EAST WEST UNIVERSITY

CSE 303: Statistics for Data Science

[Summer 2023]

Report on Assignment 01

Data Visualization

Submitted by

Student ID: 2020-2-60-213
Student Name: Md. Farhad Billah

1. Introduction

The "Food Delivery across Canada (Door Dash)" dataset is a comprehensive collection of information related to food delivery services facilitated by the popular platform Door Dash across various regions in Canada. This dataset is specifically curated to provide valuable insights into the food delivery industry, consumer preferences, restaurant performance, and more.

The dataset can be visualized using a histogram, pie chart, box plot, bar chart, violin plot, scatter plot, pair plot, and correlation heatmap, facilitating a deeper understanding of its characteristics and relationships. These visualizations offer valuable insights into the distribution, trends, and correlations present within the dataset.

2. Data Characteristics

The dataset, titled 'Food Delivery across Canada (Door Dash),' contains 3,290 rows and 10 columns. It encompasses a comprehensive collection of information related to food delivery services across various regions in Canada as facilitated by the Door Dash platform. In this data set there are 2 Numerical rows (distance, num_reviews) and 5 (restaurant, url, city, category_1, category_2) categorical rows.

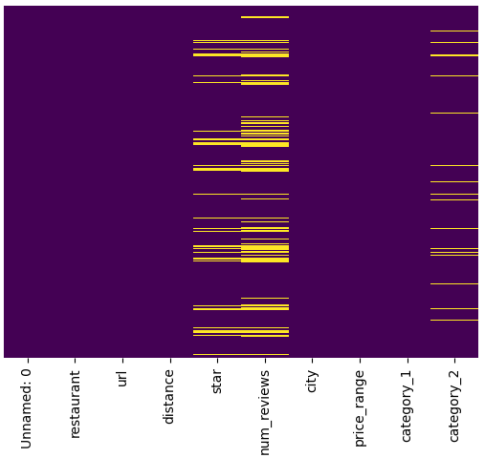

Figure 1: The heatmap for all columns.

Based on Figure 1, the heatmap representation of the dataset reveals the presence of null values in three columns. The list of columns along with their corresponding null value counts is as follows:

1. `num_reviews`: 670 null values. The number of reviews or ratings received by the restaurant.

2. `star`: 280 null values. The star rating of the restaurant is based on customer reviews.

3. `category_2`: 108 null values. The secondary category of food items, providing further classification.

Additionally, it is worth mentioning that the following columns have zero null values:

4. `Unnamed: 0`: 0 null values. A unique identification for each, serving as an index.

5. `restaurant`: 0 null values. The name of restaurant fulfilling the delivery of order.

6. `url`: 0 null values. The URL or web link associated with the restaurant's profile on Door Dash.

7. `distance`: 0 null values. The distance between the restaurant and delivery location.

8. `city`: 0 null values

9. `price range`: 0 null values

10. `category_1`: 0 null values

These findings indicate that some columns contain missing data (null values), while others are complete and do not have any null values.

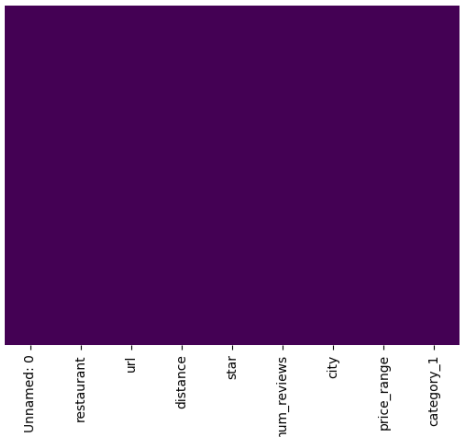

Figure 2: After cleaning data.

In Figure 2, it is evident that there are no null values present in the dataset. This outcome is a result of applying various data-cleaning methods, including removal and filling of missing data during the preprocessing stage. Now, the dataset, titled 'Food Delivery across Canada (Door Dash),' contains 3,290 rows and 9 columns.

3. Exploratory Data analysis

3.1 Histogram

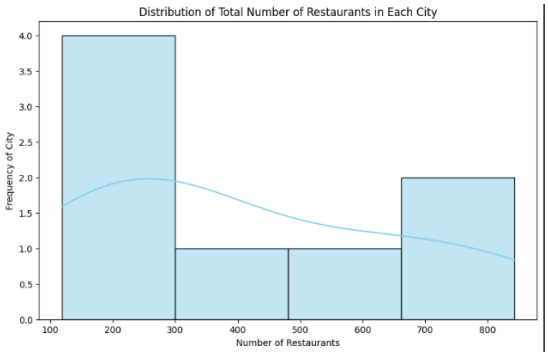

Figure 3: Distribution of Total number of restaurants in each city.

According to figure 3, The histogram is Positive Skew. Cause Mean > Median.

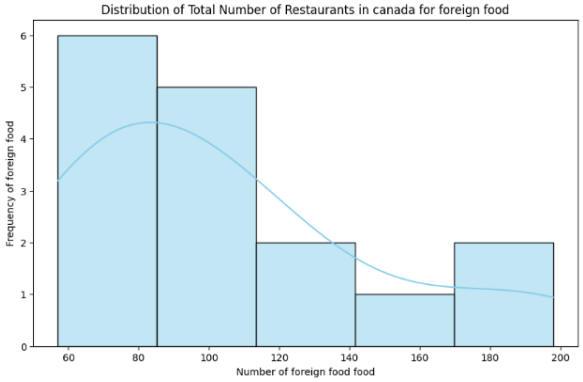

Figure 4: Distribution of total number of restaurants in Canada for foreign food.

According to figure 4, The histogram is Positive Skew. Cause Mean > Median.

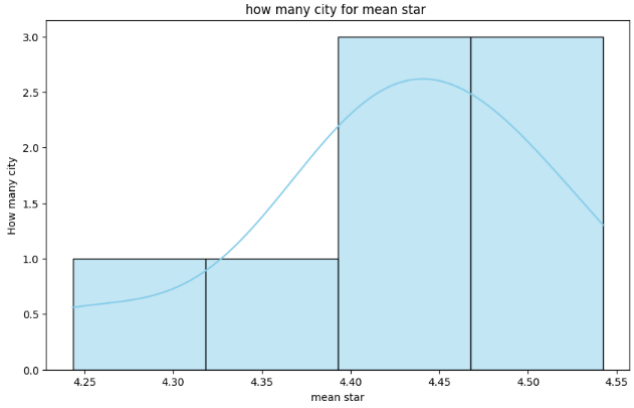

Figure 5: how many cities for mean star.

According to figure 5, The histogram is Negative Skew. Cause Mean < Median.

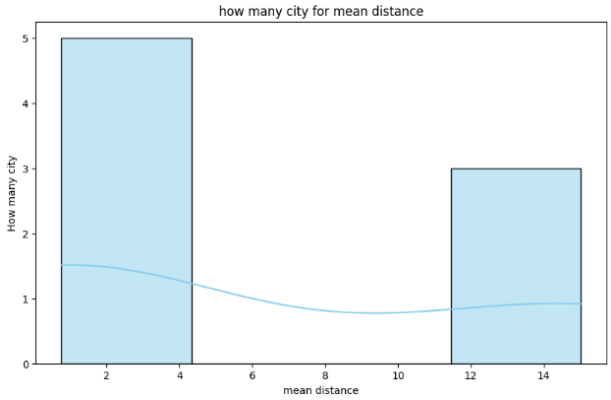

Figure 6: How many cities mean distance.

According to figure 6, The histogram is Positive Skew. Cause Mean > Median.

3.2 Bar chart

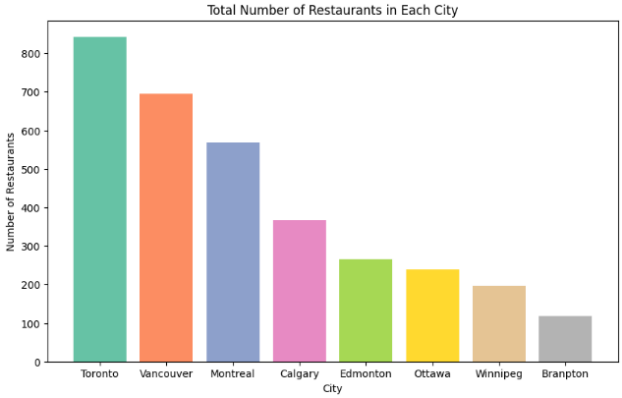

Figure 7: Total of restaurants in each city.

According to figure 7, the bar chart shows the different types of results in each city.

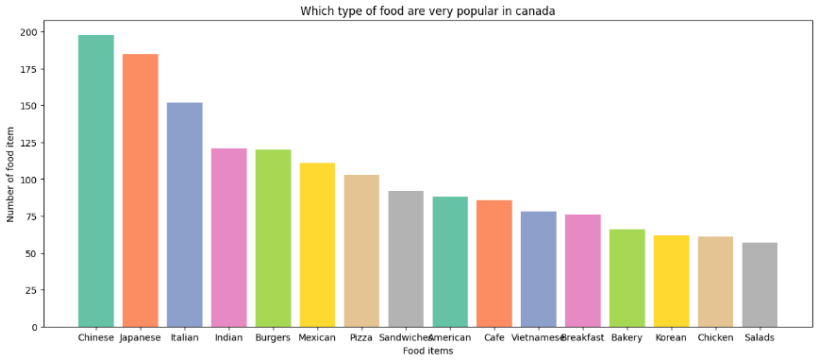

Figure 8: Which type of restaurants in Canada

According to figure 8, the bar chart shows the different types of restaurants in Canada. Also which type of restaurants in Canada.

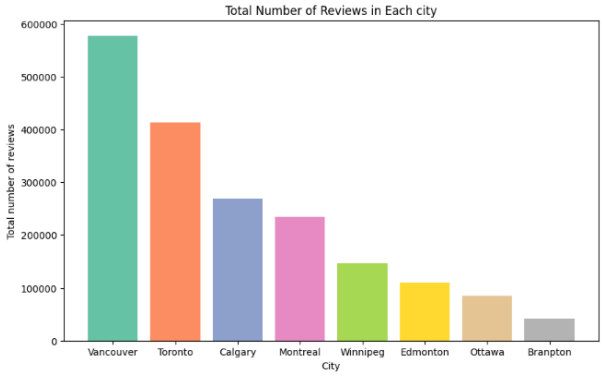

Figure 9: Total Number of reviews in each city.

According to figure 9, the bar chart shows the Total Number of Reviews in Each city.

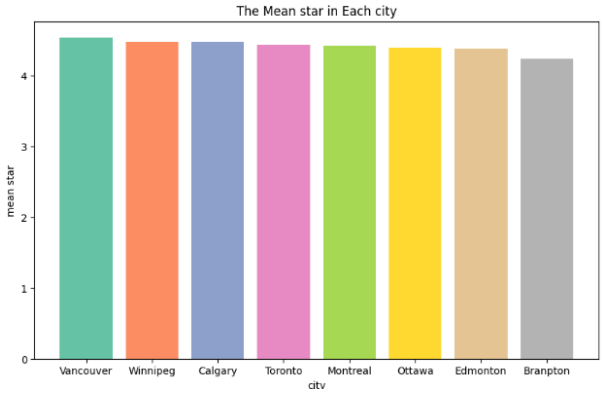

Figure 10: The mean star in each city.

According to figure 10, the bar chart shows the mean star in each city.

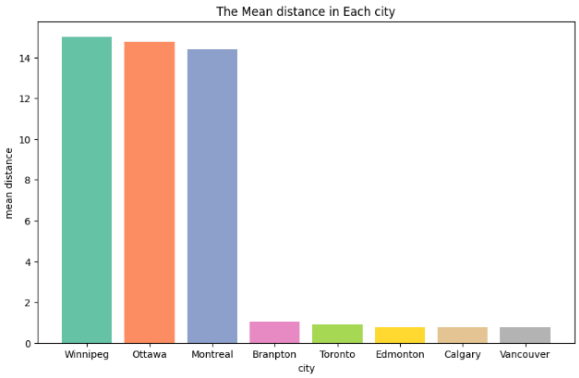

Figure 11: The mean distance in each city.

According to figure 11, the bar chart shows the mean distance in each city.

3.3 Pie chart

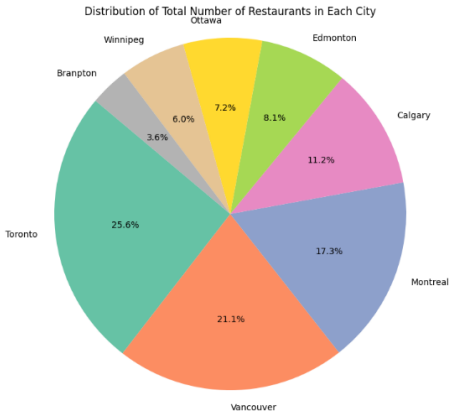

Figure 12: Distribution of Total number of restaurants in each city.

Based on Figure 12, the pie chart illustrates the distribution of different types of results across the cities. It provides a visual representation of the proportion of each result category relative to the whole dataset, making it easy to grasp the relative significance of each city's outcomes.

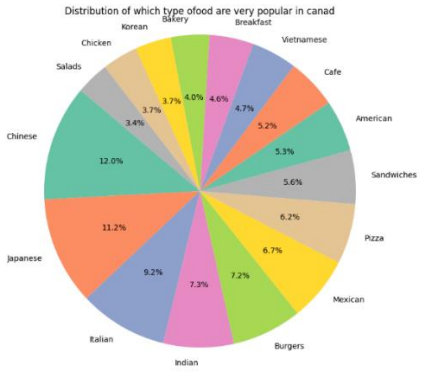

Figure 13: Distribution of which type of food are very popular in Canada.

Based on Figure 13, the pie chart illustrates the distribution of different types of food in canda. It provides a visual representation of the proportion of each result category relative to the whole dataset, making it easy to grasp the relative significance of each type of foods.

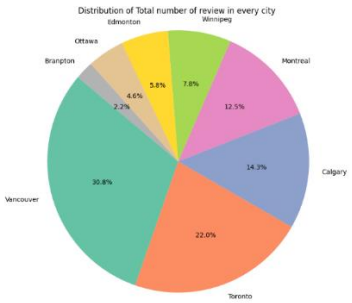

Figure 14: distribution of Total number of reviews in every city.

Based on Figure 14, the pie chart illustrates the distribution of different types of reviews across the cities. It provides a visual representation of the proportion of each

result category relative to the whole dataset, making it easy to grasp the relative significance of each city's outcomes.

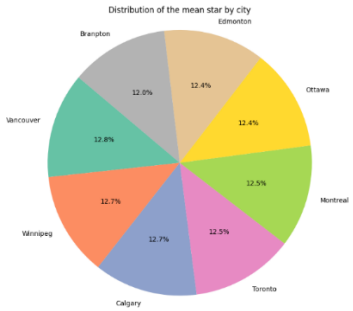

Figure 15: Distribution of the mean star by city.

Based on Figure 15, the pie chart illustrates the distribution of the mean star of results across the cities. It provides a visual representation of the proportion of each result category relative to the whole dataset, making it easy to grasp the relative significance of each city's outcomes.

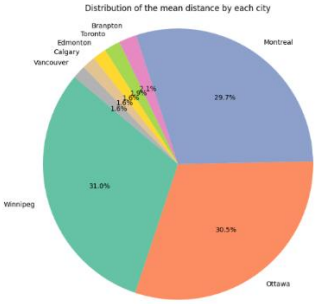

Figure 16: Distribution of the mean distance by each city.

Based on Figure 7, the pie chart illustrates the distribution of the mean distance of results across the cities. It provides a visual representation of the proportion of each result category relative to the whole dataset, making it easy to grasp the relative significance of each city's outcomes.

3.4 Scatter plot

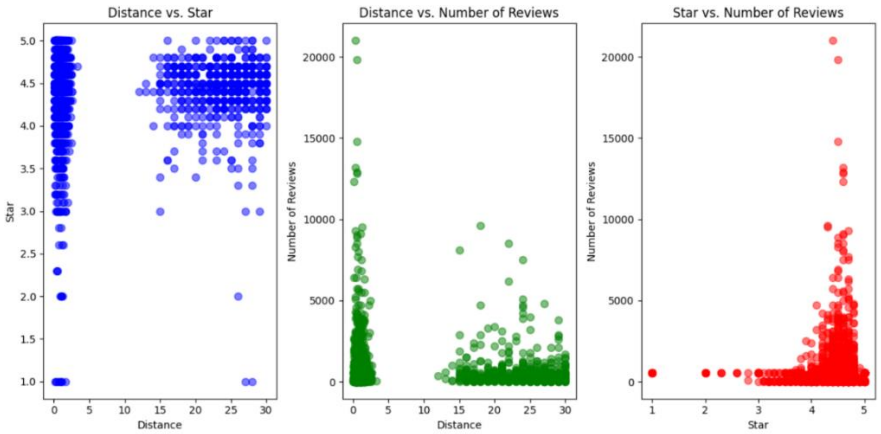

Figure 17: There scatter plot Distance vs Star, Distance vs Number of reviews and Star vs Number of Reviews

According to figure 17, the bar scatter plot (Distance vs Star, Distance vs Number, Satar vs Number of Reviews) shows the no correlation. There is no relationship between the two variables.

3.5 Box plot

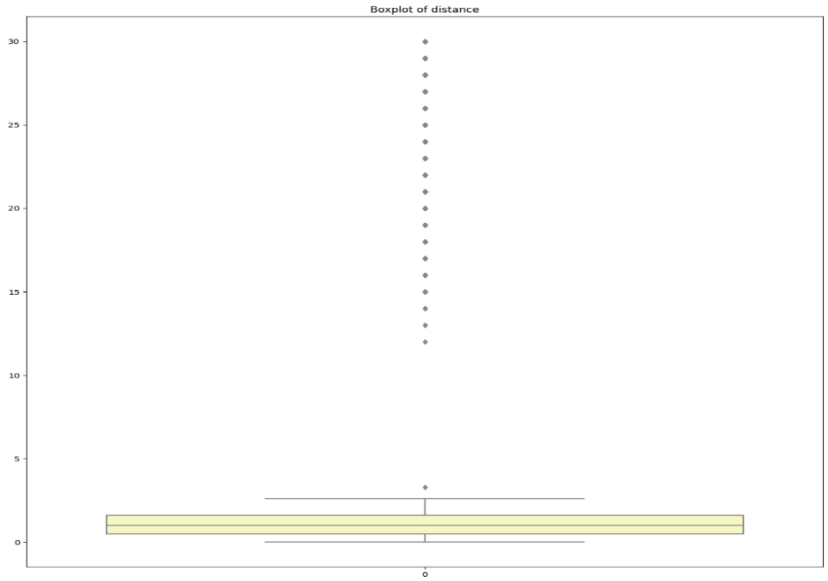

Figure 18: Box plot of distance.

Based on Figure 18, the Box Plot for the 'distance' attribute reveals the presence of outliers.

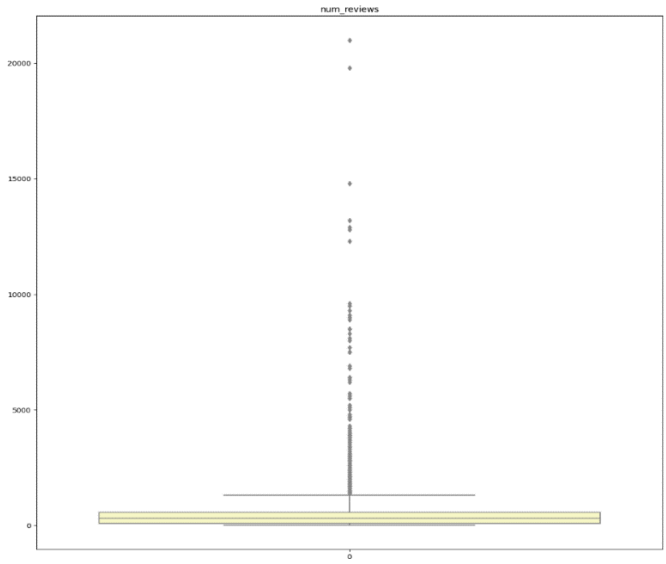

Figure 19: Box plot of num_reviews.

Based on Figure 19, the Box Plot for the 'num_reviews' attribute reveals the presence of outliers.

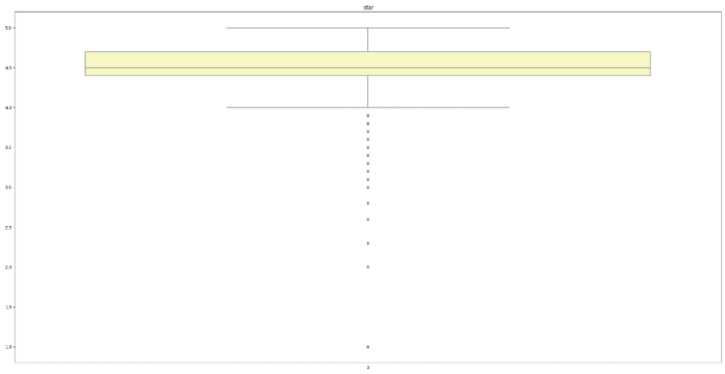

Figure 20: Box plot of Star.

Based on Figure 20, the Box Plot for the 'Star' attribute reveals the presence of outliers.

3.6 Pair Plot

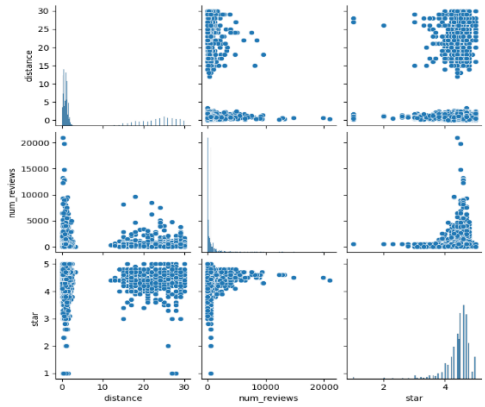

Figure 21: Pair plot of distance, num_reviews and star.

Figure 21 presents a pair plot, which displays scatter plots between different pairs of numerical attributes in the dataset. This visualization allows us to observe potential relationships and correlations.

3.7 Violin Plot

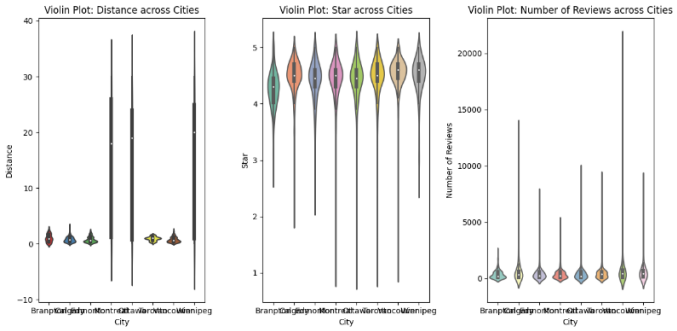

Figure 22: violence plot of Distance across cities, star across cities and number of reviews across cities.

According to Figure 22, the Violin Plot provides a comprehensive representation of the distribution and probability density of a numerical attribute, offering a visual summary of the data's characteristics, including its central tendency and spread.

3.8 Heat Map

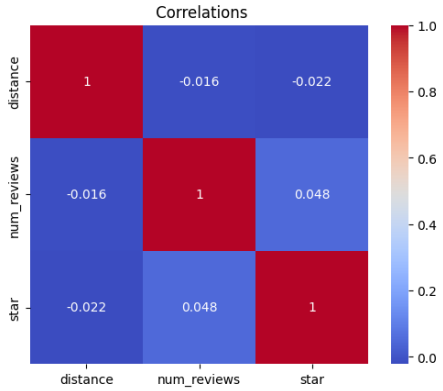

Figure 23: Heat map (star, num_reviews and distance).

As shown in Figure 23, the heat map visualizes the correlation between various numerical attributes in the dataset, using a color-coded scheme to represent the strength and direction of relationships. This allows for quick identification of patterns and dependencies, aiding in understanding the interplay between different variables.

4. Conclusion

https://colab.research.google.com/drive/1JLmCEaRtrqXBvWnt4Ay_PtZkW_Pk3qC9?usp=sharing

Upon completing this assignment, I have gained valuable insights into handling missing values and implementing appropriate techniques for data preprocessing. During the analysis of this dataset, I encountered challenges when plotting the scatter plot, as it revealed unexpected variations and patterns. Additionally, interpreting the box plot proved to be less intuitive. Moreover, dealing with the 'category_2' column, which contained numerous missing values, presented significant difficulty in the data handling process. Nevertheless, overcoming these obstacles has provided me with a valuable learning experience in effectively managing missing data and exploring various visualization methods to understand complex datasets more comprehensively.