Predicting Zomato restaurant ratings with logistic regression

MATH 1298 Analysis of Categorical Data Project Phase I

Arion Barzoucas-Evans (s3650046) & Joshua Grosman (s3494389)

02/09/2018

Contents

1 Introduction

Zomato is a restaurant search and discovery service founded in 2008. Users can access a plethora of information about restaurants listed on Zomato, including information not available on the restaurant's own website. Such information includes the type of cuisine, opening hours, photos of the menu and the restaurant, pricing, and whether the restaurant offers online delivery. Customers that visit the restaurants have the option of reviewing them and giving them a rating from 0 to 5. Higher rated restaurants receive more attention resulting in higher revenues. As such, understanding how users of Zomato rate restaurants is of great interest to restaurant owners in order to improve their business. The aim of this report is to explore the relationship of several factors like price and cuisine with Zomato ratings using publicly available data found on Kaggle. In phase I exploratory data analysis will be performed on the dataset, including data pre-processing, creation of new variables, and visual representation of the data. This will assist in observing and understanding any relationships present in the data. Following this, a logistic regression model will be fitted in phase II in order to predict the probability of receiving certain ratings according to the values of the chosen explanatory variables.

2 Data Set

The dataset used in this report was acquired from Kaggle. It contains 9,551 observations each of which corresponds to a different restaurant and contains the following information:

- *Restaurant.ID*: A unique ID assigned to each restaurant.
- *Restaurant.Name*: Name of the restaurant.
- *Country.Code*: Codes corresponding to countries listed in a separate dataset.
- *City*: Name of the city where the restaurant is located.
- *Address*: Address of the restaurant.
- *Locality*: General location of the restaurant (short description).
- *Locality.Verbose*: General location of the restaurant (long description).
- *Longitude*: Longitude of the location of the restaurant (geographic coordinates).
- *Latitude*: Latitude of the location of the restaurant (geographic coordinates).
- *Cuisines*: Type of cuisine offered by the restaurant.
- *Average.Cost.for.two*: Average cost for two people (in the countries respective currency).
- *Currency*: Currency which is used at each restaurant.
- *Has.Table.booking*: Whether the restaurant offers the option to book a table or not.
- *Has.Online.delivery*: Whether the restaurant offers delivery through the internet or not.
- *Is.delivering.now*: Whether the restaurant was delivering at the time the dataset was created or not.
- *Switch.to.order.menu*: Unclear variable with only one level (No) for all observations.
- *Price.range*: Categorised price (between 1 and 4).
- *Aggregate.rating*: Aggregated rating from user votes.
- *Rating.color*: Categorised rating into a colour code.
- *Rating.text*: Response variable. Categorised rating with 6 levels (from Poor to Excellent).
- *Votes*: Number of votes used in the aggregate rating.

3 Data Preparation

In this project, the following `R` packages were used.

```
library(mlr)
library(data.table)
library(plyr)
library(dplyr)
library(ggplot2)
library(vcd)
library(knitr)
library(kableExtra)
```

As Table 1 indicates, there are no NA values in any of the features. Additionally, according to Table 2 restaurant ID's are unique for every restaurant while it seems there are some duplicate restaurant names due to the existence of restaurant chains. Furthermore, each restaurant has multiple cuisines which causes the `cuisine` feature to have a large amount of levels. Finally, the `Switch.to.order.menu` feature only has one level ("No") for the entire dataset.

Table 1: Feature summary before data preprocessing.

name	type	na	mean	disp	median	mad	min	max	nlevs
Restaurant.ID	integer	0	9.051128e+06	8.791521e+06	6.004089e+06	8.900212e+06	53.00000	1.850065e+07	0
Restaurant.Name	factor	0	NA	9.913098e-01	NA	NA	1.00000	8.300000e+01	7446
Country.Code	integer	0	1.836562e+01	5.675055e+01	1.000000e+00	0.000000e+00	1.00000	2.160000e+02	0
City	factor	0	NA	4.269710e-01	NA	NA	1.00000	5.473000e+03	141
Address	factor	0	NA	9.988483e-01	NA	NA	1.00000	1.100000e+01	8918
Locality	factor	0	NA	9.872265e-01	NA	NA	1.00000	1.220000e+02	1208
Locality.Verbose	factor	0	NA	9.872265e-01	NA	NA	1.00000	1.220000e+02	1265
Longitude	numeric	0	6.412657e+01	4.146706e+01	7.719196e+01	1.506428e-01	-157.94849	1.748321e+02	0
Latitude	numeric	0	2.585438e+01	1.100794e+01	2.857047e+01	1.135080e-01	-41.33043	5.597698e+01	0
Cuisines	factor	0	NA	9.019998e-01	NA	NA	1.00000	9.360000e+02	1826
Average.Cost.for.two	integer	0	1.199211e+03	1.612118e+04	4.000000e+02	2.965200e+02	0.00000	8.000000e+05	0
Currency	factor	0	NA	9.412630e-02	NA	NA	20.00000	8.652000e+03	12
Has.Table.booking	factor	0	NA	1.212438e-01	NA	NA	1158.00000	8.393000e+03	2
Has.Online.delivery	factor	0	NA	2.566223e-01	NA	NA	2451.00000	7.100000e+03	2
Is.delivering.now	factor	0	NA	3.559800e-03	NA	NA	34.00000	9.517000e+03	2
Switch.to.order.menu	factor	0	NA	0.000000e+00	NA	NA	9551.00000	9.551000e+03	1
Price.range	integer	0	1.804837e+00	9.056088e-01	2.000000e+00	1.482600e+00	1.00000	4.000000e+00	0
Aggregate.rating	numeric	0	2.666370e+00	1.516377e+00	3.200000e+00	7.413000e-01	0.00000	4.900000e+00	0
Rating.color	factor	0	NA	6.087321e-01	NA	NA	186.00000	3.737000e+03	6
Rating.text	factor	0	NA	6.087321e-01	NA	NA	186.00000	3.737000e+03	6
Votes	integer	0	1.569097e+02	4.301691e+02	3.100000e+01	4.447800e+01	0.00000	1.093400e+04	0

Table 2: Variable summary for the Zomato dataset 9551 Observations of 21 Variables

Variable	Class	Cardinality	First Levels	First Values
Restaurant.ID	integer	9551	6317637, 6304287, 6300002, 6318506, 6314302, 18189371	6317637, 6304287, 6300002, 6318506, 6314302, 18189371
Restaurant.Name	character	6899	Le Petit S, Izakaya Ki, Heat - Eds, Ooma, Sambo Koji, Din Tai Fu	Le Petit S, Izakaya Ki, Heat - Eds, Ooma, Sambo Koji, Din Tai Fu
Country.Code	integer	15	162, 30, 216, 14, 37, 184	162, 162, 162, 162, 162, 162
City	factor	141	Abu Dhabi, Agra, Ahmedabad, Albany, Allahabad, Amritsar	Makati City, Makati City, Mandaluyong City, Mandaluyong City, Mandaluyong City, Mandaluyong City
Address	character	7626	Third Floo, Little Tok, Edsa Shang, Ground Flo, Building K, Building B	Third Floo, Little Tok, Edsa Shang, Third Floo, Third Floo, Ground Flo
Locality	character	1140	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM by the , Sofitel Ph	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM Megamal, SM Megamal
Locality.Verbose	character	1141	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM by the , Sofitel Ph	Century Ci, Little Tok, Edsa Shang, SM Megamal, SM Megamal, SM Megamal
Longitude	numeric	8120	121.027535, 121.014101, 121.056831, 121.056475, 121.057508, 121.056314	121.027535, 121.014101, 121.056831, 121.056475, 121.057508, 121.056314
Latitude	numeric	8677	14.565443, 14.553708, 14.581404, 14.585318, 14.58445, 14.583764	14.565443, 14.553708, 14.581404, 14.585318, 14.58445, 14.583764
Cuisines	character	417	French, Ja, Japanese, Seafood, A, Japanese, , Chinese, Asian, Eur	French, Ja, Japanese, Seafood, A, Japanese, , Japanese, , Chinese
Average.Cost.for.two	integer	140	1100, 1200, 4000, 1500, 1000, 2000	1100, 1200, 4000, 1500, 1500, 1000
Currency	factor	12	Botswana Pula(P), Brazilian Real(R$), Dollar($), Emirati Diram(AED), Indian Rupees(Rs.), Indonesian Rupiah(IDR)	Botswana Pula(P), Botswana Pula(P), Botswana Pula(P), Botswana Pula(P), Botswana Pula(P), Botswana Pula(P)
Has.Table.booking	factor	2	No, Yes	Yes, Yes, Yes, No, Yes, No
Has.Online.delivery	factor	2	No, Yes	No, No, No, No, No, No
Is.delivering.now	factor	2	No, Yes	No, No, No, No, No, No
Switch.to.order.menu	factor	1	No	No, No, No, No, No, No
Price.range	integer	4	3, 4, 2, 1	3, 3, 4, 4, 4, 3
Aggregate.rating	numeric	33	4.8, 4.5, 4.4, 4.9, 4, 4.2	4.8, 4.5, 4.4, 4.9, 4.8, 4.4
Rating.color	factor	6	Dark Green, Green, Orange, Red, White, Yellow	Dark Green, Dark Green, Green, Dark Green, Dark Green, Green
Rating.text	factor	6	Average, Excellent, Good, Not rated, Poor, Very Good	Excellent, Excellent, Very Good, Excellent, Excellent, Very Good
Votes	integer	1012	314, 591, 270, 365, 229, 336	314, 591, 270, 365, 229, 336

To rectify the identified issues, all unnecessary features were removed. This includes `Restaurant.ID` (unique identifier), `Restaurant.Name`, `Address`, `Locality`, `Locality.Verbose`, `Is.delivering.now`,

`Switch.to.order.menu`, `City`, `Currency`, `Rating.color`. The `Average.Cost.for.two` is in many different currencies and the concept of what is considered expensive would be affected by socio-economic factors in each country. For this reason, this feature was standardised by currency. Furthermore, the `cuisines` variable was separated into 18 new binary features using the most prevalent levels within the original `cuisines` variable. Each of these features indicate the presence or absence of that particular cuisine in the restaurant. This way, restaurants can have multiple cuisines. Using these new binary variables, a new feature, `Cuisine_Range`, was created as the sum of all the cuisine binary variables. The `Cuisine_Range` would indicate the number of different cuisines present in a restaurant which may be of interest to the model in deciding a restaurant's rating. `Has.Table.booking` and `Has.Online.delivery` were recoded into binary variables, the country table was joined to the Zomato dataset through `Country.Code` and then aggregated into a `Continent` feature. Finally, the levels of the response variable, `Rating.text`, originally included the level "Not rated". As it is not reasonable to predict this level, all associated instances were removed from the data. Tables 3 and 4 show the data after all pre-processing.

Table 3: Feature summary after data preprocessing.

name	type	na	mean	disp	median	mad	min	max	nlevs
Longitude	numeric	0	63.463	44.697	77.192	0.147	-157.948	174.832	0
Latitude	numeric	0	26.183	11.267	28.570	0.117	-41.330	55.977	0
Average.Cost.for.two.Std	numeric	0	0.137	1.077	-0.207	0.498	-1.328	12.384	0
Has.Table.booking	factor	0	NA	0.150	NA	NA	1111.000	6292.000	2
Has.Online.delivery	factor	0	NA	0.318	NA	NA	2355.000	5048.000	2
Price.range	integer	0	1.970	0.931	2.000	1.483	1.000	4.000	0
Aggregate.rating	numeric	0	3.440	0.552	3.400	0.593	1.800	4.900	0
Rating.text	factor	0	NA	0.495	NA	NA	186.000	3737.000	5
Votes	integer	0	202.185	479.195	60.000	74.130	4.000	10934.000	0
Seafood	numeric	0	0.023	0.150	0.000	0.000	0.000	1.000	0
Asian	numeric	0	0.335	0.472	0.000	0.000	0.000	1.000	0
European	numeric	0	0.128	0.334	0.000	0.000	0.000	1.000	0
Cafe	numeric	0	0.086	0.281	0.000	0.000	0.000	1.000	0
Fast Food	numeric	0	0.211	0.408	0.000	0.000	0.000	1.000	0
Bakery	numeric	0	0.072	0.259	0.000	0.000	0.000	1.000	0
Pizza	numeric	0	0.045	0.208	0.000	0.000	0.000	1.000	0
Desserts	numeric	0	0.081	0.274	0.000	0.000	0.000	1.000	0
Beverages	numeric	0	0.033	0.180	0.000	0.000	0.000	1.000	0
Burger	numeric	0	0.032	0.177	0.000	0.000	0.000	1.000	0
Indian	numeric	0	0.484	0.500	0.000	0.000	0.000	1.000	0
Finger Food	numeric	0	0.014	0.119	0.000	0.000	0.000	1.000	0
Continental	numeric	0	0.094	0.292	0.000	0.000	0.000	1.000	0
Street Food	numeric	0	0.054	0.227	0.000	0.000	0.000	1.000	0
Raw Meats	numeric	0	0.011	0.103	0.000	0.000	0.000	1.000	0
South American	numeric	0	0.029	0.168	0.000	0.000	0.000	1.000	0
Healthy Food	numeric	0	0.019	0.135	0.000	0.000	0.000	1.000	0
Other	numeric	0	0.074	0.261	0.000	0.000	0.000	1.000	0
Oceania	integer	0	0.009	0.093	0.000	0.000	0.000	1.000	0
Rest of World	integer	0	0.020	0.140	0.000	0.000	0.000	1.000	0
North America	integer	0	0.059	0.235	0.000	0.000	0.000	1.000	0
Asia	integer	0	0.902	0.298	1.000	0.000	0.000	1.000	0
Europe	integer	0	0.011	0.103	0.000	0.000	0.000	1.000	0
Cuisine_Range	factor	0	NA	0.573	NA	NA	70.000	3162.000	6

Table 4: Variable summary for the Zomato dataset after preprocessing 7403 Observations of 33 Variables

Variable	Class	Cardinality	First Levels	First Values
Longitude	numeric	1586	78.012, 0, 77.998, 78.008, 78.044, 78.057	78.012, 0, 78.012, 77.998, 78.008, 0
Latitude	numeric	1470	27.162, 0, 27.161, 27.196, 27.202, 27.163	27.162, 0, 27.161, 27.196, 27.202, 0
Average.Cost.for.two.Std	numeric	277	0.38, 0.129, -0.207, -0.375, 0.632, 2.311	0.38, 0.129, -0.207, -0.375, 0.632, 2.311
Has.Table.booking	factor	2	0, 1	0, 0, 0, 0, 0, 0
Has.Online.delivery	factor	2	0, 1	0, 0, 0, 0, 0, 0
Price.range	integer	4	3, 2, 4, 1	3, 2, 2, 2, 3, 4
Aggregate.rating	numeric	32	3.9, 3.5, 3.6, 4, 4.2, 4.3	3.9, 3.5, 3.6, 4, 4.2, 4
Rating.text	factor	5	Average, Excellent, Good, Poor, Very Good	Good, Good, Good, Very Good, Very Good, Very Good
Votes	integer	1008	140, 71, 94, 87, 177, 45	140, 71, 94, 87, 177, 45
Seafood	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Asian	numeric	2	0, 1	0, 0, 0, 0, 1, 0
European	numeric	2	0, 1	0, 0, 0, 0, 0, 1
Cafe	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Fast Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Bakery	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Pizza	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Desserts	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Beverages	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Burger	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Indian	numeric	2	1, 0	1, 1, 1, 1, 1, 1
Finger Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Continental	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Street Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Raw Meats	numeric	2	0, 1	0, 0, 0, 0, 0, 0
South American	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Healthy Food	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Other	numeric	2	0, 1	0, 0, 0, 0, 0, 0
Oceania	integer	2	0, 1	0, 0, 0, 0, 0, 0
Rest of World	integer	2	0, 1	0, 0, 0, 0, 0, 0
North America	integer	2	0, 1	0, 0, 0, 0, 0, 0
Asia	integer	2	1, 0	1, 1, 1, 1, 1, 1
Europe	integer	2	0, 1	0, 0, 0, 0, 0, 0
Cuisine_Range	factor	6	0, 1, 2, 3, 4, >4	1, 1, 1, 1, 2, 2

Table 5: Contingency tables for discrete and categorical variables.

Has.Table.booking	Freq
0	6292
1	1111

Has.Online.delivery	Freq
0	5048
1	2355

Price.range	Freq
1	2744
2	2711
3	1373
4	575

Rating.text	Freq
Average	3737
Excellent	301
Good	2100
Poor	186
Very Good	1079

Seafood	Freq
0	7232
1	171

Asian	Freq
0	4926
1	2477

European	Freq
0	6453
1	950

Cafe	Freq
0	6765
1	638

Fast.Food	Freq
0	5839
1	1564

Bakery	Freq
0	6868
1	535

Pizza	Freq
0	7067
1	336

Desserts	Freq
0	6800
1	603

Beverages	Freq
0	7155
1	248

Burger	Freq
0	7164
1	239

Indian	Freq
0	3822
1	3581

Finger.Food	Freq
0	7297
1	106

Continental	Freq
0	6704
1	699

Street.Food	Freq
0	7000
1	403

Raw.Meats	Freq
0	7323
1	80

South.American	Freq
0	7188
1	215

Healthy.Food	Freq
0	7266
1	137

Other	Freq
0	6858
1	545

Oceania	Freq
0	7339
1	64

Rest.of.World	Freq
0	7254
1	149

North.America	Freq
0	6968
1	435

Asia	Freq
0	727
1	6676

Europe	Freq
0	7324
1	79

Cuisine_Range	Freq
0	70
1	3162
2	2663
3	1126
4	285
>4	97

4 Data Visualisation

4.1 Univariate and Bivariate

Initially, the proportions of the different levels of the rating response variable were examined.

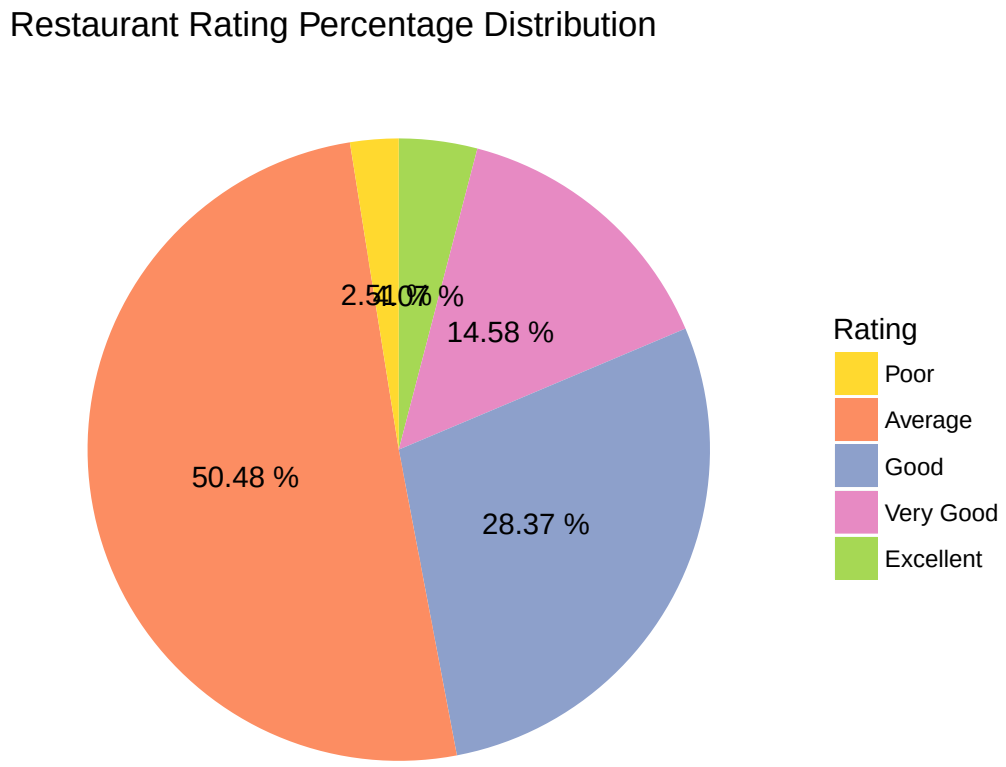

Figure 1: Distribution of the different restaurant ratings in the dataset.

Here it can be seen that the majority of restaurant ratings in the dataset were average. Only a small portion received a rating of either poor or excellent.

Next, the average cost of meal was considered both on its own and with respect to ratings. The associated visualisations are shown in Figure 2 and Figure 3, respectively.

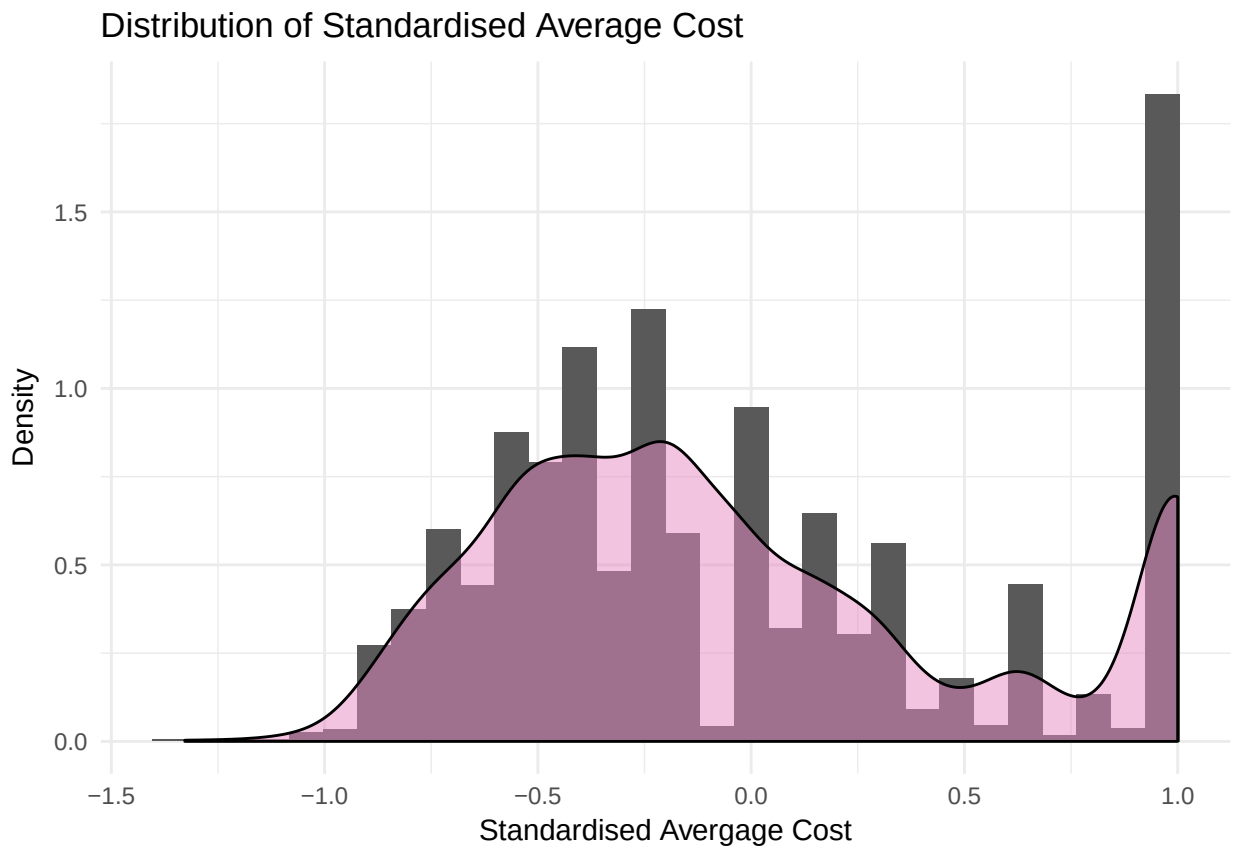

Figure 2: Distribution of the standardised average cost by currency for two people.

Figure 3: Distribution of the standardised average cost by currency for two people by rating.

Figure 2 gives a general idea of the average cost for all restaurants included in the dataset. Evidently, the distribution has normal-like qualities with respect to its bell-shaped curve. There is, however, a clear spike in restaurants which may be considered to have a very expensive average cost. From Figure 3, it is clear that all rating levels have a similar distribution to Figure 2, however the spike in expensive average meal cost can be seen to become more profound as rating goes up towards excellent. This would suggest that high rated restaurants are more likely to be more expensive on average.

Next, the binary variable describing whether or not a restaurant had the option to book a table was considered on its own and according to restaurant rating.

Figure 4: Number of restaurants offering table booking to customers.

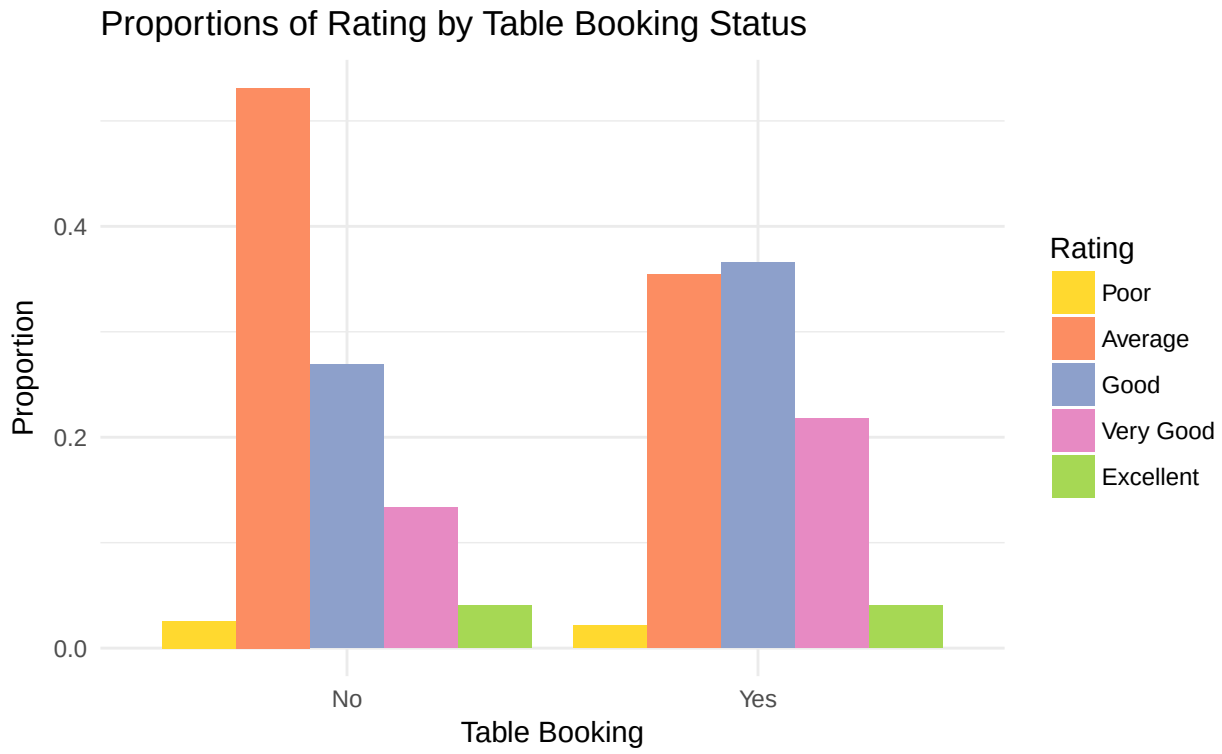

Figure 5: Proportion of different restaurant ratings by table booking availability.

Figure 4 shows that overall, there are much more restaurants which do not offer table booking in the present sample than those that do. Figure 5 bypasses this issue by examining the relative proportions of rating within each level. Here, the option of booking a table can be seen to have some effect on the rating of a restaurant, with restaurants which do offer table booking having higher proportions of good and very good ratings.

Similar visualisations were used to compare the effect of having online delivery with regard to ratings, as shown in Figure 6 and Figure 7.

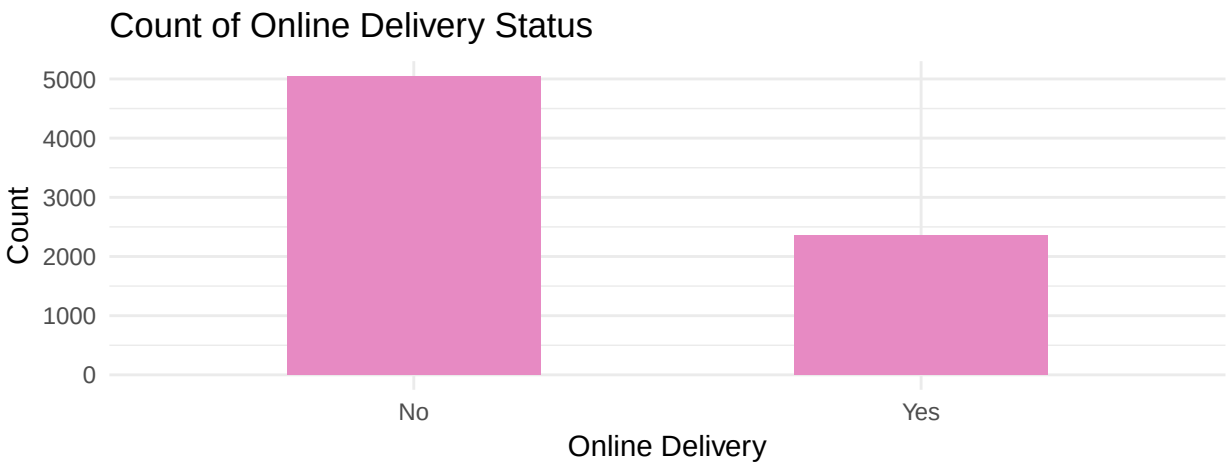

Figure 6: Number of restaurants offering online delivery to customers.

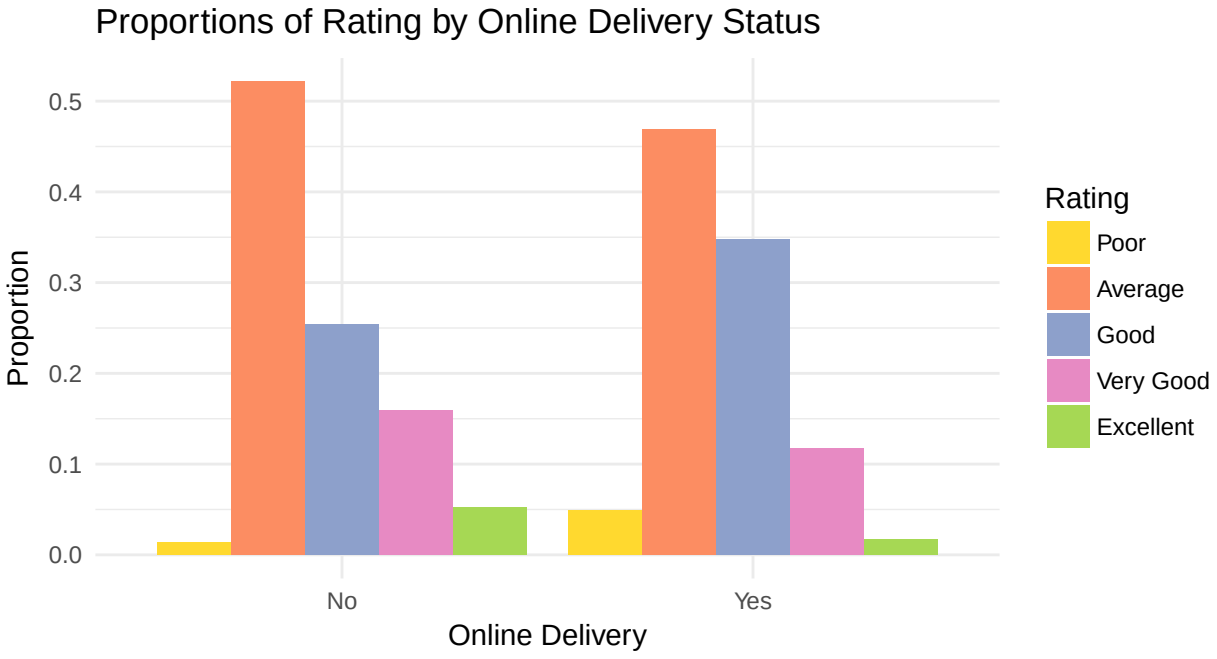

Figure 7: Proportion of different restaurant ratings by online delivery availability.

From Figure 6 it can be seen that restaurants which do not offer online delivery are more prevalent in the dataset than those which do offer it. Figure 7 suggests that having the option of online delivery has little effect on restaurant ratings, with similar proportions in both levels. There are, however, some deviations in the proportions of poor and good rated restaurants.

Next, the amount of customer votes each restaurant received was considered. Due to issues with scale, the log transformation was applied to customer votes.

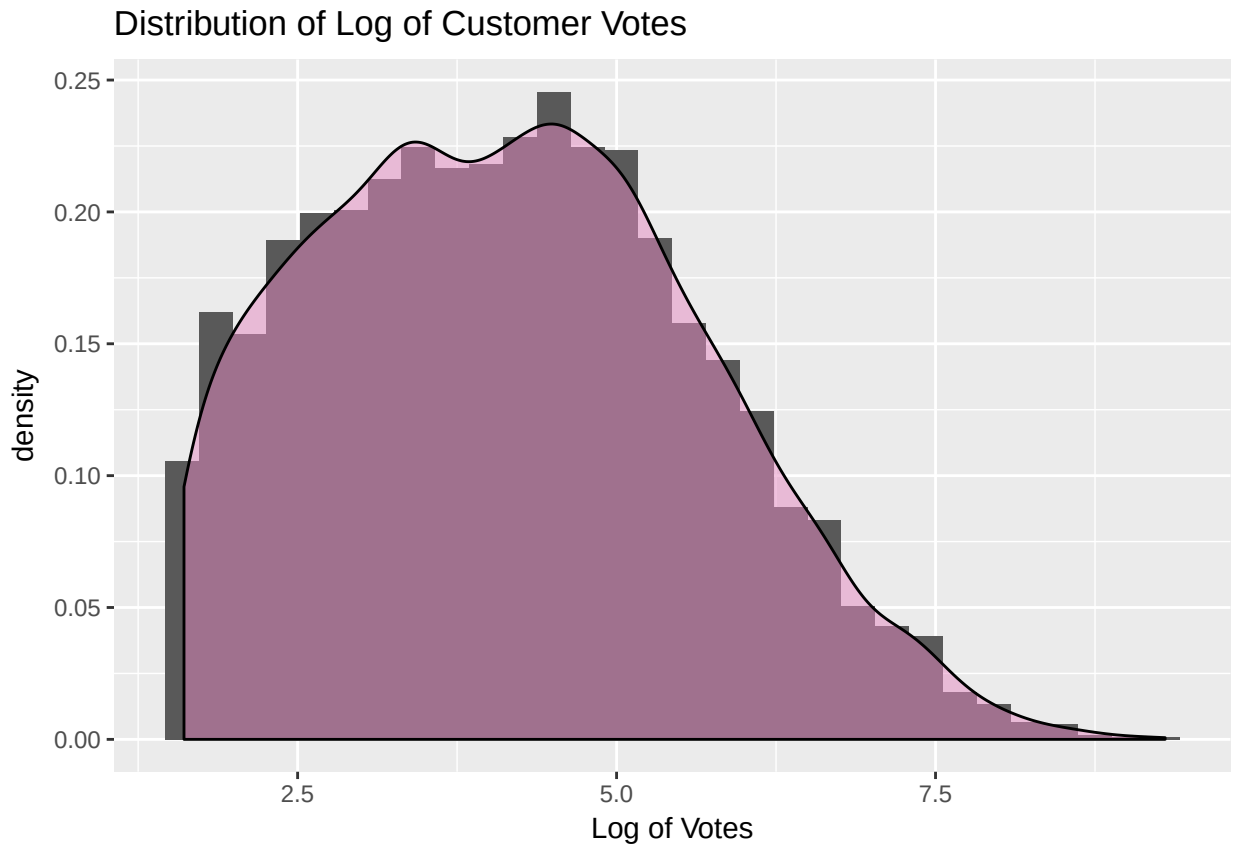

Figure 8: Distribution of the log transformed number of votes for each restaurant in the dataset.

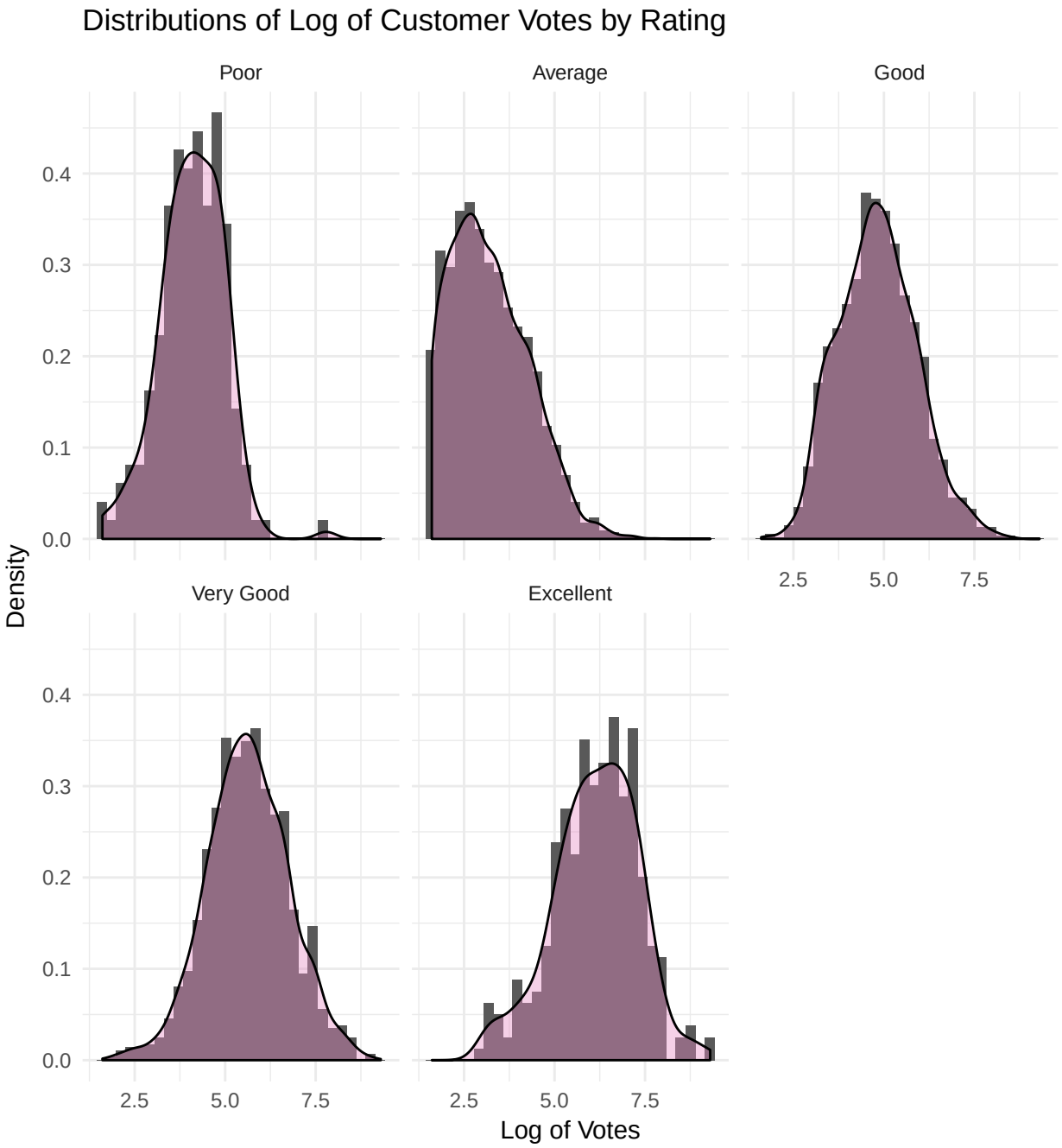

Figure 9: Distribution of the log transformed number of votes for each restaurant by rating.

Figure 8 demonstrates a clear right skew, suggesting that only a few restaurants received a large amount of votes. A trend can be observed within Figure 9, wherein higher ratings appear to be associated with more votes. This trend is disrupted however for restaurants which receive an average rating, which seem to have fewer votes than poor-rated restaurants on average. This may be due to customers being more likely to convey an underwhelming experience at a poor-rated restaurant than a mediocre experience at an average-rated restaurant.

The continent the restaurant was located in was then considered. Figure 10 below depicts a world map of the location of all restaurants within the dataset. Figure 11 then considers the proportional distribution of ratings across the continents.

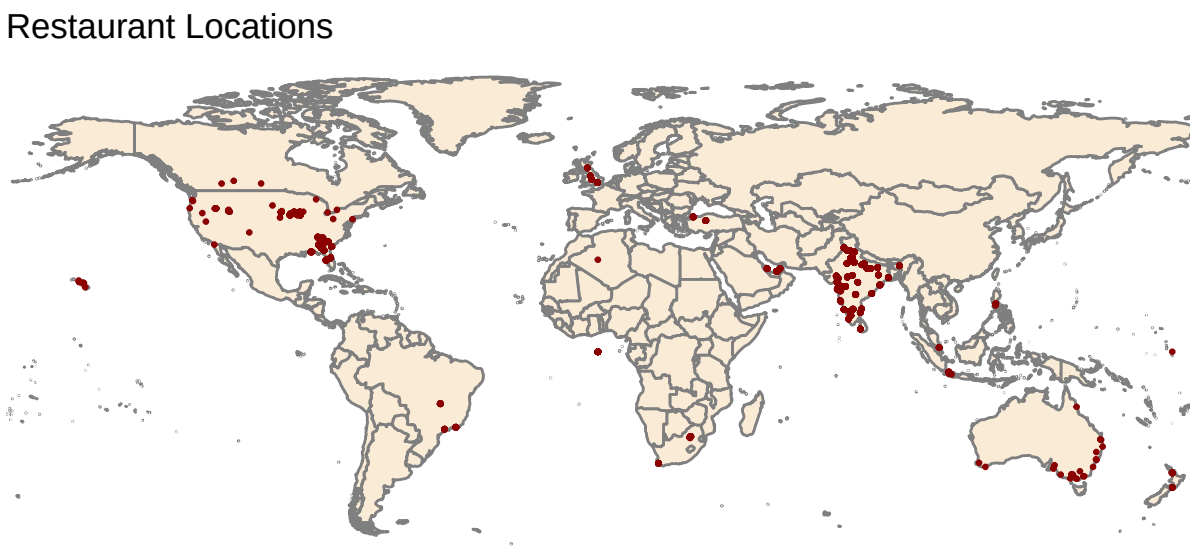

Figure 10: Restaurant locations

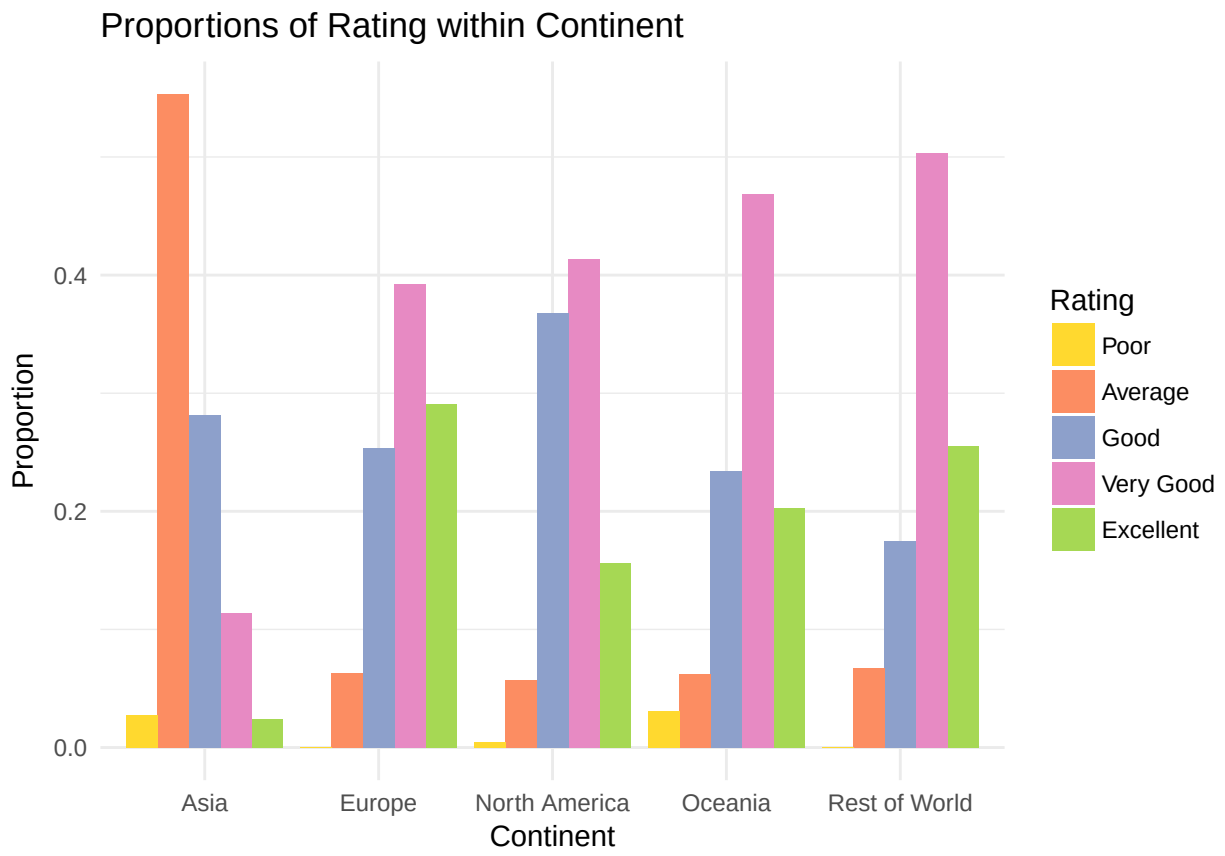

Figure 11: Proportion of different restaurant ratings by continent.

The map provides insight into the locations of the restaurants around the world. Clearly, most of the restaurants in this dataset are based in the USA, India and Australia.

Inspection of Figure 11 highlights that despite being overrepresented in the data, Asian restaurants have relatively low proportions of very good and excellent rated restaurants in comparison to other continents, but have a very high proportion of average restaurants. North America appears to have a relatively high proportion of good-rated restaurants and Oceania has the highest proportion of very good-rated restaurants.

The counts of the different restaurant cuisines were then visualised in Figure 12. Figure 13 considers the cuisine type with respect to rating, but here, the continuous aggregate rating variable was used over the discrete version for simplicity and interpretability.

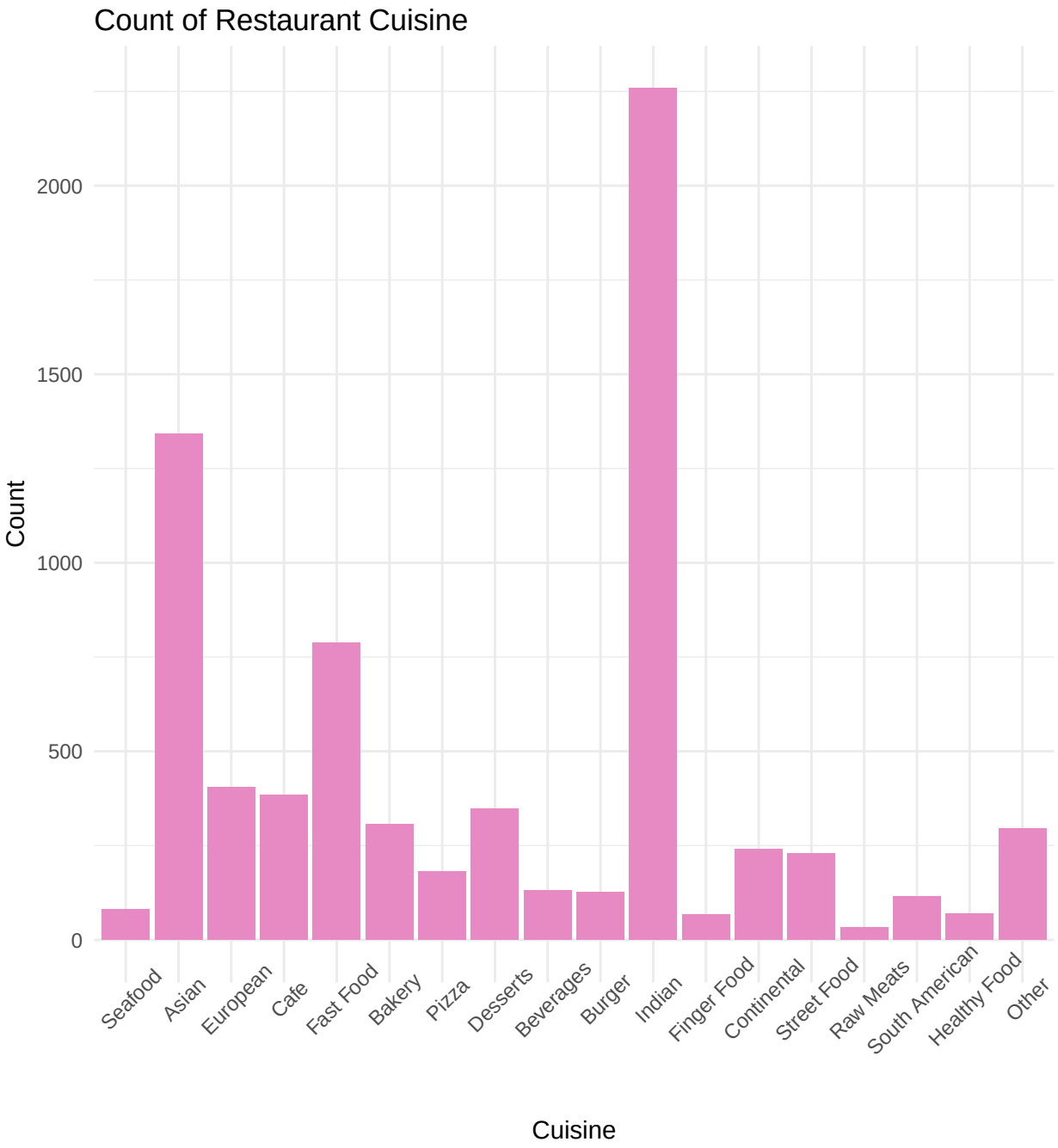

Figure 12: Different cuisines and the number of restaurants that offer them.

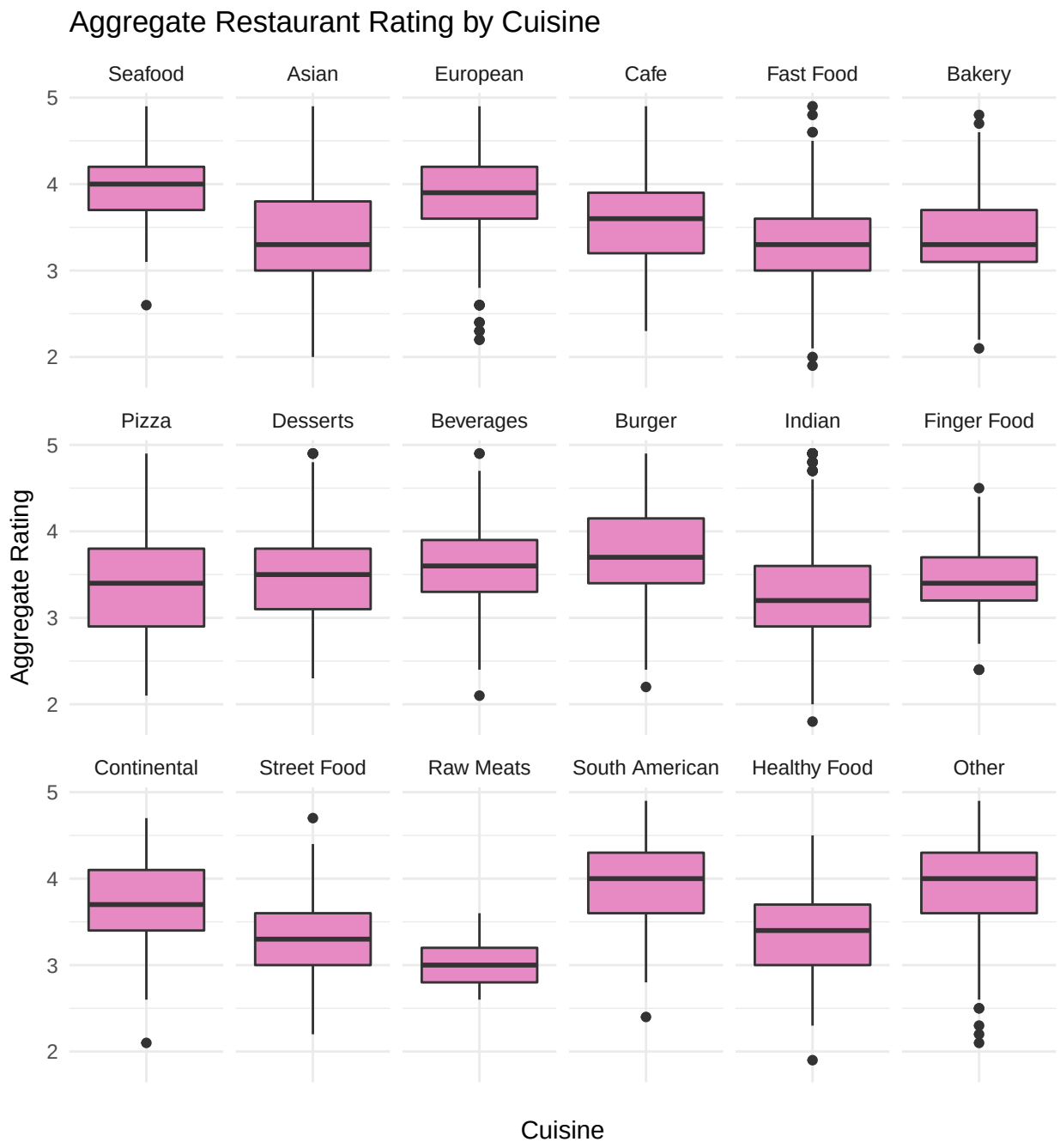

Figure 13: Rating distribution by cuisine.

From Figure 12, it is clear that the restaurants included in this sample mostly belong to either Indian or Asian cuisines. Figure 13 highlights that almost all cuisines are centred around an aggregate rating of 3 to 4, which would equate to ratings of good and very good, respectively. Furthermore, most cuisine types appear to have very little variability, however others, such as pizza and Asian cuisines have larger interquartile ranges. This suggests higher variability in the aggregate ratings of these kinds of restaurants.

4.2 Multivariate

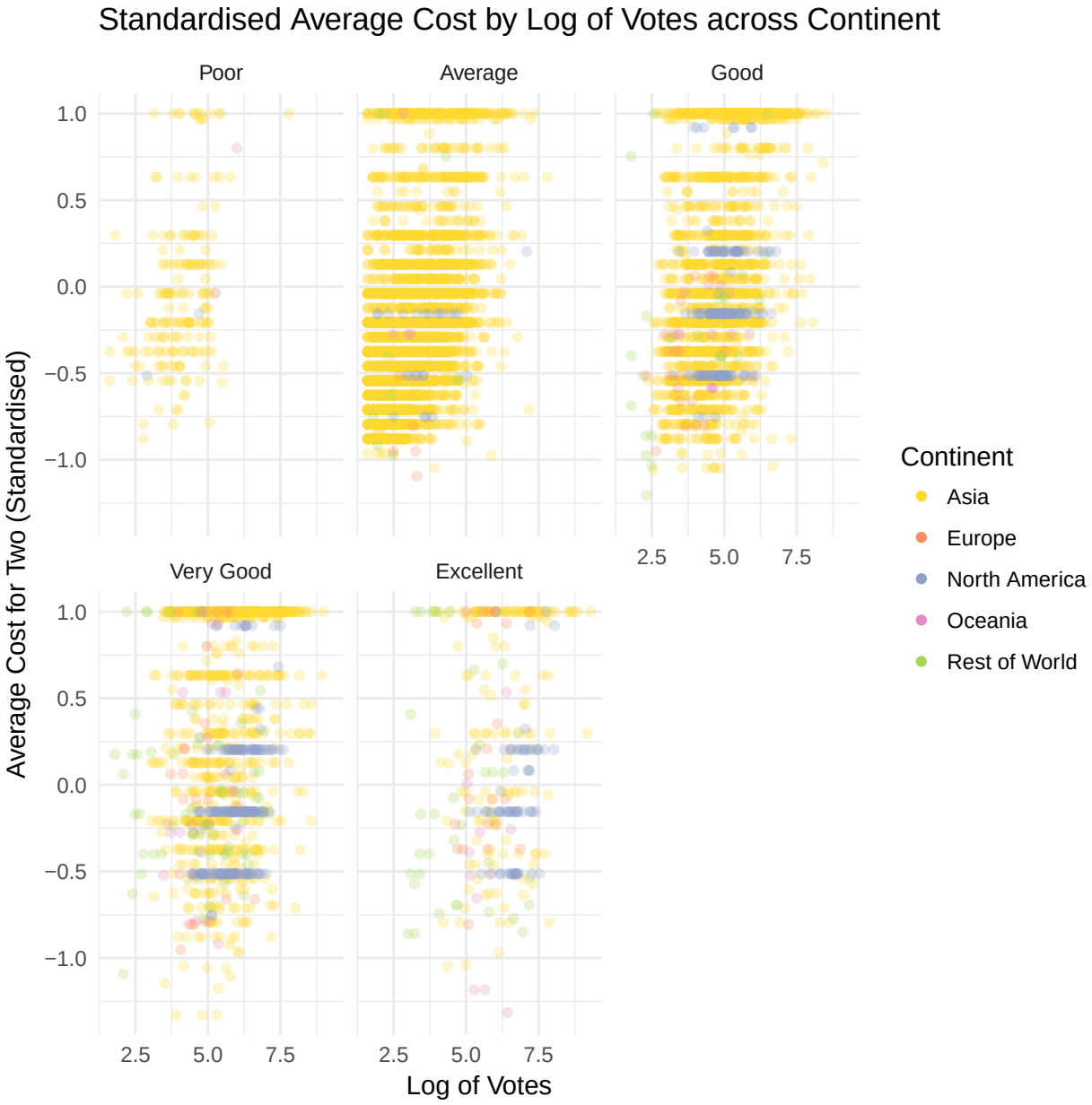

Figure 14: Relationship between cost and the log transformed number of votes accross continents and ratings.

Figure 14 serves to highlight several interesting features within the sample. Foremost, the imbalance of data across the continents as well as rating levels is very clear. The abundance of yellow again suggests that most of the restaurants are from Asia, while the density of points also shows that most of the restaurants received an average or good rating. Furthermore, the dispersion of points in all facets demonstrates that there is no apparent relationship between the cost of a meal and the amount of votes a restaurant receives. It also seems that Asian restaurants are overrepresented in the poor, average and good levels, while there is a more even distribution of the other continents across the other rating levels. On the other hand, for the excellent rating, the proportion of Asian restaurants is relatively low.

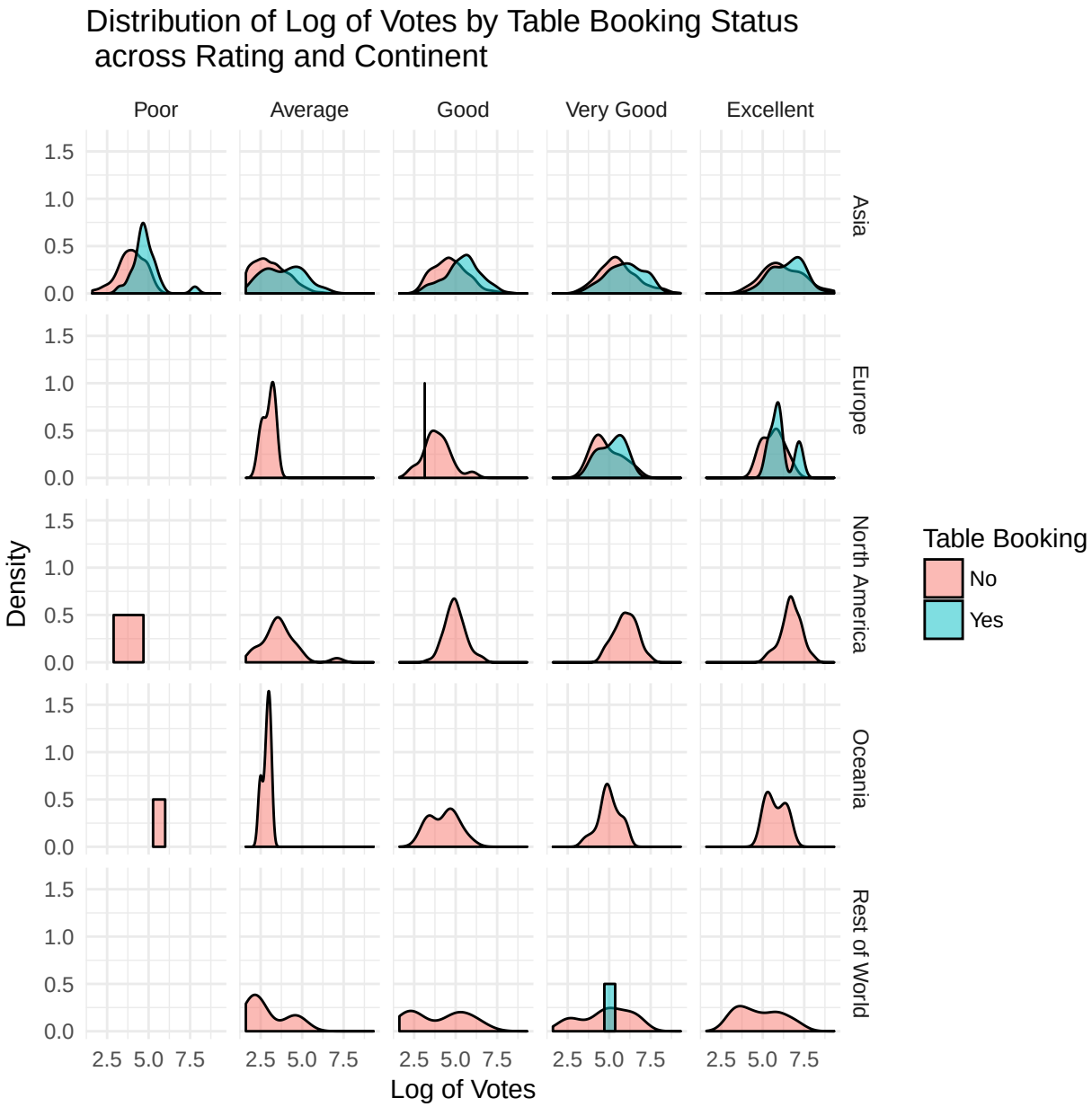

Figure 15: Distribution of the log transformed votes accross continent, rating, and whether or not the restaurant offers table booking.

Figure 15 examines how the number of votes a restaurant receives is influenced by its rating, continent and whether table booking is available. There are clearly some issues with data sparsity evidenced by the empty facets. The imbalance of the table booking variable can also be seen with restaurants in Asia being the only ones which consistently have table bookings available for each rating level. Based on this, as well as the other facets which allow for comparison for the two table booking levels, it can be seen that typically the mode of votes is higher for restaurants which offer table booking compared to those which do not, regardless of rating. This assertion should not be generalised outside of Asia however, due to the lack of data.

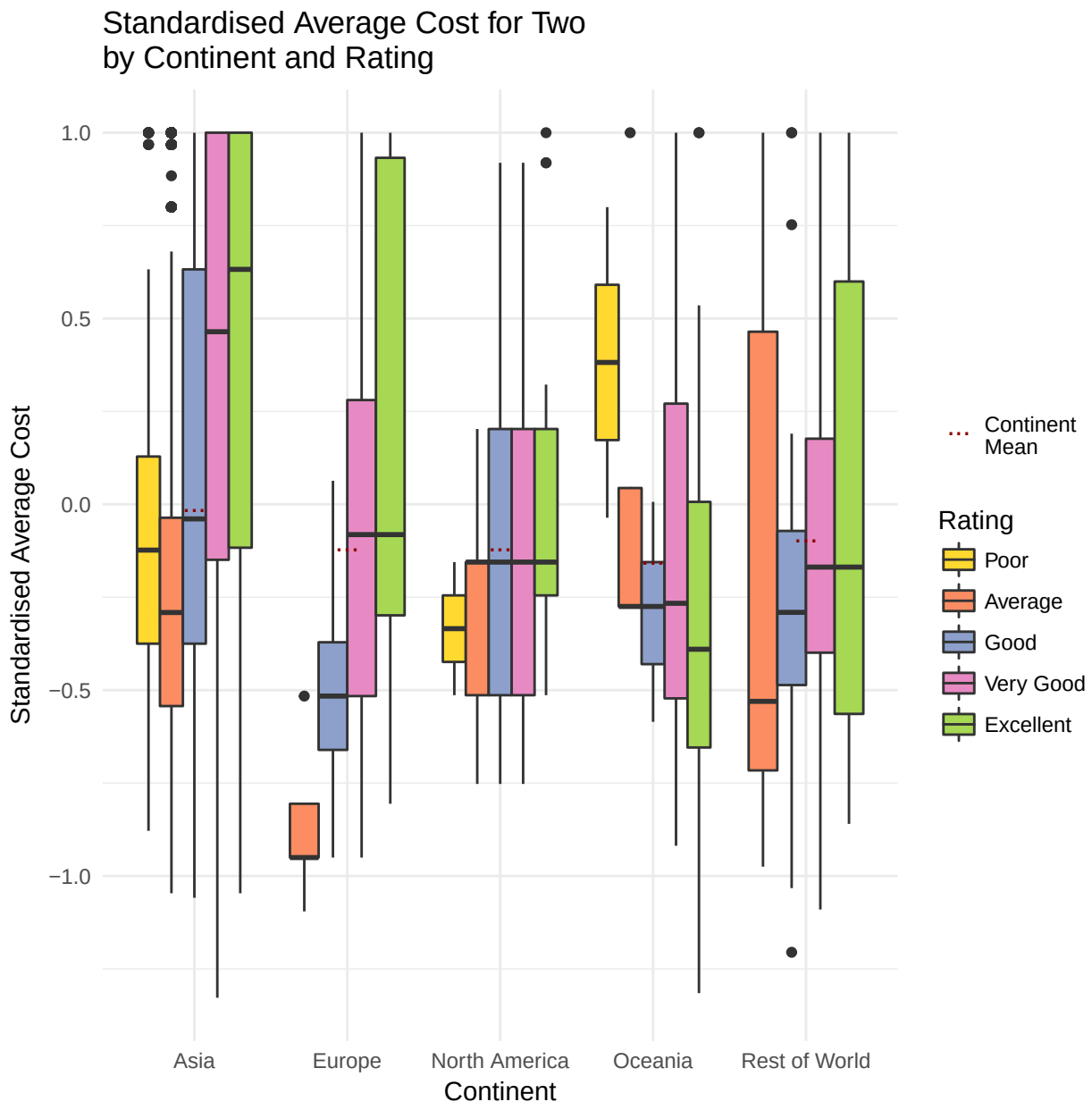

Figure 16: Distribution of the standardised average cost by currency for two people by rating and continent.

Figure 16 compares standardised average meal cost by rating and continent. The figure highlights a number of interesting features, including that there is a large amount of variation in cost when both rating and continent are considered. Very good rated restaurants appear to be the only ones which retain similar medians and variation for all continents, except for those in Asia which can be seen to be more variable and expensive. A slight trend can be seen across the continents, with cost increasing with rating type. Oceania sharply contradicts this trend however, with poorly rated restaurants being the most expensive within the continent. Finally, comparison of the mean cost reveals there is almost no difference across the continents.

5 Summary

Ultimately, this phase was concerned with preparing and exploring the Zomato restaurant data in order to ensure the most accurate and informative modelling in the next phase. Of the original 21 variables, all those which were deemed irrelevant with regards to the upcoming modelling phase were removed. This included variables such as restaurant name, ID, address, etc. Furthermore, it was found that some variables could be further improved and so were either transformed or new variables were created entirely. This included relevelling the cuisine variable, creating the new continent variable based off of country, standardising average meal cost according to currency and log transforming the number of votes each restaurant received. Not all original variables were removed however, as these will still be employed in the modelling phase to assess their relevance in predicting the response variable. Finally, the "Not rated" level of the response variable was dropped, along with all associated observations, as it would not be logical to try and predict this level. This resulted in the response variable having five distinct ordinal levels.

The features were then assessed visually to confirm their relevance for modelling. Interesting univariate insights included how average-rated restaurants made up the majority of the data and that most restaurants included in the data were from Asia and had Asian or Indian cuisines. In all bivariate visualisations, some apparent differences were observed across the levels of the response variable with regards to the other variable being assessed. For example, it was shown that Asia had a very large proportion of average-rated restaurants in comparison to other continents, the log of votes increased across ascending rating levels (expect for average-rated restaurants) and that restaurants which offered table booking were more likely be rated good or very good compared to those which did not offer it. The multivariate visualisations served to further explore the interaction between many variables simultaneously, which proved relatively strenuous on the data, with clear evidence of data sparsity. This ultimately suggests that there is not a comprehensive distribution of data across the domains of all variables. Despite this, given that the exploration of the data overarchingly demonstrated clear differences across the response levels based on the other variables, the upcoming modelling phase should prove fruitful.

References

Bischl, Bernd, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Erich Studerus, Giuseppe Casalicchio, and Zachary M. Jones. 2016. "mlr: Machine Learning in R." *Journal of Machine Learning Research* 17 (170): 1–5. http://jmlr.org/papers/v17/15-066.html.

David Meyer, Achim Zeileis, and Kurt Hornik. 2017. *Vcd: Visualizing Categorical Data*.

Dowle, Matt, and Arun Srinivasan. 2018. *Data.table: Extension of 'Data.frame'*. https://CRAN.R-project.org/package=data.table.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Muller. 2017. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2018. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=dplyr.

Zhu, Hao. 2018. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.