SMDM PROJECT

STATISTICAL METHODS FOR DECISION MAKING: FOODHUB DATA ANALYSIS

KARUNA_SMDM_GUIDED_PROJECT_12-09-2023

Karuna Pashte

INDEX

Sr.No	Title	Page no
1	About Data and Dictionary	5
2	Question 1	6
3	Question 2	6
4	Question 3	6
5	Question 4	6
6	Question 5	7
7	Question 6	7
8	Question 7	10
9	Question 8	10
10	Question 9	10
11	Question 10	10
12	Question 11	10
13	Question 12	11
14	Question 13	14
15	Question 14	14
16	Question 15	14
17	Question 16	15
18	Question 17	15

List of Figures

Sr.No	Title	Page no
1	Count plot for the cusine type.	7
2	Histogram for the cost of the order.	7
3	Box plot for the cost of the order.	8
4	Bar graph for the day of the week	8
5	Bar graph for the rating	8
6	Histogram for the food preparation time	9
7	Box plot for the food preparation time	9
8	Histogram for delivery time	9
9	Box plot for the delivery time	10
10	Box plot for the cusine vs cost of the order	11
11	Box plot for the cusine vs delivery time	11
12	Box plot for the day of the week vs delivery time	12
13	Point plot for the rating vs delivery time	12
14	Point plot for the rating vs food preparation time	13
15	Point plot for the rating vs cost of the order	13
16	Heat map for the correlation between cost of the order, food preparation time and delivery time	13

List of Tables

Sr.No	Title	Page No
1	Data types in dataset	6
2	Missing values	6
3	Statistical summary of dataset	6
4	Count of rating	7
5	Highest number of orders	10
6	IDs of top 3 most frequent customers	10
7	Revenue Generated by the restaurants	12
8	Restaurants with rating count of more than 50	14
9	Restaurants with average rating greater than 4	14
10	Net revenue generated by the company across all orders	14

About Data and Dictionary:

Context Statistical

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 a Data Scientist at Foodhub 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

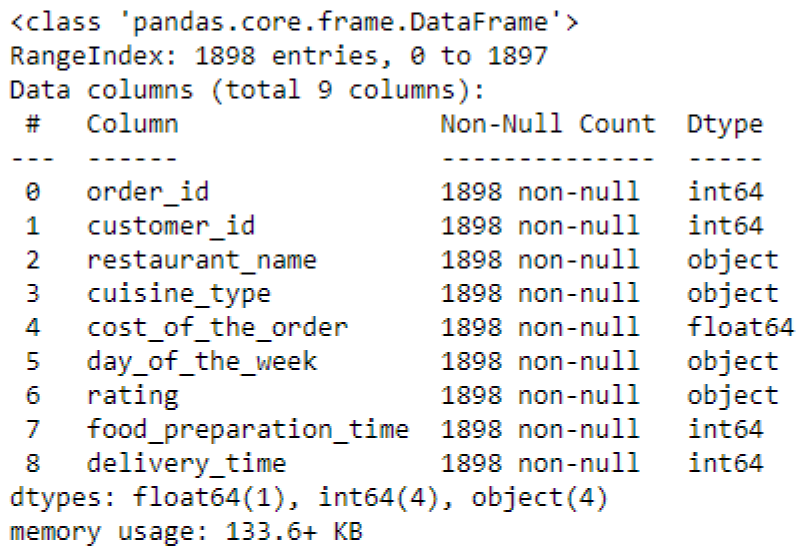
The data contains the different data related to a food order. The detailed data dictionary is given below.

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_of_the_order: 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.

Question 1: How many rows and columns are present in the data?
Answer: There are 1898 rows and 9 columns are in the Data **.**
(1898, 9)

Question 2: What are the data types of the different columns in the dataset?
Answer: In the dataset there are 1 float data type, 4 integers and 4 object
Order ID and Customer ID are categorical variables with numerical labels.

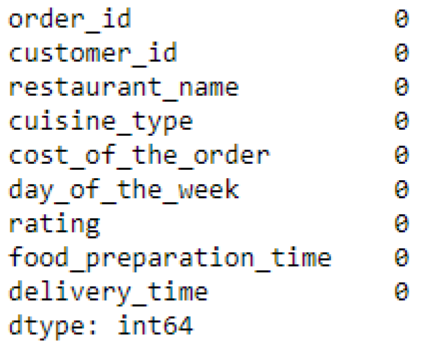


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1898 entries, 0 to 1897
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   order_id               1898 non-null   int64  
 1   customer_id            1898 non-null   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  
dtypes: float64(1), int64(4), object(4)
memory usage: 133.6+ KB
```

Table 1: Data types in dataset

Question 3: Are there any missing values in the data? If yes, treat them using an appropriate method.

Answer: There are no any missing values in dataset. Therefore no requirement of treatment.



```
order_id                 0
customer_id              0
restaurant_name          0
cuisine_type             0
cost_of_the_order        0
day_of_the_week          0
rating                   0
food_preparation_time    0
delivery_time            0
dtype: int64
```

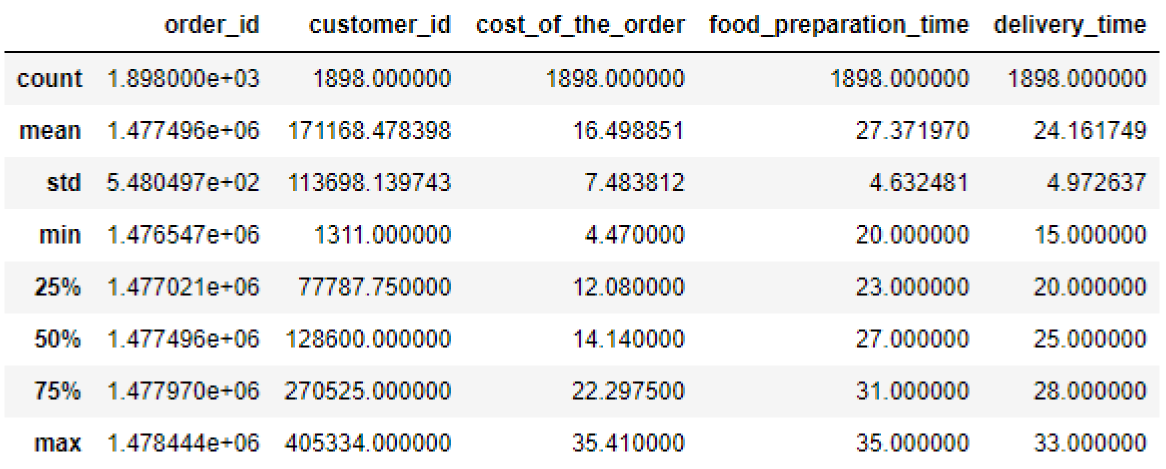
Table 2: Missing values in dataset

Question 4: 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?

Answer: Statistical summary of the data are shown in below table.

The food to be prepared once an order is placed minimum time it takes 20 minutes,

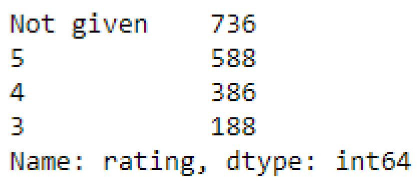
Average time is 27 minutes and maximum time is 35 minutes.



	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

Table 3: Statistical summary of dataset

Question 5: How many orders are not rated?
Answer: 736 orders were not rated.



```
Not given    736
5            588
4            386
3            188
Name: rating, dtype: int64
```

Table 4: Count of rating

Question 6: Explore all the variables and provide observations on their distributions.
Answer: Total number of unique order id’s are 1898

Total number of unique costumer id’s 1200

Total number of unique restaurant name are 178

Total number of unique Cuisine type are 14

The observations on their distributions are provide on the basis of Univariate

Analysis with use of histograms ,boxplots, count-plots, etc. are given in figures.

Cuisine type: Most popular cuisines are American, Japanese, Italian and followed with a close tie between Mexican and Indian.

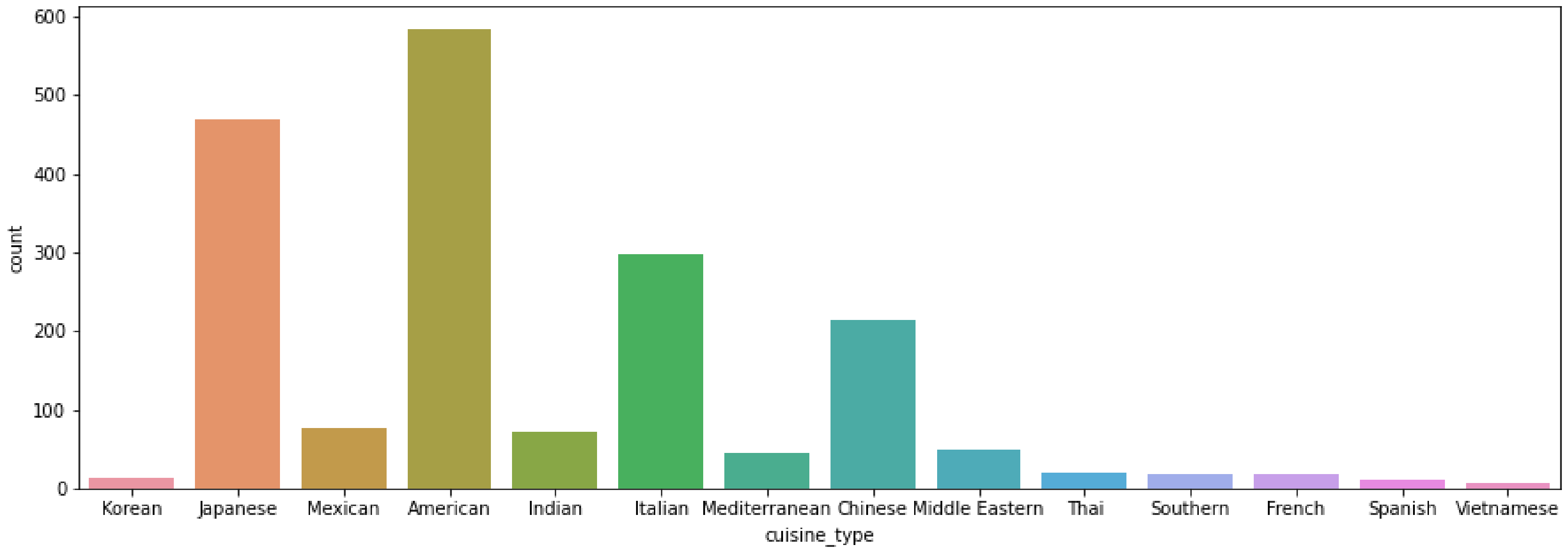


Figure 1: Countplot for cuisine type

Cost of the order

Histogram skewed to the left. That is more towards lower cost. Also note that there is slight peak at 25 dollar.

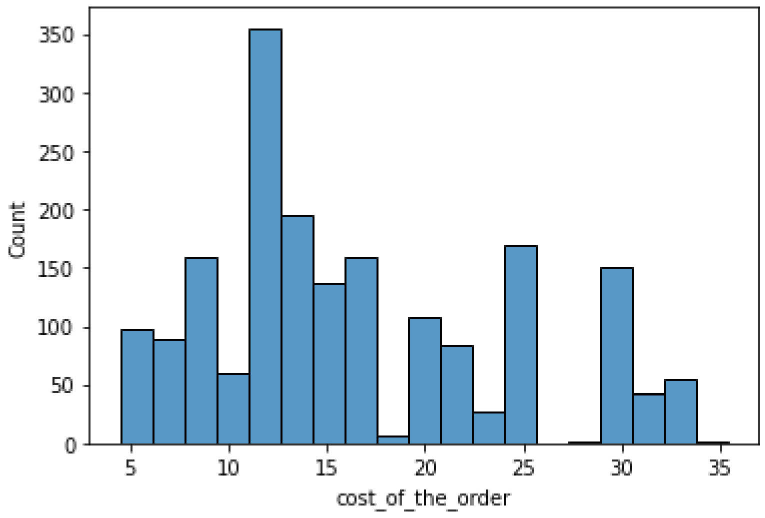


Figure 2: Histogram for the cost of the order

Cost of the order

The boxplot indicates that the median cost is about 14 dollars, with order being right skewed.

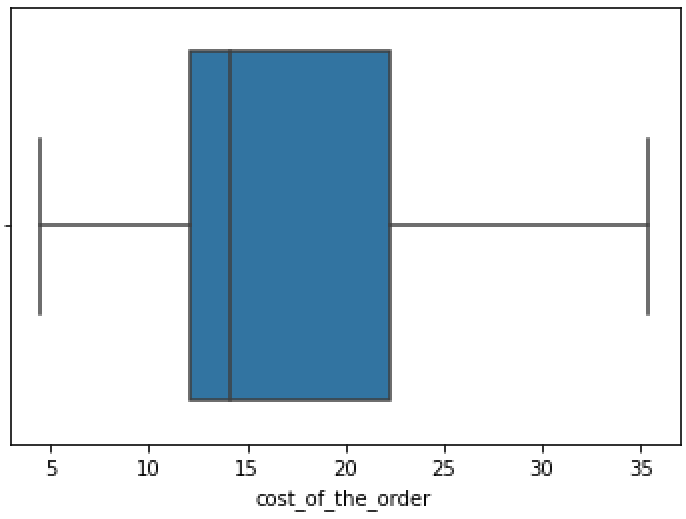


Figure 3: Boxplot for the cost of the order

Day of the week

There is higher activity on weekend than weekdays.

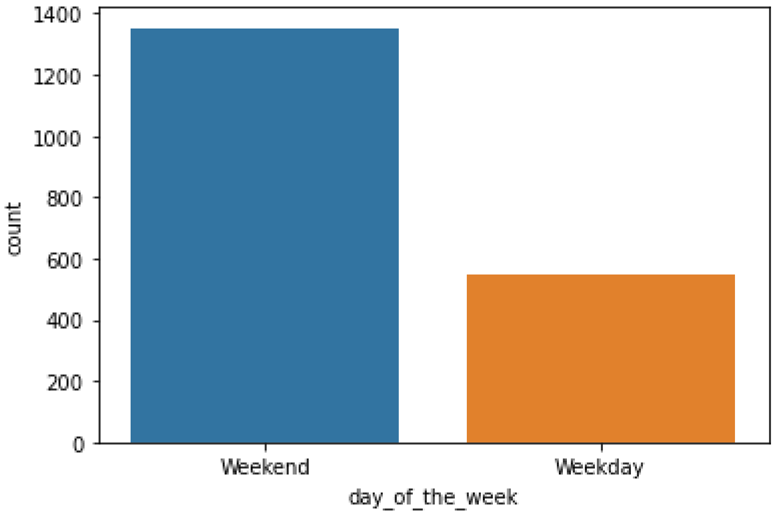


Figure 4: Bar graph for the Day_of_the_week

Rating

Here, unique value for rating are ‘Not given’, ‘5’, ‘4’, ‘3’ are shown in Bar graph. Most of the customers do not given any rating. Maximum ratings are given ’5’ followed by ‘’4’ and ‘3’.

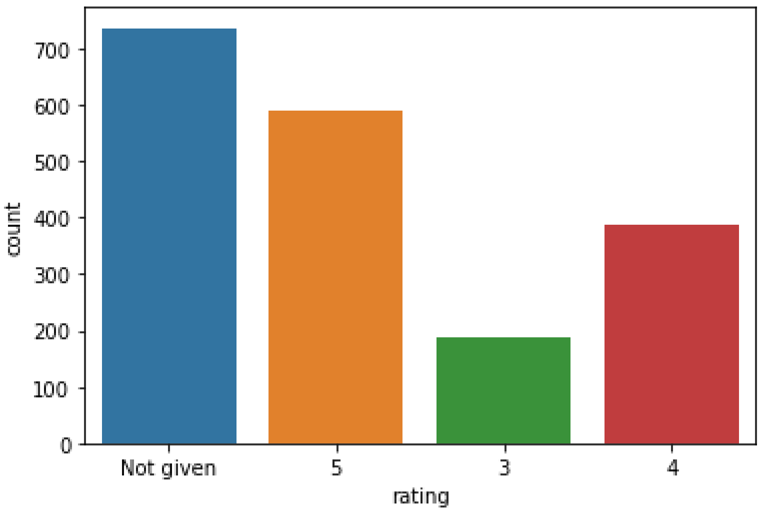


Figure 5: Bar graph for the rating

Food preparation time

For food preparation time shown by plotting the histogram and boxplot.

From the histogram it can be determine that maximum food preparation time taken between 20 and 36 minutes to be delivered.

From box plot it can be determine that median delivery time is about 27 minutes.

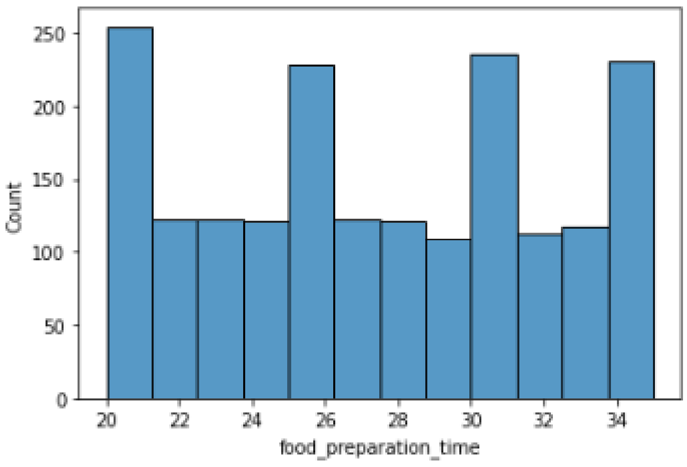


Figure 6: Histogram for the food preparation time

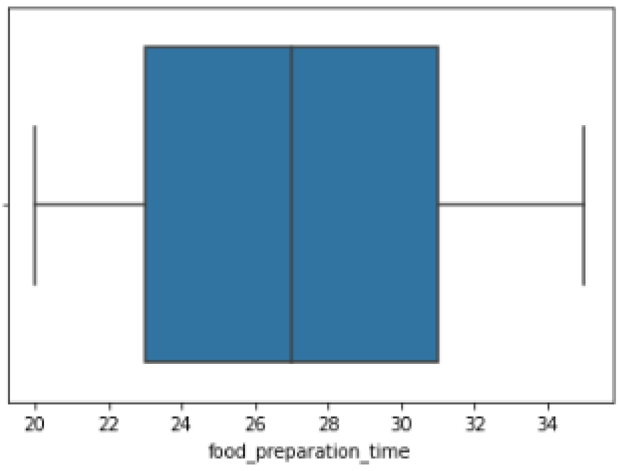


Figure 7: Boxplot for the food preparation time

Observation Delivery time

From the histogram it is observe that most orders take time to be delivered in between 25 to 38 minutes.

From box plot it is observe that median delivery time is 25 minutes.

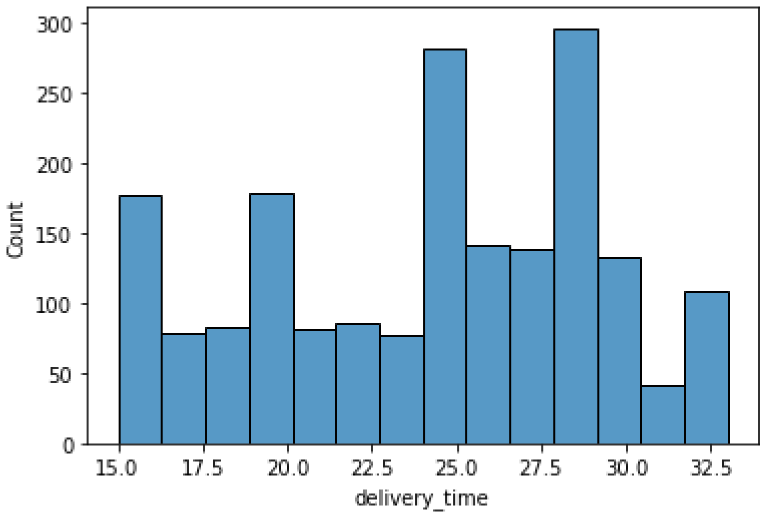


Figure 8: Histogram for delivery time

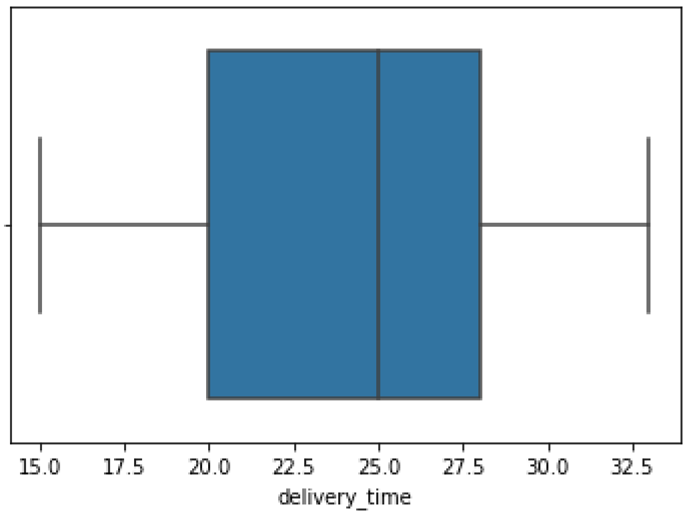
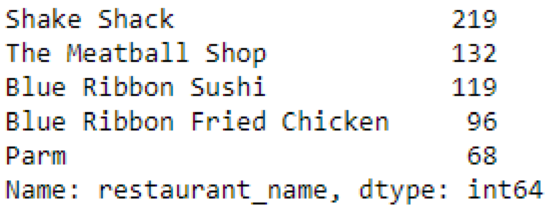


Figure 9: Boxplot for the delivery time

Question 7: Which are the top 5 restaurants in terms of the number of orders received?

Answer: As shown in table, Shake Shack is the leading restaurant with 219 orders, The Meatball Shop with 132 orders, Blue Ribbon Sushi with 119 orders,Blue Ribbon Fried Chicken with 96 orders and Parm restaurant with 68 orders.

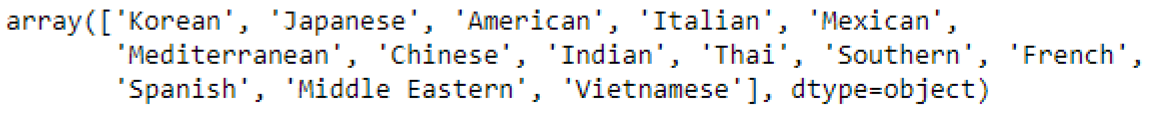


```
Shake Shack                  219
The Meatball Shop            132
Blue Ribbon Sushi            119
Blue Ribbon Fried Chicken     96
Parm                          68
Name: restaurant_name, dtype: int64
```

Table 5: Highest number of orders

Question 8: Which is the most popular cuisine on weekends?

Answer: Most popular cuisine on weekends are shown in below:



```
array(['Korean', 'Japanese', 'American', 'Italian', 'Mexican',
       'Mediterranean', 'Chinese', 'Indian', 'Thai', 'Southern', 'French',
       'Spanish', 'Middle Eastern', 'Vietnamese'], dtype=object)
```

Question 9: What percentage of the orders cost more than 20 dollars?

Answer: The number of total orders that cost above 20 dollars is: 555

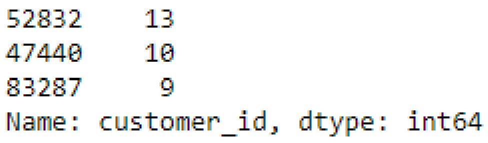
Percentage of orders above 20 dollars: 29.24 %

Question 10: What is the mean order delivery time?

Answer: The mean delivery time for this dataset is 24.16 minutes.

Question 11: 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.

Answer:



```
52832    13
47440    10
83287     9
Name: customer_id, dtype: int64
```

Table 6: IDs of top 3 most frequent customers

Question 12: 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)

Answer: To explore relation between important variables in the dataset are perform with Multivariate analysis with boxplot, pointplot, heatmap, etc.

Cuisine vs Cost of the order

Korean, Mediterranean and Vietnamese cusine having outliers.

Other cusines have different degrees of skewed costs.

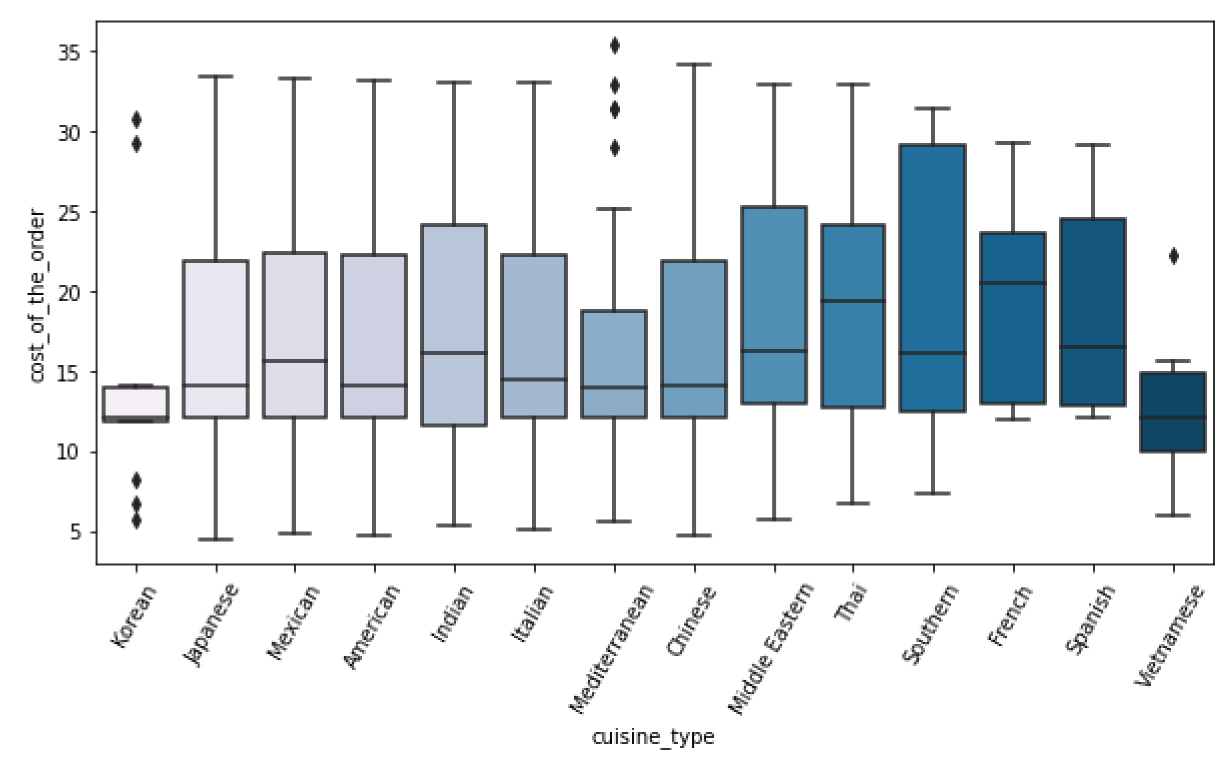


Figure 10: Boxplot for the Cuisine vs Cost of the order

Cuisine vs Food preparation time

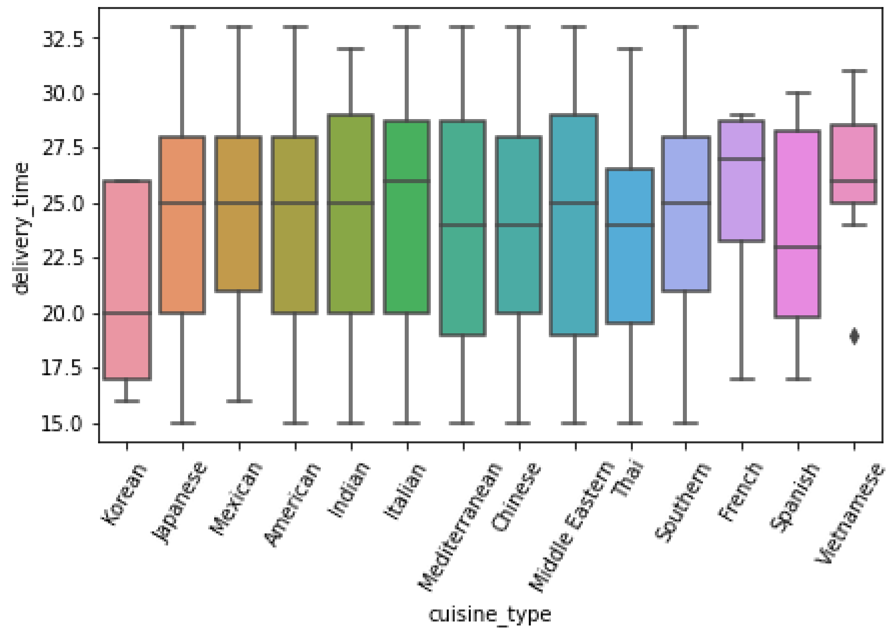


Figure 11: Boxplot for the Cuisine vs Delivery time

Day of the week vs Delivery time

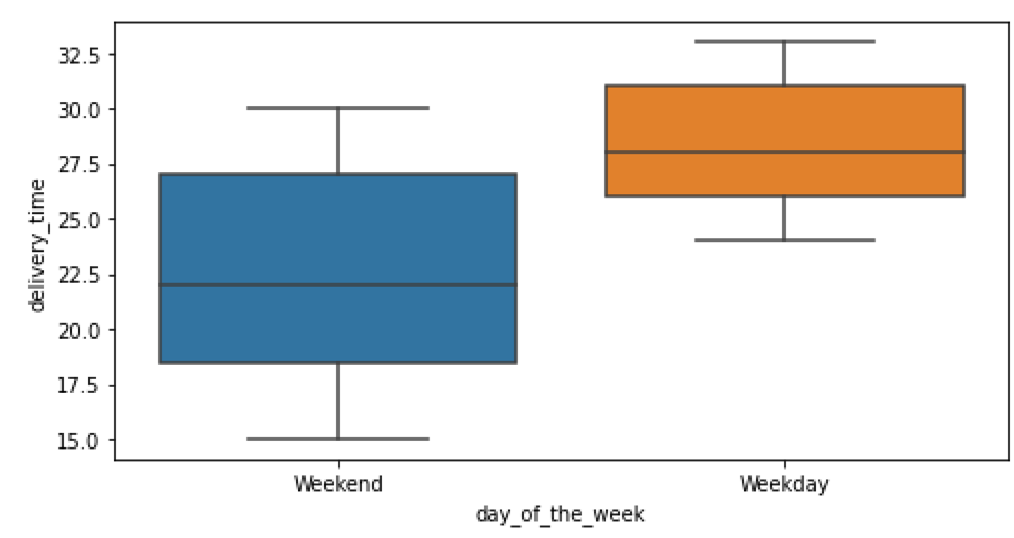
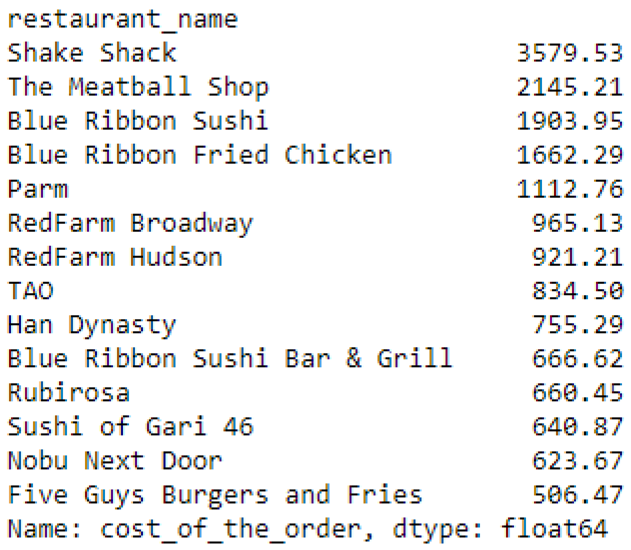


Figure 12: Boxplot for the Day of the week vs Delivery time

Revenue generated by the restaurants



```
restaurant_name
Shake Shack                      3579.53
The Meatball Shop                2145.21
Blue Ribbon Sushi                1903.95
Blue Ribbon Fried Chicken        1662.29
Parm                             1112.76
RedFarm Broadway                  965.13
RedFarm Hudson                    921.21
TAO                               834.50
Han Dynasty                       755.29
Blue Ribbon Sushi Bar & Grill     666.62
Rubirosa                          660.45
Sushi of Gari 46                  640.87
Nobu Next Door                    623.67
Five Guys Burgers and Fries       506.47
Name: cost_of_the_order, dtype: float64
```

Table 7: Revenue generated by the restaurant

Rating vs Delivery Time

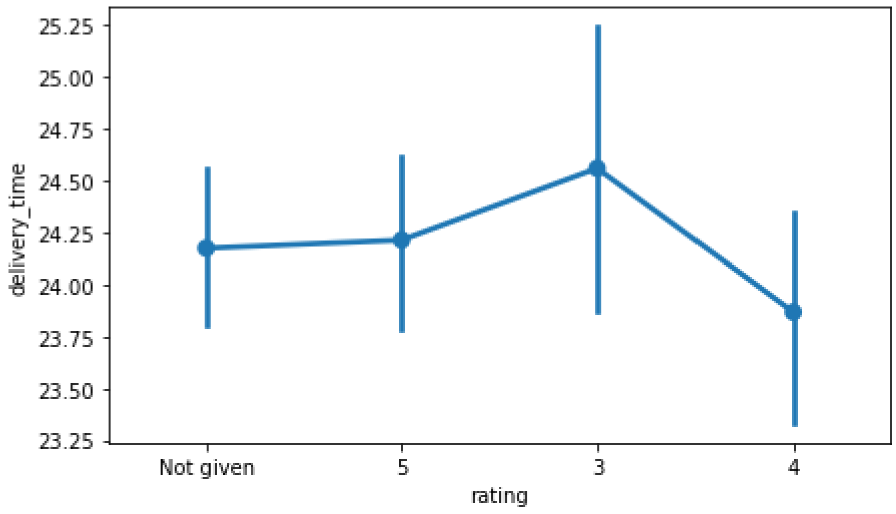


Figure 13: Point plot for the Rating vs Delivery time

Rating vs Food prepartion time

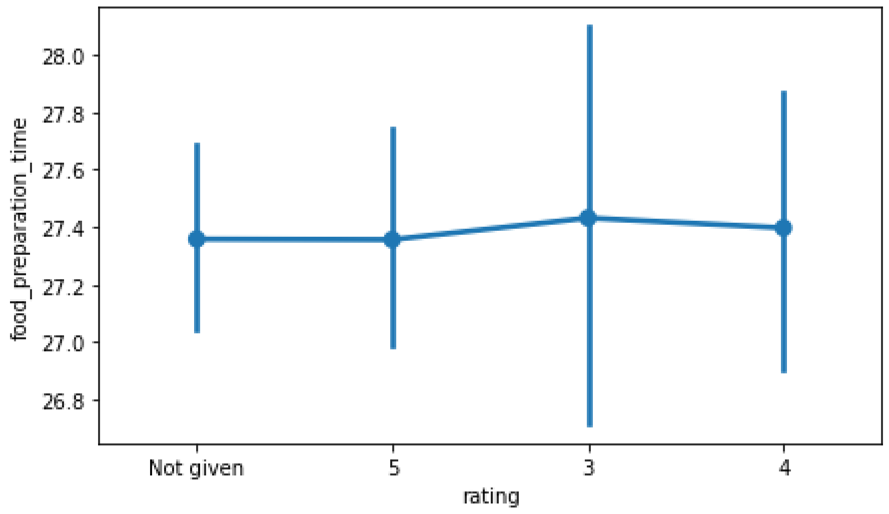


Figure 14: Point plot for the Rating vs Food preparation time

Rating vs Cost of the order

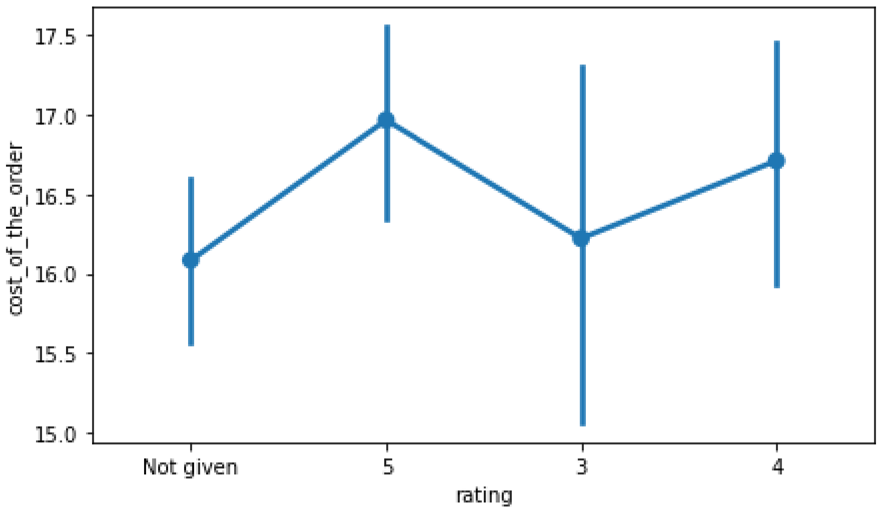


Figure 15: Point plot for the Rating vs Cost of the order

Correlation among variables

Heatmap shows very weak correlation between variables.

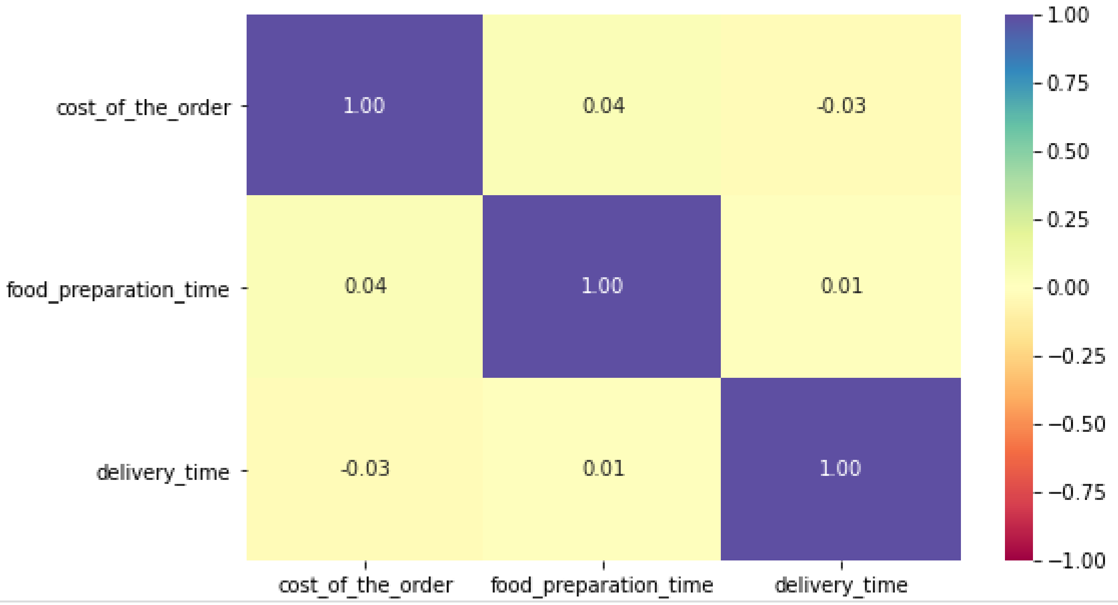
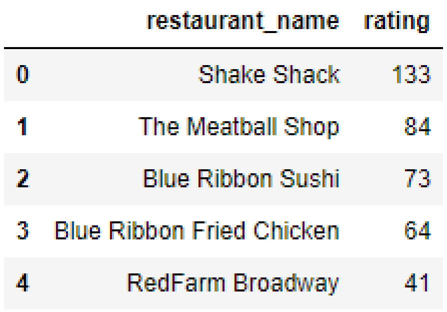


Figure 16: Heatmap for the correlation between cost of the order, food preparation time and delivery time

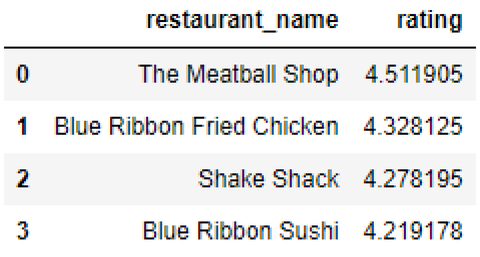
Question 13: The company wants to provide a promotional offer in the advertisement of the restaurants. The condition to get the offer is that the restaurants must have a rating count of more than 50 and the average rating should be greater than 4. Find the restaurants fulfilling the criteria to get the promotional offer.

Answer:



	restaurant_name	rating
0	Shake Shack	133
1	The Meatball Shop	84
2	Blue Ribbon Sushi	73
3	Blue Ribbon Fried Chicken	64
4	RedFarm Broadway	41

Table 8: Restaurants with rating count of more than 50



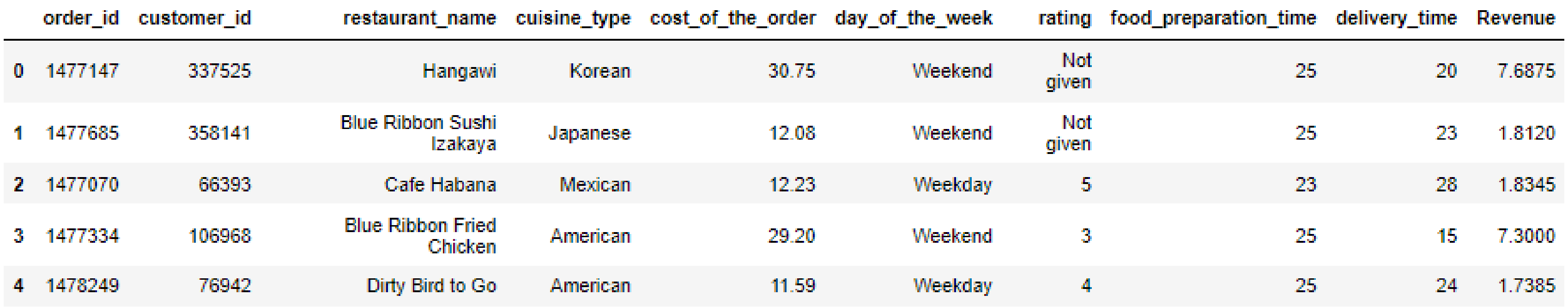
	restaurant_name	rating
0	The Meatball Shop	4.511905
1	Blue Ribbon Fried Chicken	4.328125
2	Shake Shack	4.278195
3	Blue Ribbon Sushi	4.219178

Table 9: Restaurants with average rating greater than 4

Observation: Four restaurents namely Blue Ribbon Fried Chicken, Blue Ribbon Sushi , Shake Shack, and The Meatball shop qualify for promotional offer.

Question 14: 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.

Answer:



	order_id	customer_id	restaurant_name	cuisine_type	cost_of_the_order	day_of_the_week	rating	food_preparation_time	delivery_time	Revenue
0	1477147	337525	Hangawi	Korean	30.75	Weekend	Not given	25	20	7.6875
1	1477685	358141	Blue Ribbon Sushi Izakaya	Japanese	12.08	Weekend	Not given	25	23	1.8120
2	1477070	66393	Cafe Habana	Mexican	12.23	Weekday	5	23	28	1.8345
3	1477334	106968	Blue Ribbon Fried Chicken	American	29.20	Weekend	3	25	15	7.3000
4	1478249	76942	Dirty Bird to Go	American	11.59	Weekday	4	25	24	1.7385

Table 10: Net revenue generated by the company across all orders

Observation: The net revenue is around 6166.3 dollars.

Question 15: 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.)

Answer: The number of total delivery time that above 60 minutess is: 200

Percentage of orders that have delivery time above 60 minutes: 10.54 %

Question 16: 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 weekend?

Answer: The mean delivery time on weekdays is around 28 minutes.

The mean delivery time on weekend is around 22 minutes.

Conclusion and Recommendations :

Question 17: 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.)

Answer: From the analysis some conclusions and recommendations can be made:

Conclusion:

1. Top 5 popular cuisines are American followed by Japanese, Italian, Mexican, Indian.
2. The higher activity on weekend over weekdays
3. It is slight peak at 25 dollar on cost of order.
4. The mean delivery time is 24 minutes.

Recommendations:

1. Improve the customer's response rating their orders, to reduce the rating 'NOT GIVEN' on orders.
2. Given consistent popularity of cuisines across the days of the week, better marketing can be focused on the weekday to boost sales.
3. On promotional offers the organization will have to considers a tie-breaker for cases where customers have the same score.