Restaurants near me - Uber Eats

Yunzhi Chen 32051018

2022-09-29

Introduction

Since 2020, the global demand for food delivery has grown exponentially (Research & Ltd, n.d.) as COVID-19 causes consumers to stay at home and work remotely. Correspondingly, the takeaway market has also become very popular. In 2014, Uber launched an online food ordering and delivery platform which is called Uber Eats. Users can use a mobile app, or through a web browser, to read menus, view restaurant locations and ratings, order and pay for food from participating restaurants. Besides, with Uber Eats delivery, all people favorite foods are right at doors with just a tap of phone.

This report uses **R** to focuse on the following three questions to explore the information displayed on Uber Eats and hopes to provide users with an alternative way of presenting information from Uber Eats by analyzing data from gourmet restaurants across the United States and visualizing the data.

Research questions

- How are the characteristics of the most popular restaurants?
- What is the factor that affects restaurant prices?
- How much cost can people expend for hot gourmet restaurants?

Data Wrangling

The first dataset is called “restaurants”, it comes from https://www.ubereats.com, the region is USA, and has information about the rating, price and location of the restaurants. It has a dimension of 40228 rows * 11 columns, contained both text and spatial attributes. The direct link of this data is collected from kaggle.

The second uber dataset is extracted by Great Learning from PGP-DSBA, which is based on Uber driver’s trips, and contains variables such as trip start/end times, departure and arrival locations, and trip purpose, including the data of trip for meal. The tabular data has 1156 rows and 7 columns, it has text variable as well as numeric, you can find the data here.

For the restaurant data, the variables that are relevant to the question I want to explore are restaurant score, rating, location and price range, but there are some useless columns like "position" that I sorted through the function in **R** by `select`. Then I `mutated` a new column "type" by classifying each range: `$` for "Inexpensive", `$$` for "Moderately Expensive", `$$$` for "Expensive" and `$$$$` is "Very Expensive". Finally a dataset was generated with the following variables:

Table1: Variable desription for 'restaurant' dataset

	Variable Name	Type of the variable	Description
A	id	integer	ID of restaurants
B	name	character	Restaurant names
C	score	numeric	Scores received by restaurants
D	ratings	integer	Ratings received by restaurants
E	category	character	Types of restaurants
F	price_range	character	Represent by number of dollar signs
G	full_address	character	Full address of restaurants
H	zip_code	character	Zip code of restaurants
I	lat	numeric	Latitude of restaurants
J	lng	numeric	Longitude of restaurants
K	type	character	Price type

For the uber data, I first simply used the `rename` function of **R** to rewrite the variable names into a more beautiful form, and then used the `filter` function to filter out the part of the variable that is the purpose of "Meal/Entertain" to get the uber eats related data. Next, I used the `mdy_hm` syntax from the `lubridate` package to change the start and end times from character types to datetime types to facilitate subsequent time calculations and other analysis. Finally, I added order (id) by `mutate` function and `arranged` id in a decrease order. After exclude some useless data, the final data dictionary of the dataset is as follows.

Table2: Variable desription for 'uber' dataset

	Variable Name	Type of the variable	Description
A	id	integer	ID of uber trips

	Variable Name	Type of the variable	Description
B	departure	character	Departure location
C	arrive	character	Arrive place
D	distance	numeric	Distances between departure and arrive location
E	propose	character	Propose for trip: Meal/Entertain
F	starttime	datetime	Trips start date and time
G	endtime	datetime	Trips end date and time

Data Checking

Restaurant data

I first performed the following data checking process for the restaurant data.

- By looking at the score and price_range columns I found a lot of NA and null data, I used the `na_if()` in **R** to convert the null to NA and then `removed` the NA data.
- Next use `distinc()` function with `dim` statement to check if the dimension of the data set has changed to determine if there are duplicates, but no duplicates are found.
- I first used the `scatmat` function in the `GGally` package to find out the relationship between the id, score and ratings variables to see if there was a linear relationship between any two of that variables, as shown in Figure 1 below. From the figure, we can see that “Very Expensive”(purple) has smaller values than the other three types. Also the relationship between ratings and scores for super expensive restaurants looks just the opposite of other price types of restaurants. Common sense suggests that score and rating should be positively correlated. But for the type of “Very Expensive”, the lower the rating the higher the score, which reminds that this may be a less logical part and that the amount of this data is small enough that I can rely less on information from this type of data when doing subsequent analysis.

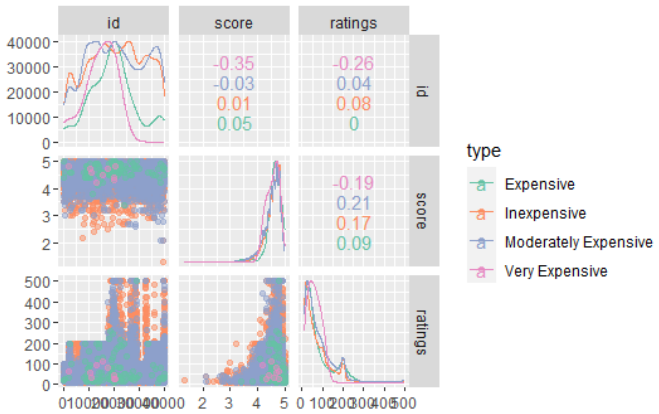

Figure1: Scatterplot matrix

- I want to check whether there is outlier data in the score and ratings columns, so I drew the distribution chart about "score" and "ratings", using the `violin plot` in `ggplot` package plus `scatter plot`, as shown in Figure 2 below. In general, the distribution of variables can be seen from the violin chart, the scores are mostly concentrated in 2-5, the ratings are widely distributed, but mostly in line with common sense. I observed that there is an outlier for one "Inexpensive" price type in the bottom left corner of the graph, and then removed this outlier by `filter`.

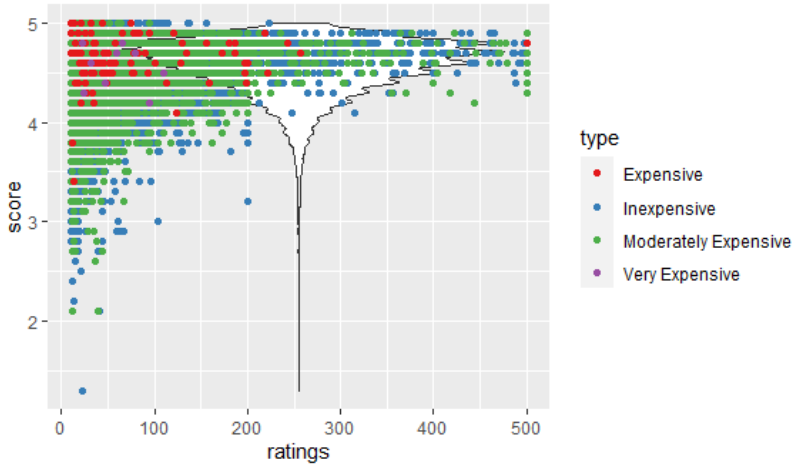

Figure2: Relationship within score and ratings

Uber data

I followed the following steps to check the uber data:

- Firstly observe the NA data and unknown data, and find that the column in "arrive" has some unknown location data, and use the `filter` to discard the above missing content.

- Similar to the restaurant data, I used `distinc` with `dim` functions to check whether the dimension of the dataset had changed to determine whether there were duplicates, the result is no duplicates were found.

- Figure 3 visualizes the scatter plot of travel distance, I tried to find the outlier and found a trip with a distance greater than 30, but this data is not considered anomalous from the practical point of view, so I kept it.

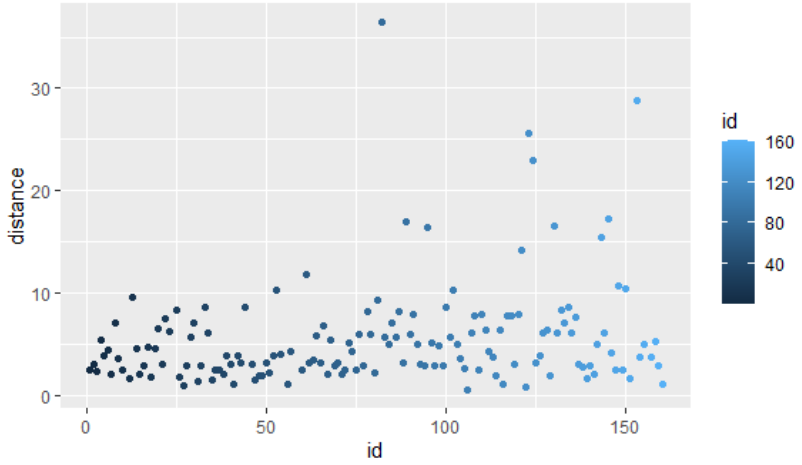

Figure3: Distance visualization

Data Exploration

Questions1: How are the characteristics of the most popular restaurants?

Regarding how to define whether a restaurant is hot or not, a score alone is not enough. After all, a restaurant may have a rating of 5 out of 5, but it may only have 10 reviews. So it's not as popular as a restaurant that scores 4.8 and 200 in the Uber Eats app. So using **R**, I processed the tidy restaurant data by filtering for scores greater than 4.5 and ratings of 150 or more. By focusing on the price type variable, I first counted the number of price types of these popular restaurants and then plotted the density graph to compare which category had the highest density.

Figure 4 reveals that the "Inexpensive" type of restaurants make up a large portion of the popular restaurants, which is also evident, who doesn't love good food at a good price?

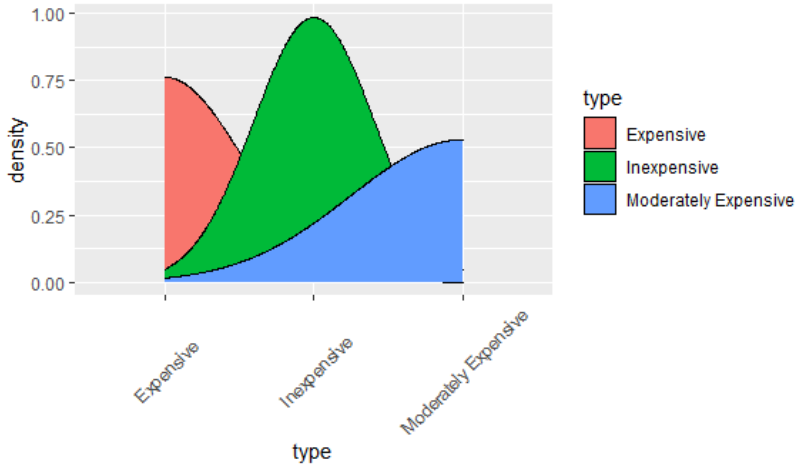

Figure4: Characteristic one - Price types of restaurants

Then I noticed that location is also an important part of popular restaurants, as some areas may concentrate on building food courts and locating them in the center of town or in some convenient places. When I grouped the zip_code using the group_by and summarise functions, I visualized the top 10 areas with the most restaurants, as figure below.

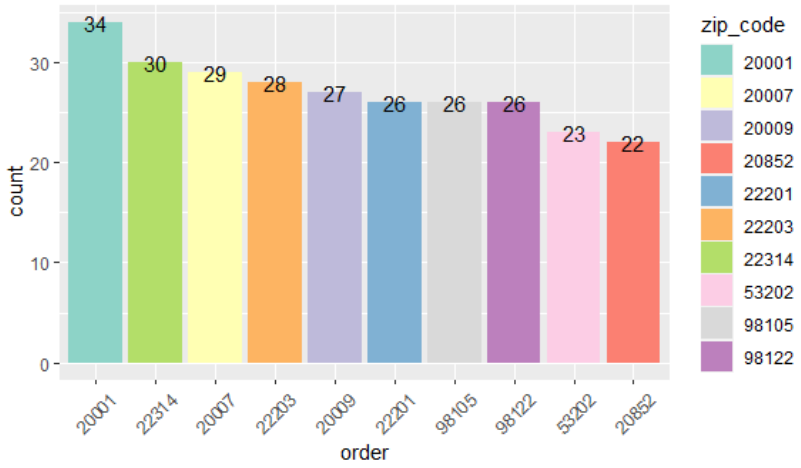

Figure5: Characteristic two - Locations of restaurants

The above figure shows that the area with zip code 20001 which is Washington has the most popular restaurants, with 34 highly rated restaurants. In recent years, Washington, the capital of the United States, has transformed into a true dining destination, with more and more great restaurants popping up all over the city (Nast, 2022).

Questions2: What is the factor that affects restaurant prices?

Restaurant prices are tied to industry changes, market price fluctuations and customer sentiment. Pricing involves considering many factors, including food costs, labor costs, what competitors are doing, and what the restaurant's target customers are willing to pay. In fact location is a very large factor in restaurant prices. Mark (2014) notes that pricing is so different across trade areas and real estate, sometimes in a counterintuitive way, that it does not always mean higher price points in higher income zip codes.

I first used the R package to filter out restaurants that were not cheap and visualized them using the latitude and longitude data contained in the dataset. Fig 6 is a presentation of the location of each price type restaurant, using facet_wrap to distinguish between the different types. Fig 7 is a plot of each type of restaurant after overlaying a map of each state in the US. Combined with Fig 6 and Fig 7, we can get that as the address changes, the price of the restaurant also changes accordingly, and the location is one of the factors affecting the price of the

restaurant. The more widely distributed types of restaurants in the country are medium expensive restaurants, and most of the restaurants with high prices are located in the capital city, Texas and other areas.

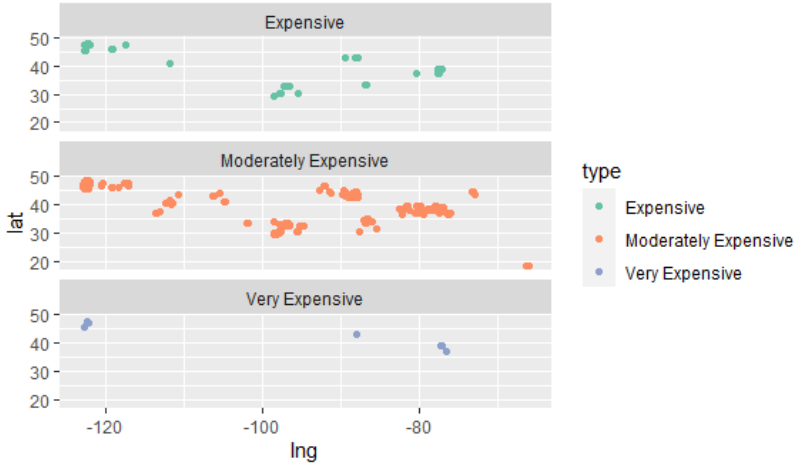

Figure6: Hot map of restaurants

Figure7: Map of high price restaurants in US

Questions3: How much cost can people expend for hot gourmet restaurants?

The uber data contains information on the time and distance spent by the user used to answer this question. I first sorted the six trips with the longest distances using the arrange function, and then presented the data in a tabular form using kabel. This is shown in Table 3.

Table3: The top 6 distances of Uber trips

id	departure	arrive	distance
82	Houston	Galveston	36.5
153	Orlando	Kissimmee	28.8

id	departure	arrive	distance
123	Sharpstown	Midtown	25.6
124	Midtown	Greater Greenspoint	23.0
145	Cary	Raleigh	17.3
89	Arabi	Metairie	17.0

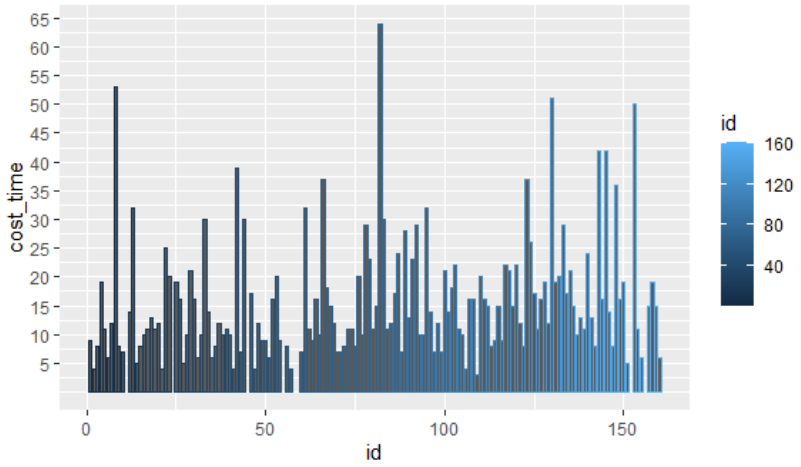

Figure8: The time cost of every trips

Figure 8 uses a method that first uses difftime() to calculate the number of minutes spent and then uses geom_col to visualize the time spent by each trip in the data set. From Table 3 and Figure 8, it can be concluded that people can spend up to 36.5 km and more than 1 hour for gourmet restaurants and delicious food.

Conclusion

By analyzing the restaurant and uber data in the above three questions, I learned the following information about Uber Eats:

- most of the popular restaurants are inexpensive and located near the capital city;
- location is an important factor in restaurant prices;
- and people can spend more than an hour and travel more than 35 km to taste the food.

Reflection

I learned a lot from this report. In the first stage, I used many tidy methods to check and organize the data, to organize the original data into a form that is easy to analyze. And in the second stage, I did a variety of data visualization to consolidate the visual data. But what I did not do well is that I tried to use regular

expressions to categorize the category column in the restaurant data and encountered a problem. The distribution of the data did not have a similar pattern, and the content of each column was not quite the same, so I could not organize it, I gave up the analysis of this column. Other areas I can improve are I can use some different type of visualizations.

Bibliography

Research, & Ltd, M. (n.d.). *Food delivery: COVID-19.* Research and Markets - Market Research Reports - Welcome. https://www.researchandmarkets.com/issues/food-deliveryon therise?utm_source=dynamic&utm_medium=GNOM&utm_code=bsnq5l&utm_campaign=1383480++Food+Delivery+Services+See+a+Surge+in+Demand+due+to+Coronavirus+Outbreak+as+Consumers+Stay+at+Home&utm_exec=joca220gnom

Nast, C. (2022, July 1). *The best restaurants in Washington, D.C., from Laotian cuisine to steakhouses.* Condé Nast Traveler. https://www.cntraveler.com/gallery/best-restaurants-in-washington-dc

Mark B. (2014, June 20). *5 factors that affect restaurant pricing strategies.* Nation's Restaurant News. https://www.nrn.com/operations/5-factors-affect-restaurant-pricing-strategies

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686 https://doi.org/10.21105/joss.01686.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. https://CRAN.R-project.org/package=kableExtra

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL https://www.jstatsoft.org/v40/i03/.

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Jason Crowley (2021). GGally: Extension to 'ggplot2'. R package version 2.1.2. https://CRAN.R-project.org/package=GGally