MA615_FinalProj

Sky Liu

12/17/2018

Overview

This project took European restaurant data from TripAdvisor as the main data sourse and conducted analysis on restaurant ratings and reviews. Major techniques applied in the analysis are data cleaning using **dplyr**, Benford's Law analysis using **benford.analysis**, data visualization using **plotly** and **ggplot**, text mining using **tidytext** and **wordcloud**, mapping using **leaflet**.

A Shiny app and a powerpoint generated by **office** are created along with this report.

Data Cleaning

The major dataset was collected from Kaggle[1]. The dataset was obtained by scrapping the ratings and reviews for restaurants across 31 European cities from TripAdvisor.

Another dataset containing geolocation information of the cities was collected from Simplemaps World Cities Database[2].

The original data contains 125527 records of restaurants from 31 European cities. After filtering out NAs, the remaining dataset contains 108178 records.

Map

The map visualization using **leaflet** provides an overview of the geographically information of the cities from the dataset and allows interaction to view the total number of restaurant, average rating and average number of reviews.

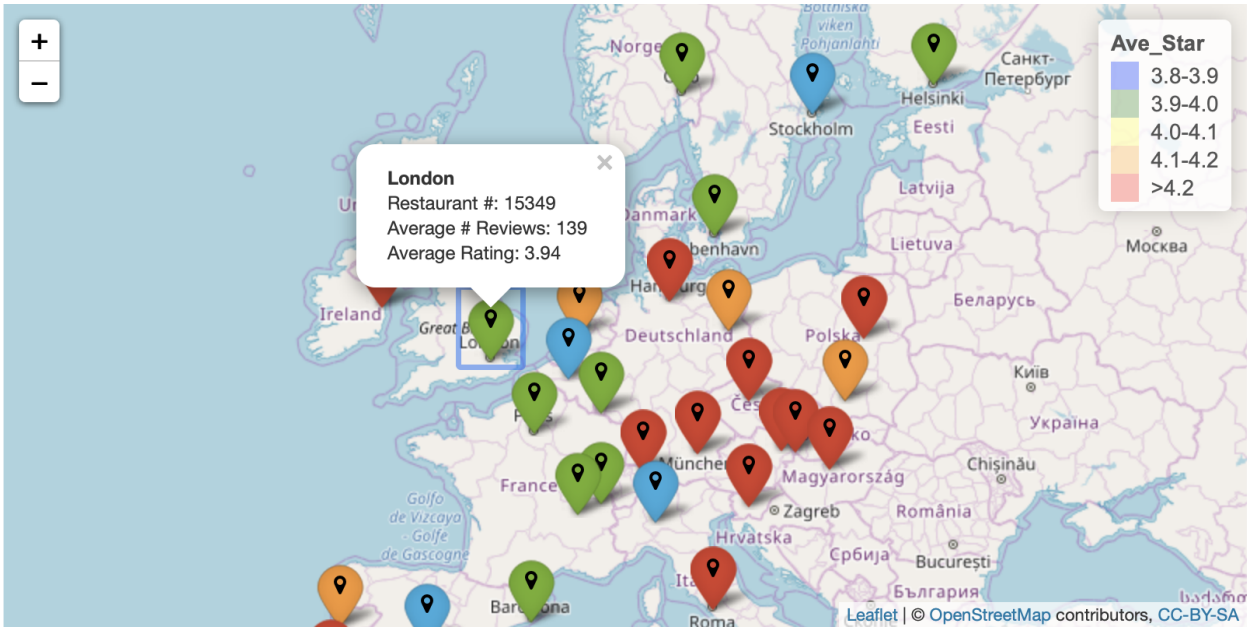
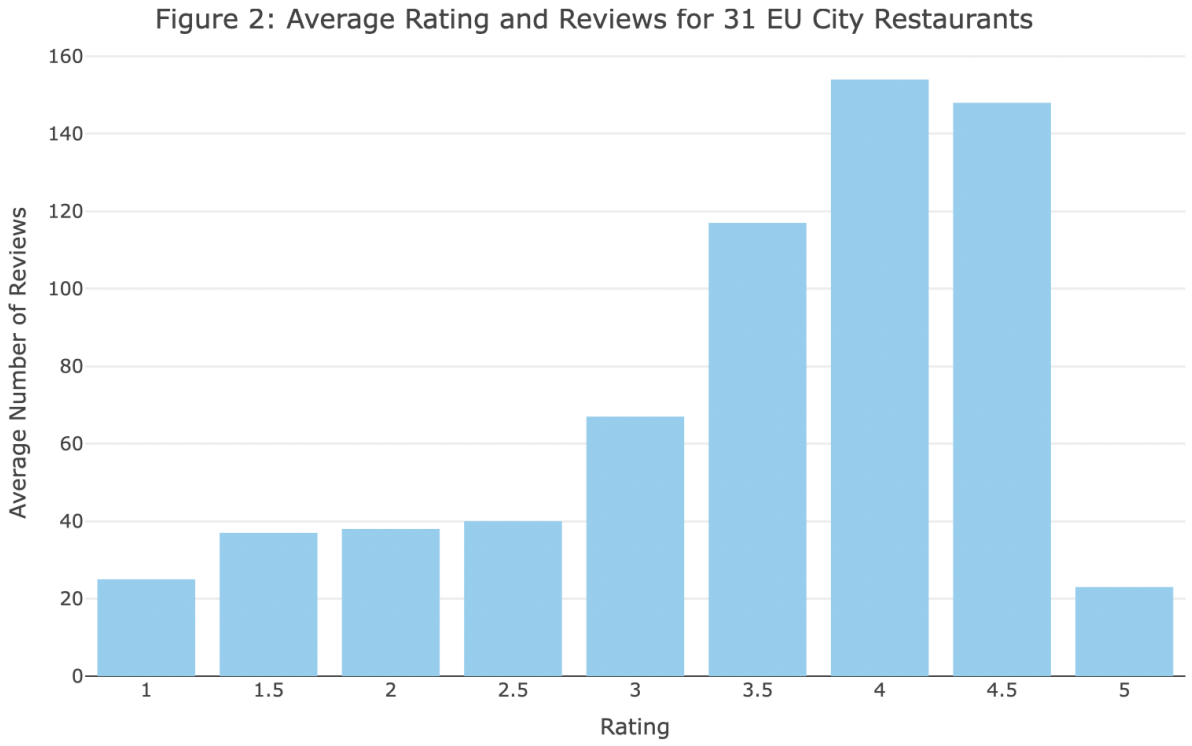


Figure 1 provides an overview of the map. Red markers are the cities with higher average restaurant rating. Please check ShinyApp for more interactive function.

Benford Analysis

From Figure 2 we could see that in general, higher ratings are associated with higher number of reviews. In reality, a restaurant with the same rating but a higher number of reviews would be prefered by custormers. (Check ShinyApp for interactive version of figure 2)



Therefore, some businesses might hire people to generate fake reviews to increase the rating and the overall number of reviews.

In order to detect possible fake reviews, Benford's Law analysis was conducted on number of reviews. Theoratically, number of reviews for each restaurant should be random and follow the Benfords Law distribution.

From Figure 3, we could see that in general the number of reviews follows One-Digit Benfords Law distribution, and no obvious susbicious reviews are detected.

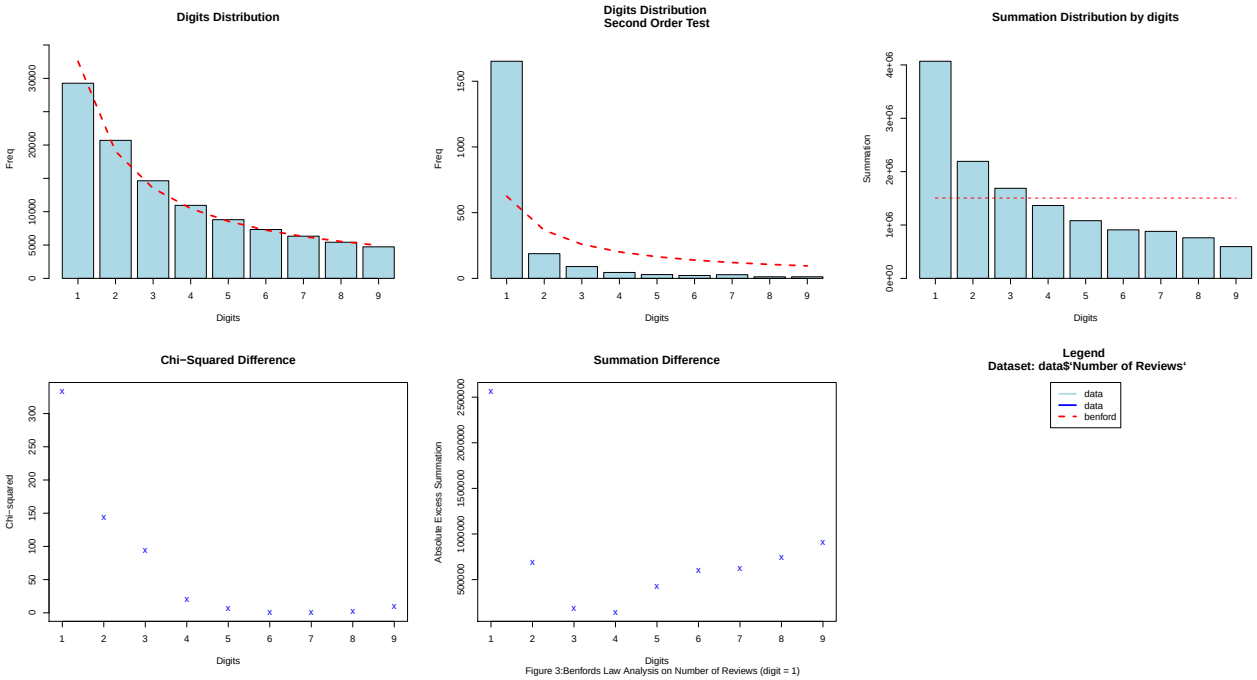


Figure 3:Benfords Law Analysis on Number of Reviews (digit = 1)

Figure 4 shows Two-Digit Benfords Law Analysis, which could not give us a clear answer, becasue in two-digit analysis, any single digit number will be treated as two-digit number. For example, a restaurant with 8 reviews would be counted as 80 reviews.

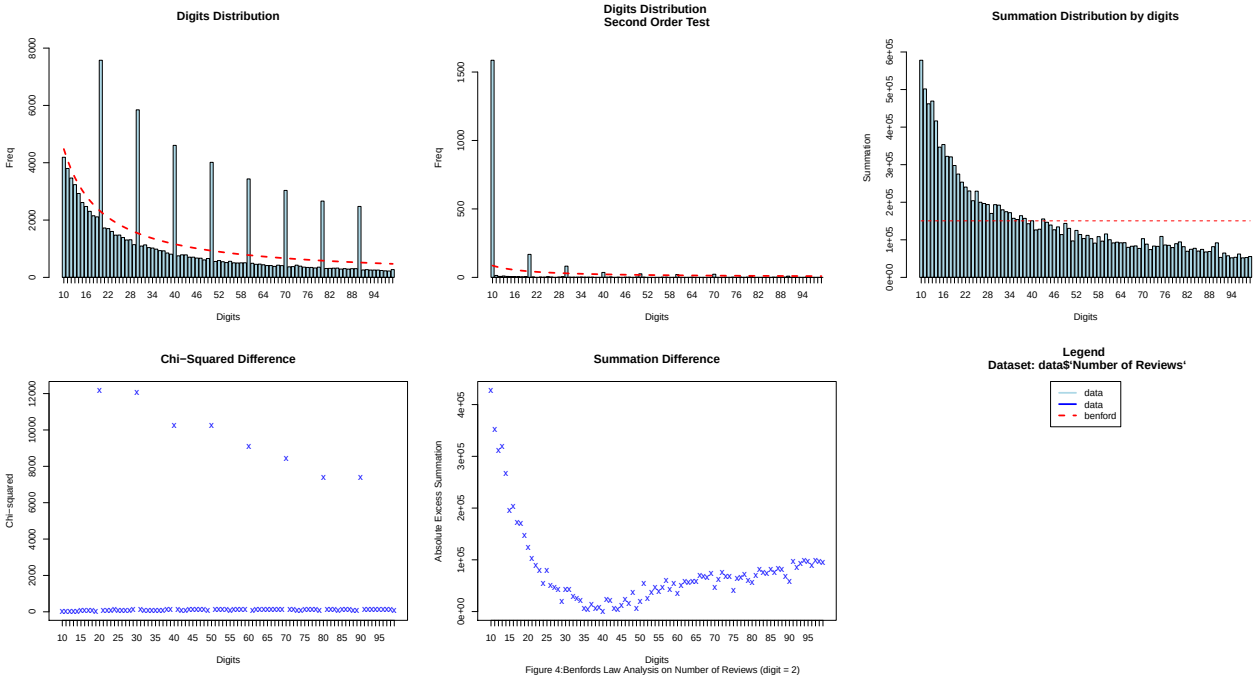


Figure 4:Benfords Law Analysis on Number of Reviews (digit = 2)

Text Analysis

Text analysis focused on the most frequently used word in reviews of each star level. Non-characters, common stopwords and customized stopwords were all deleted before the analysis.

From Figure 5 and Figure 6, we could see that the reviews for 1-star restaurants are mostly strongly negative, like “terrible”, “horrible”, “worst”, “disgusting” and etc.

Figure 5: Word Count for 1−star Restaurant

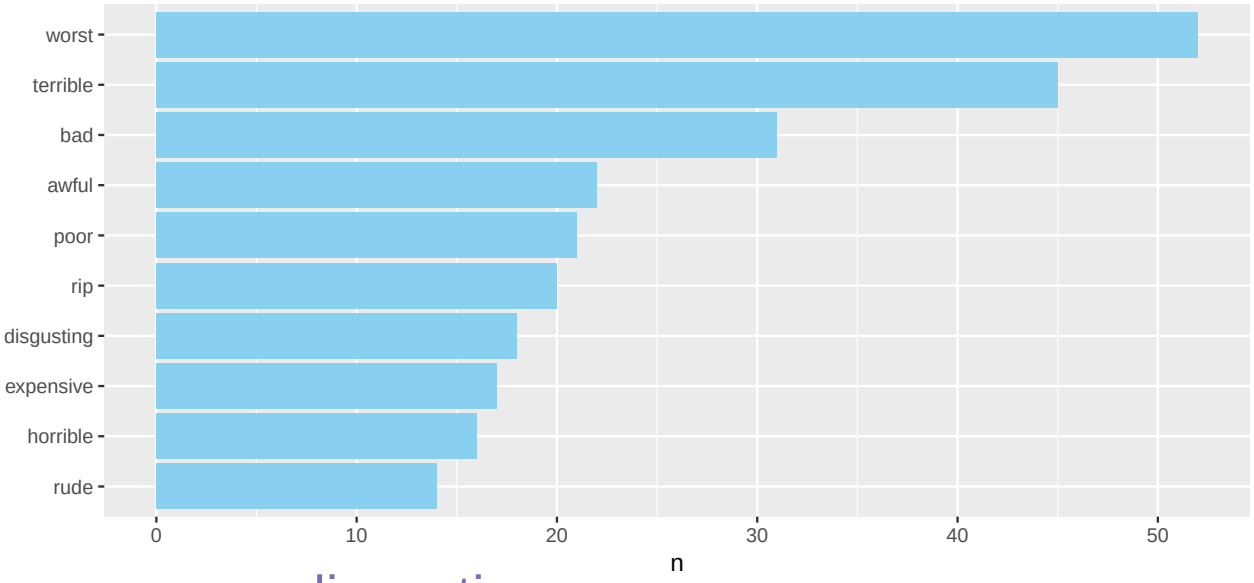


Figure 6: 1−star Restaurant Reviews

From Figure 7 and Figure 8, we could see that the reviews for 2-star restaurants are slightly less negative than the ones for 1-star restaurants. Still, negative words like “bad”, “poor” and etc. are commonly used.

Figure 7: Word Count for 2−star Restaurant

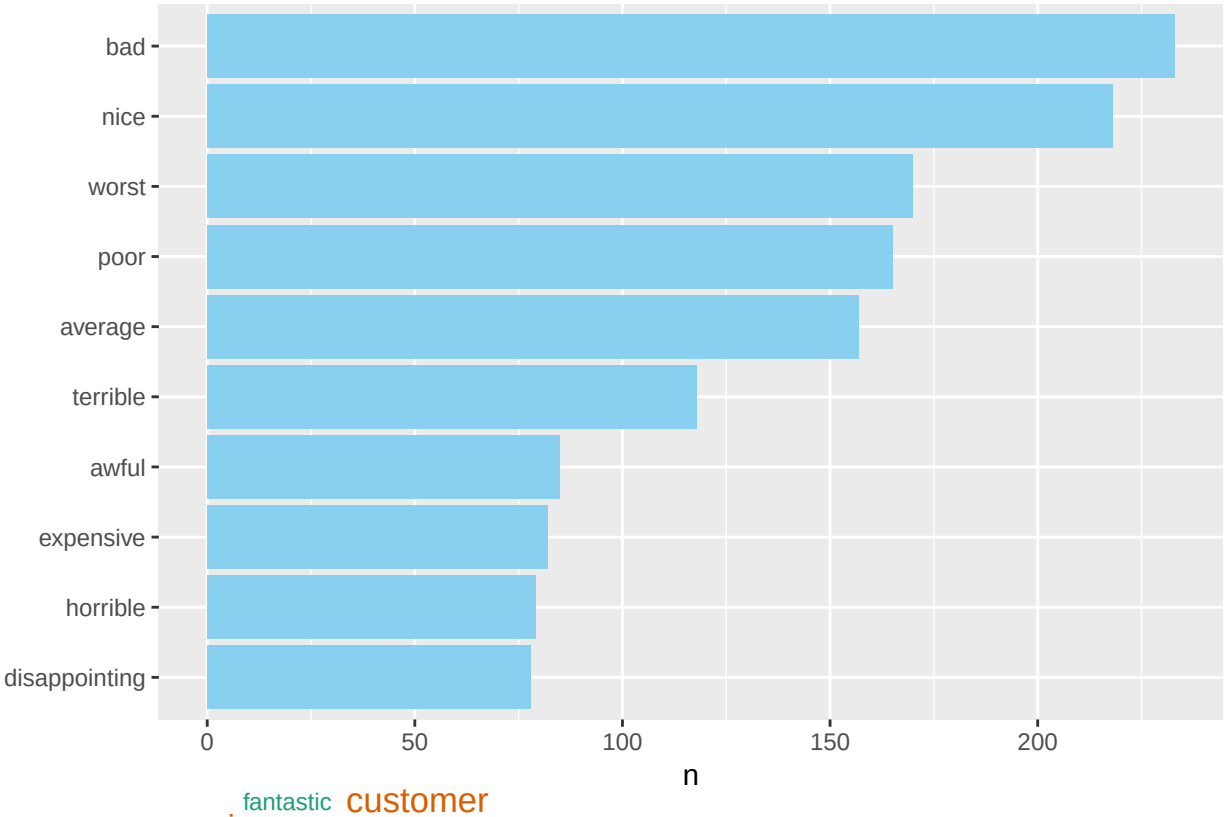


Figure 8: 2−star Restaurant Reviews

From Figure 9 and Figure 10, we could see that the reviews for 3-star restaurants look much better. More than half of most frequently used words are positive, a few negative words are used though.

Figure 9: Word Count for 3−star Restaurant

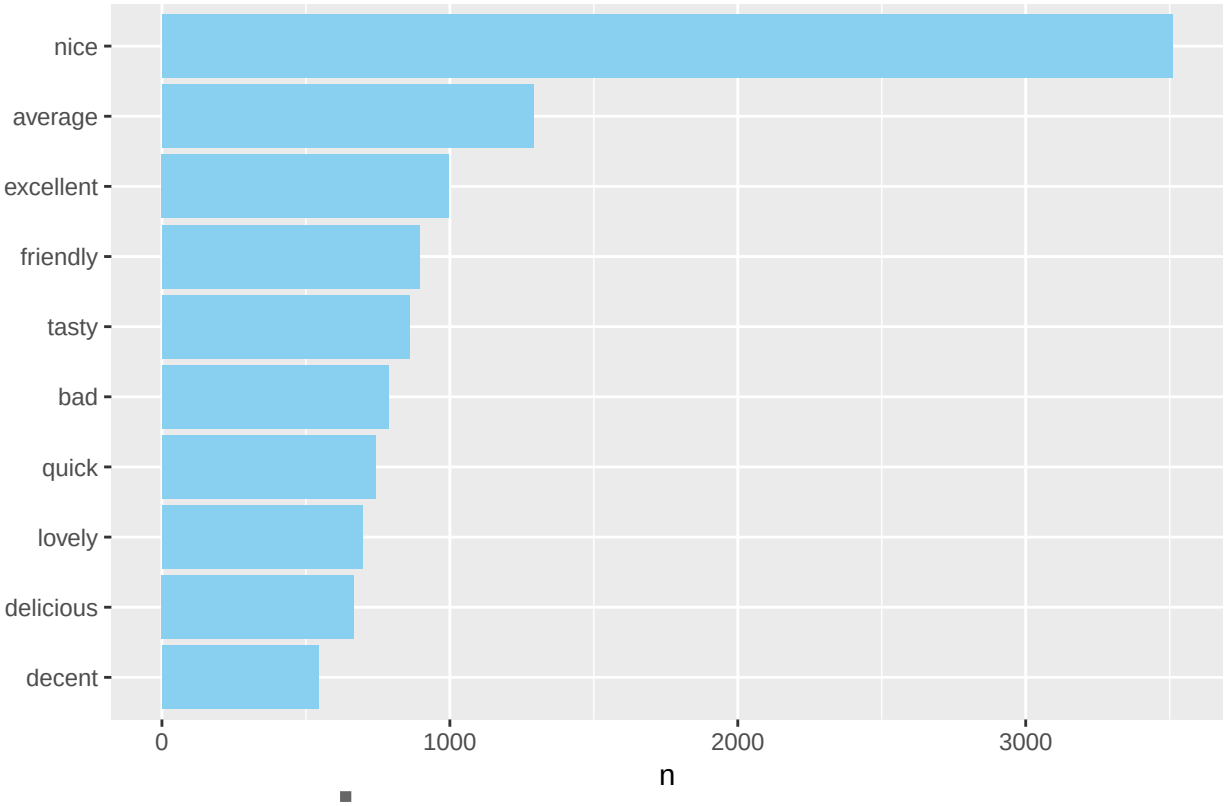


Figure 10: 3−star Restaurant Reviews

From Figure 11 and Figure 12, we could see that the reviews for 4-star restaurants are mostly positive.

Figure 11:Word Count for 4−star Restaurant

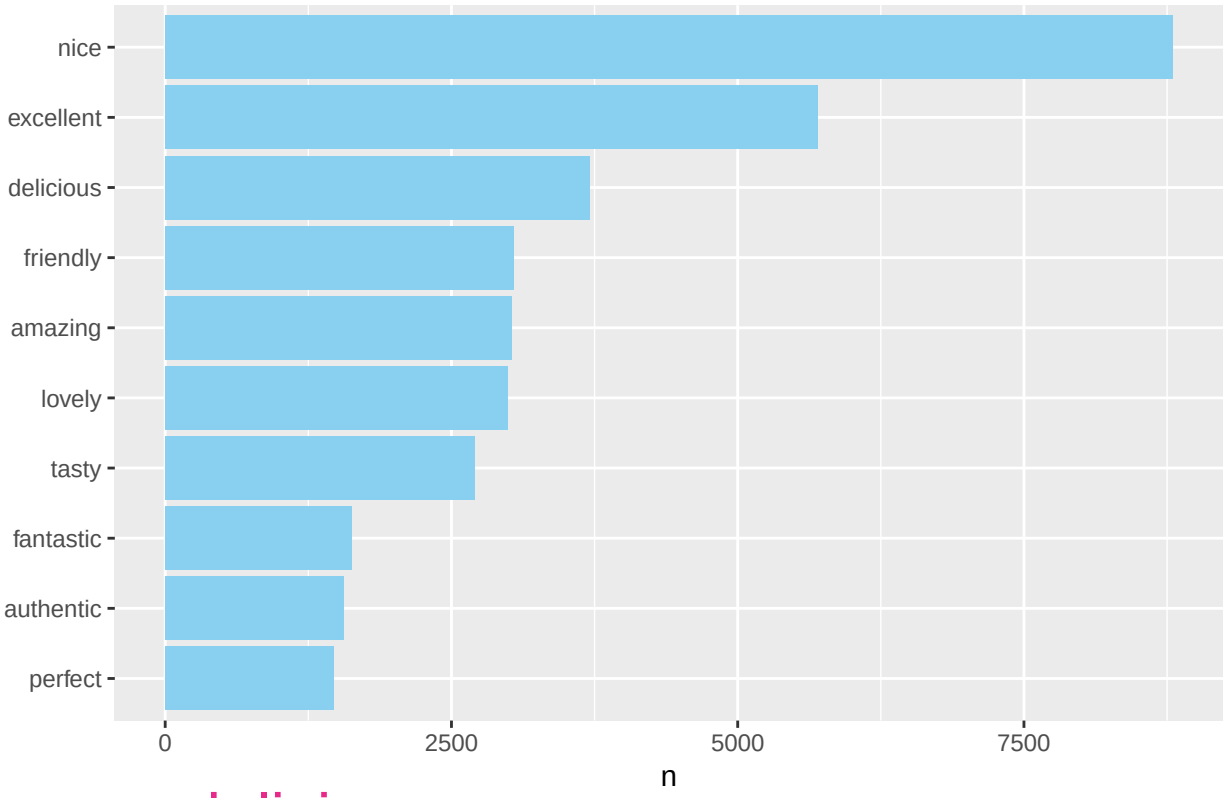


Figure 12: 4−star Restaurant Reviews

From Figure 13 and Figure 14, we could see that the reviews for 5-star restaurants are very similar to reviews for 4-star restaurants.

Figure 13: Word Count for 5−star Restaurant

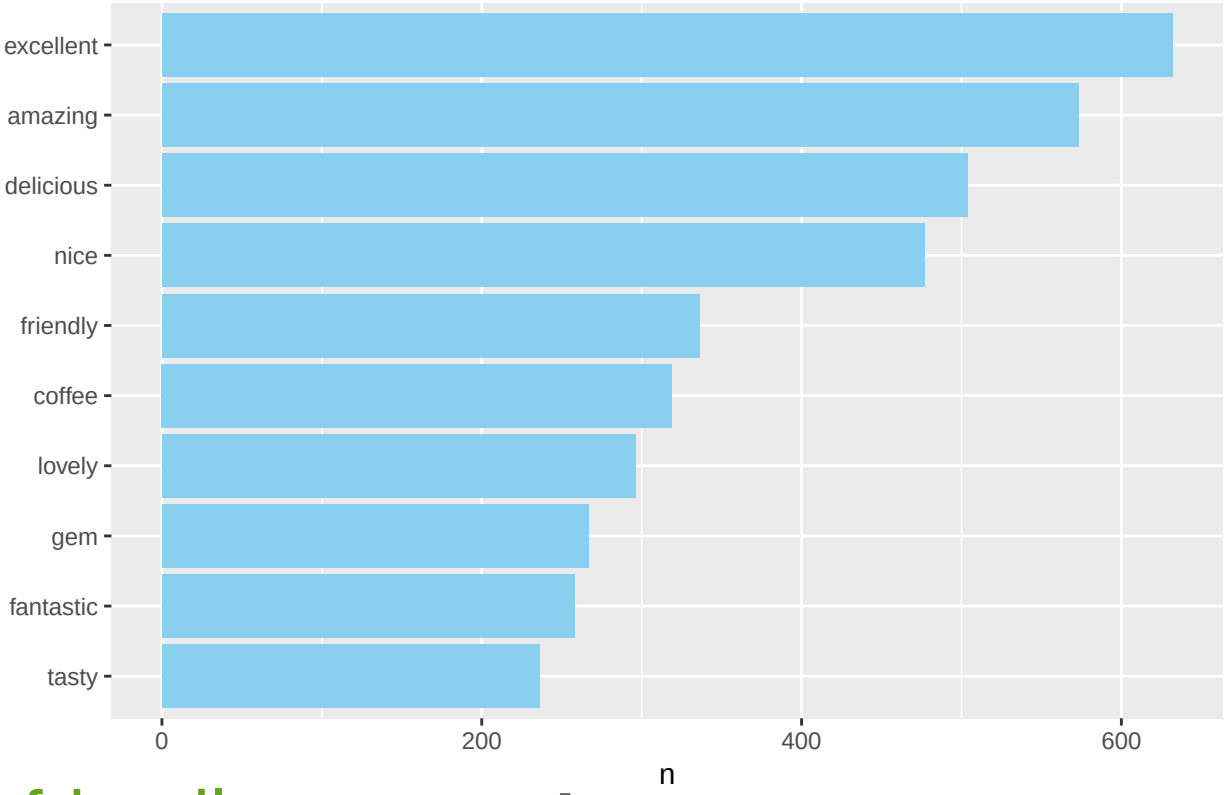


Figure 14: 5−star Restaurant Reviews

Conclusion

From the analyzed dataset, the analysis suggests that the ratings and reviews of the restaurants from the major 31 European cities look fair. No susbicious reviews are detected by using Benford's Law analysis. The review contents are consistant with ratings in general. Restaurants with comparatively higher ratings are mostly in South East Europe.

Citation

[1] TripAdvisor Restaurants Info for 31 Euro-Cities. https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw

[2] Simplemaps World Cities Database. https://simplemaps.com/data/world-cities