Youbeen Shim
Professor Jordan Rodu
STAT4310 -Data Visualization & Presentation
30 September 2017

Homework

1-1)

Data description

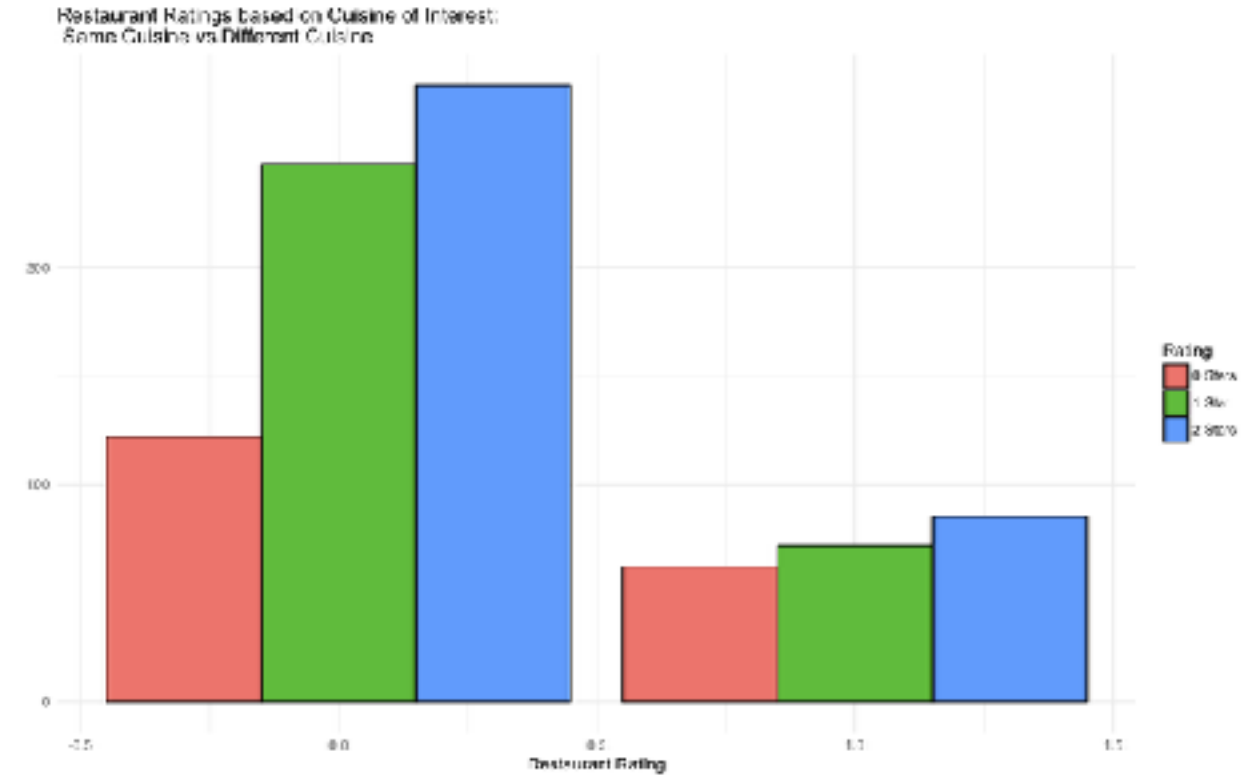
The data that was made available to us included chefmozaccepts.csv, chefmozcuisine.csv, chefmozhours4,csv, chefmozparking.csv, geoplaces2.csv, rating_final.csv, usercuisine.csv, userpayment.csv, userprofile.csv. To address the question "Do consumers rate restaurants whose cuisine is preferred differently than those whose cuisine is not preferred?", information from three different .csv files were utilized. File that indicated information on which type(s) of food that the respective restaurant served, "chefmozcusine.csv", file that indicated user's preference(s) in food, "usercuisine.csv", and file that displayed user's rating of a restaurant, "rating_file.csv".

The data was first joined using the inner_join function, using userID or placeID as a variable to join by. If the restaurant served different types of food or if the user had multiple preferences, the function replicated all other information to create another data point that accounted for the variation. The cleaning of the data revolved around determining the data entries of interest and eliminating all other redundant information. Multiple strategies were carried out, but ultimately, our final approach was as follows.

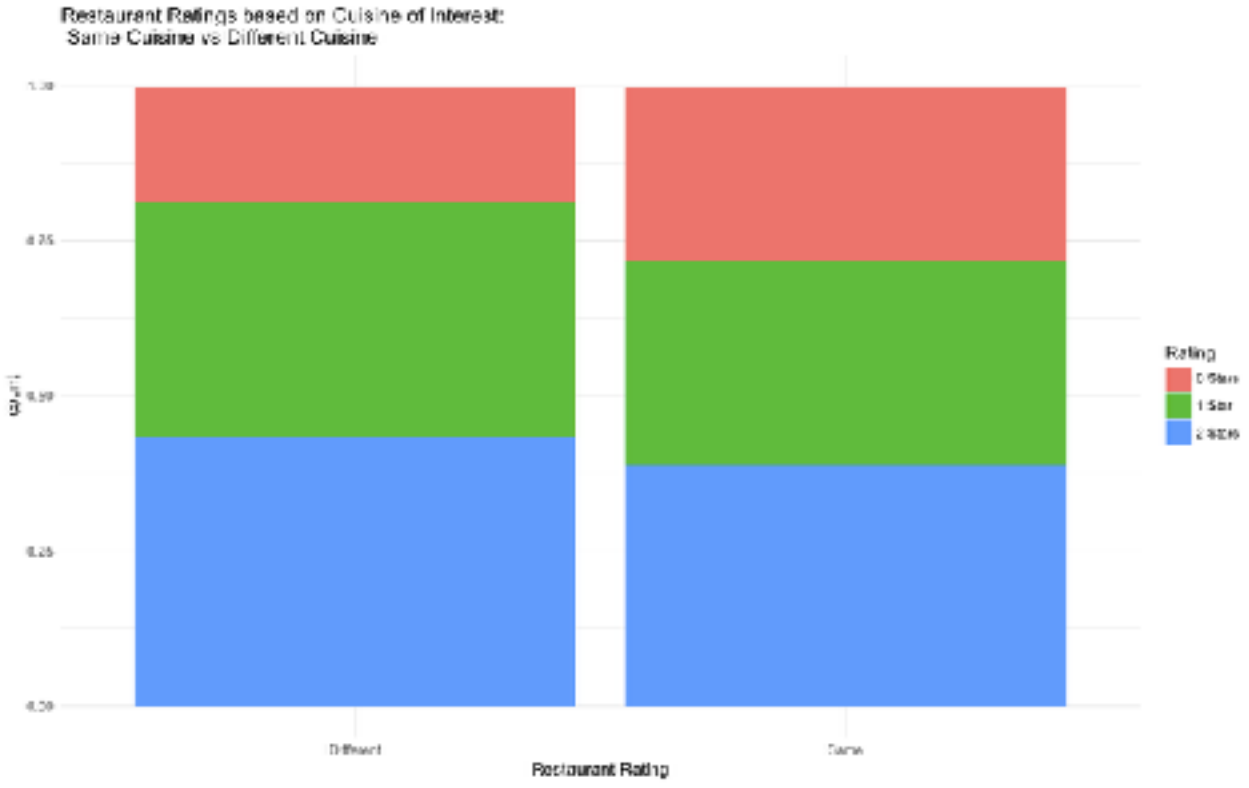
First we created a new variable via mutate( ) function. The new variable, "matching", contained a value of 1 if the user preference met the food type that the restaurant served and 0 if they did not match. The data was then separated into smaller tibbles by userId and placeID using the group_by( ) function, each tibble containing the data specific to both the user and the location. From that point, we had to choose one entry from every tibble, with a method that can obtain the data of interest for three different scenarios. (1) All possible preferences matched with all possible food type offered, or similarly none of the preferences matched with the food type offered; (2) One preference matched with one food type, but the rest did not match; (3) more than one preferences matched, but not all. We arranged the tibble in descending order based on the variable "matching" and selected the top entry. As our question concerns the difference in rating based on whether or not the cuisine offered is preferred or not, we wanted to select the incidences when the preference met the food type offered if there was a match and any random incidences of a mismatch if there were no match. Arranging in descending order and picking the top entry ensured that any match would be selected and that we would have a data representing no match when there were none.

Visualization

Reference 1-1A



Reference 1-1B



Reference 1-1A displays the bar chart of users' ratings of the restaurants. The left side shows the ratings for instances where a user rated a restaurant that did not match his/her preference and the right side shows the ratings for instances where the preference matched the food type. Reference 1-1B displays the same information, but in a relative stacked bar chart format to make the proportional comparison easier.

In interpreting the graph above, we make a few assumptions beforehand. First, we assume that the data collected and organized is robust and reliable. Second, for the sake of convenience, we assume that any visual differences that can be detected are "significant" or at least a factor to watch out for (i.e. the visual cues were not created by the random noise in the data). Lastly, we assume that the rating of 0 stars indicate that the user did not enjoy the experience at the restaurant, and the rating of 2 stars indicate that they had a fantastic experience.

Reference 1-1B appears to imply that, somewhat counterintuitively, users tend to rate the restaurants where there preferences match the type of food served more critically. Reference 1-1A also indicates that an individual visits (more specifically rates) restaurants that do not match his/her preference more frequently. However, it is important to note that the frequency that the user rates/visits the restaurants where their preferences match are much higher than what would have occurred via random selection.

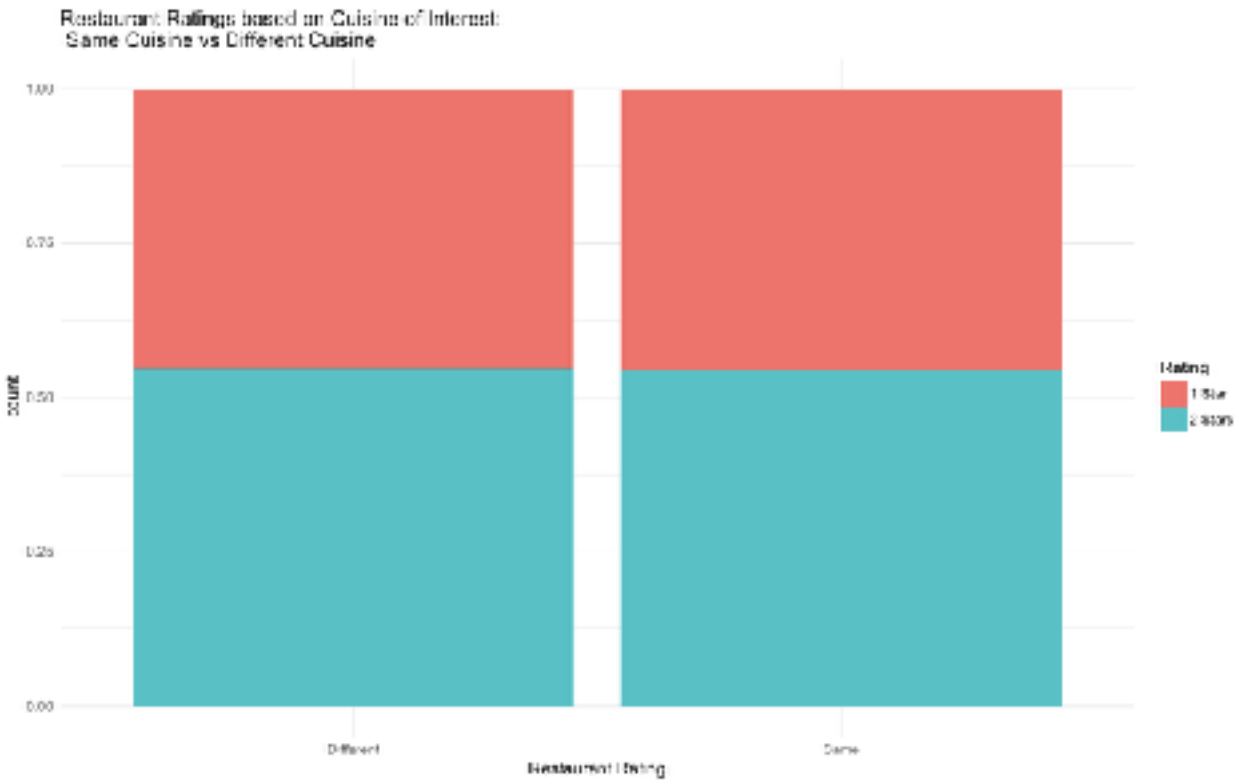
1-2)

Data description

Part 1-2 followed the exact same pipeline as part 1-1 with one additional filter, filter(rating !=0) which filtered out all the "0 star" ratings.

Visualization

Reference 1-2A



Reference 1-2A shares the assumptions made in Reference 1-1A. It is a stacked proportional bar chart that displays the proportion of 1 star and 2 star ratings, left bar is for instances where preferences do not match and the right bar is for instances where the preferences do match. Interestingly, when the rating of 0 stars are removed, it appears that the users do not rate the restaurants differently whether or not the preferences match.

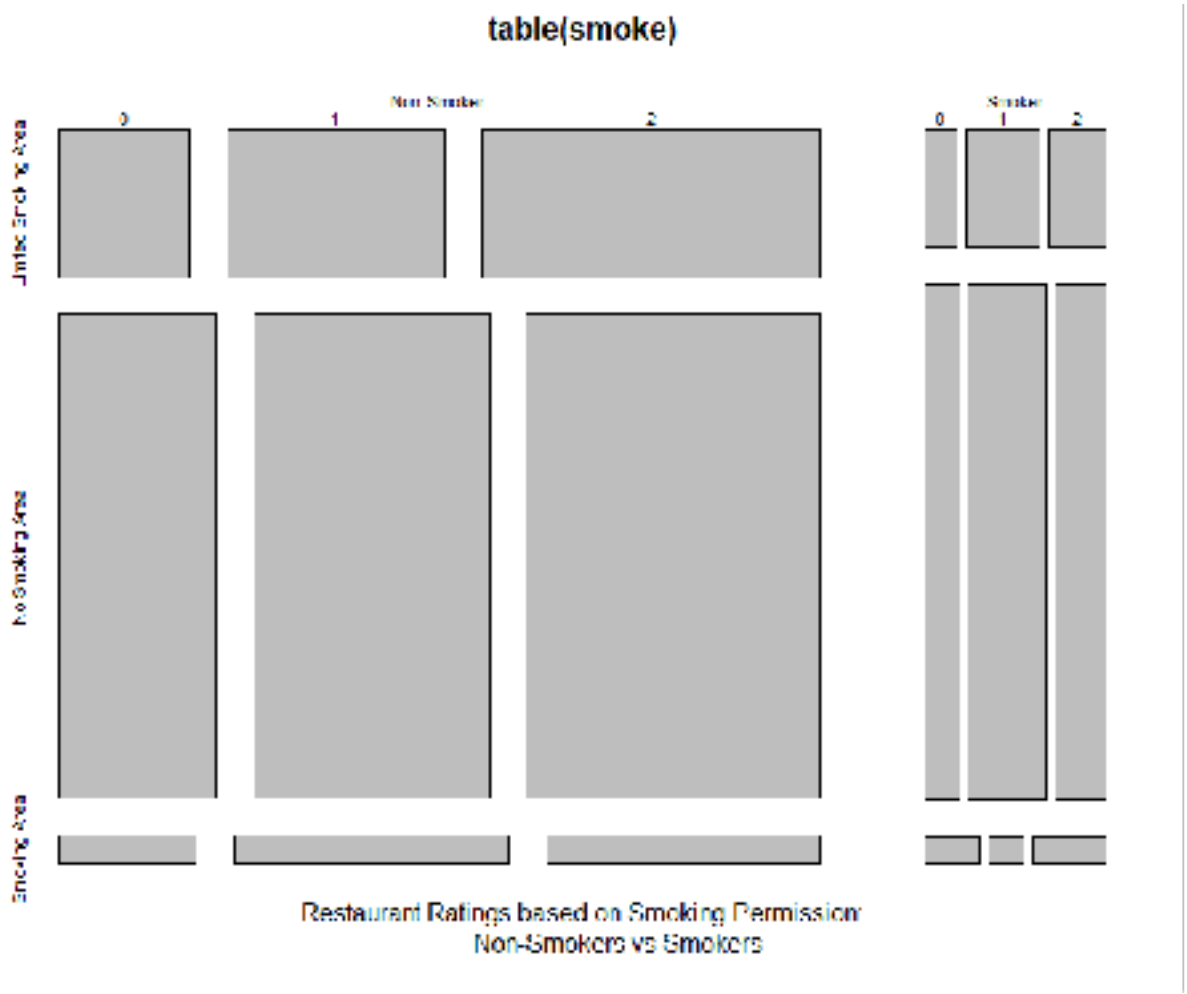
1-3)

Data description

After a bit of exploring on all the data offered, we decided to create a visualization that addressed the question "Do smokers rate restaurants differently if they were allowed to smoke during their meal?". Unlike parts 1-1 and 1-2, there were limited cleaning to be done, as all users were neatly labeled either "Smoker" or "Non-Smoker" and restaurants fit under the category of "No Smoking Area", "Limited Smoking Area" or "Smoking Area". No replication was made, all data entry after inner_join( ) were of interest. Level "none" for "smoking_area" were assumed to mean no information rather than no smoking area, and thus filtered out. Levels "only at bar" and "section" were mutated to represent "limited smoking area" as the two levels were equivalent in the context of the question.

Visualization

Reference 1-3A



Reference 1-3A shares the assumptions made in Reference 1-1A.

Reference 1-3A is a mosaic plot that displays the proportion of user's ratings, whether or not user is a smoker, and whether or not the restaurant allows smoking within the restaurant. The vertical length of the rectangles display the proportion of the ratings based on the smoking policy and type of user. The horizontal length of the rectangles display the proportion restaurants (or specifically reviews) that apply to one of "Limited Smoking Area", "No Smoking Area", or "Smoking Area".

Graph shows that there are far more restaurants that do not have a smoking area at all. Additionally, it is possible to see that the smokers have an observable preference for restaurants that allow smoking, while the non-smokers appear to be generally unaffected by the restaurants policy on smoking. Note that this may be due to individuals with a strong stance against smoking (those who are likely to rate a restaurant negatively due to smoking) avoiding the restaurant all together.

2-1)

Data description

The object of the second question was to "Create a visualization that suggests that people use the bikes to get to work.". In order to address this question, we used the dataset 201508-citibike-tripdata.csv. This dataset contained 1,179,044 entries with 15 variables, including trip duration, start/stop time, start/stop location, user information, and bike identification. The only variable of interest was the start time -while it would be interesting and informative to account for location, trip duration, and user type, the sheer size of the data both "silences" the rare and extra-circumstantial incidences and slows the process down considerably beyond the bare minimum calculation.

Using the tidyverse package "lubridate", we were able to parse the date input into something that R could understand. Then we mutated/extracted the date of the week (bikingday) and the hour at which the bike ride started (bikinghour). This data was then grouped according to bikingday and bikinghour and summarized for the number of instances per hour at each day of the week, creating 7*24=168 row tibble.

Visualization

Reference 2-1A shares the assumptions made in Reference 1-1A. Additionally, we assume that the rare occurrences, such as cross-country bike trip, are "silenced" by the size of the data and that the pattern displayed are true representation of bike use, even though all the bikes in the data are not personal bikes but citibikes.

Reference 2-1A is a line graph that displays the number of bikes used at the given time on the given day of the week. Each line represents the day of the week, x-axis represents time, y-axis displays the number of bike rides.

The graph shows two very distinct observable patterns based on the day of the week. On weekdays, there are two peaks - one around 8 am and one around 6 pm. These peaks are normal work hours for individuals, implying that bikes are being used for work. On the contrary, the curve during the weekend is round, gradual, and generally follows the activity of the sun.

Reference 2-1A