Data Analysis in R

RESTAURANT INFORMATION

As a first step, I started my analysis by uploading two libraries which are 'ggplot2' and 'dplyr' then imported given datasets as 'resturant' and 'delivery' respectively.

Analysis on based on given information on Restaurant Information:

1-Present a column chart with the top 10 neighborhoods by the number of restaurants.

I created a new table which includes restaurant neighborhood and frequency of it then assigned that table as a data frame in order to use functions. As a second step I ordered the neighborhoods by the frequency of them. As a last step created new data frame which consist of first 10 neighborhoods then created a column chart in order to demonstrate easily.

2 - Present a column chart with the top 10 neighborhoods by restaurant review score.

Firstly, I checked if there is any missing value in the rest_rating columns in order to not encounter with a problem by aggregating the reviews. Secondly, I created a new data called as data1 which includes the restaurant neighborhoods and x that represents the aggregated score for each neighborhood. Thirdly, I sorted that datasets based on aggregated score so that top 10 neighborhood can be easily seen. As a last step, I presented a column chart in order to be easily understand.

3 - Compute the top 10 biggest chains. Present the results in a tabular format.

As having done on the previous step I started analyzing this task with checking the missing value. This task and previous one is alike each other. I created a table that consist of restaurant brand name and frequency of it then assigned it as data frame in order to use arrange function. I ordered it by frequency so that I can compute the top 10 biggest chains by using head function with the parameter of n is 10.

Restaurants Delivery Times

I started this task by doing the things asked to do. Firstly, I used join function in order to merge given datasets by using primary key of restaurant id and restaurant key for doing this I renamed the rest_key columns given in the delivery-mibe dataset to rest_id.

1 - Count the number of neighborhoods where each restaurant delivers.

Abovementioned I used **inner_join** function in order to create new dataset that has both dataset's columns by 'restaurant_id'. Secondly, this dataset was grouped by both restaurant_id and neighborhood name. Then I checked if the rest_delivery_time_min has a Nan value or not if so drop them in the dataset since I thought if the delivery_time_min this could be not useful information for our analysis. In my opinion Important point for this is to not look over that there are some restaurants delivery to a neighborhood with different delivery time that can be seen in the attachment. In order to overcome this problem, I used **distinct** function that can be used to keep only unique rows from a data frame by using restaurant_id and neighborhood_name as parameters so that I obtained a dataset which does not include this problem. As

a last step, I created a new dataset which only includes restaurant_id and aggregated number of neighborhoods that given restaurant can deliver.

2 - Present in a bar chart the top 15 neighborhoods by the number of restaurants where restaurants make deliveries.

I grouped the dataset by the column of neighborhood_name and created a data frame that include neighborhood_name and the number of restaurants that can deliver to given neighborhood. Secondly, I ordered this by the column of frequency from top to down so that I obtained the top 15 neighborhoods by the number of restaurants where restaurants make deliveries.

3 - Compute the average delivery times.

I created a new dataset which includes two columns that are rest_name and averages. In order to find mean of each rest_name's I used summerise function so that I obtained the average delivery time for each restaurant.

4 - Present in tabular format the top 20 restaurants by fasted average delivery time. In the same table, present the rating score, and postcode.

I ordered the average delivery times so that I can obtain the sorted dataset from minimum to maximum. After ordered dataset I was asked to present top 20 restaurants by fasted average delivery time for this reason I operated head function with parameter of n is 20. As a final duty, I was asked to add post_code and rest_rating columns into top 20 restaurants dataframe. I order to reach this goal I created a dataframe called dataset2 which was created using inner_join function with the parameters of datasets of top_20_neighborhood and resturant which is null database imported from initiatively given dataset by 'rest name'. Finally, select function was used to select the columns which are rest_name, rest_postcode, rest_rating from the dataset2.