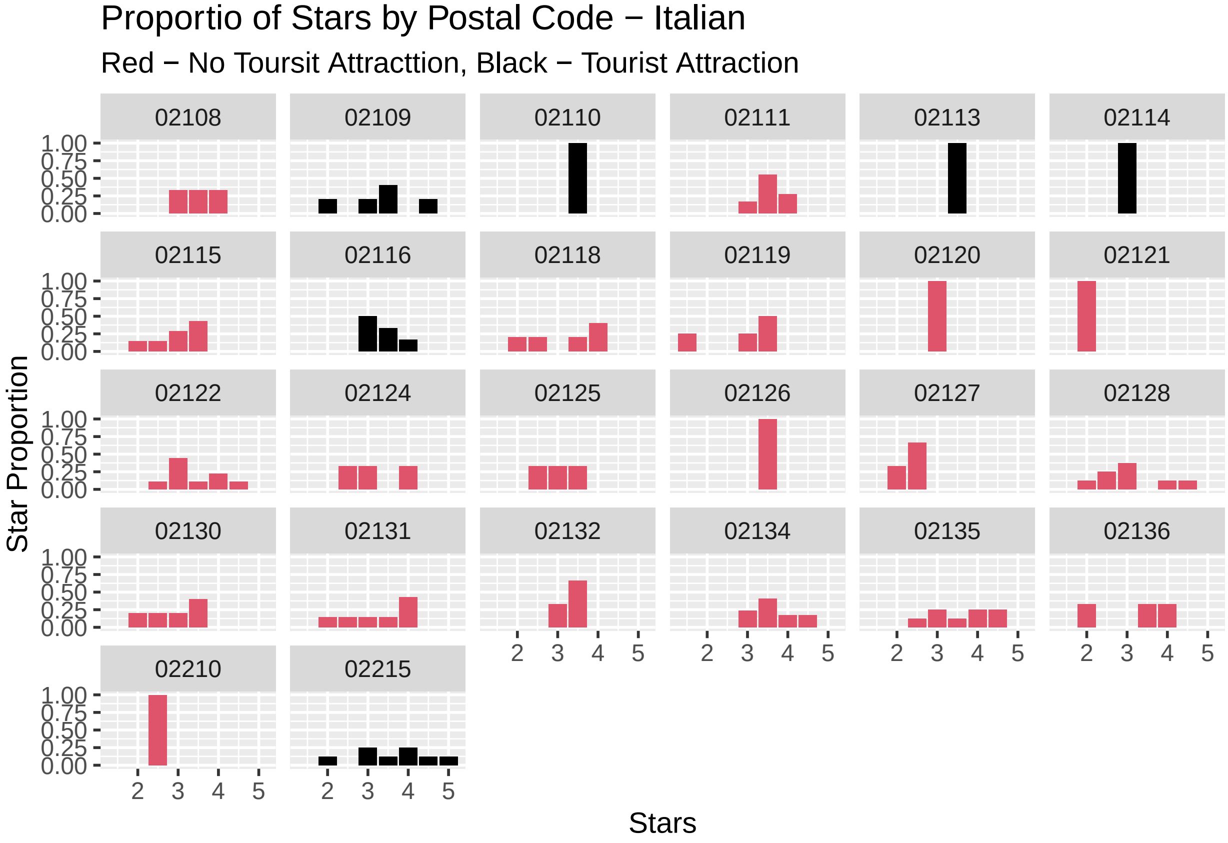
Red − No Toursit Attracttion, Black − Tourist Attraction


Proportio of Stars by Postal Code − Italian

Red − No Toursit Attracttion, Black − Tourist Attraction


Distribution of Stars by Postal Code − Chinese


Distribution of Stars by Postal Code − Chinese
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 Count
0
10
20
2
3
4
5
Stars


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
0.00
0.25
0.50
0.75
1.00
2
3
4
5
Stars