FoodHub Data Analysis

Foundations for Data Science

03.07.2024

Contents / Agenda

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

Executive Summary

Problem summary:

The goal of this analysis was to gain insights into customer demand across various restaurants to help FoodHub enhance its customer experience and business performance. FoodHub earns revenue through commissions on orders, meaning higher order volumes directly increase profits.

Therefore, the key question was: What are the customers' preferences? By answering this question, we aim to better understand how to attract and retain more customers.

To achieve this, I have analyzed data consisting of order IDs, customer IDs, restaurant names, cuisine types, order prices, order ratings, delivery periods, and preparation and delivery times. This data allowed me to form a basic understanding of customer demand regarding food ordering.

The dataset included nearly 1,900 customer orders. Customer feedback could have bene accessed in two ways:

- **Direct**: Ratings (1 – 5),
- **Indirect**: Number of orders placed (popularity).

Analyzing these factors and their relationships with the collected data provides insights into possible business development directions.

Conclusions:

- The most popular cuisine is American (30.87%), followed by Japanese, Italian, and Chinese, which together cover 83% of all orders.
- The five most popular restaurants serve American, Japanese, and Chinese food. The most popular restaurant, ‘Shake Shack’, accounts for approximately 11% of all orders.
- Customers tend to order cheaper meals, but price does not significantly impact the given ratings.
- Most orders are placed on weekends (71%), involving around 24% more restaurants than on weekdays.
- In general, orders placed on weekends are delivered faster, with an average reduction in delivery time of around 21%.
- 39% of all orders are not rated. Among the rated orders, approximately 51% received the highest score of 5. No ratings are lower than 3.
- Ratings do not seem to directly depend on delivery or preparation time.
- Since there is no significant relationship between price and ratings, we can assume that customers value other factors more, e.g. the quality of food.

Recommendations:

- The significant difference between weekend and weekday orders suggests potential for business expansion. Introducing special offers, such as happy hours or free delivery on weekdays, could encourage more orders during these times.
- One restaurant, ‘The Meatball Shop’, offers both American and Italian cuisine, with Italian being preferred. As the second most popular restaurant, this suggests potential in partnering with restaurants that offer diverse cuisines.

- The company should develop strategies to encourage more customers to provide ratings. Unrated orders offer no feedback, making it harder to meet customers' needs.
- While overall ratings provide some insight into customer satisfaction, introducing separate categories, e.g. for food quality and delivery time, would offer a more detailed understanding of customer preferences.
- Collecting information about the location of restaurants and more detailed data on the timing of orders would be beneficial. This data could help identify new customer bases, potential restaurant partners, and opportunities for targeted special offers.
- The highest average ratings are for Spanish, Thai, and Indian cuisines. Leveraging the strong ratings of these top cuisines can attract more customers, especially since these cuisines are not among the most popular.

Business Problem Overview and Solution Approach

The goal of this analysis was to gain insights into customer demand across various restaurants to help FoodHub enhance its customer experience and business performance. FoodHub earns revenue through commissions on orders, meaning higher order volumes directly increase profits.

Therefore, the key question was: What are the customers' preferences? By answering this question, we aim to better understand how to attract and retain more customers.

To achieve this, I have analyzed data consisting of order IDs, customer IDs, restaurant names, cuisine types, order prices, order ratings, delivery periods, and preparation and delivery times. This data allowed me to form a basic understanding of customer demand regarding food ordering.

The dataset included nearly 1,900 customer orders. Customer feedback could have been accessed in two ways:

- **Direct**: Ratings (1 – 5),
- **Indirect**: Number of orders placed (popularity).

Analyzing these factors and their relationships with the collected data provides insights into possible business development directions.

Data Overview

Data overview:

- The data set consists of 9 columns and 1898 rows. The data in each row corresponds to the single order placed by a customer.
- Column names are without typos with underscores, no white spaces. The name of the column with order costs is different to the one specified in the detailed data dictionary (p. 1.0.4) but still covers the meaning of the data inside the column – there is no need to correct the names of the columns.
- Looking at the last 5 row, we can see that there are some additional characters in one of restaurants name, i.e. "Chipotle Mexican Grill **$1.99 Delivery**". This needs further examination to see if it an error and whether there are more restaurant names like this.

Questions:

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

- The number of rows: 1898
- The number of columns: 9

Question 2: What are the datatypes of the different columns in the dataset?

- To check the data types I used the `info()` method which did not return any missing values. However, looking at the first and last 5 rows I could see that not every order was given a rank – "Not given", which <u>needs further examination</u>.
- The values in:
  - *order_id* and *custormer_id* columns are integers, <u>which is fine</u>.
  - *restaurant_name* and *cuisineg_type* columns are objects - <u>most likely strings, can be changed to categorical</u>.
  - *cost_of_the_order* column are floats - <u>which is fine</u>.
  - *day_of_the_week* column are objects - <u>most likely strings, can be changed to categorical</u>.
  - *rating* column are objects, should be categorical/numerical values - <u>can be changed after further examination</u>.
  - *food_preparation_time* and *delivery_time* columns are integers, <u>which is fine</u>.

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

- The current observations confirm previous findings, showing that there are no NaN values in the dataset.
- However, there are missing entries in the rating column, which constitute a significant portion (~39%) of the data. Due to their prevalence, these entries cannot be simply discarded.
- Imputing these missing ratings is not advisable, as they could have a significant impact on analyses.
- Instead, we can treat these missing ratings as a distinct category within this variable, if necessary.
- From a business perspective, these missing ratings could offer valuable insights. For instance, we could explore strategies to encourage users to provide ratings more frequently.

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?

Numeric data:

	cost_of_the_order	food_preparation_time	delivery_time
count	1898.00	1898.00	1898.00
mean	16.50	27.37	24.16
std	7.48	4.63	4.97
min	4.47	20.00	15.00
25%	12.08	23.00	20.00
50%	14.14	27.00	25.00
75%	22.30	31.00	28.00
max	35.41	35.00	33.00

Object data:

	restaurant_name	cuisine_type	day_of_the_week	rating
count	1898	1898	1898	1898
unique	178	14	2	4
top	Shake Shack	American	Weekend	Not given
freq	219	584	1351	736

- Food preparation time: **min: 20min, average: ~27.5min, max: 35min.**
- The distribution of both food preparation and delivery times appears to be symmetric.
- There is a high difference between the min and max cost of orders.
- The cost of orders exhibits a right-sewed distribution, indicating that people tend to order cheaper food.
- The most popular restaurant is 'Shake Shack', while the most popular cuisine type is American.
- There are 178 restaurants and 14 different cuisine types.
- The majority of orders are placed on weekends.
- As previously noted, a lot of people do not rate their orders.

Question 5: How many orders are not rated?

- There are **736** (39%) orders that are not rated.

Univariate Analysis

Data Exploration:

1. I have checked if data classified as *object* (restaurant_name & cuisine_type) is string or mixed type.
    - They were all strings.
2. I have changed data in the *“day_of_the_week”* and *“cuisine_type”* columns to category type.
3. I have looked for duplicated values based on the *“order_id”* column (Primary Key).
    - There were no duplicates in the data set.
4. I have checked the pairs of restaurant names and cuisine types.
    - It appeared that the “*Meatball Shop”* has assigned two types of cuisine – American and Italian. Probably it serves both types of cuisine. For clarity this should be checked by consulting it with FoodHub or by googling the restaurant.
    - I assumed that this was correct, and I left it as it was.
    - The interesting thing was that “*Meatball Shop*” customers prefer Italian food (112 orders) over American (20 orders).
5. I have checked the restaurant names, as from earlier data examination it appeared that some of them may consist non-alphanumeric signs.

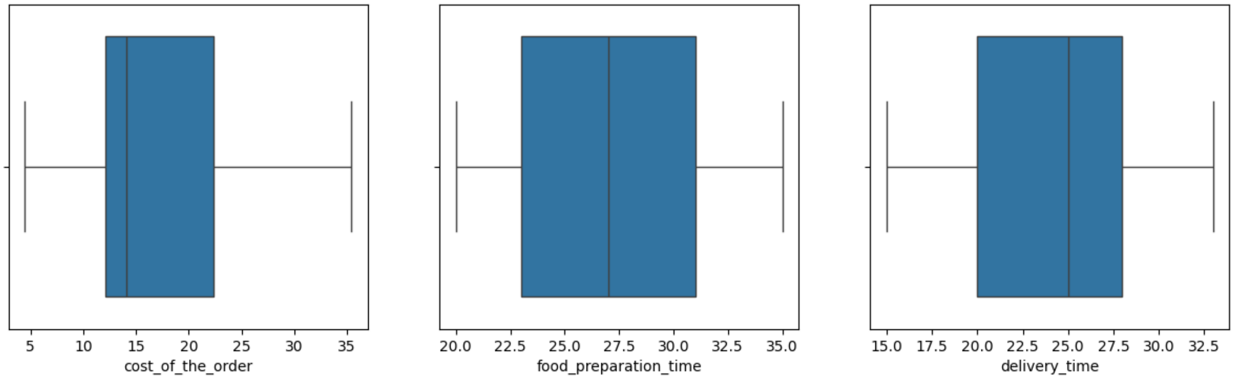
    5.1. There were several entries with non-alphanumeric characters:
    - "Chipotle Mexican Grill $1.99 Delivery",
    - "Joe's Shanghai □_À□ü£¾÷´",
    - "Big Wong Restaurant □_¤¾Ñ¼",
    - CafÌ©China,
    - DespaÌ±a

    - The decision whether to correct these entries should be taken after further investigation, e.g. asking appropriate people from FoodHub: *why these kinds of entries are present (maybe wrong formatting was used during restaurant name input)?*
    - I assumed there was no error and left them as they were.

    5.2. I have also found that there are some restaurants that are closed/archived – *“Empanada Mama (closed)”, “Dirty Bird To Go (archived)”.*

    - Sixteen restaurants, accounting for **0.84%** of the dataset, are no longer available in the market.
    - For safety, further actions should be discussed with the company, as these entries might be present for a specific reason.
    - In this analysis, I have decided to drop these entries, as they are not expected to significantly impact the overall results.

6. I have verified cuisine types
   - The names looked fine.
   - The most popular cuisine was American, the least Vietnamese.
   - There was a big difference in the number of orders between the top 4 cuisines and the rest.
7. I have looked for outliers in the order costs and food preparation/delivery times by plotting a box plot.



- There were no outliers.
- As expected, the *'cost of the order'* tends to be right-skewed.
- Food preparation time looks symmetric.
- Food delivery time is slightly left skewed.

Univariate Analysis:

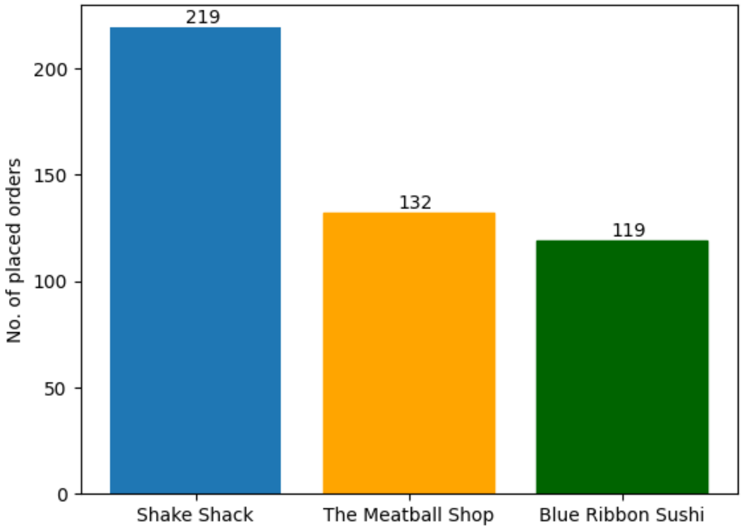
Question 6: Explore all the variables and provide observations on their distributions.

1. I have checked the count of unique values for each column.

- Each value in the *'order_id'* column is unique.
- Some customers have placed orders more than once.

2. The ‘restaurant_name’ column.

- I have checked what are the three most popular restaurants.

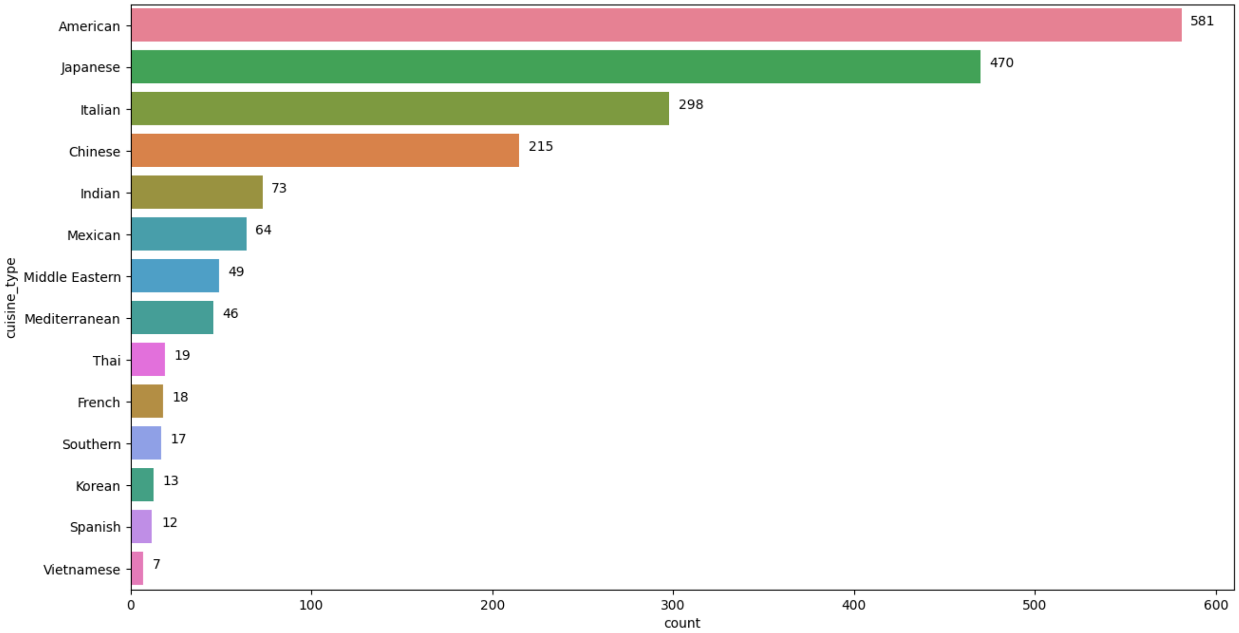


- There are 176 unique restaurants.
- The three most popular restaurants are: ‘Shake Shack’, ‘The Meatball Shop’, and ‘Blue Ribbon Sushi’.
- Nearly 12% of orders are placed at ‘Shake Shack’ restaurant.

3. The ‘cuisine_type’ column.

3.1. Popularity of cuisine types.

- In the data set there were 1882 orders.

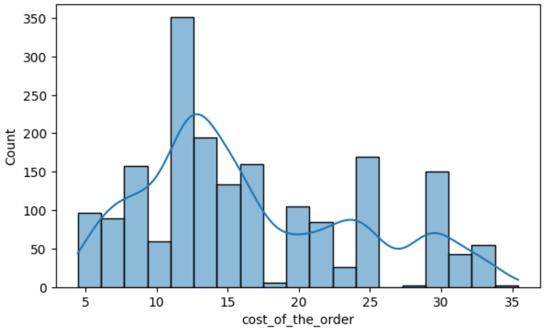


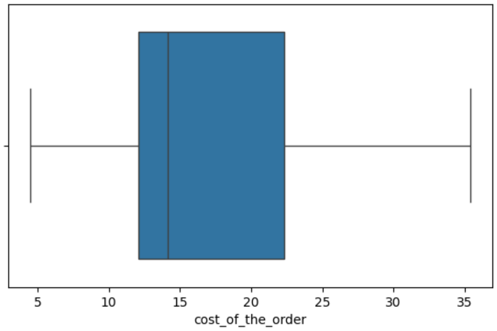
Cuisine name	No. of orders	Percentage of orders
American	581	30.87%
Vietnamese	7	0.37%
Top 4 cuisines	1564	83.1%

- The most popular cuisine is American (30.87%), followed by Japanese, Italian and Chinese.
- Collectively, these four cuisines represent **83%** of all orders.
- Vietnamese cuisine ranks as the least popular, comprising only 0.37% of total orders.

4. The ‘cost_of_the_order’ column.

- Distribution of data.



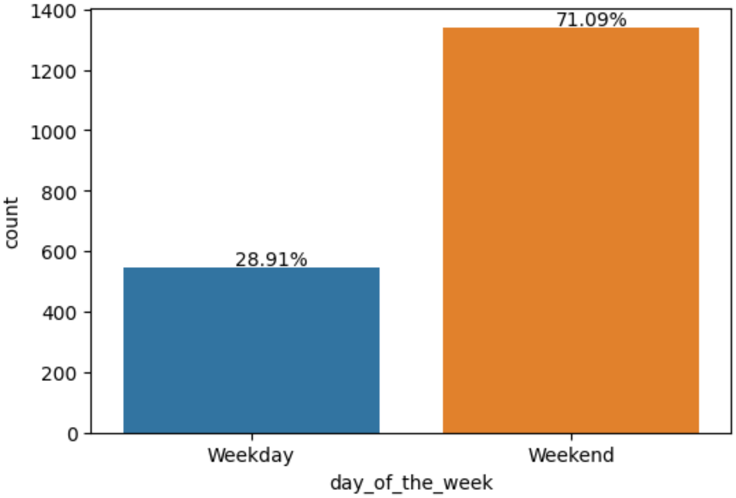


Observations:

- The distribution is right-skewed.
- Cheaper orders seem to be more popular.

5. The ‘day_of_the_week’ column.

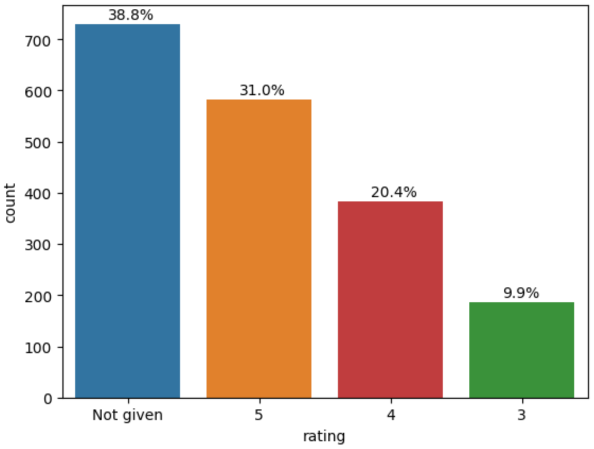
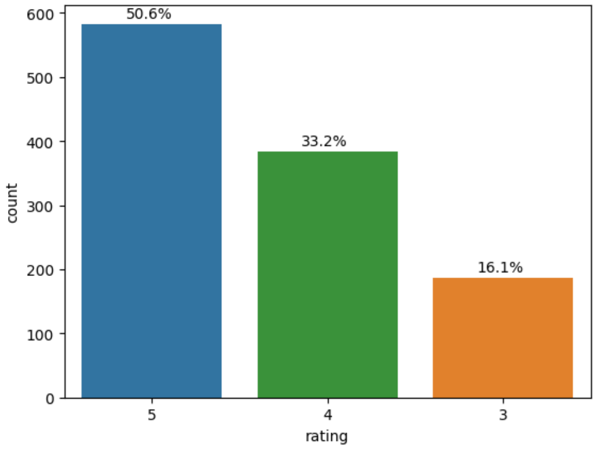
- Orders distribution weekdays vs. weekends.



Observations:

- Approximately every 3rd order out of four is placed on weekends.
- This finding is intriguing, given that weekends span only two days compared to the five weekdays, suggesting potential opportunities for business growth.

6. The *‘rating’* column.

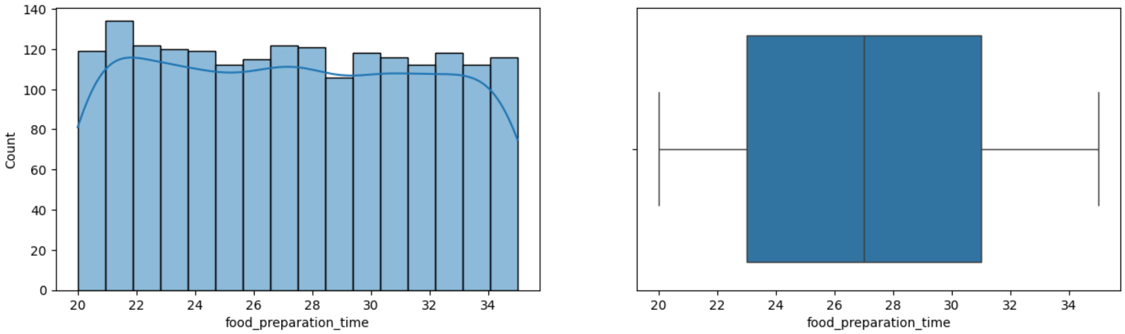
The total distribution of rating marks	Rating distribution within given marks
	

Observations:

- There are no ratings lower than 3.
- The smallest number of orders have a rating of 3 (10%).
- Many orders are not rated – 39%.
- Quite a lot of orders received the highest score (31%).

7. The ‘food_preparation_time’ column.

- Distribution of data.



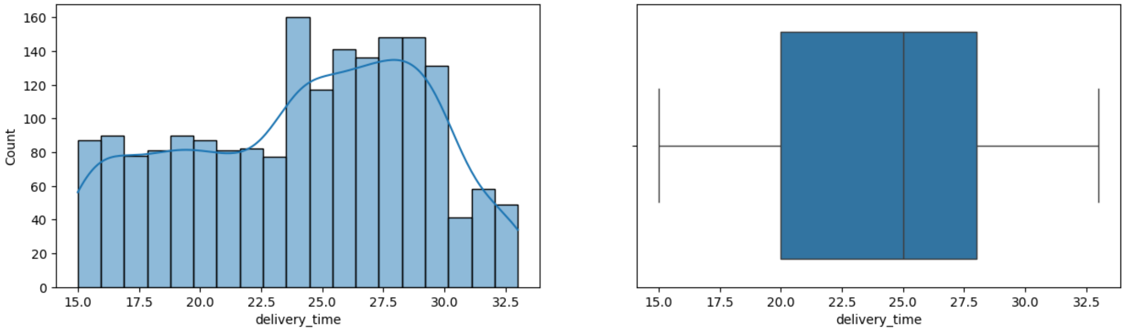
```
count    1882.000000
mean       27.376196
std         4.635327
min        20.000000
25%        23.000000
50%        27.000000
75%        31.000000
max        35.000000
```

Observations:

- The distribution is symmetrical, median and mean values are very close to each other, differences between 25th percentile and min value, as well as between 75th percentile and max value are very close.
- KDE is almost flat at the top – the distribution is uniform.

8. The ‘food_delivery_time’ column.

- Distribution of data.



- Most frequent delivery times.

Delivery time [m]	24	25	26	27	28	29	30
No. of orders	160	117	141	136	148	148	131

Observations:

- The distribution is slightly left-skewed.
- Most frequently deliveries occur within 24 to 30 minutes

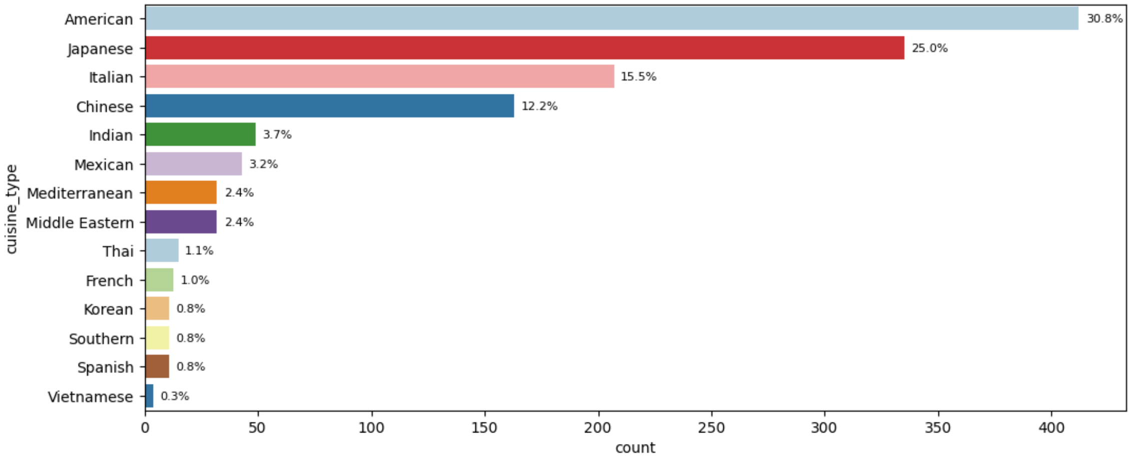
Questions:

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

Restaurant name	No. of orders
Shake Shack	219
The Meatball Shop	132
Blue Ribbon Sushi	119
Blue Ribbon Fried Chicken	96
Parm	68

- The top five restaurants: Shake Shack, The Meatball Shop, Blue Ribbon Sushi, Blue Ribbon Fried Chicken, Parm

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



- The most popular cuisine type on weekends is **American**, with 412 orders (30.8%)

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

- Around 29% of the orders cost more than $20.

Question 10: What is the mean order delivery time?

- The mean order delivery time is around 24min.

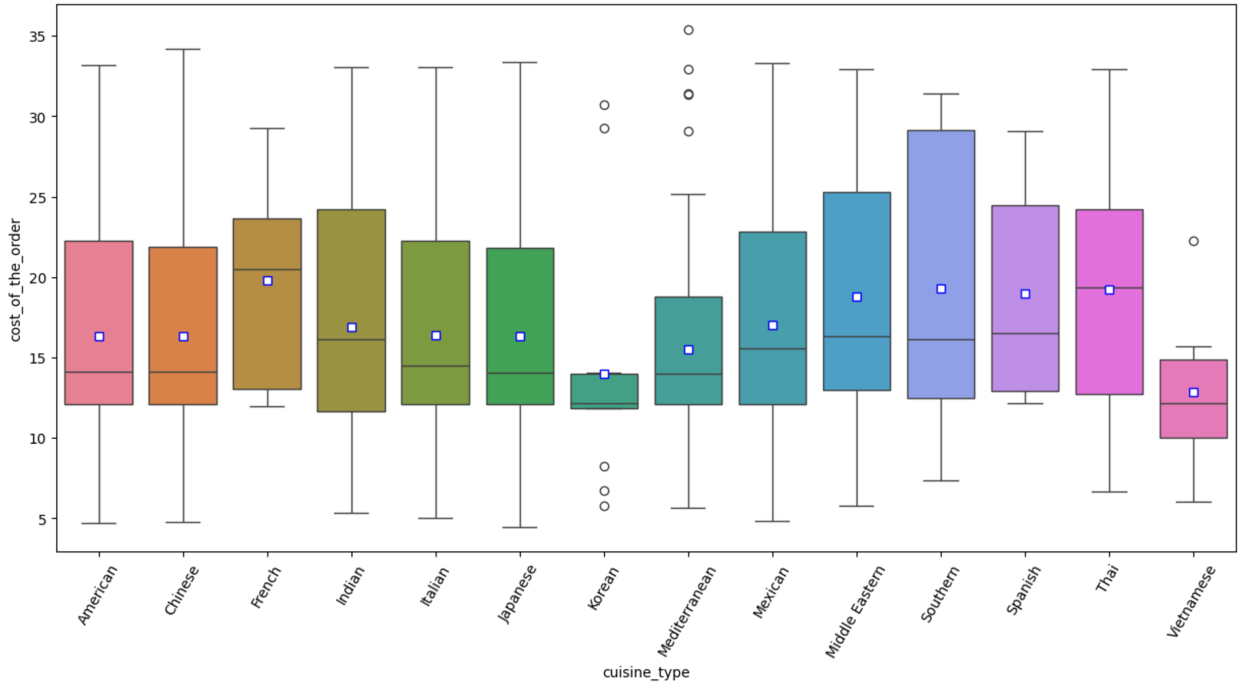
Question 11: The company has decided to give 20% discount vouchers to the top 3 most frequent customers. Find the IDs of these customers and the number of orders they placed

Top 3 customers	No. of orders
52832	13
47440	10
83287	9

- Top 3 customers IDs: 52832, 47440, 83287.
- They placed: 13, 10, and 9 orders, respectively.

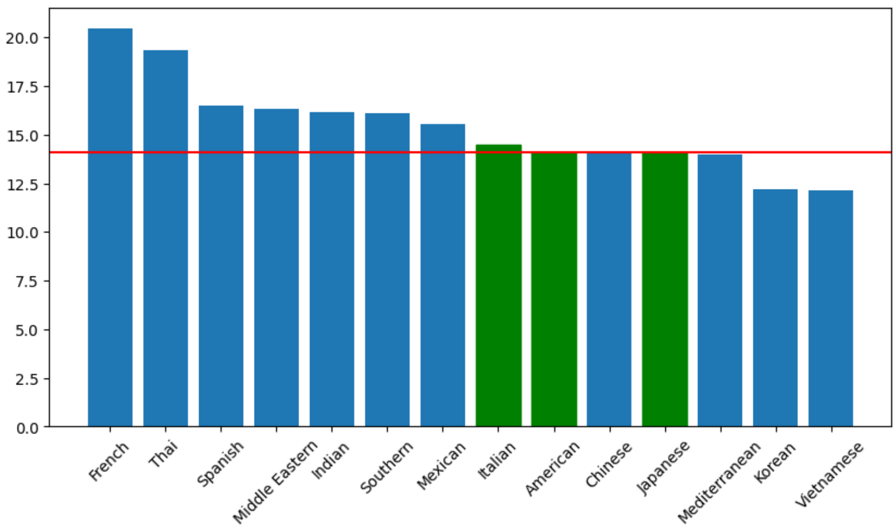
Multivariate Analysis

1. Cuisine type vs. Order cost



1.1. The most expensive cuisine, based on median.

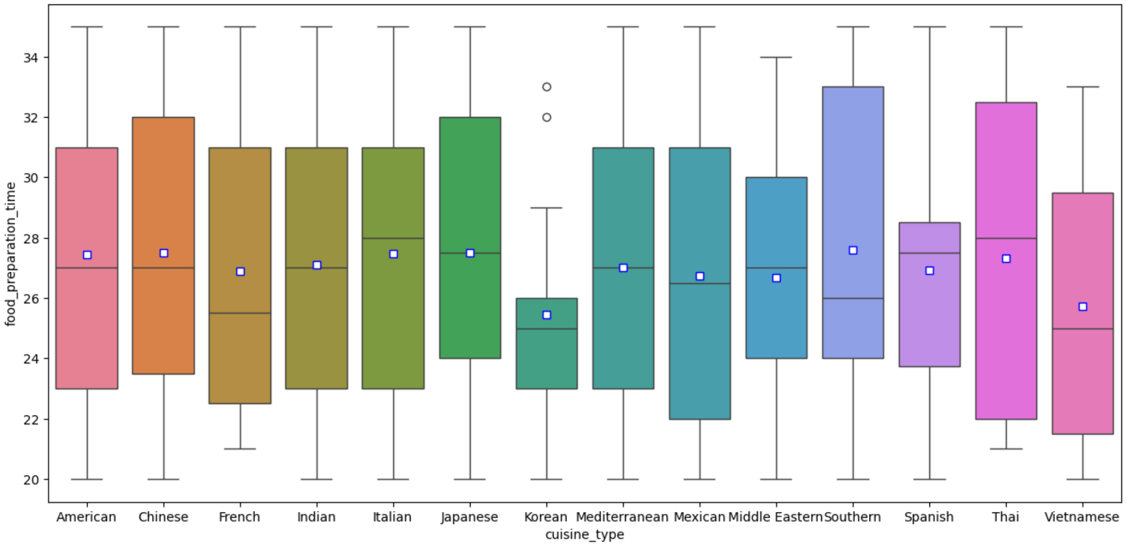
Cuisine type	Cost of the order (median)	Cuisine type	Cost of the order (median)
French	20.470	**Italian**	14.480
Thai	19.350	**American**	14.120
Spanish	16.520	**Chinese**	14.120
Middle Eastern	16.300	**Japanese**	14.070
Indian	16.150	**Mediterranean**	13.995
Southern	16.110	**Korean**	12.180
Mexican	15.570	**Vietnamese**	12.130



Observations:

- The median cost of ordered meals is highest for French cuisine, which ranks as the 10th most popular.
- The median cost of ordered meals is lowest for Vietnamese cuisine, which is also the least popular, suggesting that price may not be the most important factor for customers.
- The three most popular cuisines (American, Japanese, Italian) have median costs around the overall price median.
- The price spread of orders is highest for Mediterranean cuisine and lowest for Spanish cuisine.
- In most cases, the median value is lower than the mean value, indicating that people tend to order cheaper meals.

2. Cuisine type vs. Preparation time.



Observations:

- Considering the IQR, we can say that:
  - There is not much variation in food preparation times across different cuisines.
  - Korean food has the shortest preparation time.

3. Restaurant name vs. Income.

3.1. I have looked in which restaurants the highest orders were placed (top 5).

Restaurant name	Cuisine type	Cost of the order	Rating
Pylos	Mediterranean	35.41	4
Han Dynasty	Mexican	34.19	4
Nobu Next Door	Chinese	33.37	Not given
Blue Ribbon Sushi	Japanese	33.37	3
Tres Carnes	Japanese	33.32	4

3.2. I have also checked which restaurant generated the highest income.

Restaurant name	Cuisine type	Total income	No. of orders
Shake Shack	American	3579.53	219
The Meatball Shop	American / Italian	2145.21	132
Blue Ribbon Sushi	Japanese	1903.95	119
Blue Ribbon Fried Chicken	American	1662.29	96
Parm	Italian	1112.76	68

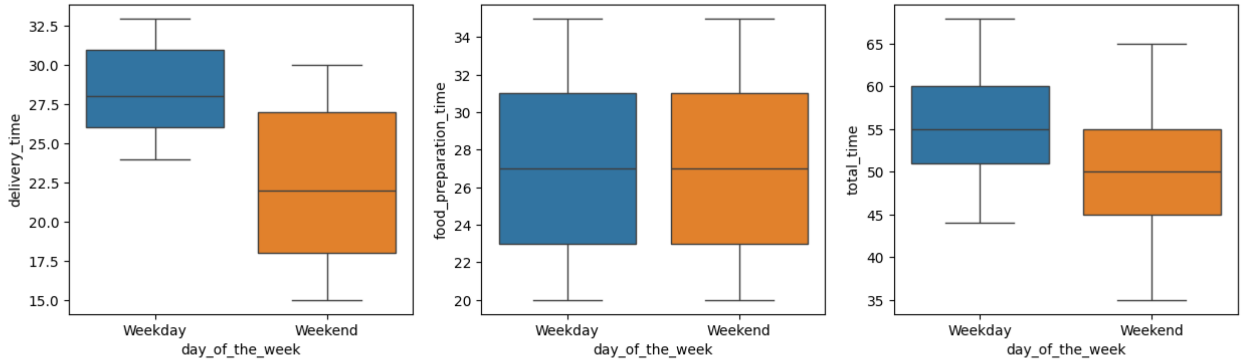
Observations:

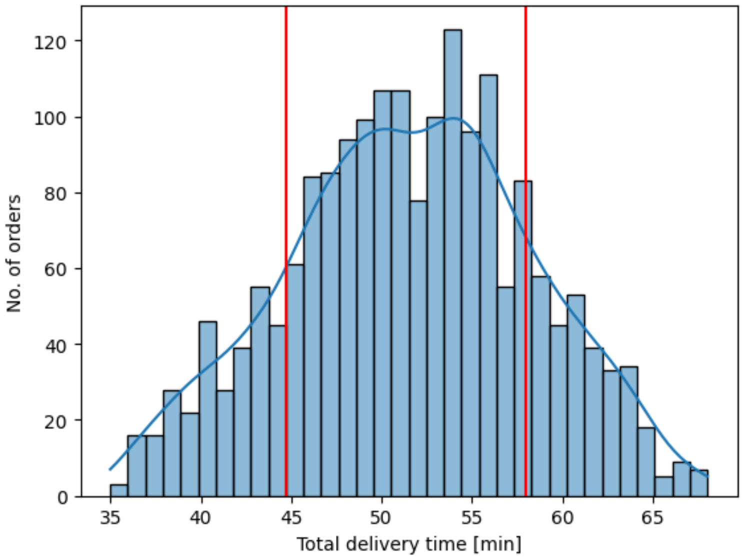
- The most expensive orders come from Mediterranean, Mexican, Chinese, and Japanese cuisines.
- Three out of the top five most expensive orders received a rating of 4.
- The most income is generated by the most popular restaurants, which is expected.

- The most popular restaurants serve American, Japanese and Italian food (3 most popular cuisine types overall).
- Among these, American cuisine is the most popular, followed by Italian and Japanese.

4. Day of the week vs. Delivery time

4.1. I have checked how preparation, delivery and total time depend on the week period.





Observations:

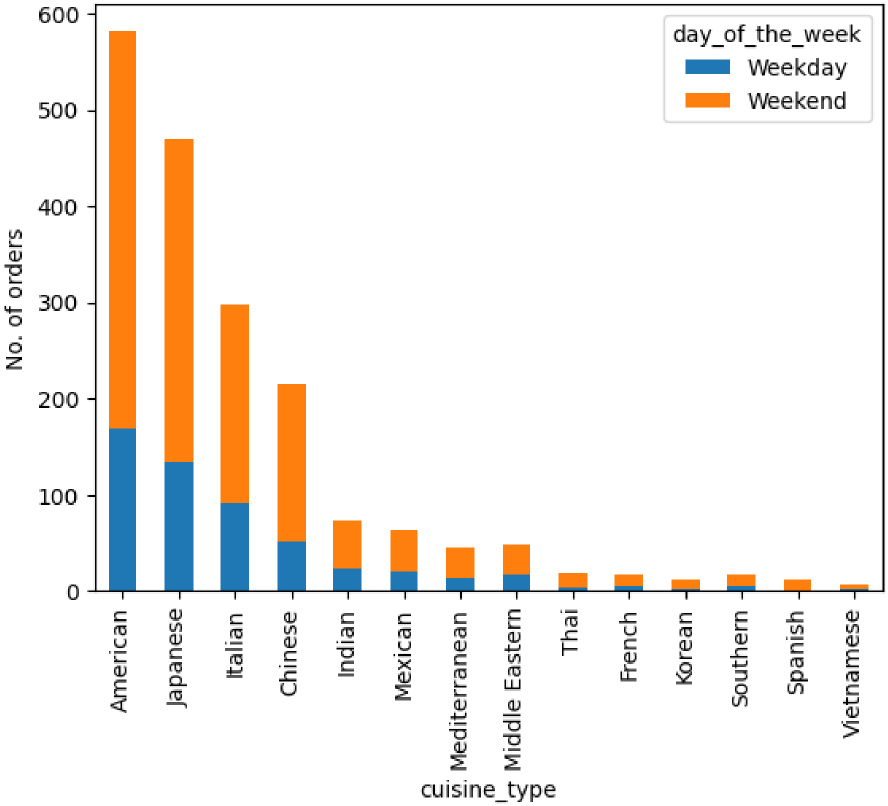
- Deliveries on weekdays tend to take longer time than on weekends.
- There is no difference between food preparation time in relation to the day of the week.
- Approximately 50% of deliveries take between 45 and 58 minutes.
- On weekdays, around 25% of food deliveries take more than 60 minutes, compared to only around 6% on weekends.
- On weekends, about 25% of orders are delivered in less than 45 minutes, while on weekdays, it is less than 1%.
- Therefore, weekend deliveries are generally faster, even though more orders are placed.

5. Day of the week vs. Orders

5.1. How many restaurants serve customers depends on the day of the week.

Weekdays	120
Weekends	157
ratio	76.43%

5.2. Number of orders placed on weekdays and weekends.



5.3. Top 4 most popular cuisines.

Cuisine type	Weekday	Weekend
American	169	412
Japanese	135	335
Italian	91	207
Chinese	52	163
Total no. of orders	544	1338

Observations:

- On weekends, orders are placed at approximately 25% more restaurants compared to weekdays.
- The four most popular cuisines on weekends (American, Japanese, Italian, and Chinese) are also the most popular on weekdays, accounting for 83% of all orders.
- Ordering food is significantly more popular on weekends than on weekdays.
- Certain cuisines, such as Spanish, Korean, and Thai, receive the majority of their orders on weekends.

6. Top 3 customers vs. Restaurants

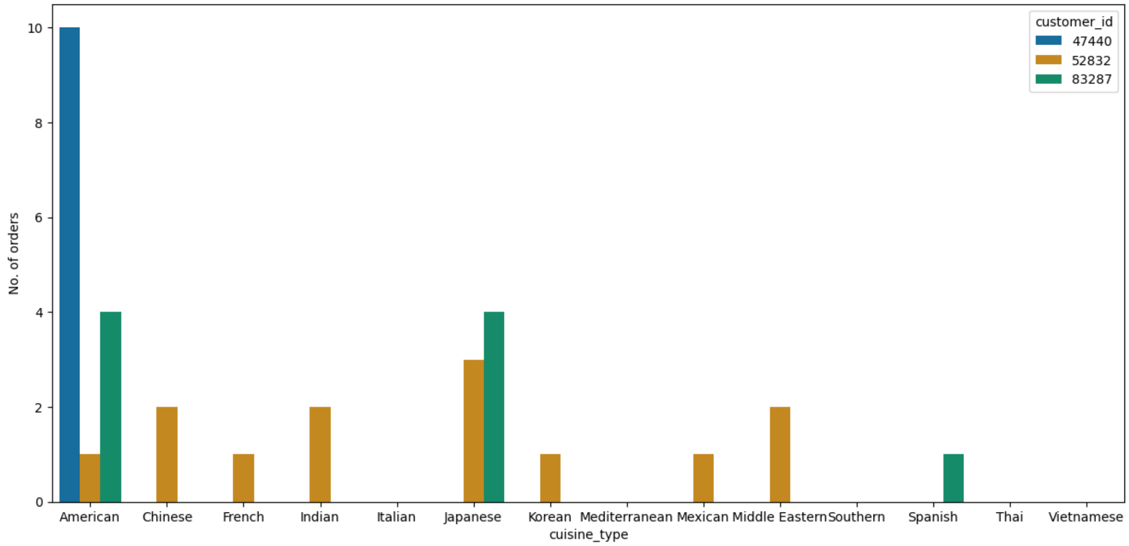
6.1. At first, I found the top 3 customers based on number of orders placed.

Top 3 customers	No. of orders
52832	13
47440	10
83287	9

6.2. Next, for these customers I have checked the prices of their orders.

Customer id	No. of orders	mean	std	min	25%	50%	75%	max
52832	13.0	17.37	8.07	6.64	12.23	15.86	24.2	31.43
47440	10.0	15.82	7.92	6.45	9.59	13.88	22.213	29.30
83287	9.0	15.48	7.29	9.02	9.41	14.5	18.24	29.10

6.3. I have also looked at cuisine types they were ordering.

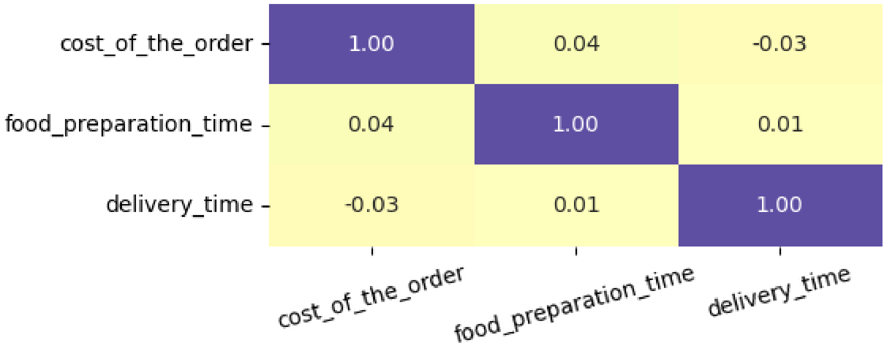


Observations:

- The top three customers placed most their orders in American cuisine.
- In total they ordered meals from eight different cuisines.
- The median order value for each of the customers is above the median for all orders, which is $14.12.

7. Cost of the order vs. Food preparation/delivery time

- I have checked whether the cost of the orders is correlated with any of the steps of the delivery.



Observations:

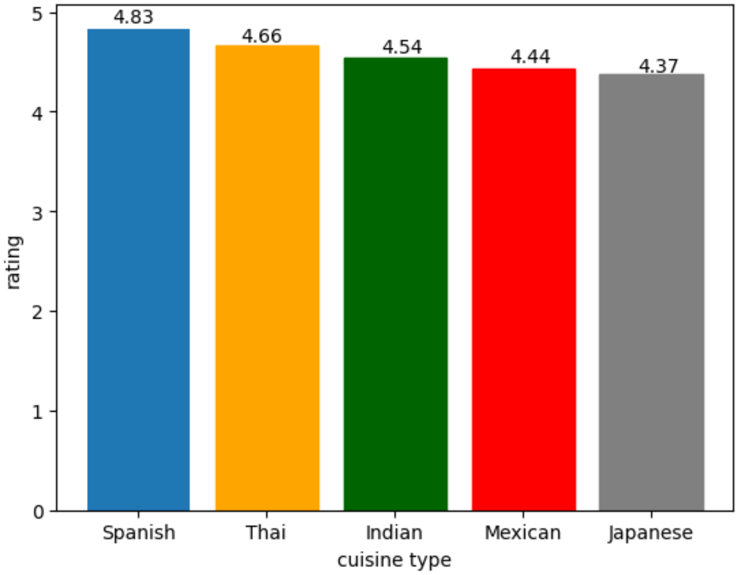
- The cost of the order is not correlated with food preparation time or delivery time.

8. Rating vs. Cuisine type

- I have checked the mean rating values for different types of cuisine.

Cuisine type	Rating	Cuisine type	Rating
Spanish	4.83	**Southern**	4.31
Thai	4.66	**French**	4.30
Indian	4.54	**American**	4.29
Mexican	4.44	**Middle Eastern**	4.23
Japanese	4.37	**Mediterranean**	4.22
Italian	4.36	**Korean**	4.11
Chinese	4.34	**Vietnamese**	4.00

- Top rated cuisines:

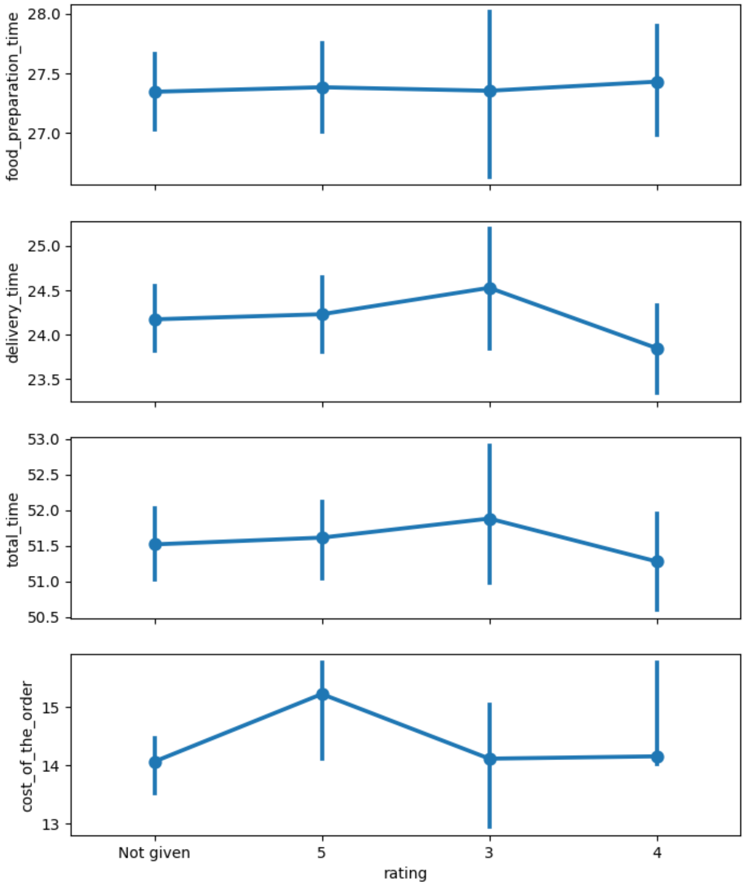


Observations:

- The highest average ratings are for Spanish, Thai, and Indian cuisines, indicating strong customer satisfaction. In contrast, Korean cuisine has the lowest average rating.
- Leveraging the strong ratings of these top cuisines can attract more customers, especially since these cuisines are not among the most popular.

9. Rating vs. Food preparation/delivery time and cost of the order

- I have examined whether rating a restaurant and the score given depend on factors such as food preparation time, food delivery time, total time, and the cost of the order.
- Below are four pointplots illustrating these relationships.
- To account for the right-skewed distribution of order costs, I have utilized median values for the rating versus order cost analysis.



Observations:

- Food preparation time appears to have no significant impact on the rating. This could be because customers are likely evaluating the overall delivery time (preparation + delivery) rather than solely the preparation time. Therefore, an analysis of the overall delivery time was also conducted.

- For ratings of 3, the confidence interval is widest and overlaps with the ranges for other ratings. While this suggests a potential relationship, it indicates that other factors may hold greater importance for customers.
- The highest median cost of orders is associated with a rating of 5. This suggests that cost is not the most important factor for customers, as it is known and accepted at the time of order. Instead, customers may value food quality more, which they reflect in their ratings. For example, if the food quality is worth the price, the rating increases; if not, the rating decreases.
- Similar behavior is observed for orders with a rating of 4. Although the median cost is similar to that of ratings 3 and 'Not given', the confidence interval is skewed upwards, indicating a tendency toward higher prices.

Questions:

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.

Restaurant name	Rating
The Meatball Shop	4.51
Blue Ribbon Fried Chicken	4.33
Shake Shack	4.28
Blue Ribbon Sushi	4.22

Observations:

- Four restaurants meet the criteria of having more than 50 ratings with average grater than4:
  - Blue Ribbon Fried Chicken
  - Blue Ribbon Sushi
  - Shake Shack
  - The Meatball Shop
- Therefore, the promotional offer should target these four restaurants.
- Notable points:
  - *‘Shake Shack’* is also the most popular restaurant (based on previous analysis).
  - *‘The Meatball Shop’* (the highest rated) serves two types of cuisine – Italian and American, from which Italian is preferred (based on previous analysis).

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.

- The net revenue generated by the company across all orders is **$6121.03**.

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]

- The mean delivery time on **Weekends** is **22:27min**
- The mean delivery time on **Weekdays** is **28:21min**
- On average, weekend deliveries are around 21% faster than weekday deliveries.

Conclusions and Recommendations

Conclusions:

- The most popular cuisine is American (30.87%), followed by Japanese, Italian, and Chinese, which together cover 83% of all orders.
- The five most popular restaurants serve American, Japanese, and Chinese food. The most popular restaurant, ‘Shake Shack’, accounts for approximately 11% of all orders.
- Customers tend to order cheaper meals, but price does not significantly impact the given ratings.
- Most orders are placed on weekends (71%), involving around 24% more restaurants than on weekdays.
- In general, orders placed on weekends are delivered faster, with an average reduction in delivery time of around 21%.
- 39% of all orders are not rated. Among the rated orders, approximately 51% received the highest score of 5. No ratings are lower than 3.
- Ratings do not seem to directly depend on delivery or preparation time.
- Since there is no significant relationship between price and ratings, we can assume that customers value other factors more, e.g. the quality of food.

Recommendations:

- The significant difference between weekend and weekday orders suggests potential for business expansion. Introducing special offers, such as happy hours or free delivery on weekdays, could encourage more orders during these times.
- One restaurant, ‘The Meatball Shop’, offers both American and Italian cuisine, with Italian being preferred. As the second most popular restaurant, this suggests potential in partnering with restaurants that offer diverse cuisines.
- The company should develop strategies to encourage more customers to provide ratings. Unrated orders offer no feedback, making it harder to meet customers' needs.
- While overall ratings provide some insight into customer satisfaction, introducing separate categories, e.g. for food quality and delivery time, would offer a more detailed understanding of customer preferences.
- Collecting information about the location of restaurants and more detailed data on the timing of orders would be beneficial. This data could help identify new customer bases, potential restaurant partners, and opportunities for targeted special offers.
- The highest average ratings are for Spanish, Thai, and Indian cuisines. Leveraging the strong ratings of these top cuisines can attract more customers, especially since these cuisines are not among the most popular.