Project Python Foundations: FoodHub Data Analysis- Business Report

By Subhadeep Banerjee Chowdhury

Marks: 50

Context

The number of restaurants in New York is increasing day by day. Lots of students and busy professionals rely on those restaurants due to their hectic lifestyles. Online food delivery service is a great option for them. It provides them with good food from their favorite restaurants. A food aggregator company FoodHub offers access to multiple restaurants through a single smartphone app.

The app allows the restaurants to receive a direct online order from a customer. The app assigns a delivery person from the company to pick up the order after it is confirmed by the restaurant. The delivery person then uses the map to reach the restaurant and waits for the food package. Once the food package is handed over to the delivery person, he/she confirms the pick-up in the app and travels to the customer's location to deliver the food. The delivery person confirms the drop-off in the app after delivering the food package to the customer. The customer can rate the order in the app. The food aggregator earns money by collecting a fixed margin of the delivery order from the restaurants.

Objective

The food aggregator company has stored the data of the different orders made by the registered customers in their online portal. They want to analyze the data to get a fair idea about the demand of different restaurants which will help them in enhancing their customer experience. Suppose you are hired as a Data Scientist in this company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

Data Description

The data contains the different data related to a food order. The detailed data dictionary is given below.

Data Dictionary

- order_id: Unique ID of the order
- customer_id: ID of the customer who ordered the food
- restaurant_name: Name of the restaurant
- cuisine_type: Cuisine ordered by the customer
- cost: Cost of the order
- day_of_the_week: Indicates whether the order is placed on a weekday or weekend (The weekday is from Monday to Friday and the weekend is Saturday and Sunday)
- rating: Rating given by the customer out of 5
- food_preparation_time: Time (in minutes) taken by the restaurant to prepare the food. This is calculated by taking the difference between the timestamps of the restaurant's order confirmation and the delivery person's pick-up confirmation.

- delivery_time: Time (in minutes) taken by the delivery person to deliver the food package. This is calculated by taking the difference between the timestamps of the delivery person's pick-up confirmation and drop-off information

Understanding the structure of the data

The following output shows the first 5 rows of the dataset, containing the aforementioned columns

Output:

	order_id	customer_id	restaurant_name	cuisine_type	cost_of_the_order	day_of_the_week	rating	food_preparation_time	delivery_time
0	1477147	337525	Hangawi	Korean	30.75	Weekend	Not given	25	20
1	1477685	358141	Blue Ribbon Sushi Izakaya	Japanese	12.08	Weekend	Not given	25	23
2	1477070	66393	Cafe Habana	Mexican	12.23	Weekday	5	23	28
3	1477334	106968	Blue Ribbon Fried Chicken	American	29.20	Weekend	3	25	15
4	1478249	76942	Dirty Bird to Go	American	11.59	Weekday	4	25	24

Observations:

The DataFrame has 9 columns as mentioned in the Data Dictionary. Data in each row corresponds to the order placed by a customer.

Approach used

Python programming language along with it's libraries such as Numpy, Pandas, Matplotlib and Seaborn have been used to fetch all the answers to the questions below.

Question 1: How many rows and columns are present in the data? [1mark]

Output:

```
The number of rows are: 1898
The number of columns are: 9
```

The dataset consists of 1898 rows and 9 columns which shows the data contains information related to 1898 orders placed

Question 2: What are the datatypes of the different columns in the dataset? (The info() function can be used) [1 mark]

The following output shows the datatypes of the different columns. The .info() function was used here.

Output:

```
#   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   order_id               1898 non-null   int64
 1   customer_id            1898 non-null   int64
 2   restaurant_name        1898 non-null   object
 3   cuisine_type           1898 non-null   object
 4   cost_of_the_order      1898 non-null   float64
 5   day_of_the_week        1898 non-null   object
 6   rating                 1898 non-null   object
 7   food_preparation_time  1898 non-null   int64
 8   delivery_time          1898 non-null   int64
```

Observations:

The dataset contains datatypes such as integer, float and object with no null values. There 4 integer variables, 1 float variable and 4 object variables.

Question 3: Check the statistical summary of the data. What is the minimum, average, and maximum time it takes for food to be prepared once an order is placed? [2 marks]

The following output shows the statistical summary of the attributes.
The .describe() function was used here.

Output:

	order_id	customer_id	cost_of_the_order	food_preparation_time	delivery_time
count	1.898000e+03	1898.000000	1898.000000	1898.000000	1898.000000
mean	1.477496e+06	171168.478398	16.498851	27.371970	24.161749
std	5.480497e+02	113698.139743	7.483812	4.632481	4.972637
min	1.476547e+06	1311.000000	4.470000	20.000000	15.000000
25%	1.477021e+06	77787.750000	12.080000	23.000000	20.000000
50%	1.477496e+06	128600.000000	14.140000	27.000000	25.000000
75%	1.477970e+06	270525.000000	22.297500	31.000000	28.000000
max	1.478444e+06	405334.000000	35.410000	35.000000	33.000000

Observations:

The minimum, average and maximum food preparation time is 20 mins, 27.37 mins and 35 mins. The average delivery time is lower than the average food preparation time but the former is more variable than the latter as observed from their respective standard deviations. Out of all the orders, 25% of the orders cost $12.08 or less, 50% cost less than or equal to $14.14 and 75% cost less than or equal to $22.29.

Question 4: How many orders are not rated? [1 mark]

The following output shows the number of orders that are not rated. The .count() function was applied after applying the condition df['rating']=='Not given'.

Output:

```
736
```

Observations:

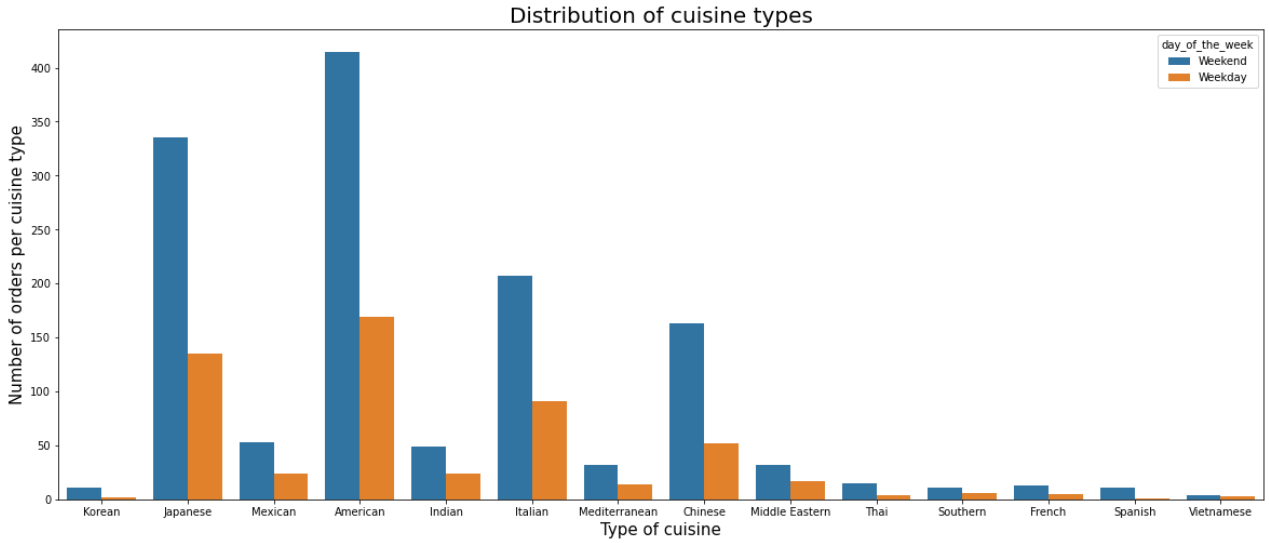
There are 736 orders that are not rated.

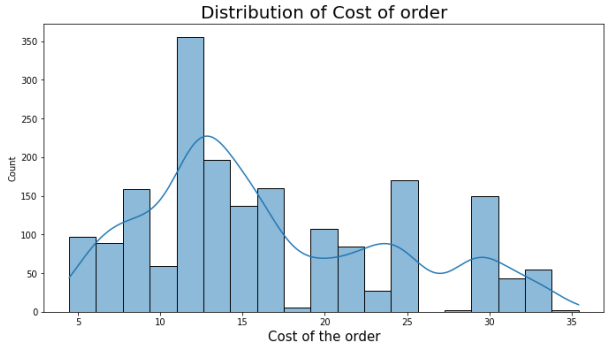
Exploratory Data Analysis (EDA)

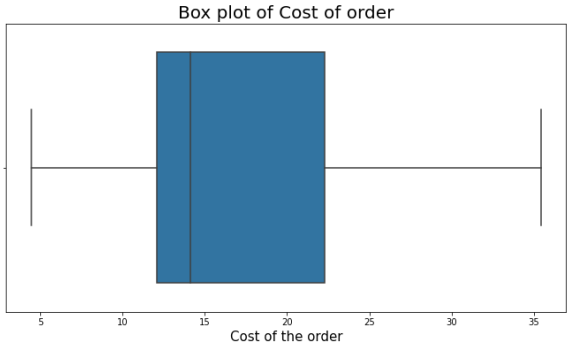
Univariate Analysis

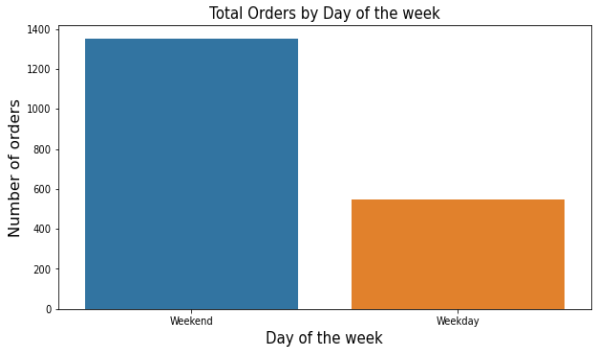
Question 5: Explore all the variables and provide observations on their distributions. (Generally, distinct number of values, histograms, boxplots, countplots, etc. are used for univariate exploration.) [9 marks]

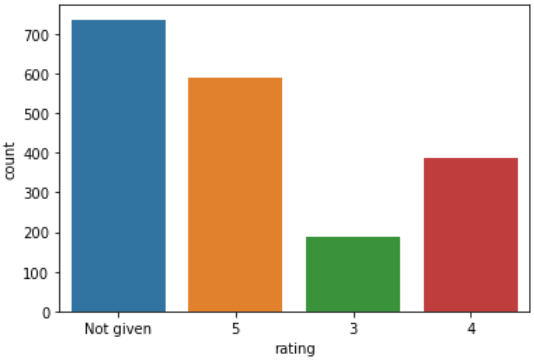
Countplots, Histogram and Boxplots have been used to conduct the exploratory data analysis of all the relevant variables such as: Cuisine type, Cost of Order, Day of the Week, Rating, Food preparation time and delivery time. Python's Seaborn and Matplotlib libraries have been used here. The outputs of the EDA are shown below.

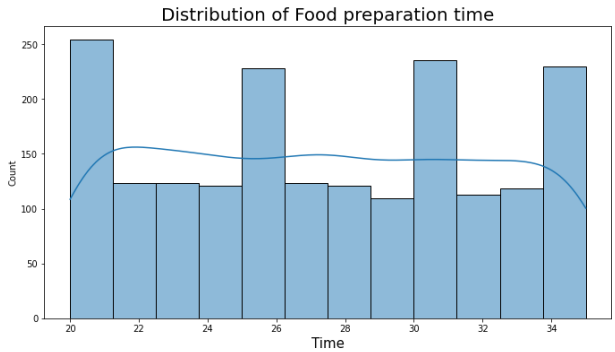

Distribution of cuisine types
day_of_the_week
Weekend
Weekday
Number of orders per cuisine type
Korean
Japanese
Mexican
American
Indian
Italian
Mediterranean
Chinese
Middle Eastern
Thai
Southern
French
Spanish
Vietnamese
Type of cuisine

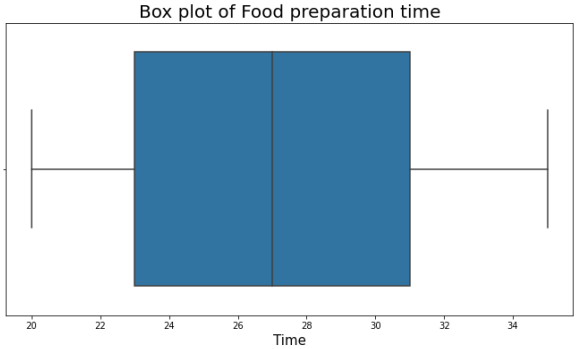

Distribution of Cost of order
Count
Cost of the order

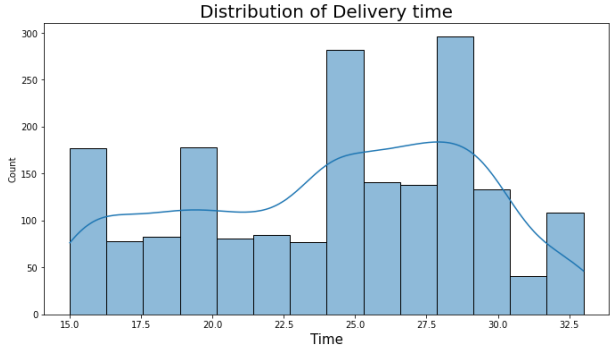

Box plot of Cost of order
Cost of the order

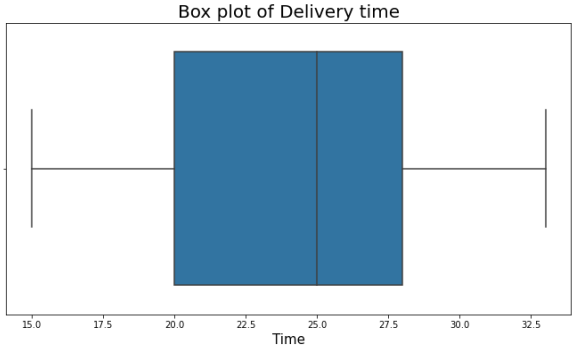

Total Orders by Day of the week
Number of orders
Weekend
Weekday
Day of the week


count
Not given
5
3
4
rating


Distribution of Food preparation time
Count
Time


Box plot of Food preparation time
Time





Observations

The American cuisine is the most ordered cuisine type among all the cuisines irrespective of the day of the week. This is closely followed by Japanese, Italian and Chinese. The cost of order has positively skewed distribution, with more concentration of values towards the left. The same is visible from the boxplot. Most orders are unrated, however amongst the rated orders, the rating of 5 has the highest occurrence. More orders are placed on the weekends. The food preparation time is evenly distributed with 50% of the orders lying within the range of 23-31mins and has a median of 27 mins. The delivery time on the other hand is negatively skewed with a median time of 25 mins and 50% of the orders taking preparation time in the range of 20 mins to 28mins.

Question 6: Which are the top 5 restaurants in terms of the number of orders received? [1 mark]

The following output gives the list of the top 5 restaurants in terms of the number of orders received. The .value_counts() method was used here to calculate the number of orders per restaurant and .head() was used to fetch the top 5 names of the restaurants.

Output:

```
restaurant_name
Shake Shack                  219
The Meatball Shop            132
Blue Ribbon Sushi            119
Blue Ribbon Fried Chicken     96
Parm                          68
```

Observations:

The top 5 restaurants in terms of the number of orders received are Shake Shack, The Meatball Shop, Blue Ribbon Sushi, Blue Ribbon Fried Chicken and Parm. It can be seen that Shake Shack tops the list with 219 orders.

Question 7: Which is the most popular cuisine on weekends? [1 mark]

The following output displays the most popular cuisine type on the weekends. The .value_counts() method was used here and .head(1) was used to display the top value with the highest orders.

Output:

```
American    415
```

Observations:

The American cuisine is the most popular cuisine on weekends with 415 orders.

Question 8: What percentage of the orders cost more than 20 dollars?(use .round function to round the final percentage) [2 marks]

The following output displays the percentage of orders that cost more than 20 dollars

Output:

```
29
```

Observations:

29% of all the orders cost more than $20.

Question 9: What is the mean order delivery time? [1 mark]

The following output displays the average time it takes to deliver the order to the customer. The .mean() function was used here.

Output:

```
24.161749209694417
```

Observations:

The average time to deliver an order is 24.16 minutes

Question 10: The company has decided to give 20% discount vouchers to the top 3 most frequent customers. Find the IDs of these customers and the number of orders they placed. [1 mark]

The following output displays the top 3 most frequent customers. The .value_counts() method was used along with .head(3) to display the top 3 customer IDs.
Output:

```
customer_id
52832          13
47440          10
83287           9
```

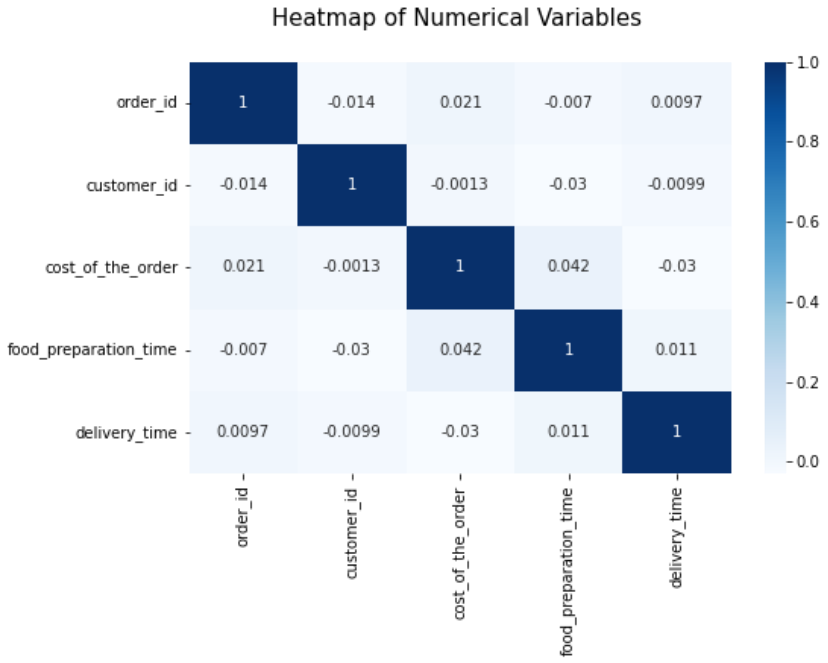
Observations:

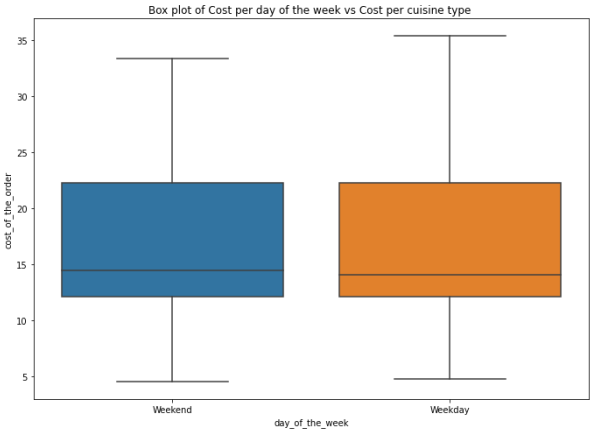
Customer IDs 52832, 47440 and 83287 are the top 3 most frequent customers with number of orders 13, 10 and 9 respectively

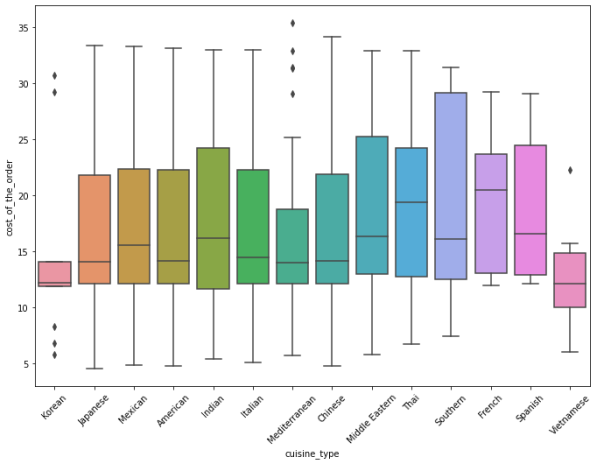
Multivariate Analysis

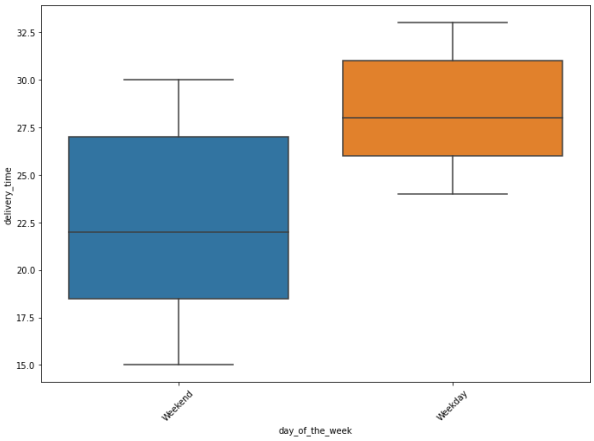
Question 11: Perform a multivariate analysis to explore relationships between the important variables in the dataset. (It is a good idea to explore relations between numerical variables as well as relations between numerical and categorical variables) [8 marks]

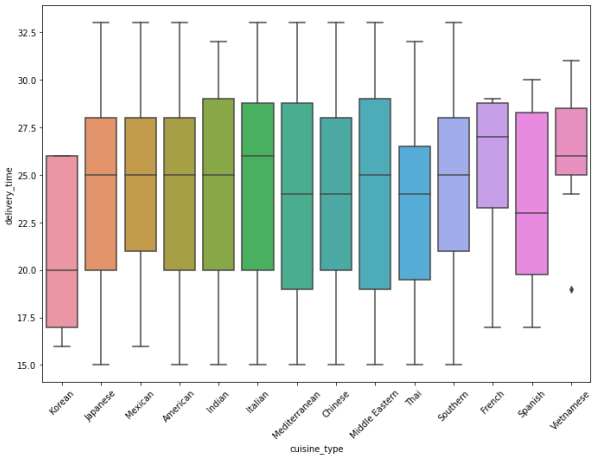
For the multivariate analysis, Heatmap from has been used to display the correlation matrix showing correlation between the numeric variables. Boxplots have been used to display the distribution of numeric variables category wise. Finally pairplot has been used to display the relationship between the numeric variables combined. All the following graphs have been generated using python's seaborn library.

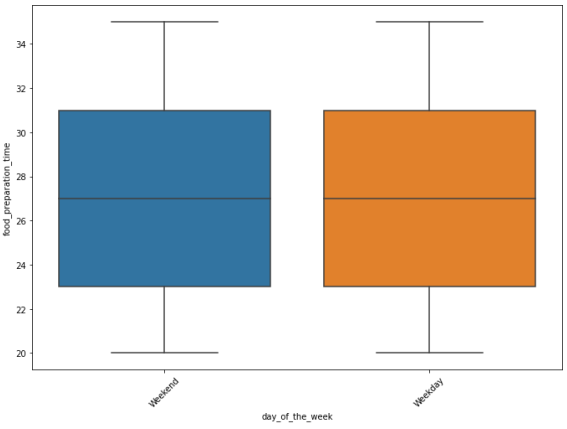


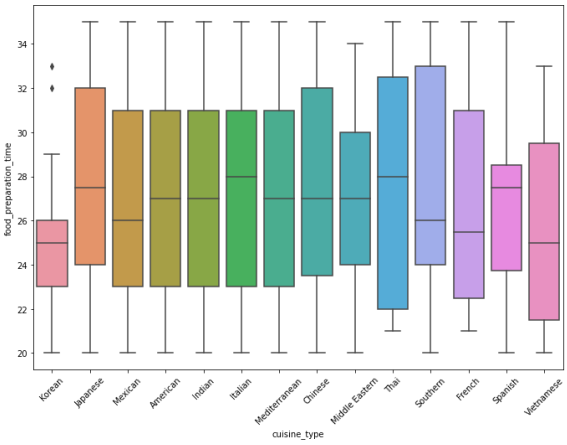

Box plot of Cost per day of the week vs Cost per cuisine type
cost_of_the_order
Weekend
Weekday
day_of_the_week

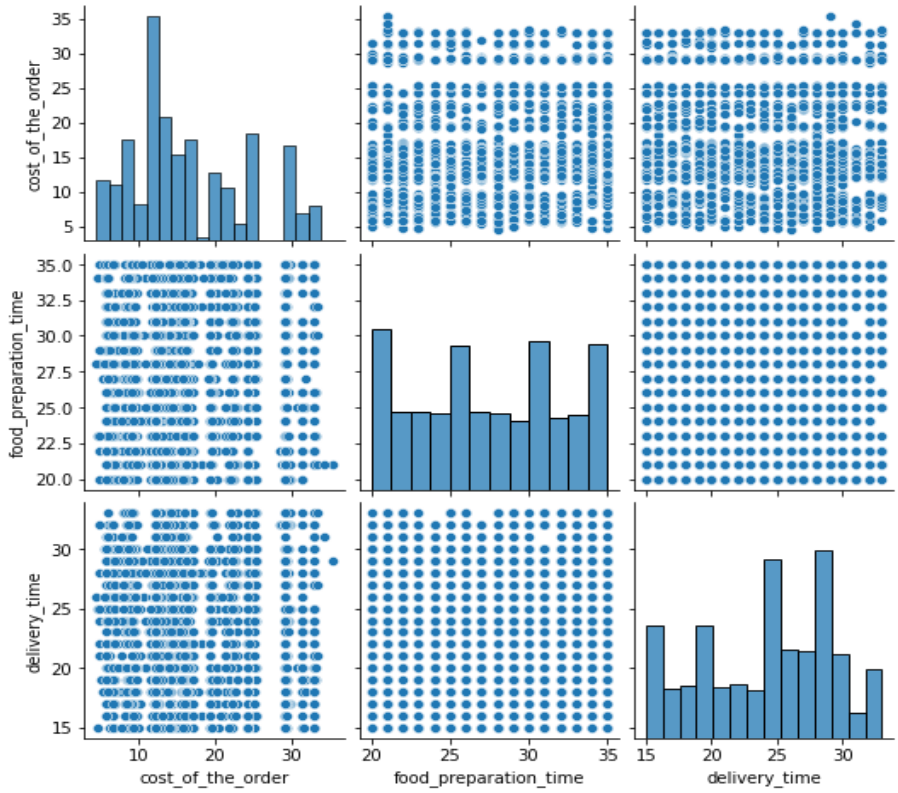

cost_of_the_order
Korean
Japanese
Mexican
American
Indian
Italian
Mediterranean
Chinese
Middle Eastern
Thai
Southern
French
Spanish
Vietnamese
cuisine_type


delivery_time
Weekend
Weekday
day_of_the_week


delivery_time
Korean
Japanese
Mexican
American
Indian
Italian
Mediterranean
Chinese
Middle Eastern
Thai
Southern
French
Spanish
Vietnamese
cuisine_type


food_preparation_time
Weekend
Weekday
day_of_the_week


food_preparation_time
Korean
Japanese
Mexican
American
Indian
Italian
Mediterranean
Chinese
Middle Eastern
Thai
Southern
French
Spanish
Vietnamese
cuisine_type



Observations

The numeric variables have low correlation among them. The median delivery time is lower on the weekends. Korean, Mediterranean and Vietnamese cuisine tend to have more outlier cost values.

Question 12: The company wants to provide a promotional offer in the advertisement of the restaurants. The condition to get the offer is that the restaurants must have a rating count of more than 50 and the average rating should be greater than 4. Find the restaurants fulfilling the criteria to get the promotional offer. [3 marks]

The following output shows the names of the restaurants must have a rating count of more than 50 and the average rating should be greater than 4. First, the groupby method was used to aggregate the values by the restaurant name and then .count() and .mean() to fetch the number of orders and average rating respectively. Finally, pd.merge was used to combine the 2 dataframes to get the names of the restaurants who fulfil both the criteria.

Output:

	restaurant_name
0	Blue Ribbon Fried Chicken
1	Blue Ribbon Sushi
2	Shake Shack
3	The Meatball Shop

Observations:

The restaurants that have a rating count of more than 50 and the average rating greater than 4 are Blue Ribbon Fried Chicken, Blue Ribbon Sushi,Shake Shack and The Meatball Shop.

Question 13: The company charges the restaurant 25% on the orders having cost greater than 20 dollars and 15% on the orders having cost greater than 5 dollars. Find the net revenue generated by the company across all orders. [2 marks]

The following output shows the net revenue generated by the company across all orders after applying the given margins. The .sum() function was used.

Output:

```
6166.303
```

Observations:

The net revenue generated by the company across all orders is $6166.303

Question 14: The company wants to analyze the total time required to deliver the food. What percentage of orders take more than 60 minutes to get delivered from the time the order is placed? (The food has to be prepared and then delivered.) (Use .round function to round value to nearest zero) [2 marks]

The following output shows the percentage of orders take more than 60 minutes to get delivered from the time the order is placed. The total time has been calculated by adding the food preparation time and delivery time.

Output:

```
11
```

Observations:

About 11% of orders take more than 60 minutes to get delivered from the time the order is placed

Question 15: The company wants to analyze the delivery time of the orders on weekdays and weekends. How does the mean delivery time vary during weekdays and weekends? [2 marks]

The mean delivery time and the standard deviations on weekdays and weekends are given below

Output:

	Weekday	Weekend
Mean Delivery Time	28.34	22.47
Standard Deviation	2.89	4.63

Observations:

The mean delivery time vary during weekdays is 28.34 mins and during weekends is 22.47 mins. The Standard Deviation of the delivery time on the weekdays is 2.89 mins and on the weekends is 4.63.

Conclusion and Recommendations

Question 16: What are your conclusions from the analysis? What recommendations would you like to share to help improve the business? (You can use cuisine type and feedback ratings to drive your business recommendations.) [4 marks]

Conclusions:

From the above exploratory data analysis of the FoodHub data set, the following conclusions can be derived:

- The American cuisine seems to be the most popular among all the cuisines irrespective of the day of the week. This is closely followed by Japanese, Italian and Chinese.
- The cost of order values are positively skewed, with the majority of the costs clustered towards the left of the chart. This can be inferred from both the histogram and the boxplot
- More orders are placed on weekends as compared to weekdays
- Most of the orders are not rated. Among the ratings, the rating of 5 occurs the highest time. This shows that even though most of the orders are not rated, most of those that are rated have a high rating showing that it is mostly satisfied customers that are inclined to rate the experience
- The food preparation time is fairly evenly distributed with the values of time ranging between 20 minutes and 36 minutes with a median of about 27 minutes

- The delivery time of the order is slightly negatively skewed (skewed to the left). The box plot shows the interquartile range which shows that 50% of the orders take between 20mins to about 28 mins to deliver with the median delivery time being 25 mins.
- There is a very low correlation between the numerical variables.
- The food preparation time remains consistent irrespective of the day of the week, however the median delivery time is much lower during the weekends than the weekdays.
- In terms of the cost, Korean, Mediterranean and Vietnamese cuisine tend to have a lot of extreme values (outliers)
- Although the delivery is faster during the weekend, the variation in delivery time is also higher on the weekends.

Recommendations:

- Given the fact that American Cuisine is the most popular, the company should promote it more in order to generate more orders and thereby revenue. This could be done by offering periodic discounts and offering new dishes in the same category. The same could be done for other popular cuisines like Japanese, Italian and Chinese.
- More low to mid cost dishes should be introduced since most customers tend to order food in the same range, which can be inferred from the positively skewed cost distribution
- There is higher demand on the weekends and as a result there should be more delivery agents and related staff working on the weekends and lesser on the weekdays
- Most orders are not rated. The company should examine the causes and provide incentives to customers for providing feedback/ratings. More ratings would provide more insights into the customer preferences.
- The company should try to reduce the delivery time during the weekdays in order to provide a more consistent experience to customers