FoodHub Case Study Project

Python Foundations: FoodHub Data Analysis

September 23, 2022

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- Data Overview
- EDA - Univariate Analysis
- EDA - Multivariate Analysis
- Appendix

Executive Summary (Insights)

- We analyzed a dataset of 1898 customer food orders containing 9 variable columns for FoodHub in New York. Main focus is on customer satisfaction. The dataset is a ratings-based dataset so we have, focused on factors that contribute to customer satisfaction: customer repeat business, and cost of the orders.
- FoodHub's revenue is from a surcharge, fixed margin of the delivery order from the restaurants. Therefore, we'll focus on some concepts related to increase revenue including customer order size and frequency.
- Key Questions:
  - "What are the different variables that influence the number of" orders?
  - Which factor affects the number of" orders the most? What could be the possible reasons for that?
  - What are your recommendations to FoodHub management to capitalize on customer satisfaction?

Executive Summary (Conclutions)

- **Conclusions**,

We have been able to conclude that:

American is the most desired cuisine. Adding Japanese, Italian, and Chinese is nearly approximately 80% of the revenue.

The number of orders are more than double on the weekend.

Delivery time contributes to customer rating. Food preparation time may not.

Lower food cost contributes to lower satisfaction.

Executive Summary (Recommendations)

- **Recommendations to buisness**

  Recommendation focused on increase order frequency, increase order size, and adding new customers. Using cuisine type and feedback rating.

  Focus on offerings at American Japanese, Italian, Italian, and Chinese.

  The number of orders are more than double on the weekend. Capitalize on that.

- **Further Analysis** that can be done

  Dig deeper to explore the main factors.

  Understand the diverse demographics of ethnicities and uncover overlap in demand and satisfaction.

Business Problem Overview and Solution Approach

- Context
- Objective
- Solution Approach
- Methodologies

Business Problem Overview and Solution Approach

- **Context** - ***FoodHub*** is a food aggregator company that helps streamline the logistics of the food ordering processes by connecting customers, restaurants and delivery personnel. Competitors include **GrubHub, Uber Eats,** and **DoorDash**. The customer can rate the order. Revenue is generated by collecting a fixed margin of the delivery order from the restaurants. The customer order data is captured by the company and provided to us.

- **Objective** – FoodHub needs us to analyze the data to help them (1) understand the **demand of different restaurants** which will help them in (2) enhancing their **customer experience**. We will perform the data analysis to find answers to these questions that will help the company to improve the business.

- We will be focusing on (1) **customer ratings** and (2) **demand of different restaurants**.
  - Variable that effect customer ratings
  - Variable that drive demand of restaurants.

- **Key Questions** - “What are the different variables that influence the number of” orders?
  Which factor affects the number of” orders the most? What could be the possible reasons for that?
  What are your recommendations to FoodHub management to capitalize on customer satisfaction?

Business Problem Overview and Solution Approach

- **Solution Approach**
  - Solution approach will rely on statistical analysis, variable analysis, and visualization methodologies.
- **Methodologies**
  - Understanding the **structure of the data**. –Size, shape, data type, statistical summary.
  - **Visualization of the data** – to recognize patterns, tendencies, trends, and correlations.
  - Perform Data Analysis
    - **Univariate** Data Analysis – Make observations on the **statistical distributions**. Identify relevant variables.
    - **Multivariate** Data Analysis – Identify **relationships** between the important variables
  - Conclusion and Recommendations – Use observations and insights collected at each step.

Data Overview

- **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) t

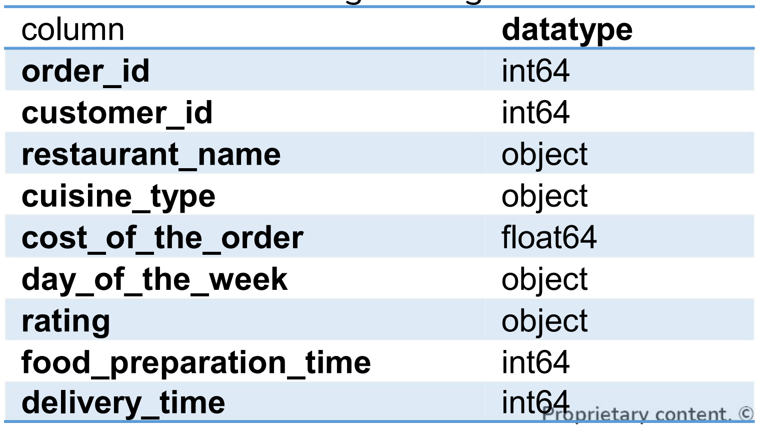
Data Overview - Question 1 and 2

- **Question 1:** How many rows and columns are present in the data? [0.5 mark] The dataframe contains **1898 rows and 9 columns**.

- **Question 2**: What are the datatypes of the different columns in the dataset? [0.5 mark] The dataset contains datatypes **float, integer and object**.

- Five (5) columns are numerical. Four (4) columns are object type.

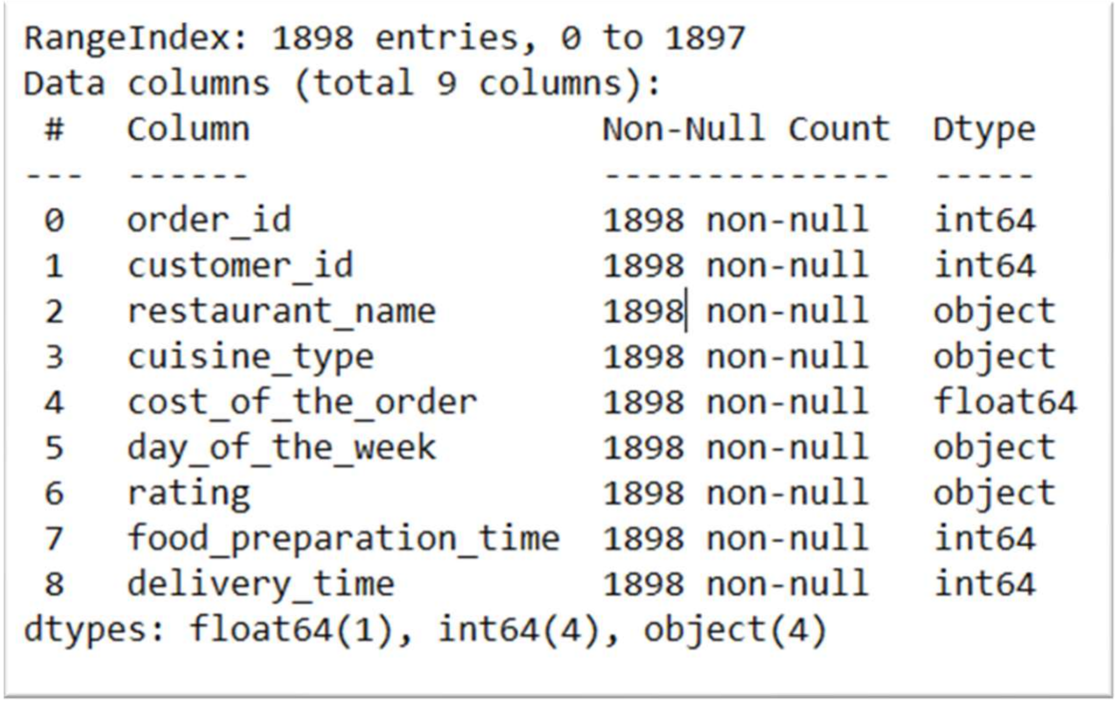
    - The ratings column was given consideration for type conversion from categorical to integer. Decided to keep as categorical for initial analysis and change zero to "not given". Later we will convert when we calculate average rating.



column	**datatype**
order_id	int64
customer_id	int64
restaurant_name	object
cuisine_type	object
cost_of_the_order	float64
day_of_the_week	object
rating	object
food_preparation_time	int64
delivery_time	int64

Data Overview - Question 3

- **Question 3:** Are there any missing values in the data? If yes, treat them using an appropriate method. [1 Mark]
- **No missing values** were identified.
- The row count matched the Non-null count, 1898, for each column.
- **Note:** *rating* is categorical and contains "not given“. This is addresses further in Question 5.

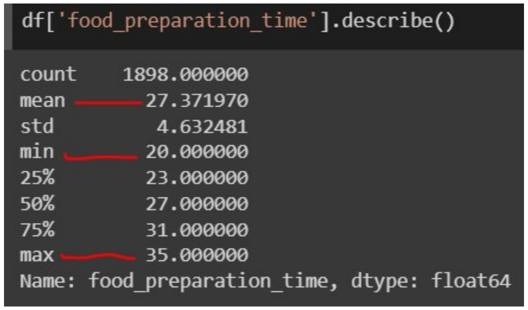


```
RangeIndex: 1898 entries, 0 to 1897
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   order_id               1898 non-null   int64
 1   customer_id            1898 non-null   int64
 2   restaurant_name        1898 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)
```

Data Overview - Question 4

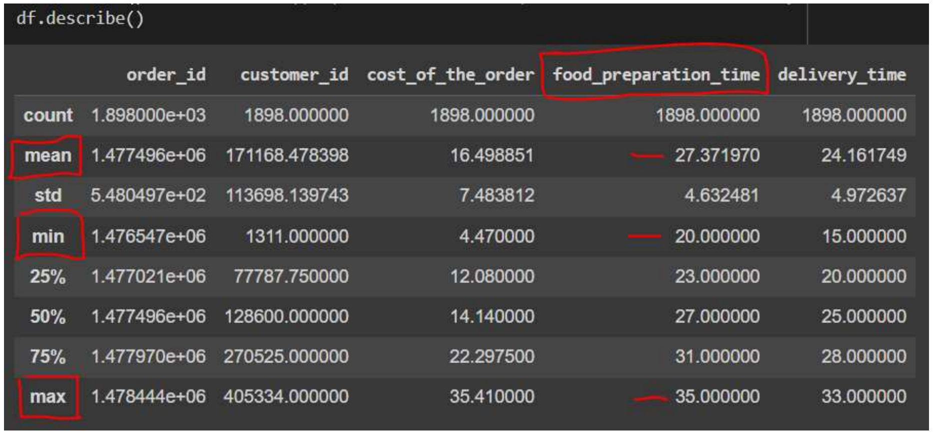
- Question 4: Check the **statistical summary** of the data. What is the **minimum**, **average**, and **maximum** time it takes for food to be prepared once an order is placed? [2 marks]

food_preparation_time	
Minimum	20.0
Average	27.37197
Maximum	35.0



```
df['food_preparation_time'].describe()

count    1898.000000
mean       27.371970
std         4.632481
min        20.000000
25%        23.000000
50%        27.000000
75%        31.000000
max        35.000000
Name: food_preparation_time, dtype: float64
```

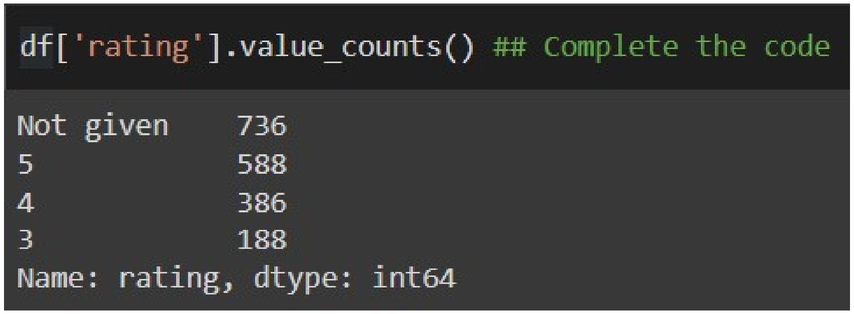


```
df.describe()

            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
```

Data Overview - Question 5

- **Question 5:** How many orders are not rated? [1 mark]
- **Answer 5:** Of the 1898 total orders, **736** are not rated. The text string contains "Not given".



```
df['rating'].value_counts() ## Complete the code
```

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

- **Note:** The ratings column was given consideration for type conversion from categorical to integer. Decided to keep as categorical for initial analysis and change zero to "not given“. Later we will convert when we calculate average rating.

Univariate Analysis (General)

- Please mention regarding univariate analysis for all columns
- Please add answers for all question from 6 till 11

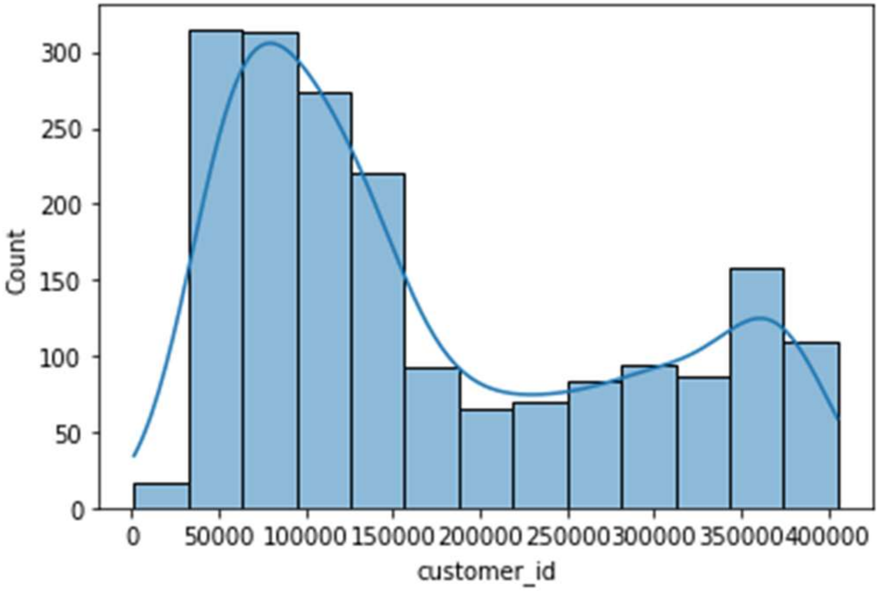
***Note**: You can use more than one slide if needed*

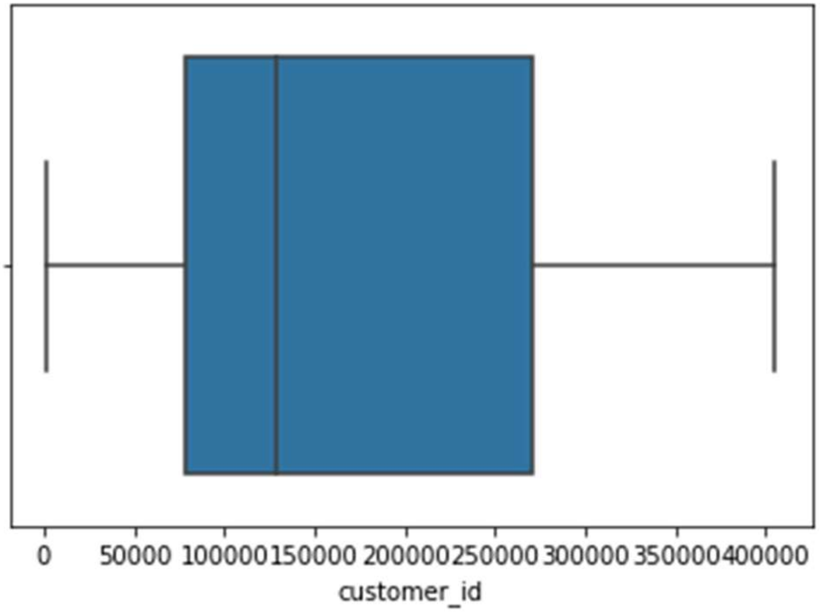
Univariate Analysis (observations on value count)

- Variable unique value count is X. Any observation?
- Order ID count is 1898. All values are unique.
- Customer ID count is 1200. At least 698 customers have ordered twice.
- Restaurant name count is 173. Is this a lot? What area is covered? Over what period of time was the data collected?
- Cuisine type count is 14. Many observations. Ex: The top five Cuisine types takeup 33% of the market share
- Cost of the order count is 1898. All values are unique
- Day of the week count is 2. Raw data was grouped according to weekday vs weekend.
- Rating count is 4. Categorical.
- Food Preparation time count is 1898. All values are unique
- Delivery time count is 1898. All values are unique

Univariate Analysis - Question 6 (Order ID and Customer ID)

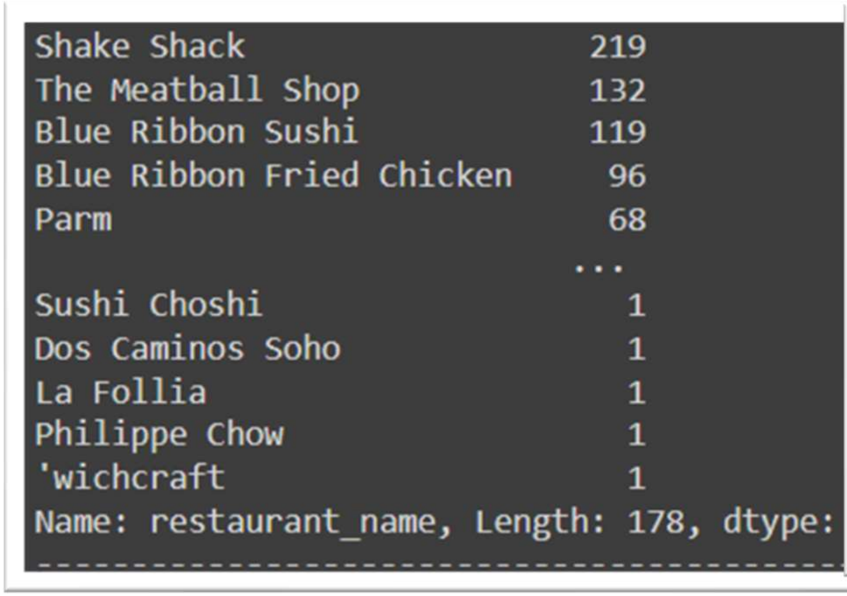
- Order ID count is 1898. Observed all are unique. Graphs are omitted.
- Customer ID count is 1200. Observation is at least 698 customers have ordered twice.



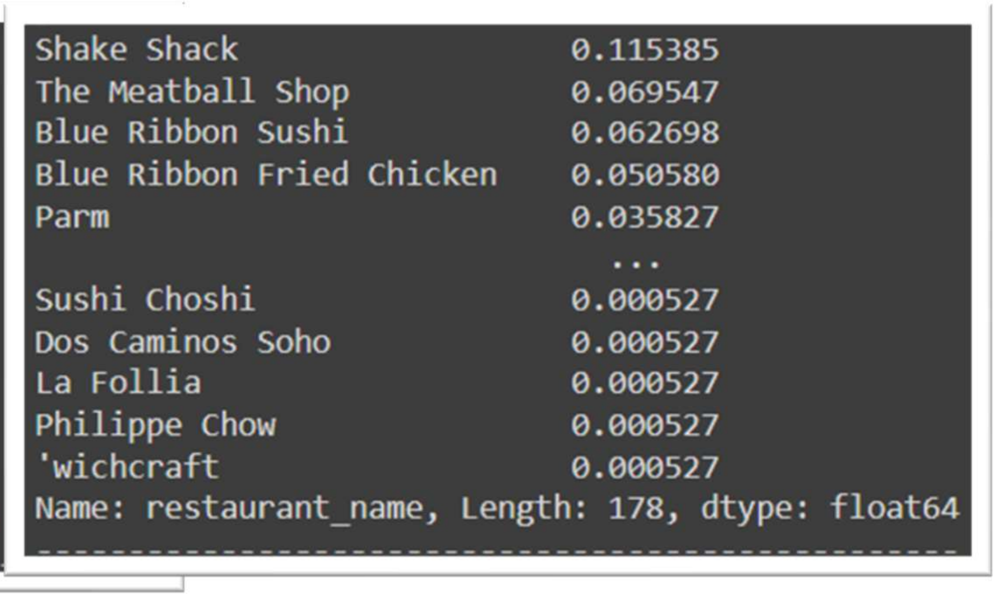


Univariate Analysis - Question 6 (Restaurant name)

- Restaurant name count is 173. Observed Shake Shack is highest (count, %)



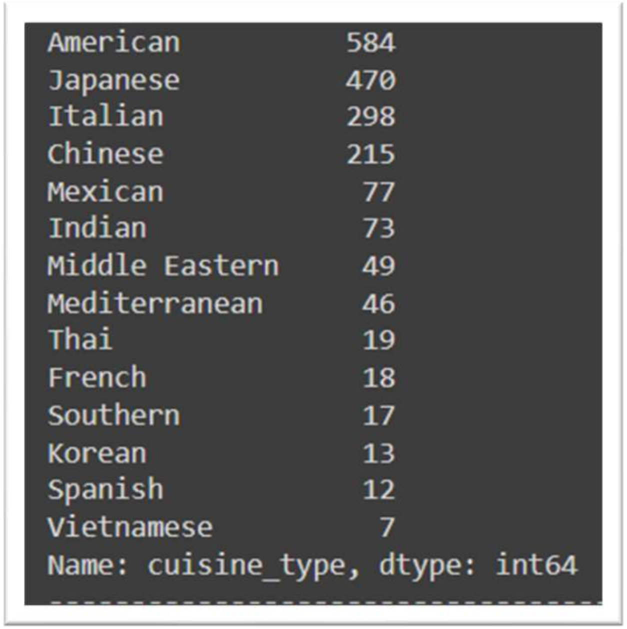
```
Shake Shack                  219
The Meatball Shop            132
Blue Ribbon Sushi            119
Blue Ribbon Fried Chicken     96
Parm                          68
                            ...
Sushi Choshi                   1
Dos Caminos Soho               1
La Follia                      1
Philippe Chow                  1
'wichcraft                     1
Name: restaurant_name, Length: 178, dtype:
------------------------------------------
```



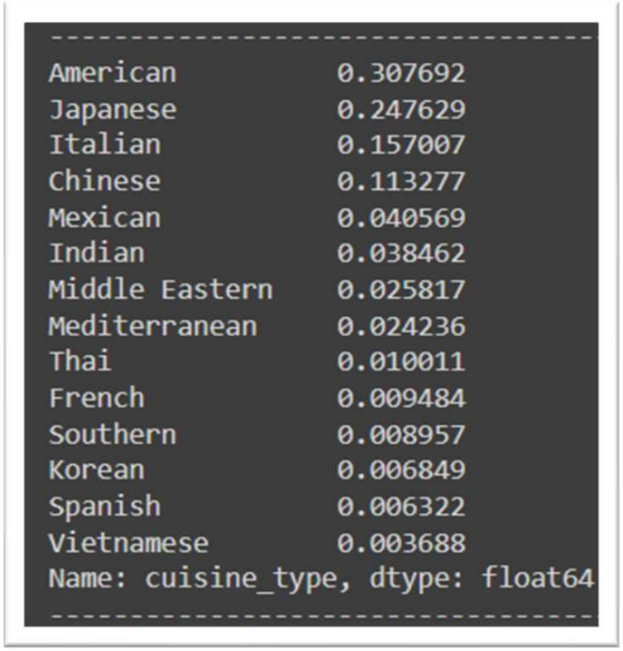
```
Shake Shack                  0.115385
The Meatball Shop            0.069547
Blue Ribbon Sushi            0.062698
Blue Ribbon Fried Chicken    0.050580
Parm                         0.035827
                               ...
Sushi Choshi                 0.000527
Dos Caminos Soho             0.000527
La Follia                    0.000527
Philippe Chow                0.000527
'wichcraft                   0.000527
Name: restaurant_name, Length: 178, dtype: float64
--------------------------------------------------
```

Univariate Analysis - Question 6 (Cuisine type)

- Cuisine type count is 14.
- Observed American and Japanese cuisine making up more than 55%.
- The top four cuisine types take up 80% of the market share

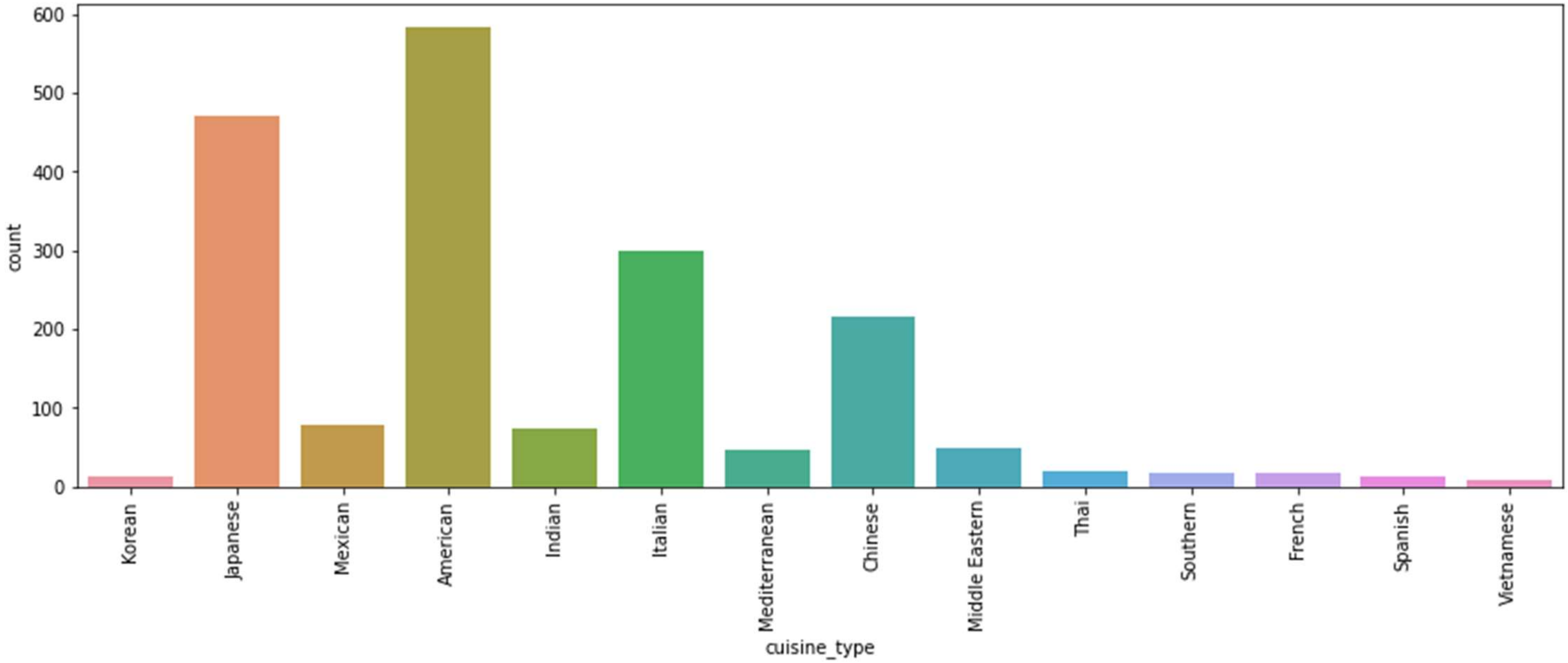


```
American          584
Japanese          470
Italian           298
Chinese           215
Mexican            77
Indian             73
Middle Eastern     49
Mediterranean      46
Thai               19
French             18
Southern           17
Korean             13
Spanish            12
Vietnamese          7
Name: cuisine_type, dtype: int64
----------------------------------
```



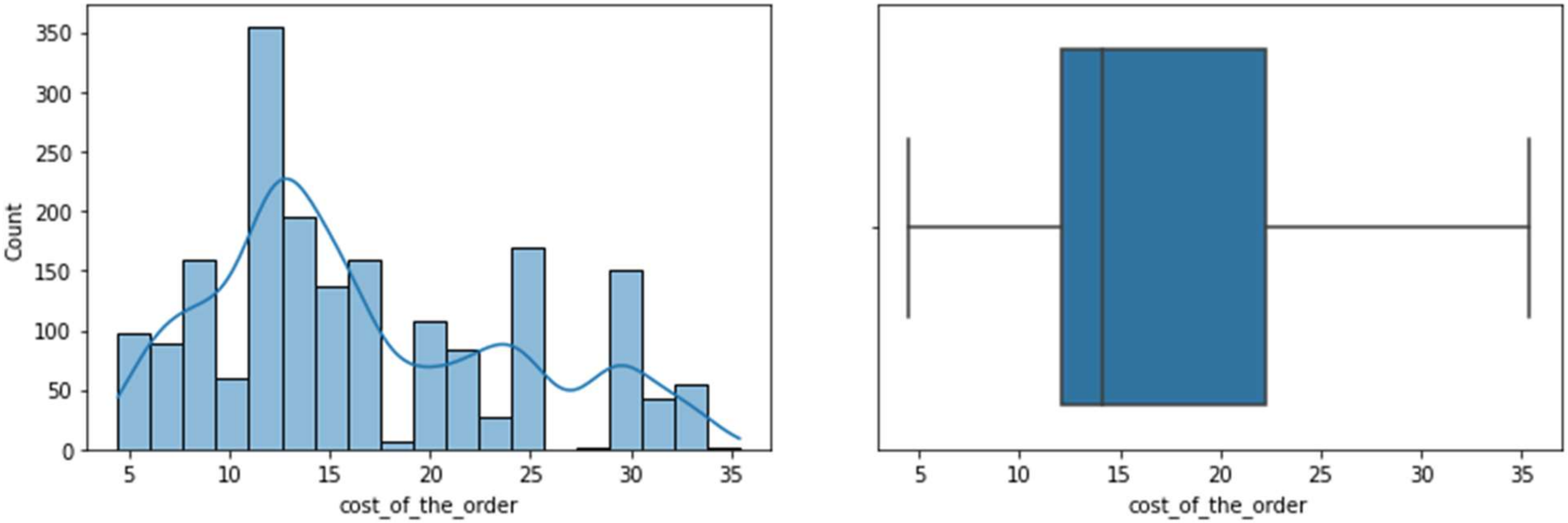
```
----------------------------------
American          0.307692
Japanese          0.247629
Italian           0.157007
Chinese           0.113277
Mexican           0.040569
Indian            0.038462
Middle Eastern    0.025817
Mediterranean     0.024236
Thai              0.010011
French            0.009484
Southern          0.008957
Korean            0.006849
Spanish           0.006322
Vietnamese        0.003688
Name: cuisine_type, dtype: float64
----------------------------------
```

Univariate Analysis - Question 6 (Cuisine type)



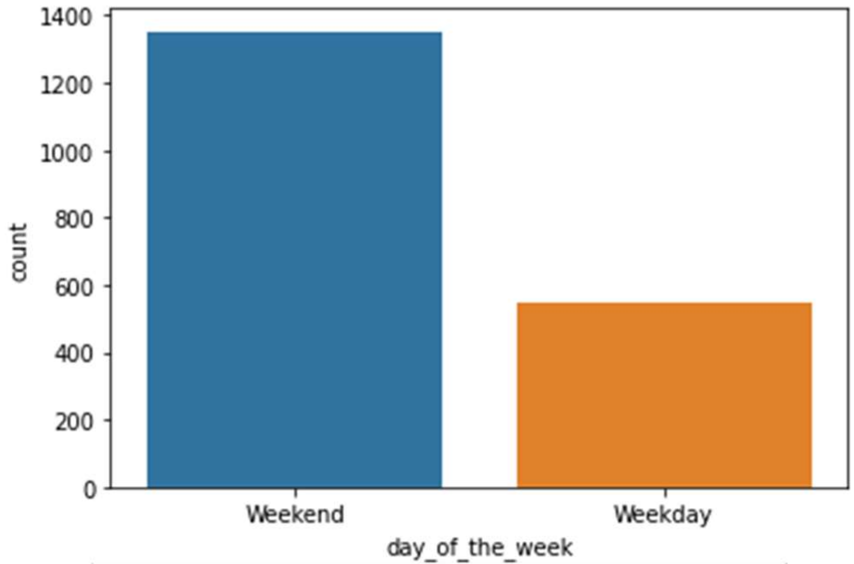
Univariate Analysis - Question 6 (Cost of the order)

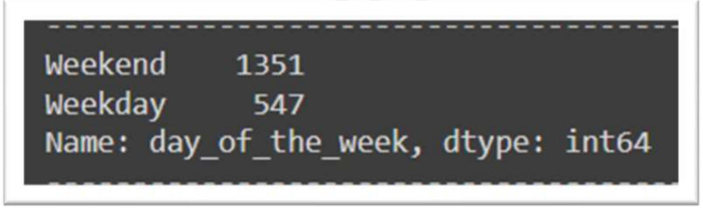
- Cost of the order . Right skew.



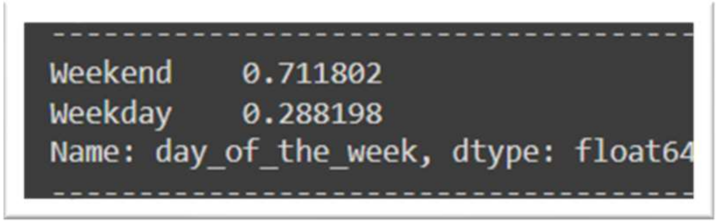
Univariate Analysis - Question 6 (Day of the week)

- Day of the week. Weekends are more than twice as busy.





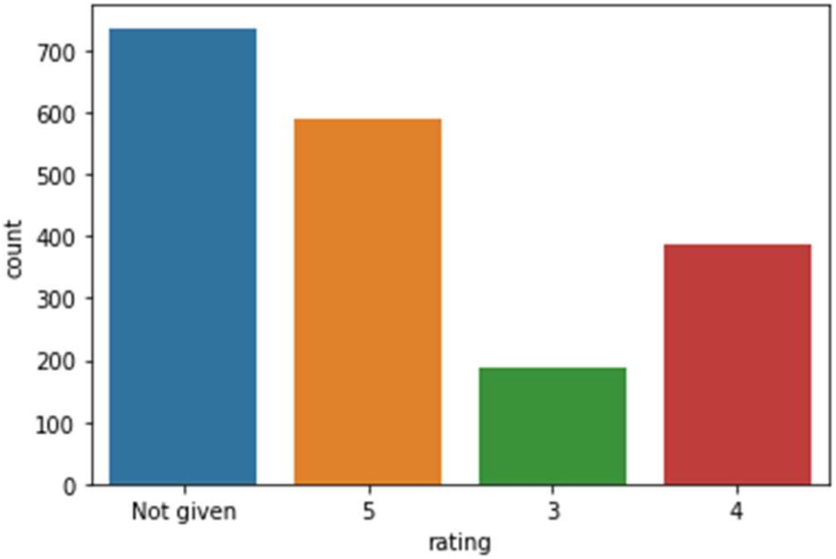
```
------------------------------------
Weekend     1351
Weekday      547
Name: day_of_the_week, dtype: int64
------------------------------------
```

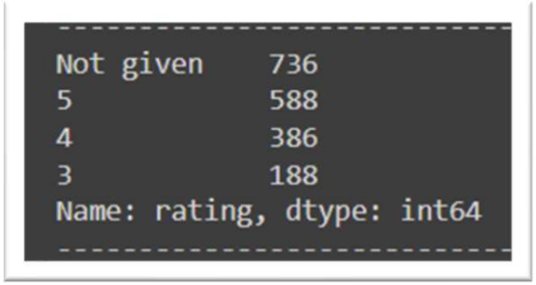


```
------------------------------------
Weekend     0.711802
Weekday     0.288198
Name: day_of_the_week, dtype: float64
------------------------------------
```

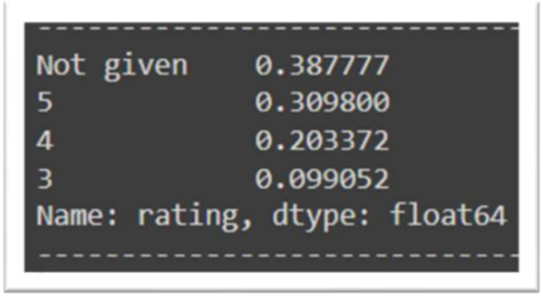
Univariate Analysis - Question 6 (Rating)

- Rating. Break out “not given”.





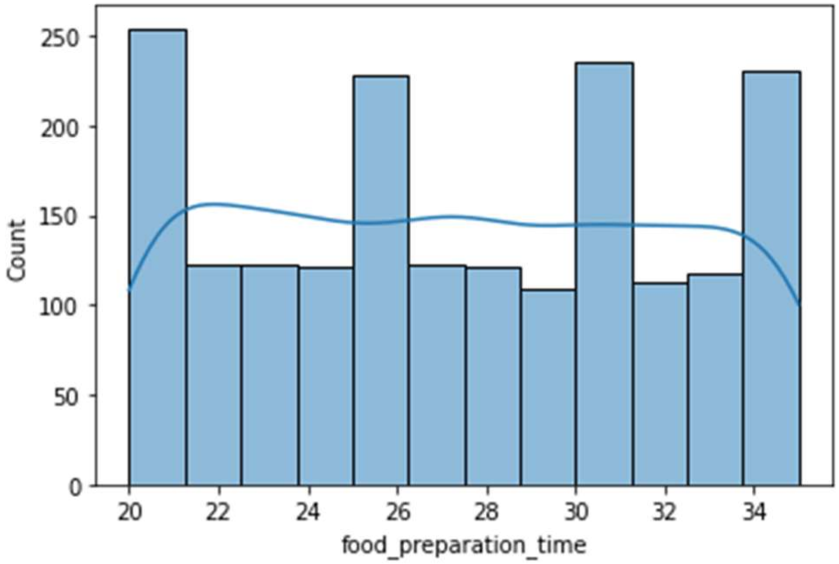
```
Not given     736
5             588
4             386
3             188
Name: rating, dtype: int64
```

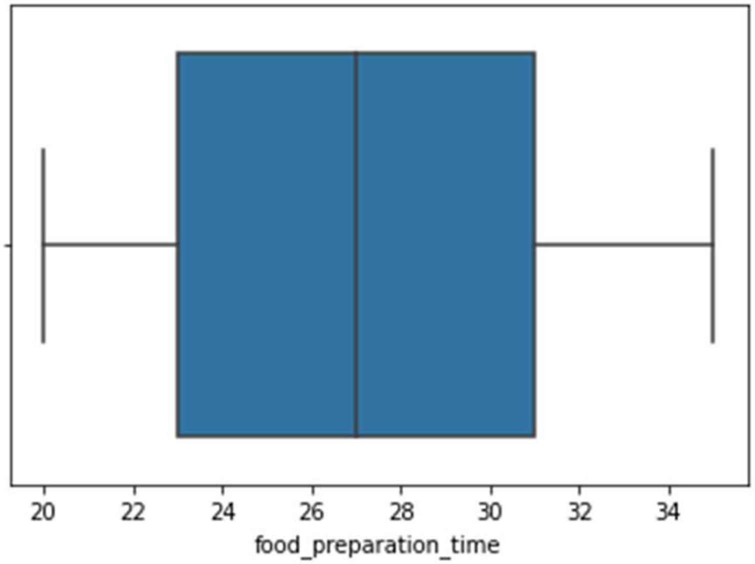


```
Not given     0.387777
5             0.309800
4             0.203372
3             0.099052
Name: rating, dtype: float64
```

Univariate Analysis - Question 6 (Food Preparation time)

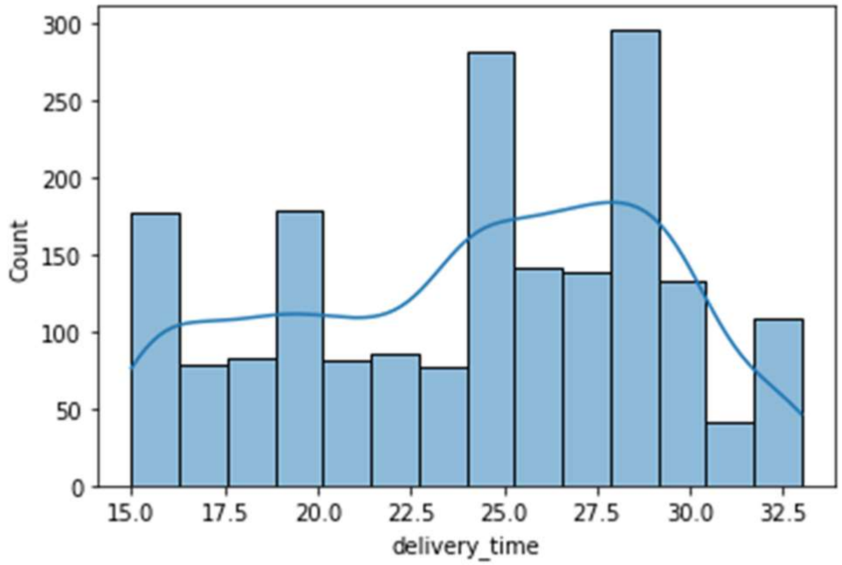
- Food Preparation time. Four, quad-modal. Look at KDE. Normal distribution.

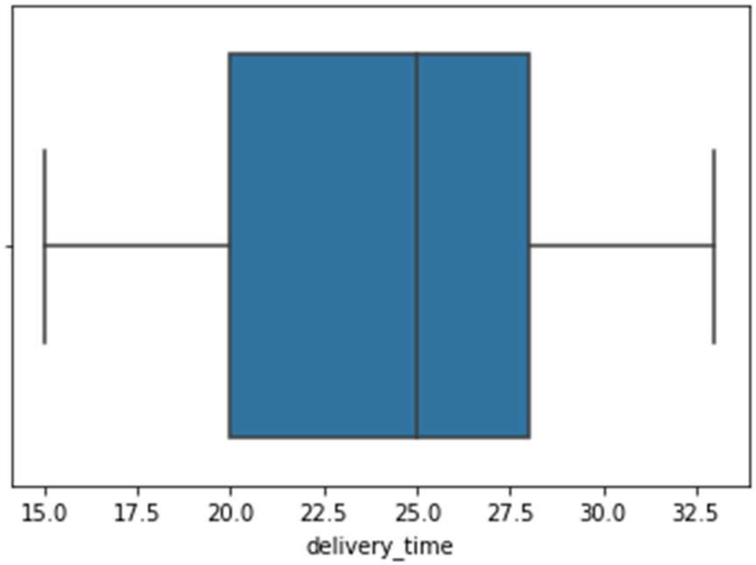




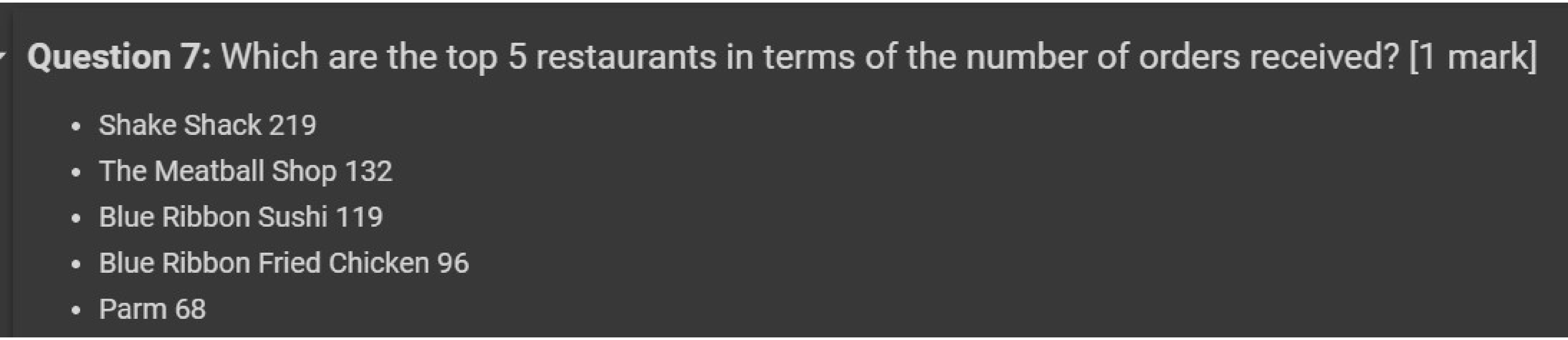
Univariate Analysis - Question 6 (Delivery time)

- Delivery time. Four, quad-modal. Slight left skew.



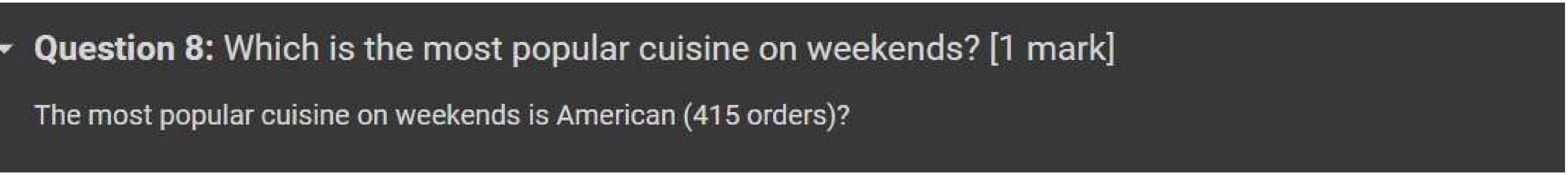


Univariate Analysis - Question 7, 8, 9



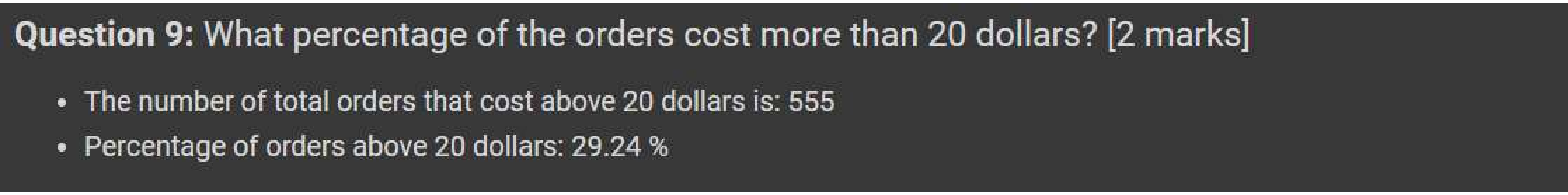
Question 7: Which are the top 5 restaurants in terms of the number of orders received? [1 mark]

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



Question 8: Which is the most popular cuisine on weekends? [1 mark]

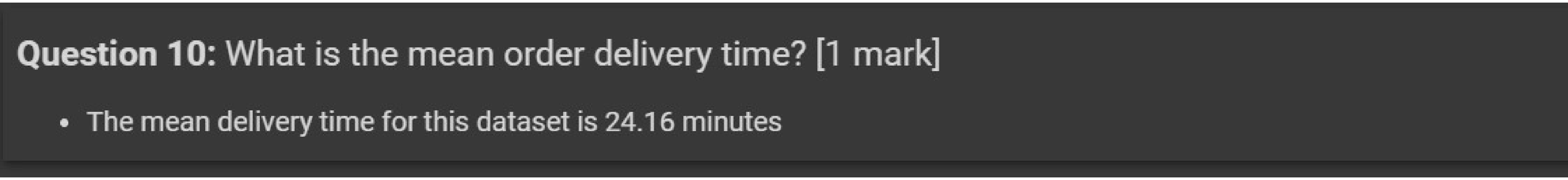
The most popular cuisine on weekends is American (415 orders)?



Question 9: What percentage of the orders cost more than 20 dollars? [2 marks]

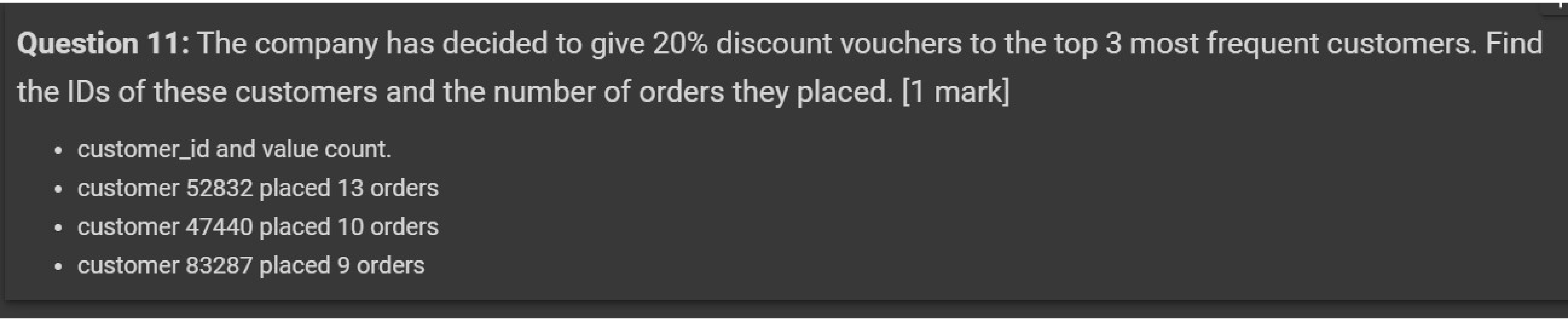
- The number of total orders that cost above 20 dollars is: 555
- Percentage of orders above 20 dollars: 29.24 %

Univariate Analysis - Question 10, 11



Question 10: What is the mean order delivery time? [1 mark]

- The mean delivery time for this dataset is 24.16 minutes



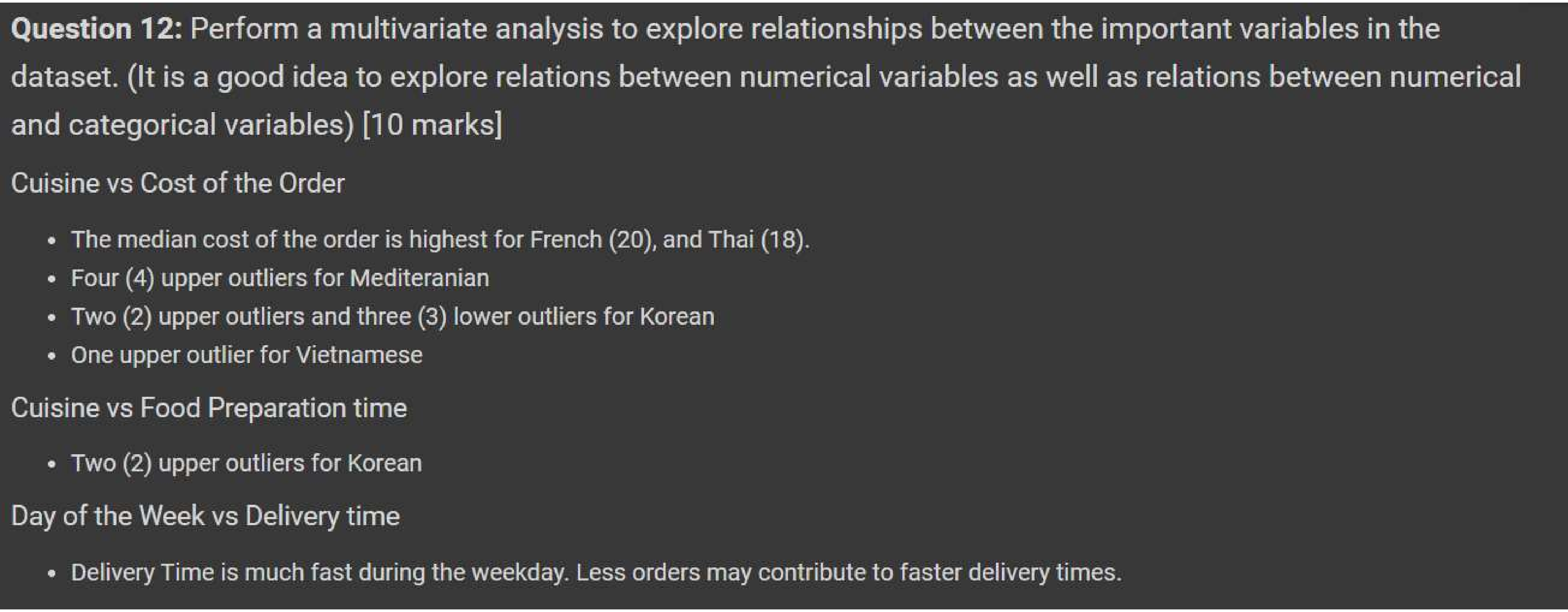
Question 11: The company has decided to give 20% discount vouchers to the top 3 most frequent customers. Find the IDs of these customers and the number of orders they placed. [1 mark]

- customer_id and value count.
- customer 52832 placed 13 orders
- customer 47440 placed 10 orders
- customer 83287 placed 9 orders

Multivariate Analysis – Question 12

- 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) [10 marks]
    - Cuisine vs Cost of the order
    - Cuisine vs Food Preparation time
    - Day of the Week vs Delivery time
    - Revenue generated by the restaurants.
    - Rating vs Delivery time
    - Rating vs Food preparation time
    - Rating vs Cost of the order
    - Correlation among variables

Multivariate Analysis – Question 12



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) [10 marks]

Cuisine vs Cost of the Order

- The median cost of the order is highest for French (20), and Thai (18).
- Four (4) upper outliers for Mediteranian
- Two (2) upper outliers and three (3) lower outliers for Korean
- One upper outlier for Vietnamese

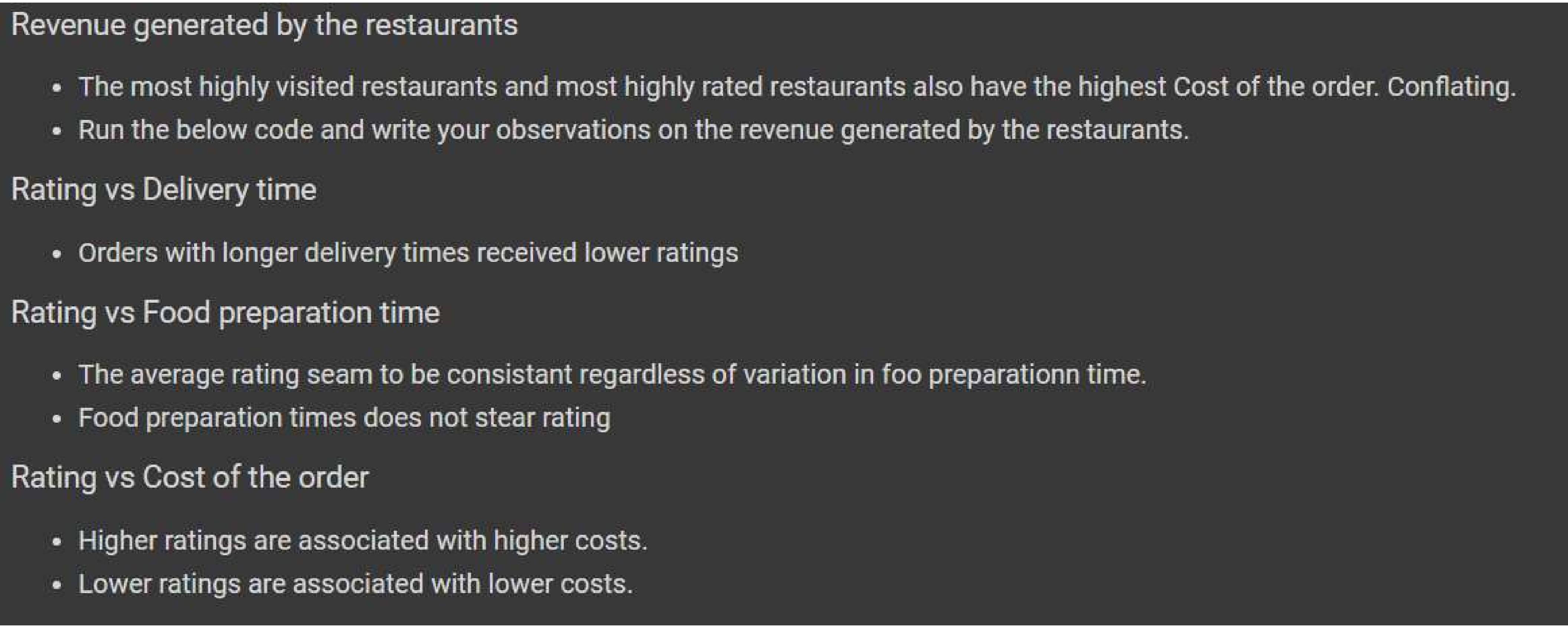
Cuisine vs Food Preparation time

- Two (2) upper outliers for Korean

Day of the Week vs Delivery time

- Delivery Time is much fast during the weekday. Less orders may contribute to faster delivery times.

Multivariate Analysis – Question 12



Revenue generated by the restaurants

- The most highly visited restaurants and most highly rated restaurants also have the highest Cost of the order. Conflating.
- Run the below code and write your observations on the revenue generated by the restaurants.

Rating vs Delivery time

- Orders with longer delivery times received lower ratings

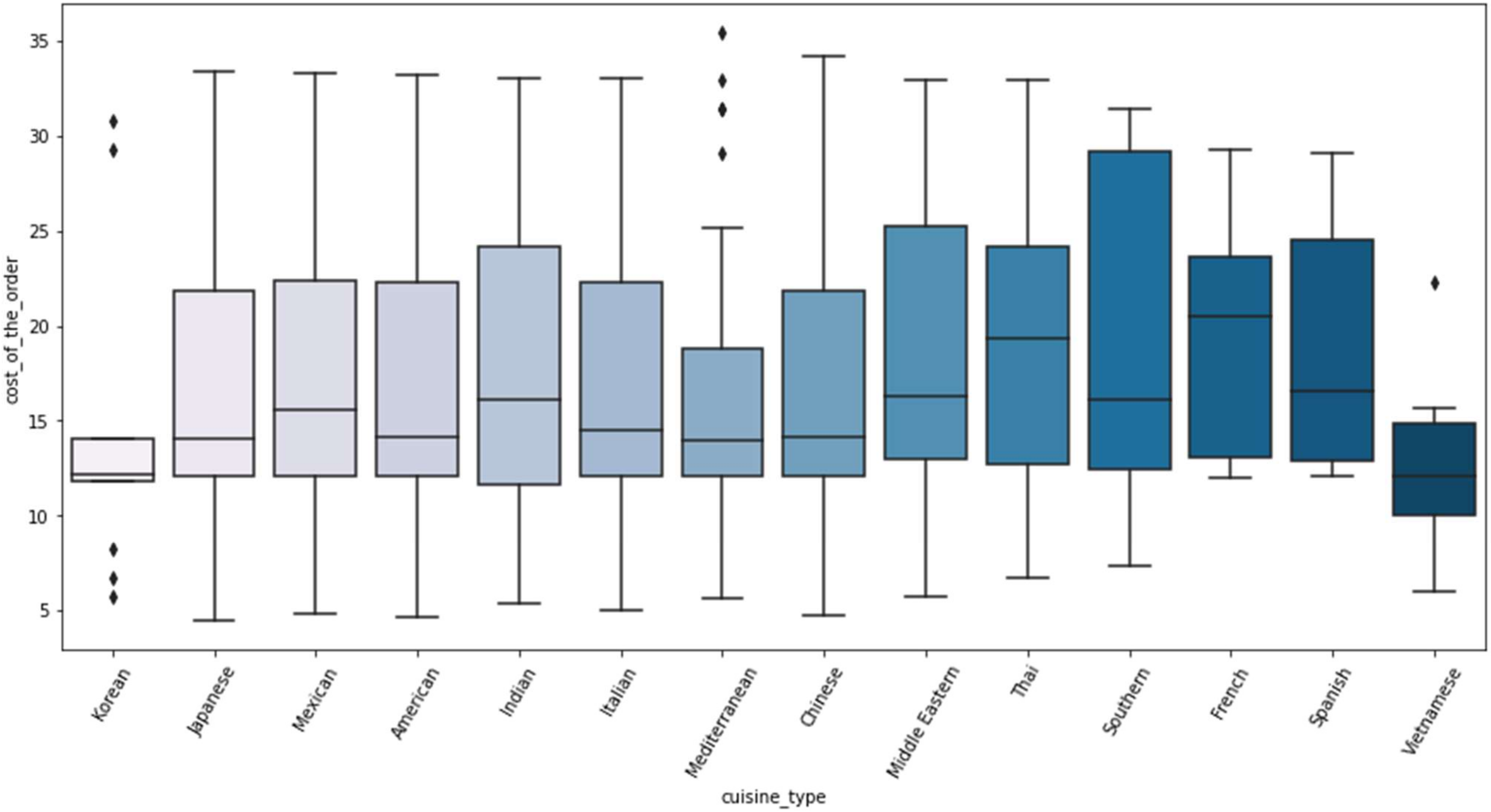
Rating vs Food preparation time

- The average rating seam to be consistant regardless of variation in foo preparationn time.
- Food preparation times does not stear rating

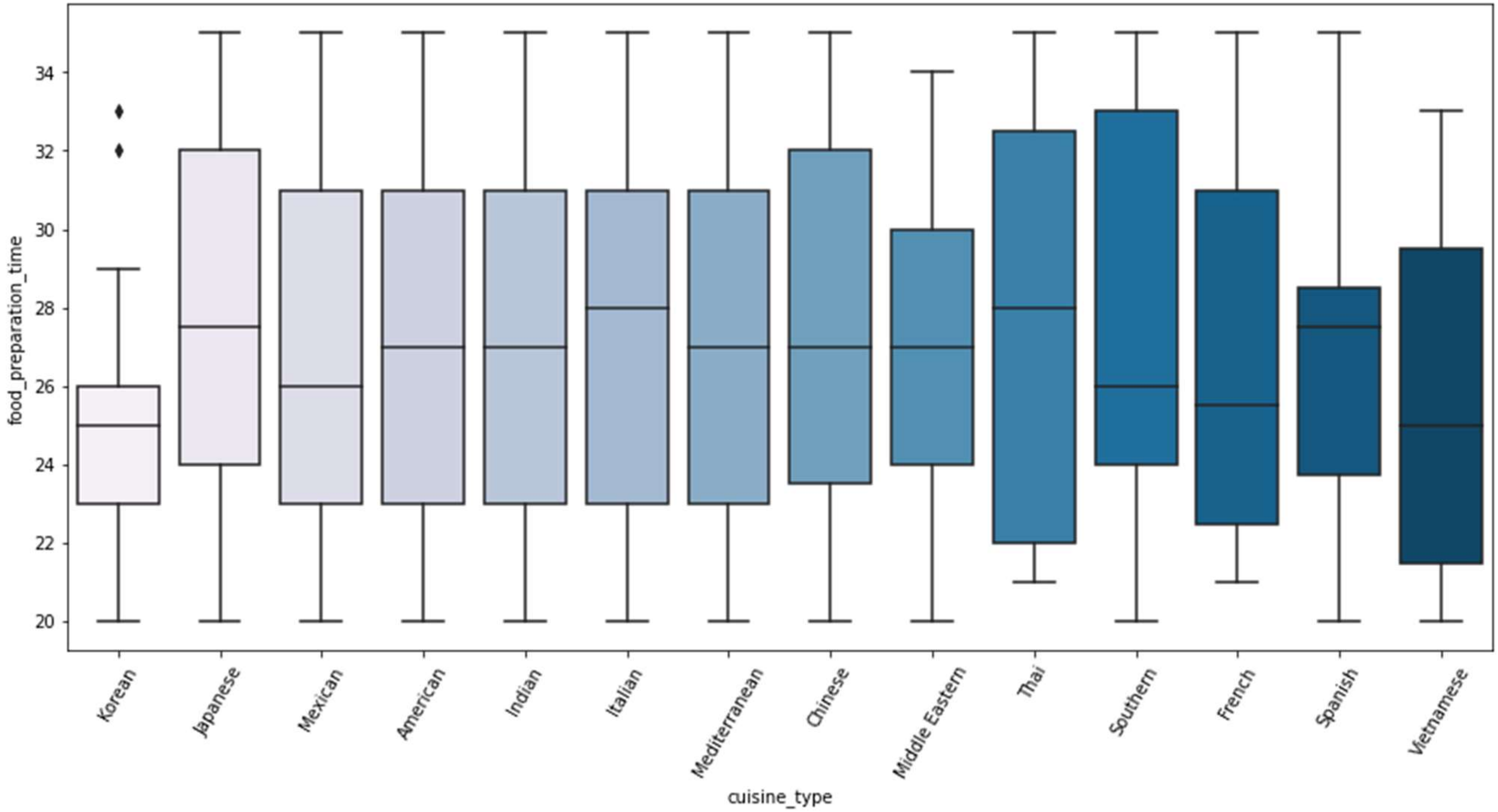
Rating vs Cost of the order

- Higher ratings are associated with higher costs.
- Lower ratings are associated with lower costs.

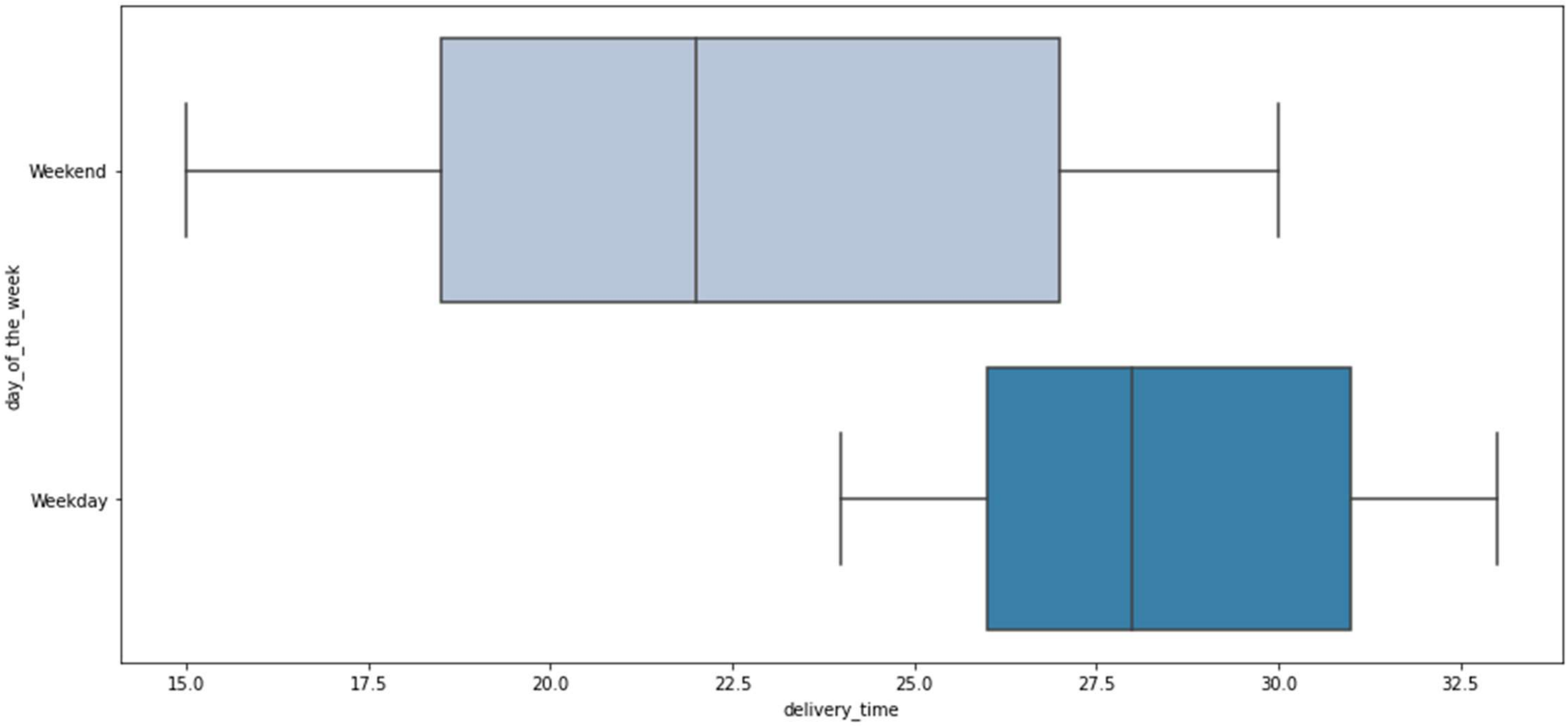
Multivariate Analysis – Cuisine vs Order Cost



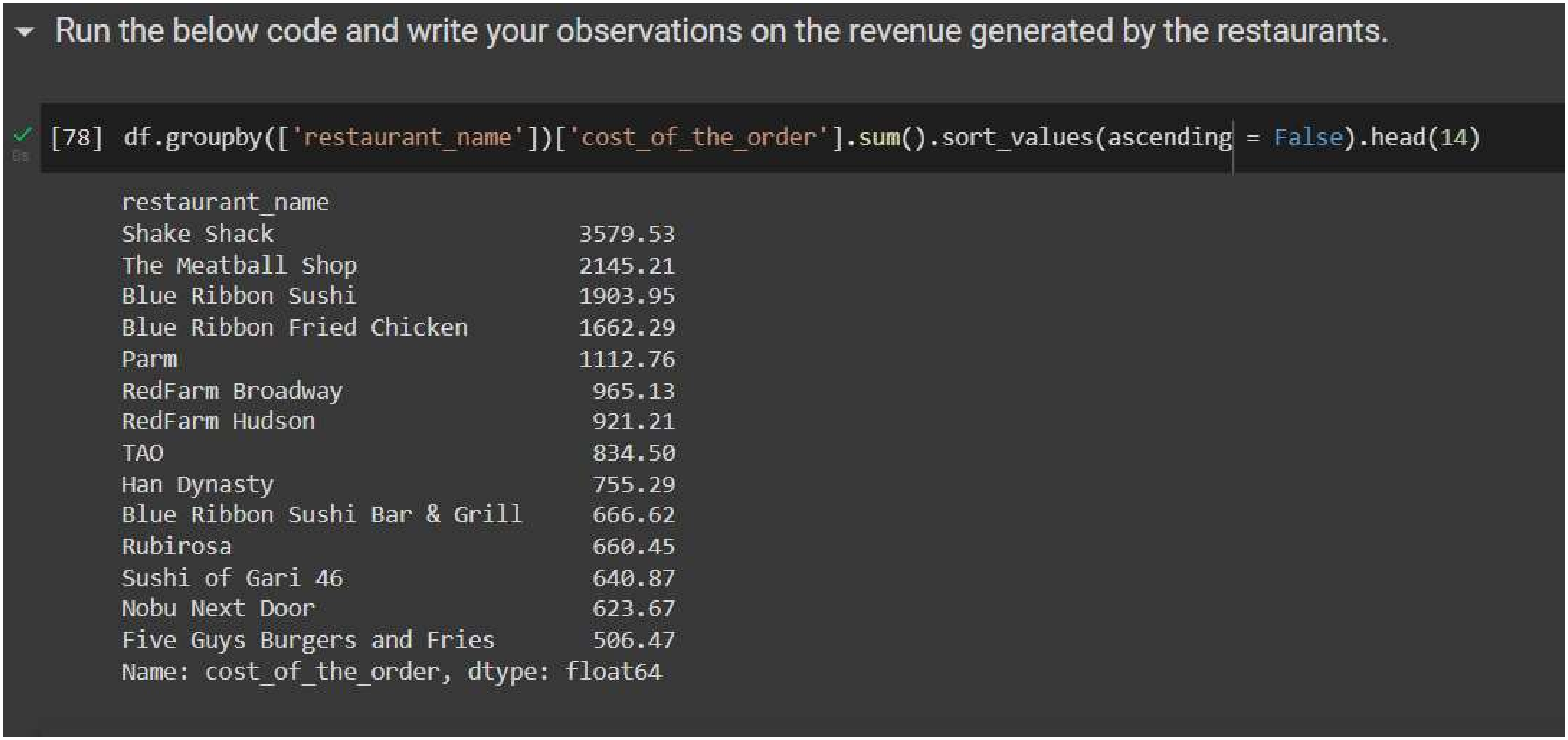
Multivariate Analysis – Cuisine vs Food Preparation time



Multivariate Analysis – Day of the Week vs Delivery time



Multivariate Analysis – Revenue by the restaurant



Run the below code and write your observations on the revenue generated by the restaurants.

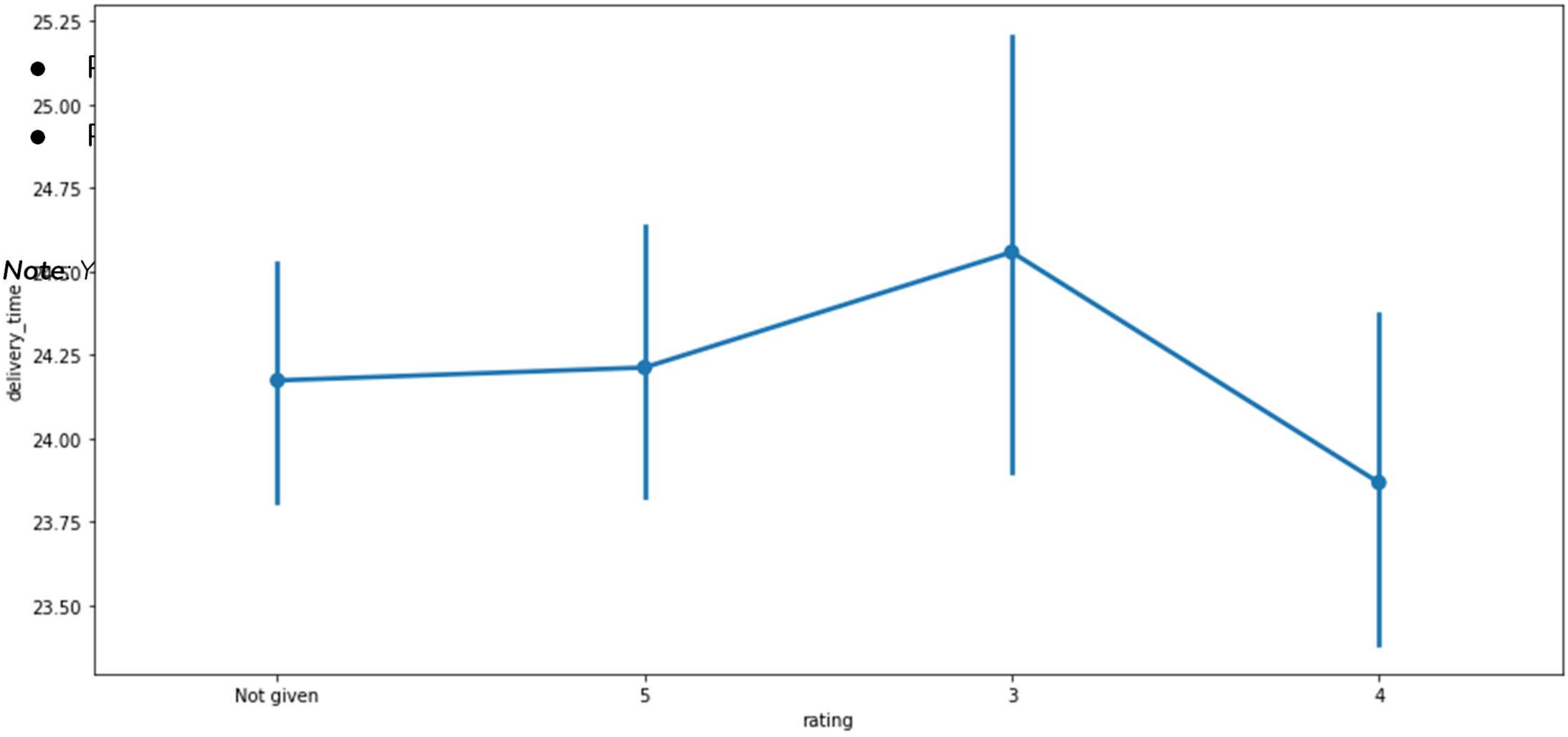
```
[78] df.groupby(['restaurant_name'])['cost_of_the_order'].sum().sort_values(ascending = False).head(14)
```

```
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
```

Multivariate Analysis – Rating vs Delivery time

- F
- F

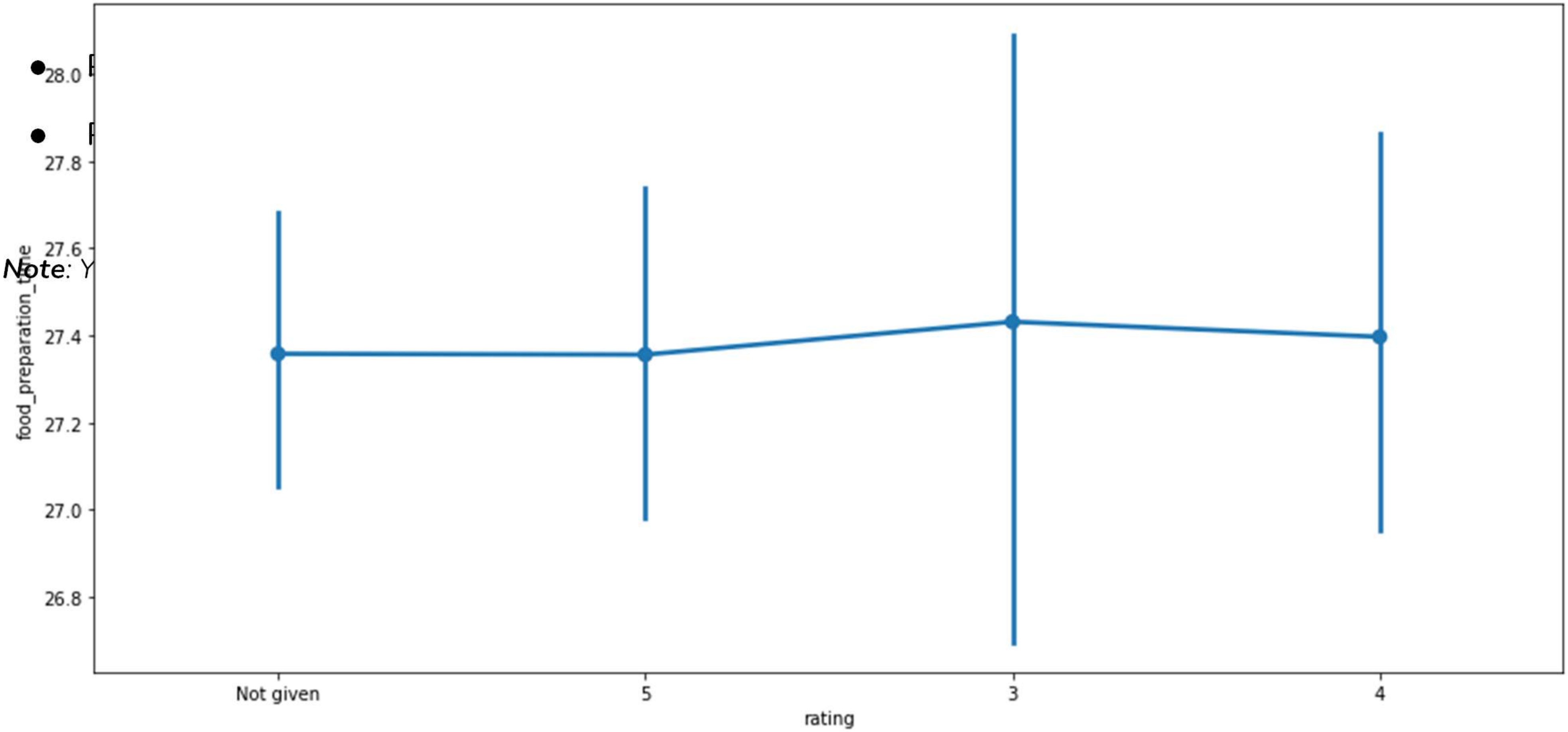
Note: Y



Multivariate Analysis – Rating vs Food preparation time

- F
- F

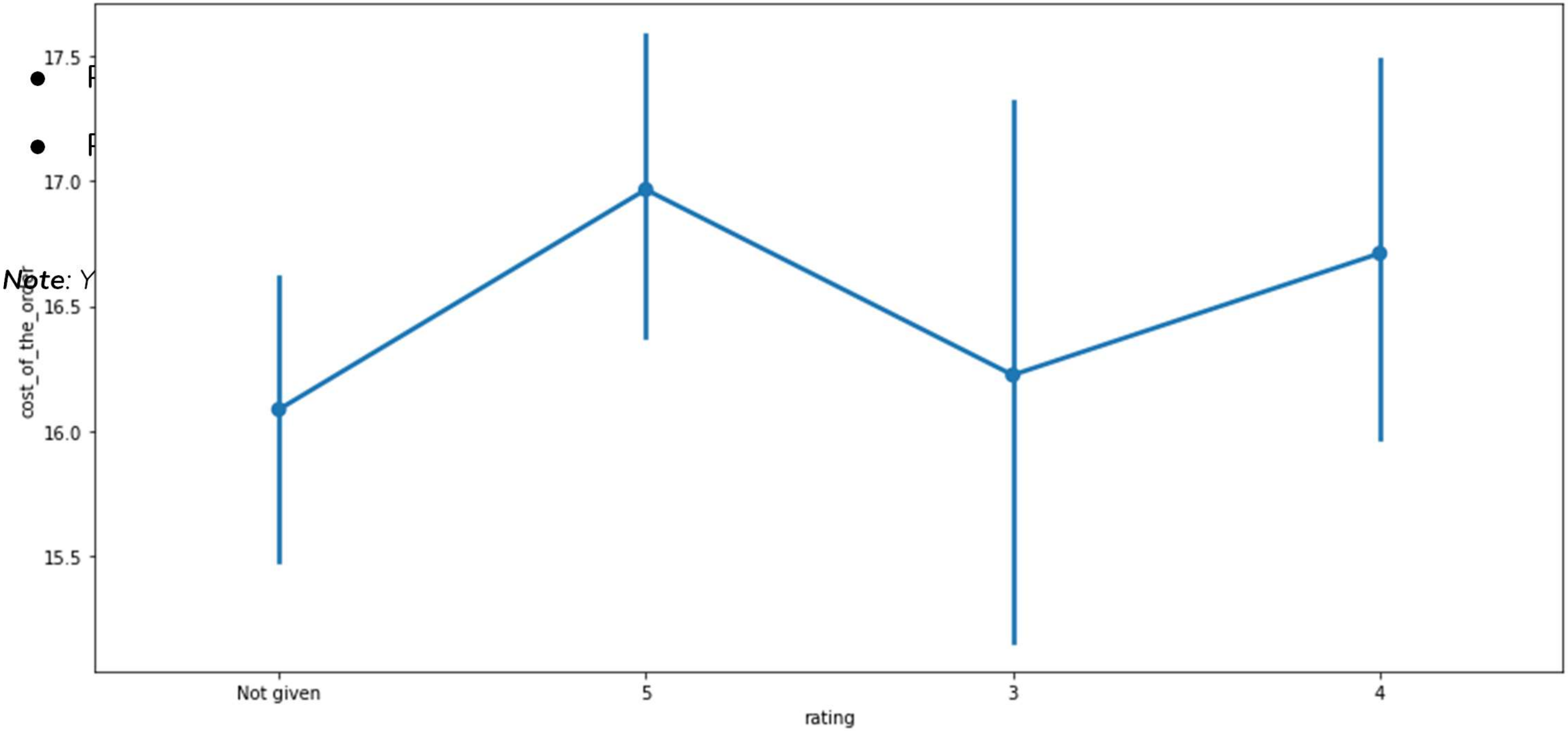
Note: Y



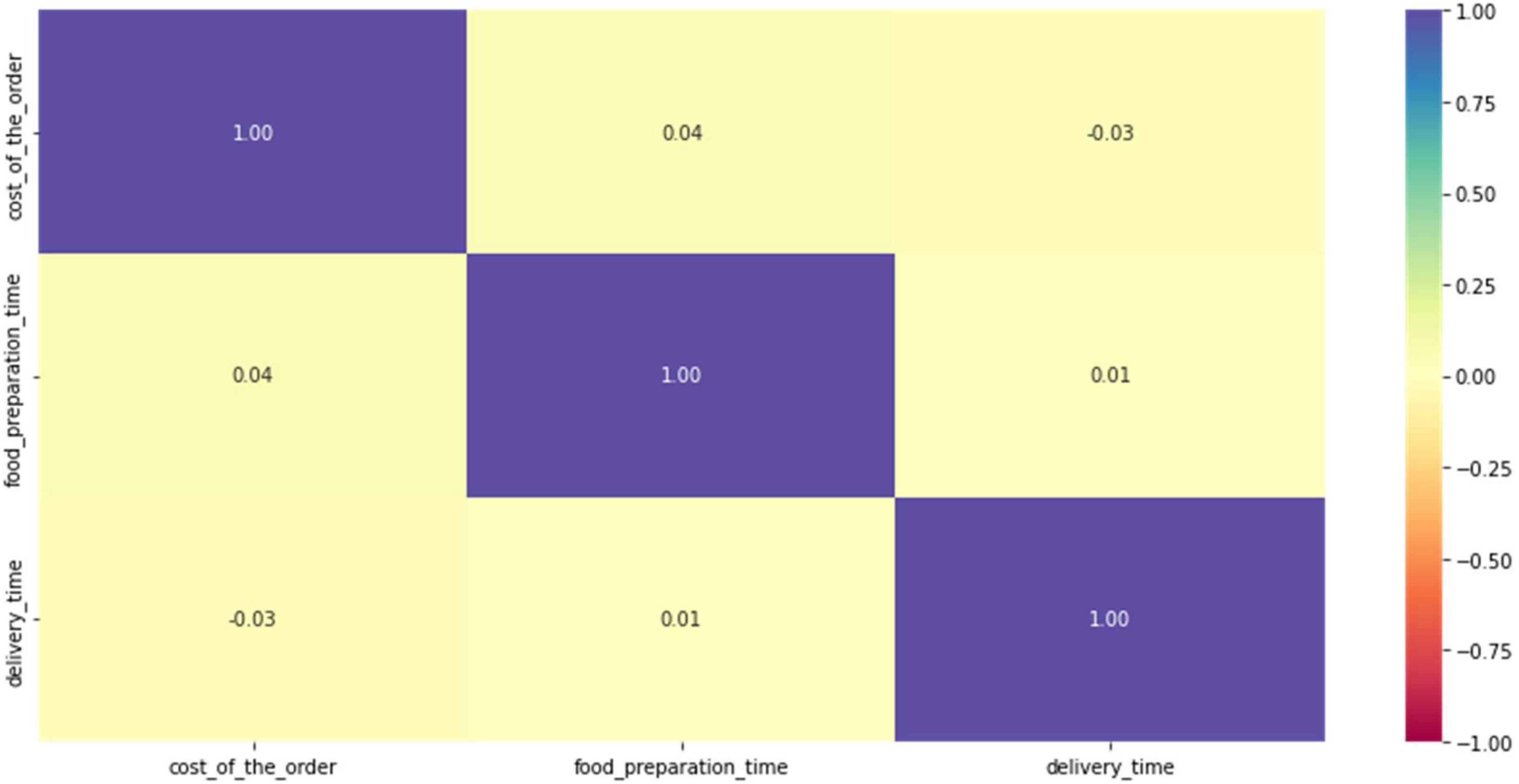
Multivariate Analysis – Rating vs Cost of the order

- F
- F

Note: *Y*



Multivariate Analysis – Correlation among variables

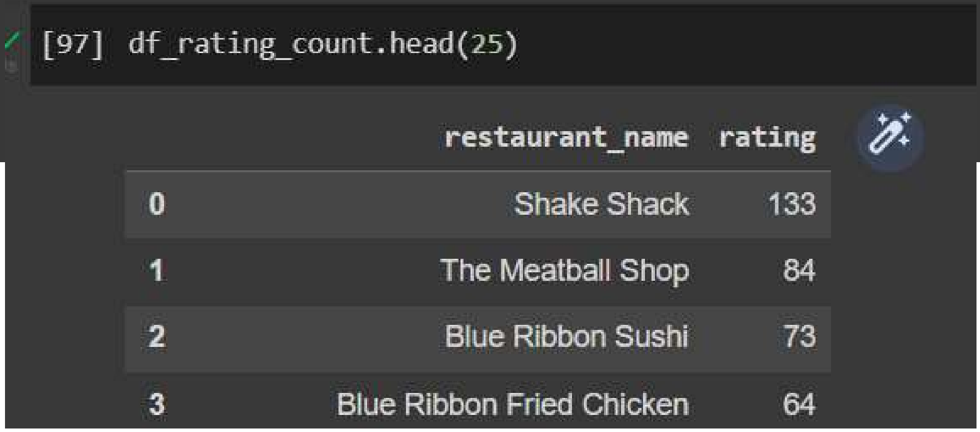


Multivariate Analysis – Question 13

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. [3 marks]

Answer 13: The restaurants fulfilling the criteria to get the promotional offer are

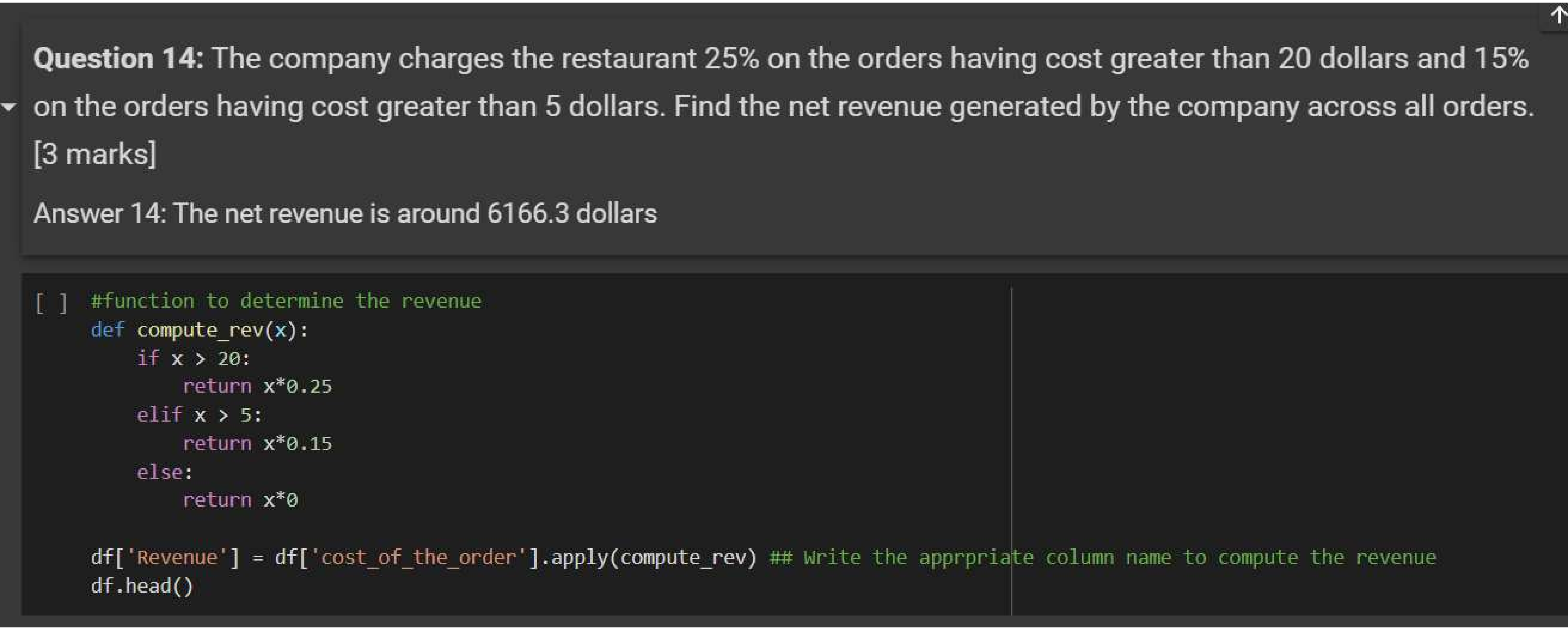
- Shake Shack, 133, 4.278195489
- The Meatball Shop 84, 4.511904762
- Blue Ribbon Sushi 73, 4.219178082
- Blue Ribbon Fried Chicken 64,4.328125



```
[97] df_rating_count.head(25)
```

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

Multivariate Analysis – Question 14



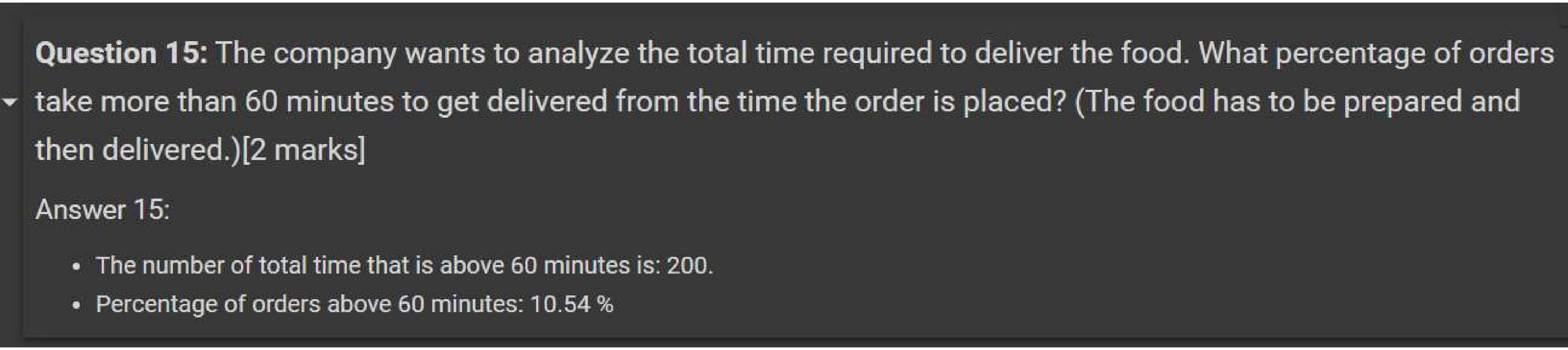
Question 14: The company charges the restaurant 25% on the orders having cost greater than 20 dollars and 15% on the orders having cost greater than 5 dollars. Find the net revenue generated by the company across all orders. [3 marks]

Answer 14: The net revenue is around 6166.3 dollars

```
#function to determine the revenue
def compute_rev(x):
    if x > 20:
        return x*0.25
    elif x > 5:
        return x*0.15
    else:
        return x*0

df['Revenue'] = df['cost_of_the_order'].apply(compute_rev) ## Write the apprpriate column name to compute the revenue
df.head()
```

Multivariate Analysis – Question 15

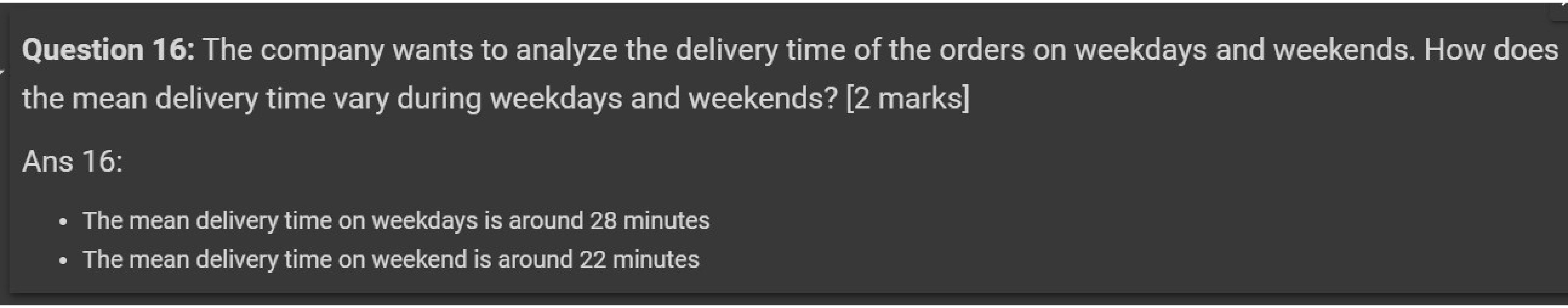


Question 15: The company wants to analyze the total time required to deliver the food. What percentage of orders take more than 60 minutes to get delivered from the time the order is placed? (The food has to be prepared and then delivered.)[2 marks]

Answer 15:

- The number of total time that is above 60 minutes is: 200.
- Percentage of orders above 60 minutes: 10.54 %

Multivariate Analysis – Question 16



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 weekends? [2 marks]

Ans 16:

- The mean delivery time on weekdays is around 28 minutes
- The mean delivery time on weekend is around 22 minutes

APPENDIX

Appendix – Full Context Given

- Full Context Given

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.

Appendix – Full Objective Given

- Full Objective Given

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.

Appendix – Full Data Information Given

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

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

Great Learning

Happy Learning !