Introduction

For part one of this project, we analyzed a dataset that describes the tips given to waitresses at a certain restaurant over a period in 1990. This dataset had 244 records and each record had 7 attributes. Each record represents a single group of people who come into the restaurant and purchase meals. These attributes are total bill, tip, sex, smoker, day, time, and size. The total bill attribute describes the total bill paid by the party, and the tip attribute describes the total tip given by that party. The sex attribute describes the gender of the tipper, and the smoker attribute is a binary attribute that describes whether the tipper smokers or not. The day attribute describes the weekday that the record occurred on, and the time attribute describes whether that record was during breakfast, lunch, or dinner. Finally, the size attribute describes how many people are involved in one record. We used these attributes to answer four questions. First, we determined the average customer tip. Then we determined which gender of customer tipped more. We then grouped the dataset by the day of the week and determined which day of the week the restaurant was the busiest. Finally, we determined if customers tipped more on certain days of the week.

In part two of this project, we chose to do exploratory data analysis on Ramen Ratings with more than 2500 data. The dataset was collected from Kaggle which was an exported version of "*The Big List*" [1]. It contains information about the ratings of different types of ramen on a scale of one to five. Each record in the dataset represents single ramen review. The attributes of the dataset are as followed: review number, brand, variety, style, country, stars, and top ten. The review number served as a unique numerical identifier for each row. The brand indicated the name of the company that made the ramen. The variety represented the name of the individual product, and the style indicated how the ramen was packaged or served. It could be either in a cup, pack, bowl, tray, or box. The country column described the country where the ramen was made, and the

star attribute was a numerical description of the score of the ramen out of five. The top ten attribute was null in all cases except where the ramen it described had attained a position among the top ten ramens of that year. In that case, it contained the year of the award and placement among the top ten. With this information, we wanted to determine which brands and which ramen styles were the most popular based on their ratings.

Background

We performed our analysis using the python programming language. This was done in a Jupyter Notebook to provide interactivity and flexibility. We used the *numpy* library for handling arrays, the *pandas* library to represent and manipulate our dataset, and the *pyplot* library from matplotlib in order to visualize our results.

We decided to use the Ramen Ratings dataset because both researchers had personal familiarity with eating ramen. We also thought that useful information could be gleaned from this dataset.

Experimental Setup and Results

In part one of the assignment, we followed these steps:

1. Firstly, we loaded the data into a dataframe. We renamed the time and size columns to meal and party size respectively. We obtained some basic statistics about the dataset with the describe function.

2. Next, we considered the average tip amount. We obtained this with a built-in pandas function, and then displayed some basic information about this average, as well as a box plot. The average tip amount was approximately 2.998.
3. Then we split the dataset by gender. We displayed the relative amounts of male versus female, as well as boxplots and scatter plots about the tips given by each gender. On average, men tipped slightly more than women, with men tipping 3.09 and women tipping 2.83 on average.
4. Next, we grouped the dataset by weekday. We showed a line graph about the number of parties on each given weekday. The day with the least number of records was Friday, while Saturday had the most records.
5. We also showed which weekdays most of the tips fell on using a pie chart. Over 35% of all tips fell on Saturday.

In part two we followed these steps:

1. Import the dataset "**Ramen Ratings.csv**" in a pandas DataFrame named *df* by using ***pd.read_csv ()*** function. Here, we changed the data type of 'Stars' from string to float and assigned "Unrated" for the null values. We also set the "Review#" column as an index.
2. Print out the name of the attributes in the dataset.
3. ***head ()*** function was used to view the part of the data.
4. Determined the total number of observations this dataset has by using ***shape*** attribute.
5. ***describe ()*** function was used to get the basic statistics of the numerical attributes i.e., max, min, mean, standard deviation and the interval values required for box plot.
6. We also determined the basic statistics of all attributes.

7. Investigated total number of null values per attribute and from this analysis we found that "Top Ten" has most nulled values of 2539 out of 2580 i.e., this attribute is not important for further data analysis.
8. In last step, we dropped the "Top Ten" column.

After completing the above steps, we headed to precise data analysis by investigating aforementioned questions.

Q1. What brand has the most stars on average?

I. Get the average amount of stars per brand by using groupby () function.

II. On a scatter plot, this data was plotted where x_axis represents "Brand" and y_axis represents "Average Stars".

III. Get basic statistics about the 'Stars' column using the describe function. This helps give an idea about the min, max, median values of tip in addition to other basic statistics.

IV. Boxplot of Stars column was plotted to depict the visual representation and identify the outliers.

V. In the final step, we grouped the companies in a pie char as 'nostar', 'onestar', 'twostar', 'threestar', 'fourstar' and 'fivestar' based on their average ratings. Here, we also made sure the pie chart shows as a perfect square.

Q2. Which one is the most popular style of ramen?

To answer this question, we took the following actions:

I. ***describe ()*** function was used on the Style columns to get basic statistics like total number of values, number of unique styles, top-most style name and the frequency of most popular style.

II. Determined the name of most popular style of ramen which is Pack.

III. Created a bar chart that shows the comparisons among the number of styles each country produces. Assigned labels to the x-axis = 'Countries', y-axis = 'Total Number of Styles' and title = 'Country-Wise Ramen Style'. From this analysis we found that Japan is the top-most country for ramen production.

Conclusion

In conclusion, the tips dataset informed us of many things about the typical tipping habits. For example, we determined that men tip more than women, and that both genders tip close to around 3.00. From the ramen rating dataset, we recognized the popularity of the ramen pack, and that Japan produces more ramen in comparison with the rest of the world. These are just some of the things we can learn by applying data science tools to readily available data.

References

[1] Bilogur, Aleksey. "Ramen Ratings." *Kaggle*, 11 Jan.2018, https://www.kaggle.com/residentmario/ramen-ratings