Data-Driven Solutions for Delivery Efficiency and Customer Satisfaction

Data Analysis Project - FoodHub Customer Satisfaction

2024-11-15

Contents / Agenda

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

Executive Summary - Conclusions

Weekend vs. Weekday Patterns:

- There are significantly more orders on weekends than on weekdays. This suggests that customers are more likely to order food during the weekend, possibly due to having more leisure time, for example social gatherings and events or watching movies, sports events, or other at-home entertainment, where ordering food complements the experience
- Delivery times are generally faster on weekends. Identifying peak times and optimizing delivery logistics can enhance efficiency.
- The average order cost is marginally higher on weekends, which could be attributed to customers ordering larger or more expensive meals.
- Targeting high-value orders on weekends with focused marketing can boost revenue.

Customer Ratings and Delivery Time:

- Faster and more consistent delivery times generally align with higher ratings. However, the difference in delivery times across ratings (3, 4, and 5) is relatively small with very low correlation. This indicates that while faster deliveries are appreciated, consistency and reliability or other factors outside of this dataset might be more valued than speed alone.
- After imputing missing values, we observed that ratings of 5 were still the most common, suggesting overall customer satisfaction with the service.

Ratings and Order Cost:

- Orders with higher ratings (4 and 5) also had slightly higher average costs, indicating that customers who spend more may have higher expectations, but also tend to be more satisfied. However the data show low correlation between cost and rating.
- While higher costs are somewhat correlated with higher ratings, the differences aren't large enough to conclude that spending more guarantees a better rating.

Popular Cuisines:

- American and Japanese cuisines were among the most ordered types, particularly on weekends. This insight could guide promotional efforts for these cuisines during peak times.

Executive Summary - Recommendations

Promotions:

- Since weekends have more orders and higher ratings, the business could promote higher-priced or premium meals, perhaps bundling items or creating special weekend-only deals to maximize revenue. For example; Run targeted promotions or loyalty programs for customers who frequently place orders above $20 to enhance high-margin revenue.
- Offering weekday promotions or loyalty incentives may help balance order volume and drive more weekday business.
- Higher ratings tend to have a slightly higher costs, consider implementing a loyalty program for customers who consistently rate their experience highly. For example, offer a discount or reward after a certain number of high-rated purchases. This could incentivize repeat business and encourage customers to invest in higher-value items that tend to receive better ratings.

Gather and Act on Customer Feedback:

- Given that ratings seem to align with consistent delivery times, maintaining reliable delivery times may be more beneficial than pushing for absolute speed, focusing on consistency over speed. Customer feedback could help refine this approach, identifying areas to improve consistency without sacrificing quality.
- To capture more feedback, the business could incentivize customers to rate their orders, especially on weekdays. This will provide more data to better understand weekday vs. weekend experiences. For example encourage restaurants to boost their rating counts by providing incentives for customer reviews

Enhance Advertising Strategy:

- Advertise weekends as a relaxed, high-quality experience with special meals and emphasize efficient, quick weekday service. Tailoring the messaging could attract both leisurely weekend diners and busy weekday customers looking for convenience.

Popular Cuisine Promotions:

- With American and Japanese cuisines being the most popular, these can be promoted more prominently, particularly on weekends. Offering special deals on these cuisines could increase customer engagement and drive higher ratings.

Final Summary

This analysis highlights that FoodHub excels on weekends, with faster delivery times, high customer satisfaction, and increased order volume. To capitalize on this momentum, FoodHub could implement targeted upselling strategies, such as combo offers or premium meal promotions, specifically on weekends. This approach would encourage customers to add higher-margin items to their orders, maximizing revenue per transaction.

On weekdays, there is an opportunity to further streamline operations to increase order volume and enhance delivery consistency. By offering weekday-specific discounts, targeted promotions, or loyalty incentives, FoodHub can make weekdays more appealing to a broader customer base. Ensuring reliable and consistent delivery times across all days will help maintain high satisfaction levels and foster long-term customer loyalty.

Moreover, this analysis shows that factors such as food preparation time, delivery time, and order cost have minimal direct correlation with customer ratings, indicating that other external factors might play a significant role in customer satisfaction. Potential influences may include customer service quality or for example food packaging standards. These areas present opportunities for further investigation, as improvements here could positively impact the overall customer experience.

By leveraging high weekend demand and strategically enhancing weekday engagement, FoodHub can not only drive customer satisfaction but also increase revenue and customer retention throughout the week.

Business Problem Overview and Solution Approach

Context:

FoodHub connects busy New Yorkers to multiple restaurants via a single app, managing orders and delivery through dedicated drivers and earning a fixed margin per transaction.

Defining the Problem:

To stay competitive in this dynamic environment, FoodHub wants to improve customer satisfaction and operational efficiency across its delivery services. Key metrics that affect both the business and customer experience—such as delivery time, consistency, order cost, and customer ratings—show variability and reveal potential areas for enhancement. Understanding the patterns and factors influencing customer satisfaction, especially in terms of delivery and cost factors on weekdays vs. weekends, will be critical to optimizing service quality.

The company specifically aims to:

- Gain insights into demand for different restaurants and cuisines to align with customer preferences.
- Assess the influence of delivery time, cost, and ratings on customer satisfaction.
- Identify opportunities to improve operational efficiency, focusing on trends across weekdays and weekends.

Business Problem Overview and Solution Approach

Solution Approach / Methodology

- **Data Analysis Framework**:
  - Conduct an **Exploratory Data Analysis (EDA)** to identify patterns, trends, and anomalies in key variables such as **delivery time**, **order cost**, and **ratings** across different time periods.
  - Compare metrics such as **delivery and food preparation times on weekdays vs. weekends** to identify potential bottlenecks and efficiency gains.
  - Assess **correlation between delivery time and customer ratings** to identify areas for improvement in the customer experience.
- **Methodology**:
  - **Data Cleaning**: Address missing values and standardize data formats (e.g., handling "Not Given" ratings).
  - **Univariate & Bivariate Analysis**: Explore individual variables and relationships between variables to uncover trends.
  - **Visualization**: Use charts (e.g., box plots, bar charts) to make insights clear and actionable.
  - **Recommendations**: Based on insights, develop specific, data-backed recommendations for operational improvements.

Data Overview

Dataset Dimensions:

- The dataset consists of **1898 rows** and **9 columns**.
- This gives a substantial amount of data to analyze customer behavior, delivery, and food preparation times.

Data Types of Columns:

- The dataset includes columns with data types numerical and categorical

- 1 datatype of float(64)
- 4 datatype of int64
- 4 datatype of object
- This means there are 5 numerical columns and 4 object type columns

---

- There are no missing values
- Observation: rating is read as object type column (should be numerical)

.

```
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
```

.

Data Overview

Missing Values:

- No missing values per se in the datasets columns since the value 'Not given' in the variable 'rating' was read as a string, making the column a object.
- After converting 'Not given' as 'NaN', there are now 736 missing values in 'rating'. Approximately 39% of the customers don't leave a rating, which is considered a high number.
- The 'rating' has the values: nan, 5, 4 and 3.

	0
order_id	0
customer_id	0
restaurant_name	0
cuisine_type	0
cost_of_the_order	0
day_of_the_week	0
rating	736
food_preparation_time	0
delivery_time	0

Data Overview - Statistical summary

	count	mean	std	min	25%	50%	75%	max
order_id	1898.000	1477495.500	548.050	1476547.000	1477021.250	1477495.500	1477969.750	1478444.000
customer_id	1898.000	171168.478	113698.140	1311.000	77787.750	128600.000	270525.000	405334.000
cost_of_the_order	1898.000	16.499	7.484	4.470	12.080	14.140	22.297	35.410
rating	1162.000	4.344	0.741	3.000	4.000	5.000	5.000	5.000
food_preparation_time	1898.000	27.372	4.632	20.000	23.000	27.000	31.000	35.000
delivery_time	1898.000	24.162	4.973	15.000	20.000	25.000	28.000	33.000

Description of the data frame

- **'order_id'** is currently treated as numerical, but it would be more appropriate as a categorical identifier, as it uniquely distinguishes each order without any inherent quantitative value. For this analysis, it will remain as is, since it doesn't provide significant insight at this stage.
- Similarly, **'customer_id'** is a unique identifier and would be better treated as categorical rather than numerical, as it doesn't hold numerical meaning. Analyzing customer behavior patterns, such as examining correlations between customer ID and ratings or order costs, **could** offer insights into customer satisfaction trends and help identify loyal customers or those who might need targeted improvements. This is however not considered a priority at this point.

Data Overview - Statistical summary

- **Cost of the order:** The average cost of an order is 16,5 dollars with a standard deviation of 7.5. This means that, on average, the cost of orders varies by about 7.5 dollars from the mean, suggesting there is a reasonable spread in the data, indicating variability in customer spending. The minimum cost for an order is 4.47 dollars while 25% of the orders are under 12 dollars. The median, at 50 percentile order is 14 dollars, and 75% of orders cost up to 22.3 dollars. The maximum order cost is 35.4 dollars. This distribution shows that most orders are clustered in the range between 12 and 22, with a few higher-priced orders pulling up the maximum value. The spread in the data suggests that while the majority of customers tend to place orders within a middle range (around 12 to 22 dollars), there is a notable amount of higher-value orders that are likely increasing the average. The high maximum could be due to bulk or larger family orders, more expensive restaurants, or add-ons in the order. It could be interesting to investigate whether certain restaurants or cuisine types contribute more to these higher-value orders, or whether weekend orders tend to have a higher average cost.

	count	mean	std	min	25%	50%	75%	max
order_id	1898.000	1477495.500	548.050	1476547.000	1477021.250	1477495.500	1477969.750	1478444.000
customer_id	1898.000	171168.478	113698.140	1311.000	77787.750	128600.000	270525.000	405334.000
cost_of_the_order	1898.000	16.499	7.484	4.470	12.080	14.140	22.297	35.410
rating	1162.000	4.344	0.741	3.000	4.000	5.000	5.000	5.000
food_preparation_time	1898.000	27.372	4.632	20.000	23.000	27.000	31.000	35.000
delivery_time	1898.000	24.162	4.973	15.000	20.000	25.000	28.000	33.000

Data Overview - Statistical summary

	count	mean	std	min	25%	50%	75%	max
order_id	1898.000	1477495.500	548.050	1476547.000	1477021.250	1477495.500	1477969.750	1478444.000
customer_id	1898.000	171168.478	113698.140	1311.000	77787.750	128600.000	270525.000	405334.000
cost_of_the_order	1898.000	16.499	7.484	4.470	12.080	14.140	22.297	35.410
rating	1162.000	4.344	0.741	3.000	4.000	5.000	5.000	5.000
food_preparation_time	1898.000	27.372	4.632	20.000	23.000	27.000	31.000	35.000
delivery_time	1898.000	24.162	4.973	15.000	20.000	25.000	28.000	33.000

- **Rating:** There are 736 missing values, approximate 36% missing values in this column. The average is 4,3, min 3 and max 5. At 25 percentile a 4, at 50 percentile 5, and at 75 percentile is 5. At first glance one can tell that this indicates that most customers give high ratings, and the distribution of ratings is relatively narrow, as the standard deviation is only 0.741. The minimum rating is 3, which suggests that overall, customers are quite satisfied, as there are no very low ratings in the data. This means that the distribution of ratings is centered around higher values, and few customers give ratings below 4. When analyzing the missing ratings (39% of the rows), it may be worthwhile to check if the missing ratings are associated with specific restaurants, times, or days of the week. This can provide insights into why some orders are missing ratings and help decide how to best handle these missing values.
- A deeper dive into ratings will be covered in future slides

Data Overview - Statistical summary

	count	mean	std	min	25%	50%	75%	max
order_id	1898.000	1477495.500	548.050	1476547.000	1477021.250	1477495.500	1477969.750	1478444.000
customer_id	1898.000	171168.478	113698.140	1311.000	77787.750	128600.000	270525.000	405334.000
cost_of_the_order	1898.000	16.499	7.484	4.470	12.080	14.140	22.297	35.410
rating	1162.000	4.344	0.741	3.000	4.000	5.000	5.000	5.000
food_preparation_time	1898.000	27.372	4.632	20.000	23.000	27.000	31.000	35.000
delivery_time	1898.000	24.162	4.973	15.000	20.000	25.000	28.000	33.000

Description of the data frame

- **Food preparation time:** The average time it takes for food to be prepared is 27.4 minutes, with a relatively small standard deviation of 4.63. The minimum preparation time is 20 minutes, which makes sense considering food cannot be prepared instantaneously. At the 25th percentile, food is prepared in 23 minutes, at the 50th percentile it takes 27 minutes, and at the 75th percentile, 31 minutes. The maximum recorded preparation time is 35 minutes. Overall, the food preparation time seems to be quite consistent with little variation, which suggests that most restaurants follow a standard time range for preparation.

- **Delivery time:** The average time for food delivery is 24.16 minutes, with a slightly larger standard deviation of 4.97 minutes compared to the preparation time. The minimum delivery time is 15 minutes, and the maximum is 33 minutes. At the 25th percentile, deliveries are completed in 20 minutes, the median time is 25 minutes, and at the 75th percentile, it's 28 minutes. The delivery times show a little more variability than the preparation times, likely due to external factors like traffic, distance, or delivery efficiency. However, the variation is still within a reasonable range, and there don't appear to be any extreme outliers in the data.

Data Overview - Clarification ‘rating’

A deeper dive is considered necessary regarding ‘rating’ and a few points may need clarification. The dataset only includes ratings of 3, 4, and 5, which suggests that either:

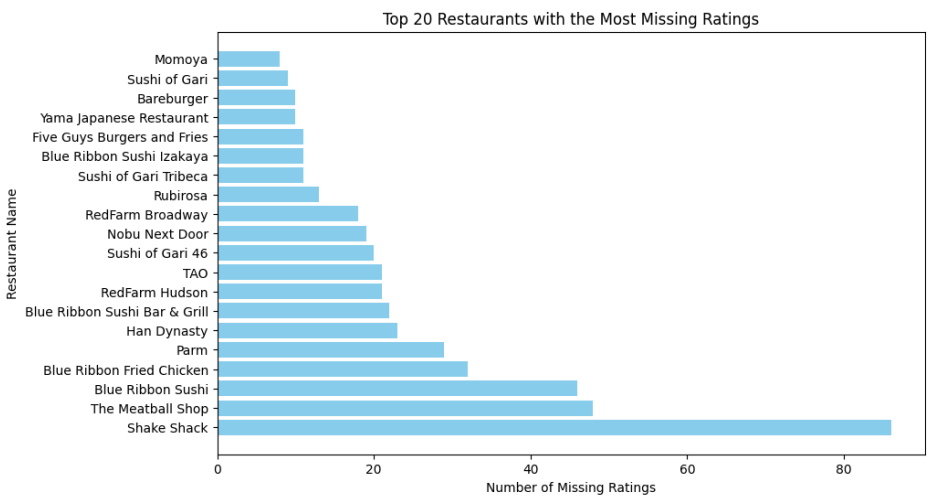
	count	mean	std	min	25%	50%	75%	max
order_id	1898.000	1477495.500	548.050	1476547.000	1477021.250	1477495.500	1477969.750	1478444.000
customer_id	1898.000	171168.478	113698.140	1311.000	77787.750	128600.000	270525.000	405334.000
cost_of_the_order	1898.000	16.499	7.484	4.470	12.080	14.140	22.297	35.410
rating	1162.000	4.344	0.741	3.000	4.000	5.000	5.000	5.000
food_preparation_time	1898.000	27.372	4.632	20.000	23.000	27.000	31.000	35.000
delivery_time	1898.000	24.162	4.973	15.000	20.000	25.000	28.000	33.000

- **Sample dataset** The dataset could possibly be a sample of the whole dataset. This particular data might only include higher ratings if the platform actively filters or excludes lower ratings from the dataset.
- **Ratings below 3** (1 or 2) were not available for customers to select. The narrow range with no ratings below 3 indeed indicates a high level of customer satisfaction overall. However, it's also possible that a limited range (3-5) restricts nuanced feedback. If customers can only give a minimum of 3, this would prevent a more complete understanding of truly negative experiences.
- **Missing Ratings:** Since approximately 39% of the data has missing ratings, exploring patterns in these missing values is valuable. Identifying any trends—such as certain restaurants, peak times, or specific days—associated with missing ratings could reveal areas where the company may need to prompt for ratings more effectively. For example, certain times of the day or specific restaurants might correlate with fewer customer ratings, which could indicate less engaged customers or other factors influencing feedback collection.

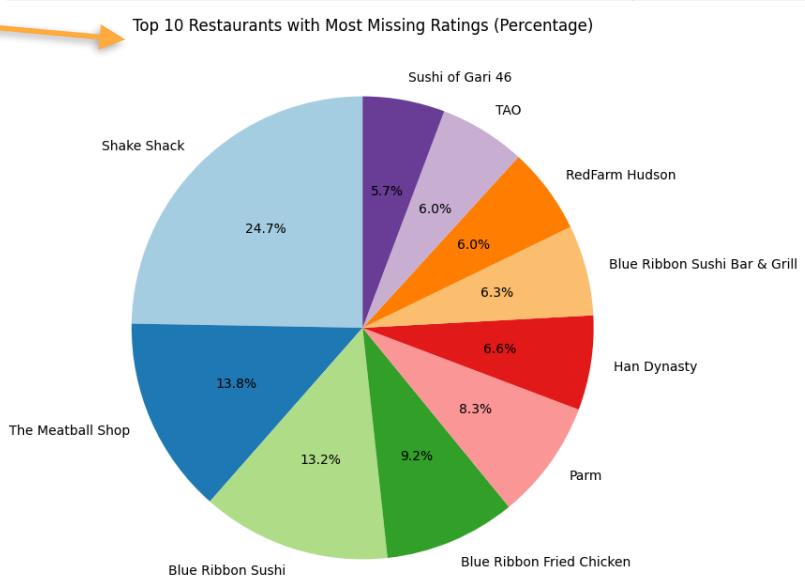
Overall, this analysis suggests that customers are largely satisfied, but the limited range and missing values provide opportunities for improvement in understanding customer experience more completely. For this analysis it I plan to proceed with the assumption that customers had the option to rate from 1 to 5, but the dataset only contains ratings from 3 to 5, a sample dataset. Further investigation could help confirm whether ratings below 3 were possible but simply underrepresented, or if they were excluded or filtered out for some reason.

Data Overview - missing ‘rating’

- Overview for restaurants with most missing values, top 20 and top 10



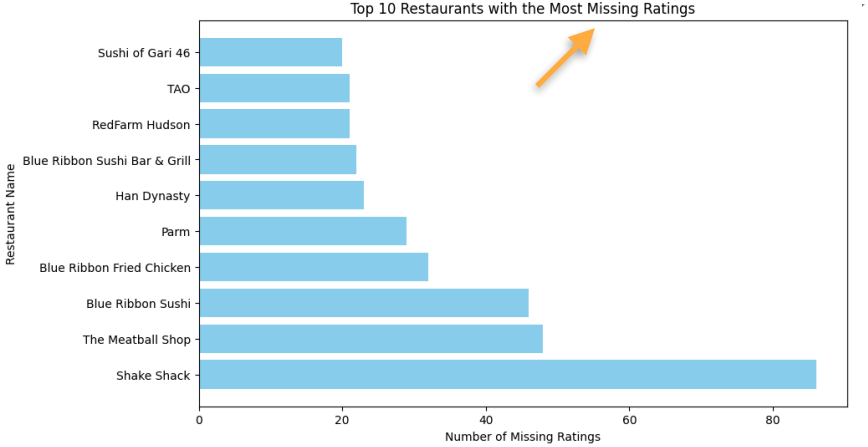
Shake Schack accounts for 11.68% of all missing ratings

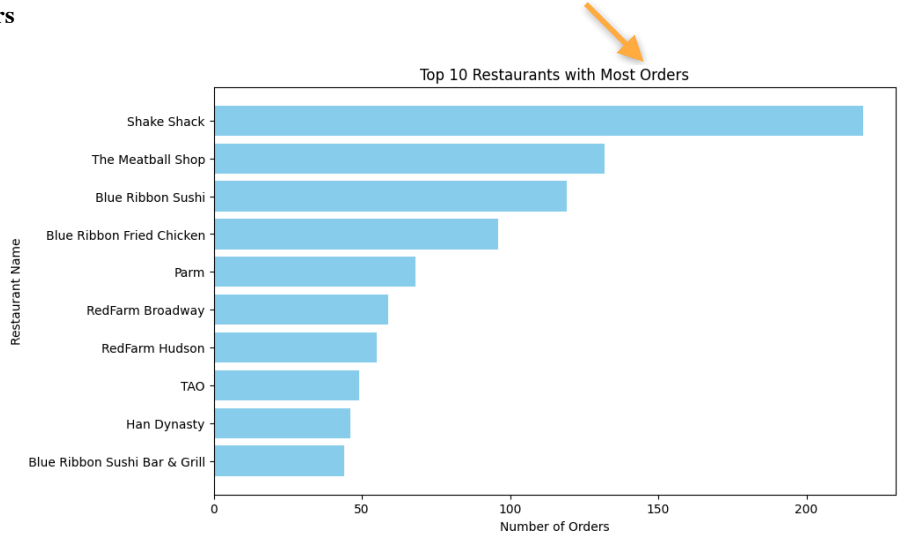


Piechart showing which top 10restaurants have most missing ratings

Data Overview - missing ‘rating’

- Narrowing down to top 10 **missing ratings** vs top 10 **restaurants with most orders**





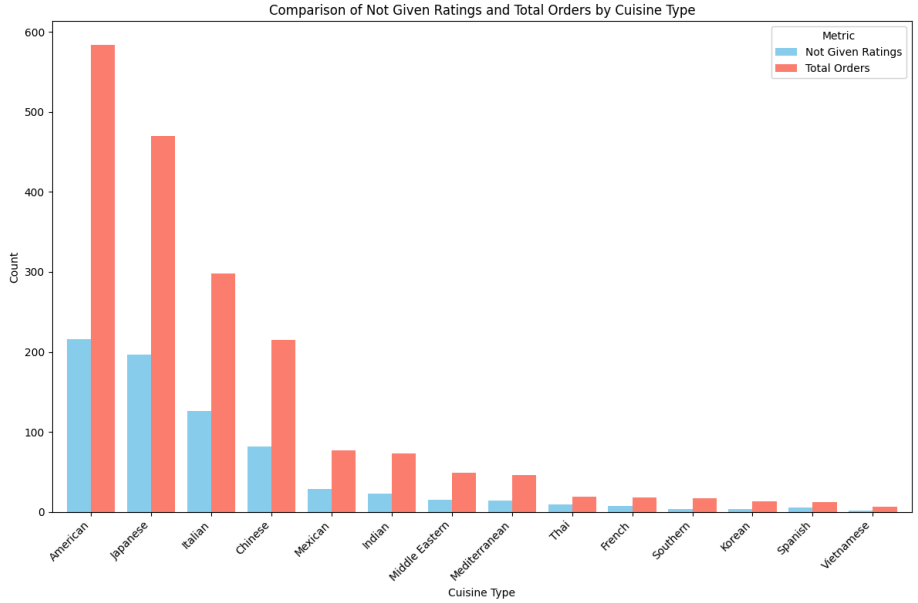
The similarity in patterns between the most frequently ordered restaurants and those with the highest number of missing ratings suggests that if these missing ratings were available, they might reflect the general trend seen in rated orders.

In other words, these restaurants are likely to have ratings similar to those of other high-traffic restaurants, which generally receive high ratings. This assumption provides a basis for treating these missing values as likely positive, potentially indicating customer satisfaction even in the absence of explicit ratings.

Why/How to imputate the missing values

Previous slides showed patterns between the restaurants with the most missing ratings and those with the highest order frequency. This graph, however, show the distribution of 'Not Given' ratings and total orders by cuisine type.

Since the missing ratings more or less follow the same pattern as the total orders by cuisine type, we can assume that customer satisfaction for the missing ratings likely reflects the same trends as the provided ratings. This suggests that higher order volumes in certain cuisines may simply result in some customers not leaving a rating, without it necessarily indicating dissatisfaction.



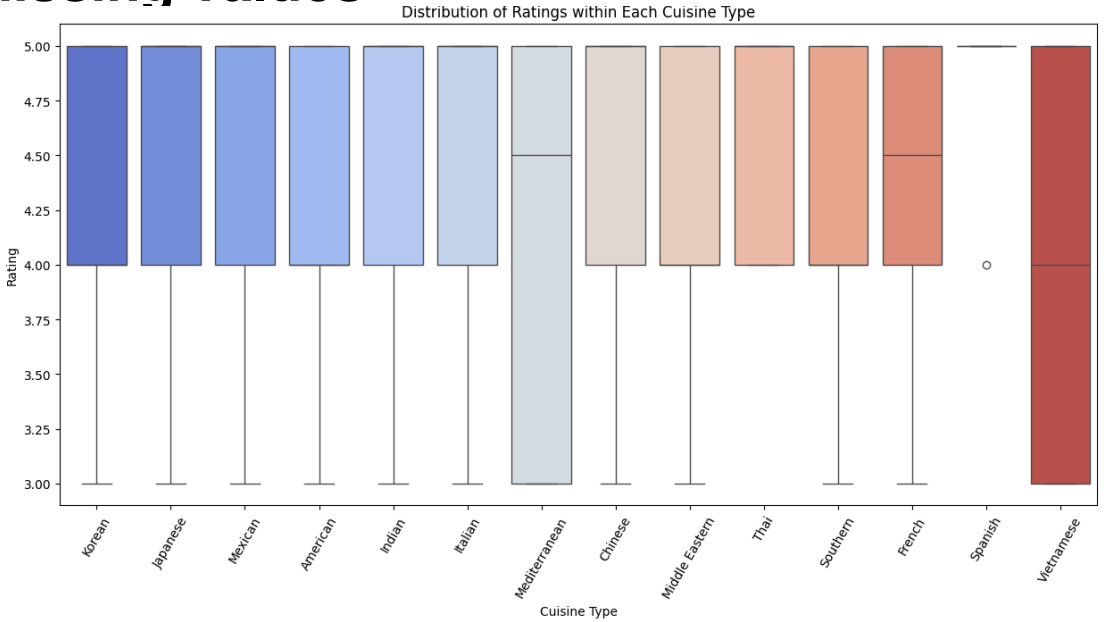
	cuisine_type	not_given_count
0	American	216
1	Japanese	197
2	Italian	126
3	Chinese	82
4	Mexican	29
5	Indian	23
6	Middle Eastern	15
7	Mediterranean	14
8	Thai	10
9	French	8
10	Spanish	6
11	Korean	4
12	Southern	4
13	Vietnamese	2

Why/How to imputate the missing values

When looking into the distribution of ratings within each cuisine type, the analysis reveals insights.

Analysis of Ratings Distribution Across Cuisines:

- **Outliers and Skewness:** The Spanish cuisine shows an outlier, suggesting a distinct rating that deviates from the majority.
- **Consistency and Clustering:** Vietnamese and mediterranean cuisines show tight clustering in ratings, as indicated by the lack of whiskers. This suggests consistent customer satisfaction for these cuisines. In contrast, cuisines with longer whiskers, like almost all other cuisines, show more variability, reflecting a wider range of customer feedback. The box plot for cuisines with a visible lower whisker, indicates that there is some variability in ratings, though it's primarily on the lower end. This lower whisker suggests that some customers rated these cuisine slightly lower than the median range.
- **High Ratings Overall:** 5 cuisines have median 4, 2 cuisines have median 4.5 and 7 cuisines have median 5.




```
Median Ratings for Each Cuisine Type:
cuisine_type
Chinese           5.000
Indian            5.000
Italian           5.000
Japanese          5.000
Mexican           5.000
Spanish           5.000
Thai              5.000
French            4.500
Mediterranean     4.500
American          4.000
Korean            4.000
Middle Eastern    4.000
Southern          4.000
Vietnamese        4.000
Name: rating, dtype: float64
```

Chosen Imputation Strategy

Chosen strategy - Choice of mean or median:

- For cuisines with more skewed distributions or outliers, using the median is preferable, as it avoids the influence of extreme values that could misrepresent the typical experience for that cuisine. **Imputation using median for missing ratings in Spanish cuisine**
- For **cuisines with consistent ratings** (tight distributions), the **mean is appropriate** as it accurately reflects the general customer sentiment without distortion. Imputation using mean for missing ratings in the rest of the cuisines, in each cuisines separately

```
[31] # Define the function to impute missing ratings using the mean for specified cuisines
     def impute_rating(x):
         # List of cuisines to impute using mean
         cuisines_to_impute_with_mean = [
             'American', 'Chinese', 'Indian', 'Korean',
             'Mexican', 'Middle Eastern', 'Southern',
             'Mediterranean', 'French', 'Italian',
             'Japanese', 'Vietnamese', 'Thai'
         ]

         # If the cuisine is in the specified list, fill missing values with the mean
         if x.name in cuisines_to_impute_with_mean:
             return x.fillna(x.mean())
         return x

     # Apply the function to impute missing values for the specified cuisines
     df['rating'] = df.groupby('cuisine_type')['rating'].transform(impute_rating)
```

```
[25] # Impute missing ratings in the 'rating' column for Spanish cuisine
     # using the median rating for Spanish cuisine, ensuring index alignment.
     def impute_spanish_rating(x):
         if x.name == 'Spanish':
             return x.fillna(x.median())
         return x

     df['rating'] = df.groupby('cuisine_type')['rating'].transform(impute_spanish_rating)
```

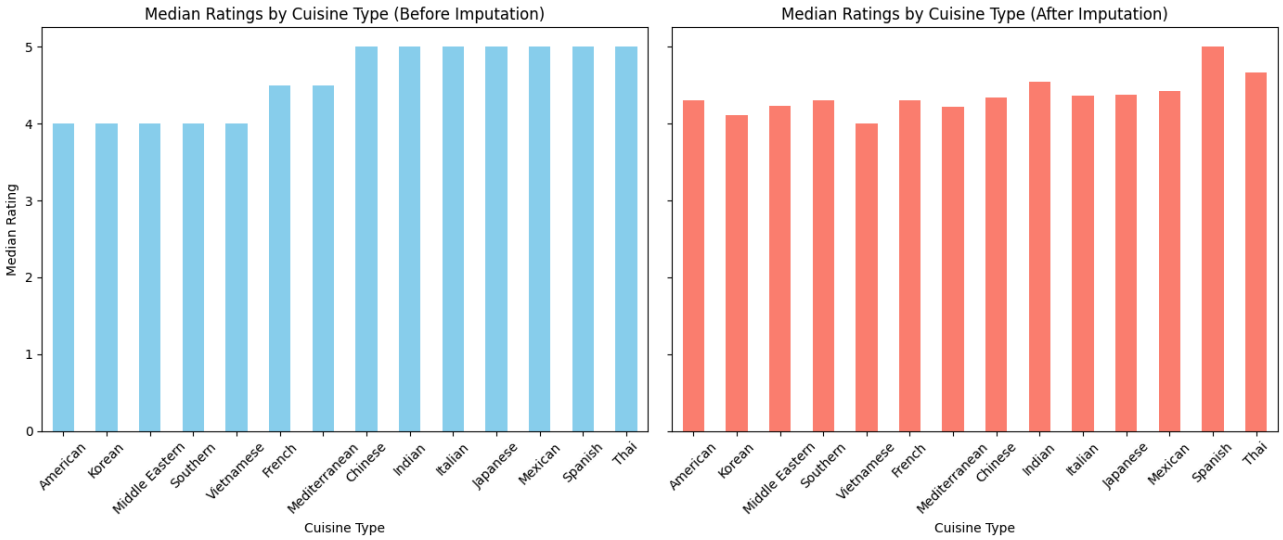
The median rating before and after imputation

The plots show that after imputation, the median ratings across cuisines got more variety. This imputation likely provides a more accurate reflection of customer satisfaction, as the imputed values help to fill gaps while maintaining the rating patterns for each cuisine type.

Low Impact: The imputation process has had a low effect on the mean ratings across most cuisines. This suggests that the missing values, once imputed, align well with the existing ratings, indicating consistent customer satisfaction trends across cuisines.

Slight Adjustments: Some cuisines, like Chinese, Indian, Italian, Japanese, Mexican and Thai, show a small shift in median rating post-imputation. This may reflect that imputed values for these cuisines differed slightly from their existing average ratings.

Overall, the imputation maintained the integrity of the data, keeping the average ratings stable across cuisines



Reasoning for Imputation Strategy

In this dataset, the 'rating' column includes some missing values, with these gaps distributed across different cuisine types. To ensure that these missing ratings don't distort the analysis, it's essential to fill them in a way that maintains the integrity of the data.

- Cuisine-Specific Imputation:

  - Customer expectations and satisfaction may vary by cuisine type. To retain this nuance, we impute missing values based on each cuisine's unique characteristics. Instead of using a single global value across all cuisines, we fill missing ratings using the mean or median rating within each cuisine. This approach respects the diversity in customer satisfaction across different cuisines.

- Minimizing Bias and Maintaining Consistency:

  - By using cuisine-specific averages (mean or median), this imputation method minimizes bias, ensuring that missing values align with existing customer preferences and ratings trends. This helps avoid artificially inflating or deflating the ratings for any cuisine.

- Data-Driven Decision-Making:

  - This approach also enables the business to perform more accurate analysis on customer satisfaction, as we're filling in missing data based on historical trends rather than assumptions.

Using mean or median imputation by cuisine type preserves the data's original structure and trends, allowing for more reliable and insightful analysis. By respecting the unique distribution of ratings within each cuisine, this strategy ensures that our insights remain as close to the true customer experience as possible.

Univariate Analysis of Key Variables

Exploring Each Variable's Distribution

1. Overview of Univariate Analysis:

- **Objective**: To explore each variable individually and understand their distributions, central tendencies, and any significant patterns.
- **Methodology**: Used histograms, box plots, and count plots to visually assess each variable.

Univariate Analysis of Key Variables

- There are 1898 unique order ID: Every order sold once, no duplicates nor multiple orders on same OrderId
- There are 1200 unique Customer ID: There are customers that have ordered more than once during this timeframe in the dataset.
- There are a total of 178 unique restaurant names

```
[ ] # check unique order ID
    df['order_id'].nunique()
```

```
1898
```

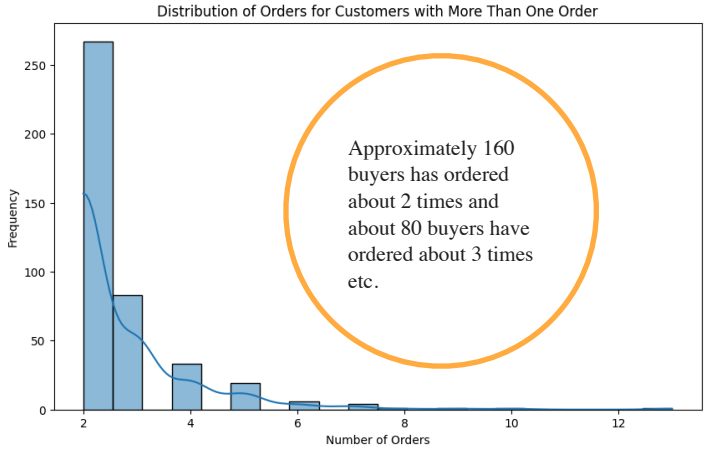
```
[ ] # check unique customer ID
    df['customer_id'].nunique()
```

```
1200
```

restaurant_name	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

178 rows × 1 columns

dtype: int64



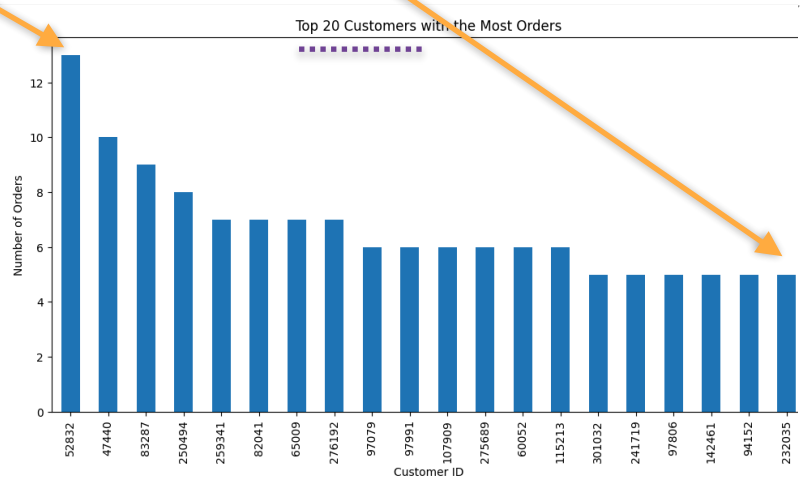
Univariate Analysis of Key Variables

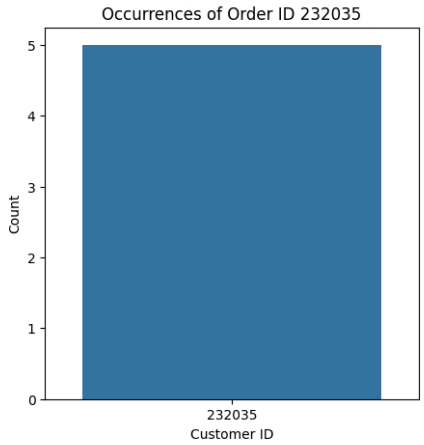
The top 20 customers with the most orders.

Customer Id 52832 has order 13 times, and the OrderId 232035 has ordered 5 times

	customer_id	order_count
0	52832	13
1	47440	10
2	83287	9
3	250494	8
4	259341	7

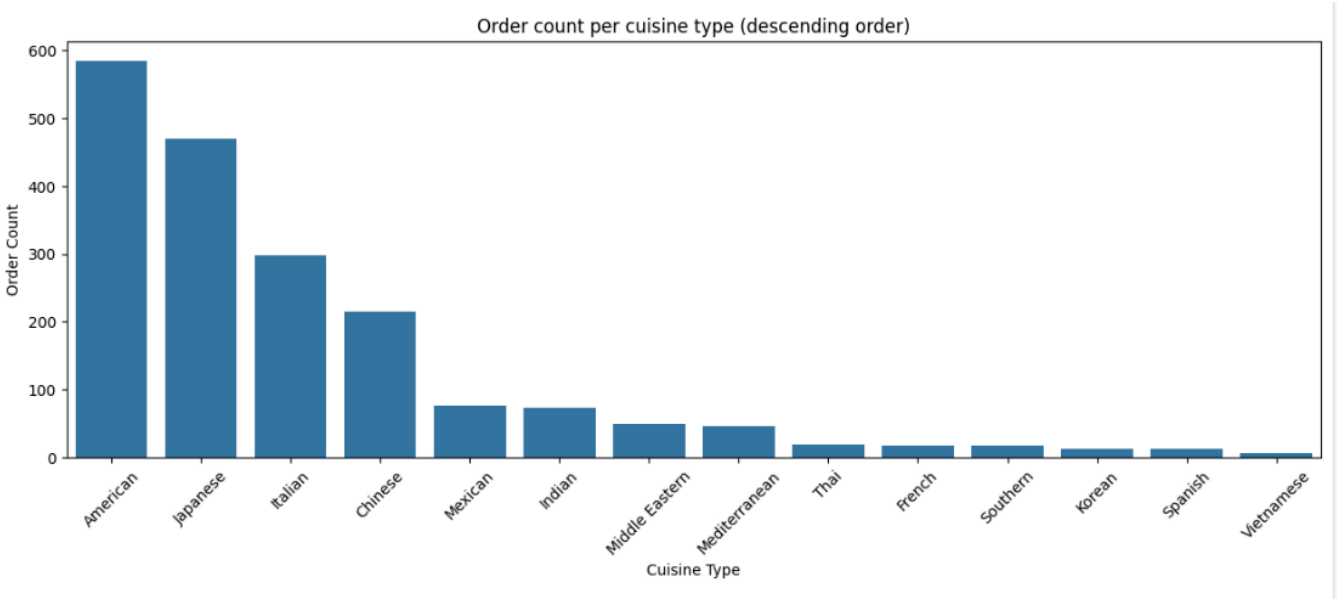
Multiple orders top 5 descending





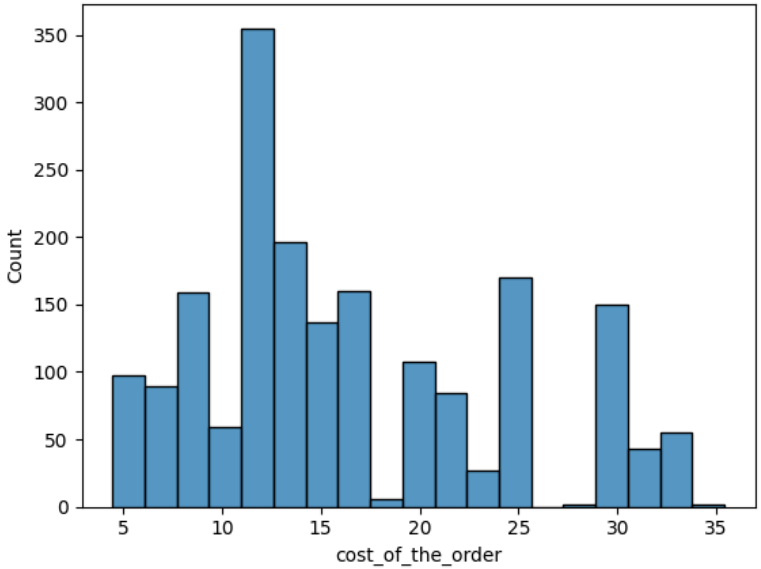
Univariate Analysis of Key Variables

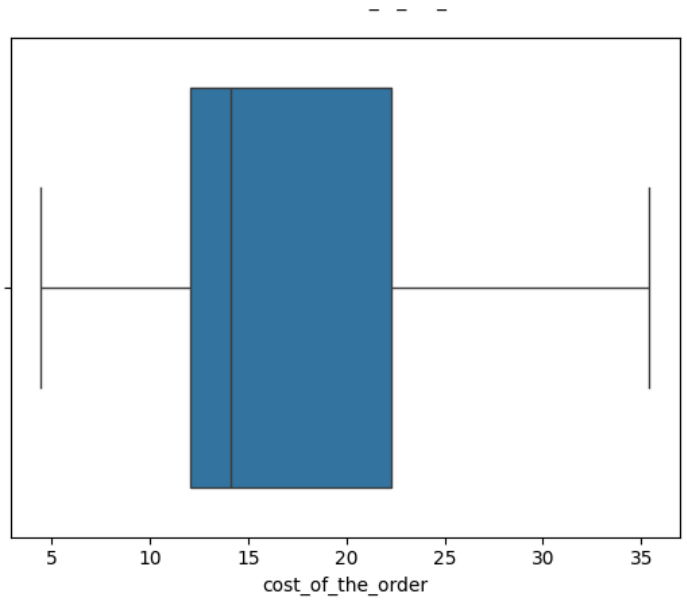
- There are 14 unique cuisines: American, Japanese, Italian, Mexican, Indian, Middle Eastern, Mediterranean, Thai, French, Southern, Korean, Spanish, Vietnamese
- How many orders per cuisine? The top 3: American; 600 orders, Japanese; 450 orders, Italian; 300 orders



Univariate Analysis of Key Variables

- Most orders fall between about $12 to $22. The data is right-skewed and 75% of the data falls within the upper half of the range, with a concentration
around higher order costs, suggesting a trend towards mid-to-high priced orders
- There are no extreme outliers. The whiskers extend to the minimum (5 dollars) and maximum (35 dollars) values, which implies the data captures the full range without any unexpected extreme values.

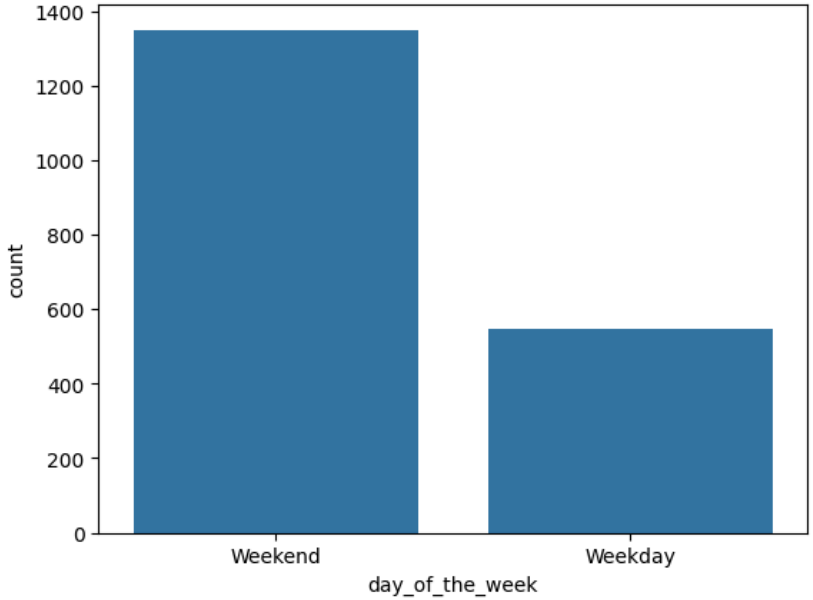




Univariate Analysis of Key Variables

Day of the week
The dataset contains Weekday or Weekend and there is a significant difference in order count between weekends and weekdays indicates that customers are placing a much higher number of orders on weekends compared to weekdays.

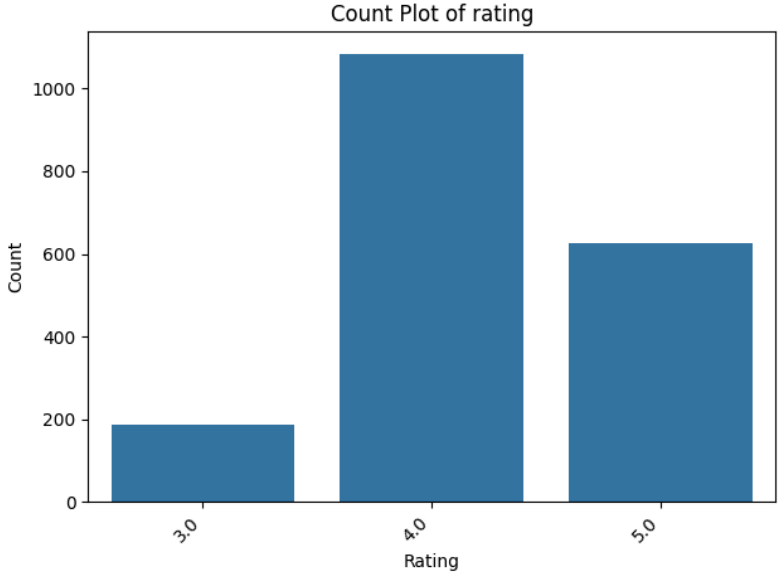
This suggests **increased demand on weekends**: There is a noticeable spike in orders over the weekend, which could be due to customers' increased availability and interest in ordering in during their days off. This trend could be useful for operational planning.



Univariate Analysis of Key Variables

Count of Rating

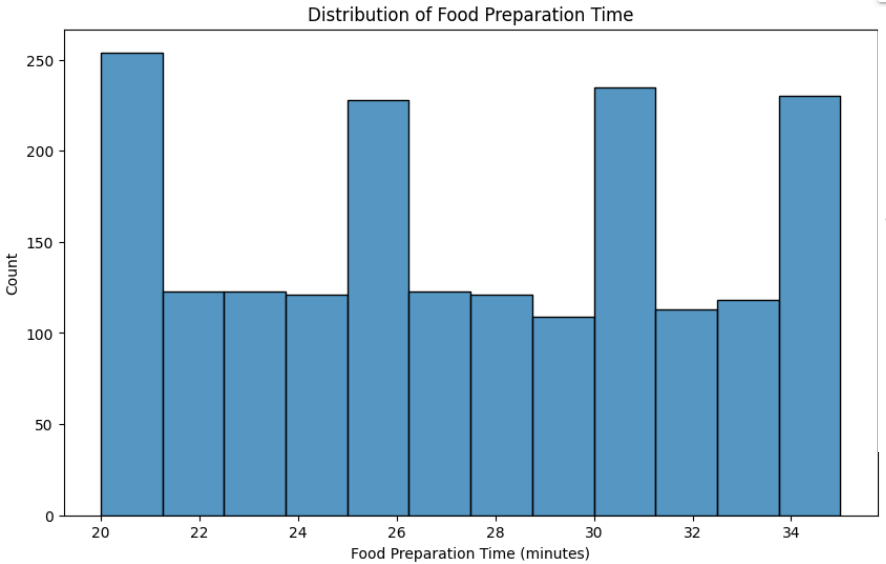
The dataset contain rating of 3, 4 or 5 and there is significant more count of the rating 4. Over 1000 of rating 4, around 600 counts on the rating 5, and the least for rating 3.

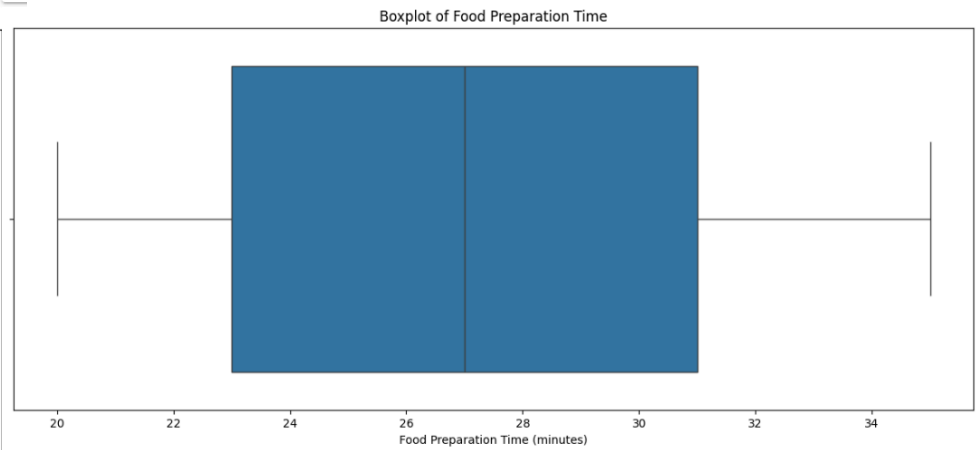


Univariate Analysis of Key Variables

Food Preparation Time

These plots illustrate the distribution of Food Preparation Time across orders. The histogram shows that food preparation times are fairly spread out, with a few peaks, especially around 20, 26, 31 and 34 minutes, indicating common preparation durations. The box plot confirms a relatively even spread without extreme outliers, with the interquartile range (middle 50% of data) centered between 23 and 31. This suggests that while there is some variation in preparation times, most orders fall within a consistent time frame, providing a predictable experience for customers.

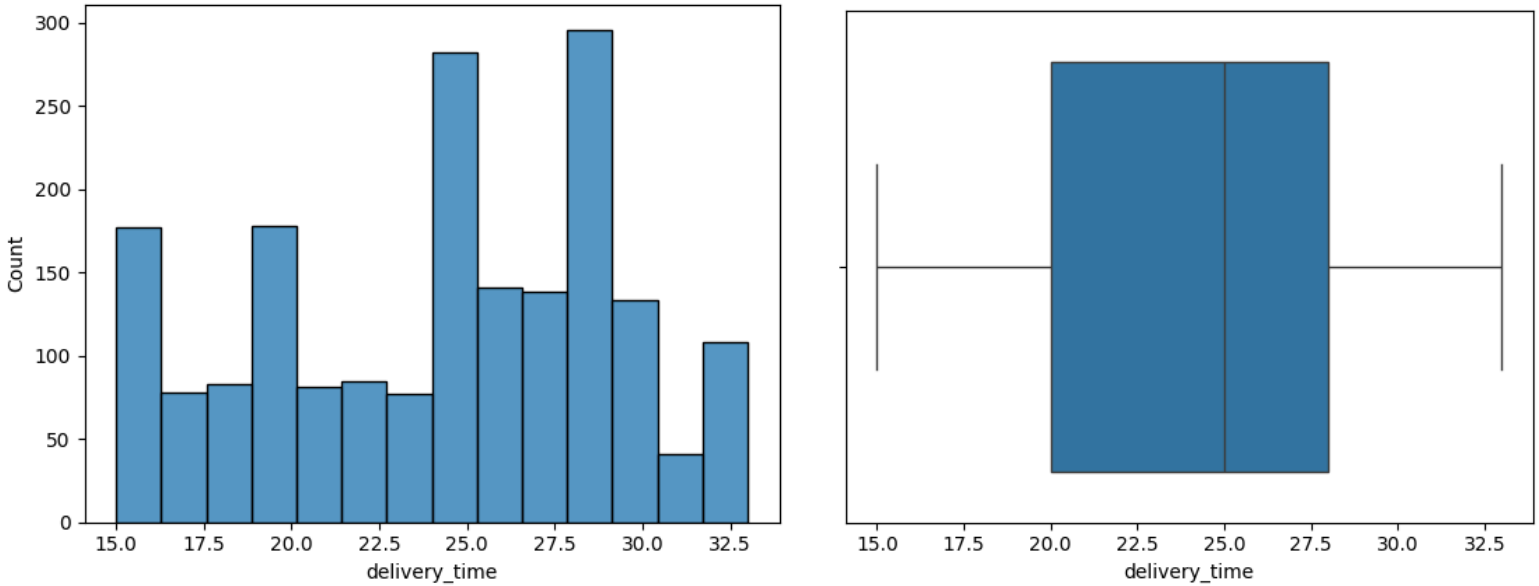




Univariate Analysis of Key Variables

Delivery Time

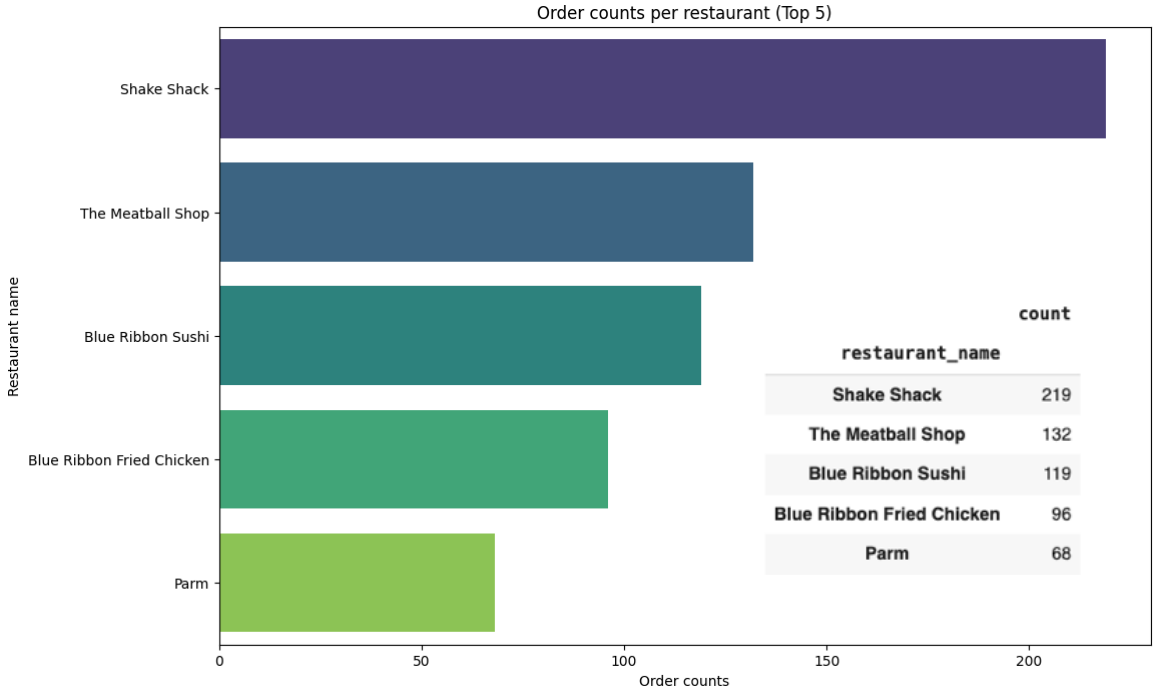
These plots illustrate Delivery Time Distribution. The histogram shows a left-skewed distribution, with most delivery times around 25-28 minutes. The box plot confirms this skewness, as the lower interquartile is longer than the upper interquartile, highlighting that most deliveries are under 25 minutes. This pattern suggests a generally efficient delivery process, but there could be room for improvement in minimizing the upper whisker.



Univariate Analysis of Key Variables

Top 5 Restaurants:

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

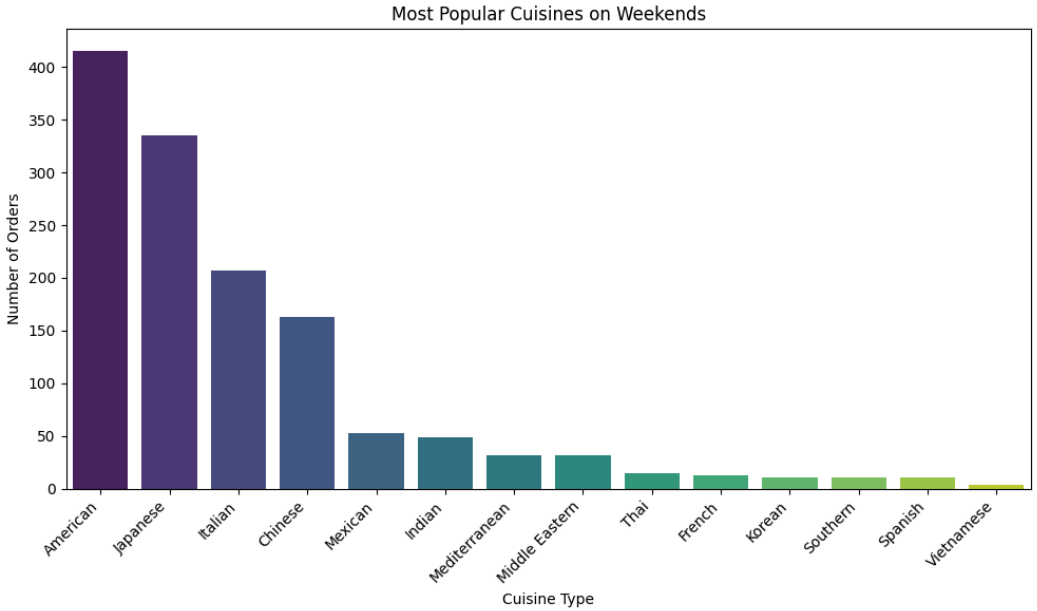


restaurant_name	count
Shake Shack	219
The Meatball Shop	132
Blue Ribbon Sushi	119
Blue Ribbon Fried Chicken	96
Parm	68

These restaurants consistently receive high order volumes, making them key players in the platform's success.

Univariate Analysis of Key Variables

Most popular cuisine type on weekends
The histogram shows that the American cuisine is the most popular cuisine on weekends with 415 counts, with Japanese cuisine just below at approximately 340 counts and the third most popular is Italian cuisine approximately 200 counts. Chinese cuisines also see high demand, while cuisines like Vietnamese and Spanish have the lowest order counts.



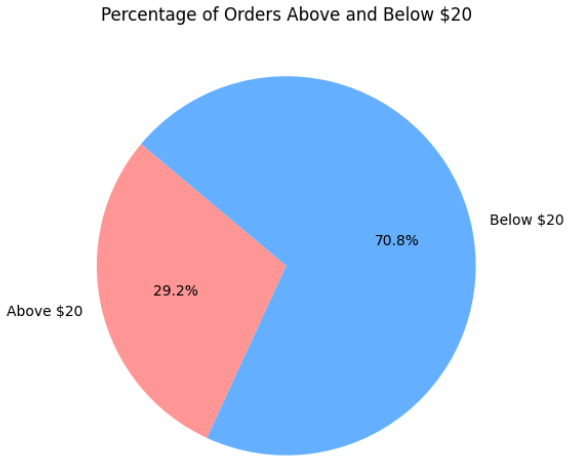
These cuisines consistently receive high order volumes, making them key players in the platform's success.

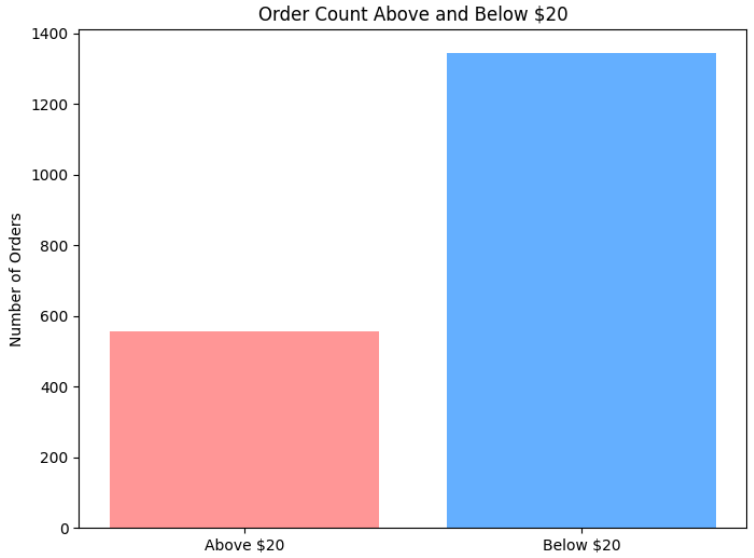
Univariate Analysis of Key Variables

Percentage of Orders Above $20

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

This analysis shows that 29.24% of orders have a cost exceeding $20, indicating a significant proportion of high-value orders. The visual breakdown highlights that while most orders are below $20, there is a substantial market of higher-spending customers. This insight suggests an opportunity for the business to target these high-value customers with tailored promotions or loyalty programs, potentially enhancing customer retention and increasing order frequency.



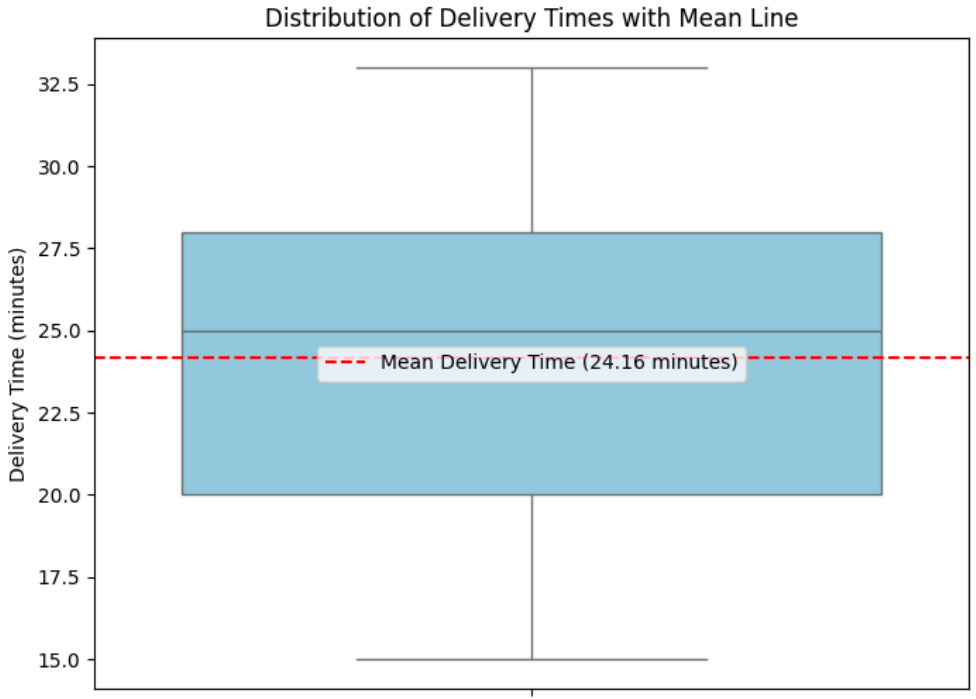


Univariate Analysis of Key Variables

While previous slide showed the median of delivery, this plot shows the **mean delivery time.**

Mean Delivery Time: 24.16 minutes

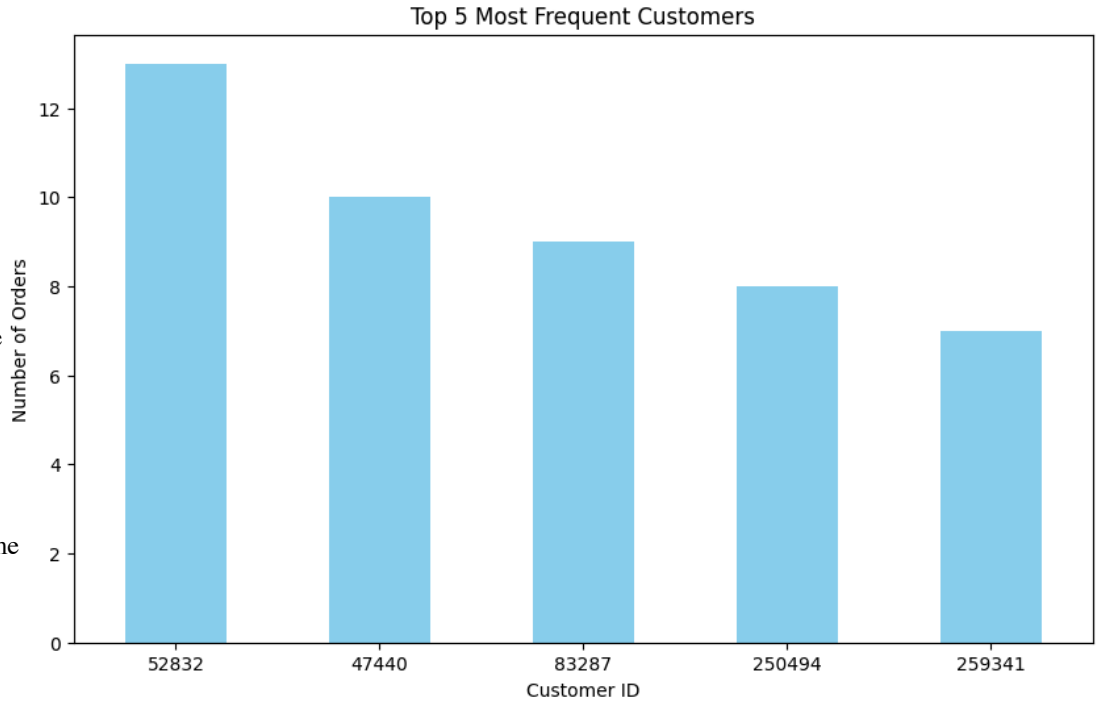
The mean delivery time is 24.16 minutes, as shown in the box plot, which almost mirrors the median at 25 minutes.
The distribution indicates that most delivery times are centered around this mean with minimal outliers, suggesting overall consistency. This average can serve as a benchmark for operational efficiency.



Univariate Analysis of Key Variables

Top 5 Most Frequent Customers (Eligible for 20% Discount)

- **Customer IDs and Order Counts**:
  - Customer 52832: 13 orders
  - Customer 47440: 10 orders
  - Customer 83287: 9 orders
  - Customer 250494: 8 orders
  - Customer 25934: 7 orders



This bar chart shows the top 5 most frequent customers. The data suggests that these customers are highly engaged with the
platform. Offering a 20% discount to these loyal customers could be a strategic move to increase their retention and encourage even
more frequent orders, potentially enhancing customer lifetime value.

Multivariate Analysis of Key Relationships

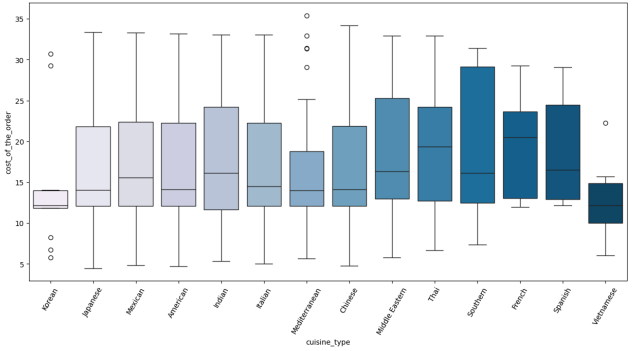
Exploring Relationships Between Variables

1. Multivariate Analysis Overview

- **Objective**: To explore the relationships between key variables and identify patterns or dependencies that can drive business insights.
- **Methodology**:
  - Analyzed numerical relationships (e.g., order cost vs. delivery time).
  - Explored the interactions between numerical and categorical variables (e.g., day of the week vs. delivery time, cost distribution by ratings, cost vs cuisine).

Observation of Multivariate Analysis of Key Relationships

Cost of the Order vs. Cuisine type (slide 1/3):
This box plot and scatterplot visualizes the relationship between the **cost of the order** and different **cuisine types**.

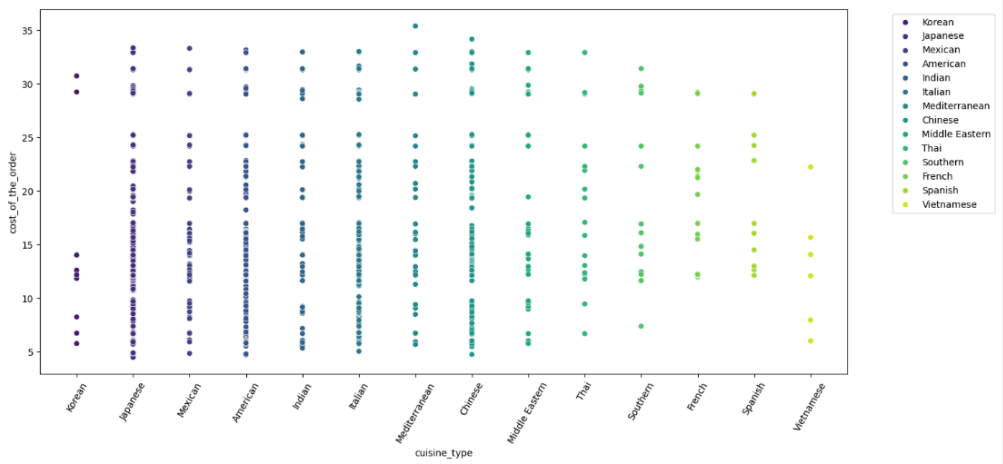


Cost Range:

- Most cuisines have a broad range of order costs, with wide interquartile ranges (the height of each box).
- **Korean** and **Vietnamese** cuisines have much narrower cost ranges compared to others. Korean cuisine, in particular, has the lowest median and a compact distribution, indicating that orders from Korean restaurants tend to be more affordable and consistent in price, although a few outliers.

Median Cost:

- **Middle Eastern, Thai, Southern**, and **French** cuisines have higher median costs, implying that these cuisines are generally more expensive.
- **Korean** and **Vietnamese** cuisines, on the other hand, show the lowest median costs, making them the most budget-friendly options on average.



Observation of Multivariate Analysis of Key Relationships

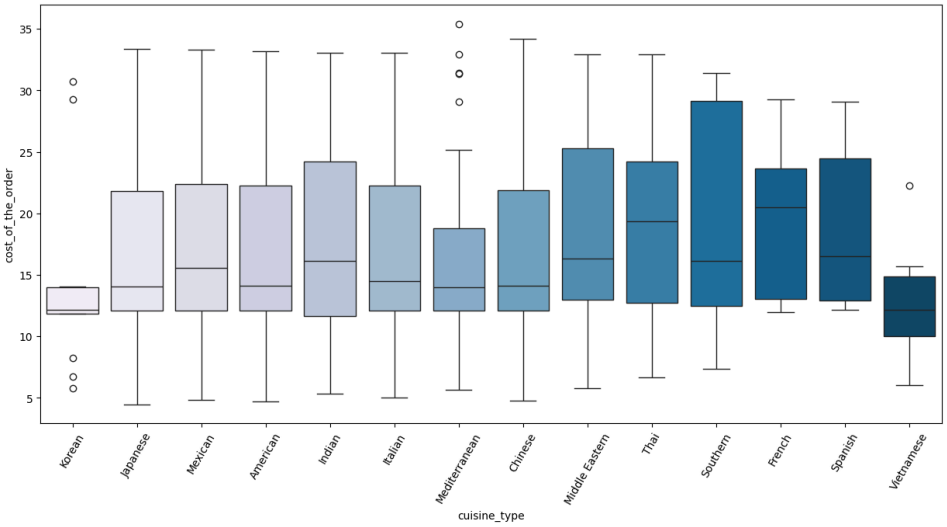
Cost of the Order vs. Cuisine type (slide 2/3):

Outliers:

- Outliers are visible in **Mediterranean** and **Korean** cuisines, and **Vietnamese**. These outliers represent orders that are significantly different from the usual cost range—either much lower or higher.
- **Mediterranean** cuisine, in particular, has several outliers above the interquartile range, indicating some high-cost orders that are uncommon for this cuisine type. **Korean** cuisine have some outliers both under and above the interquartile range but the outliers lies within the range of a resonable cost range, no further investigation needed at this point.
- For the rest of the cuisines there are no outliers, suggesting a more consistent range of cost of the orders.

General Distribution:

- **American, Japanese, and Mexican** cuisines show relatively consistent distributions with similar median costs, positioning these cuisines in a middle price range with less extreme cost variation.

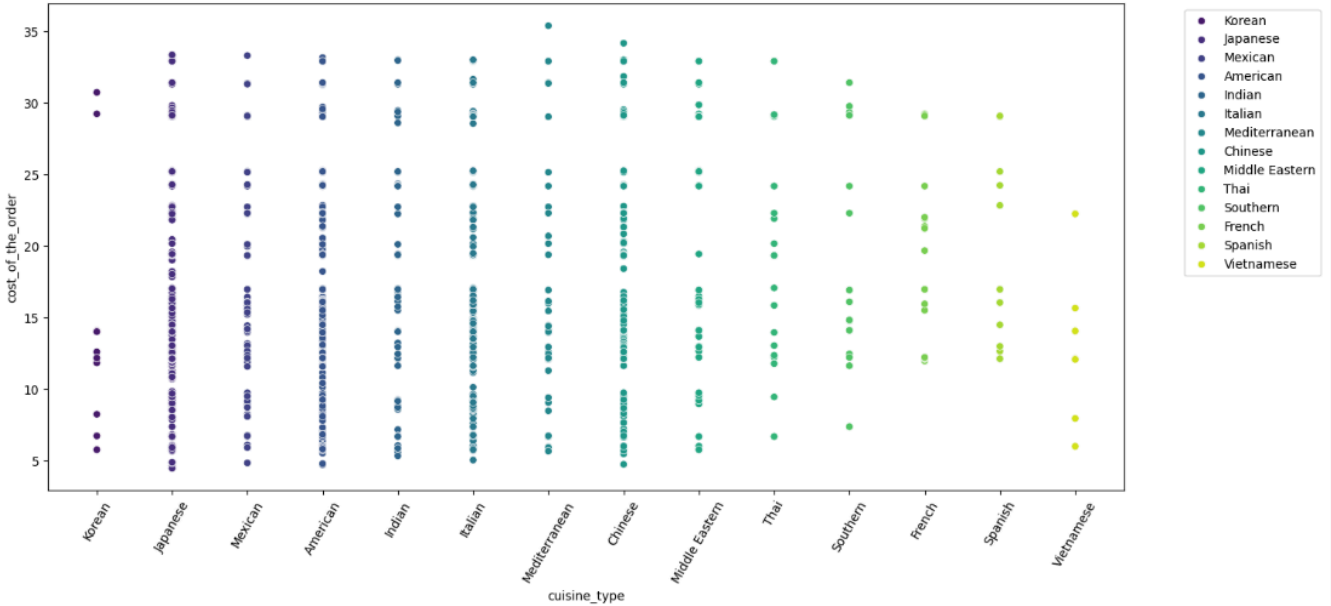


Observation of Multivariate Analysis of Key Relationships

Cost of the Order vs. Cuisine type (slide 3/3):

This analysis suggests:

- **Korean** and **Vietnamese** cuisines are generally more affordable, with narrow and consistent cost ranges.
- **Middle Eastern, Southern, and French** cuisines appear to have higher costs on average, appealing to customers seeking a more premium dining experience.
- The wider cost ranges in most of the cuisines, for example **Chinese, Thai and Southern** cuisines indicate a mix of affordable and higher-priced options, potentially appealing to a broader customer base.
- The presence of outliers in **Mediterranean** cuisine highlights some variability, with occasional high-cost orders that differs from the norm.



This information provides insights into which cuisines may be more budget-friendly, premium, or varied in price, which can guide recommendations and marketing strategies.

Observation of Multivariate Analysis of Key Relationships

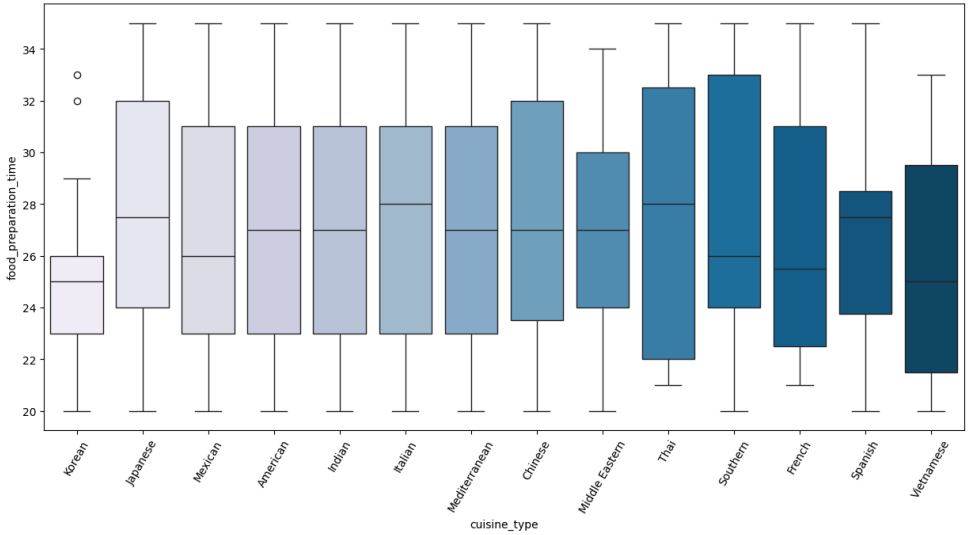
Food preparation time vs. Cuisine type (1/2):
This box plot illustrates the relationship between **food preparation time** and different **cuisine types.**

Quick-Service Cuisines:

- **Korean** and **Vietnamese** cuisines have the shortest preparation times. Korean cuisine, in particular, has the lowest median preparation time, and a narrow interquartile range suggests consistently short preparation times. Koreans also the only cuisine with outlier, but very few indicating longer food preparation time is probably not a common event.
- These cuisines appear well-suited for customers looking for faster food options, as they generally involve shorter preparation times.

Longer Preparation Time Cuisines:

- **Thai, Japanese,** and **Italian** cuisines have the longest preparation times, with higher medians and wider interquartile ranges. This implies that these cuisines may involve more complex or time-intensive dishes, making them potentially less suitable for customers in a hurry.



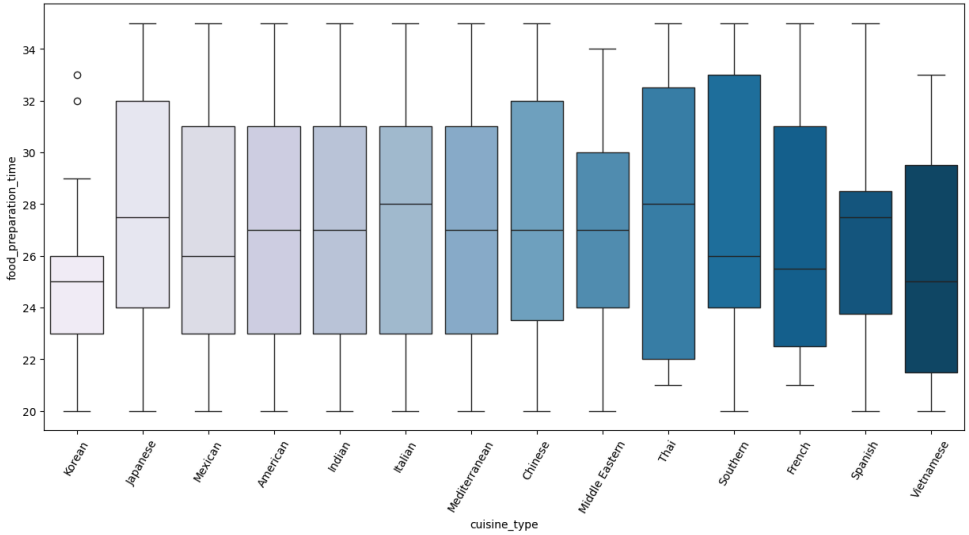
Observation of Multivariate Analysis of Key Relationships

Food preparation time vs. Cuisine type (2/2):

Outliers:

- **Korean** cuisine is the only one that shows visible outliers. These outliers suggest some unusually long preparation times, but is still within a reasonable time range compared to the other cuisines.
- No other cuisines show visible outliers, meaning the preparation times for these cuisines are relatively consistent without extreme variations.

- **Moderate and Consistent Preparation Times**:
  - **Japanese, Thai, Indian, American, Italian, Mediterranean,** and **Chinese** cuisines exhibit moderate preparation times with consistent ranges, indicating a balanced approach to food preparation. **Middle Eastern** cuisine shows a more compact box plot preparation time. Its interquartile range is narrower, indicating consistent preparation times that have a narrower spread in the data.



Observation of Multivariate Analysis of Key Relationships

Day of the Week vs. Delivery Time (1/2):

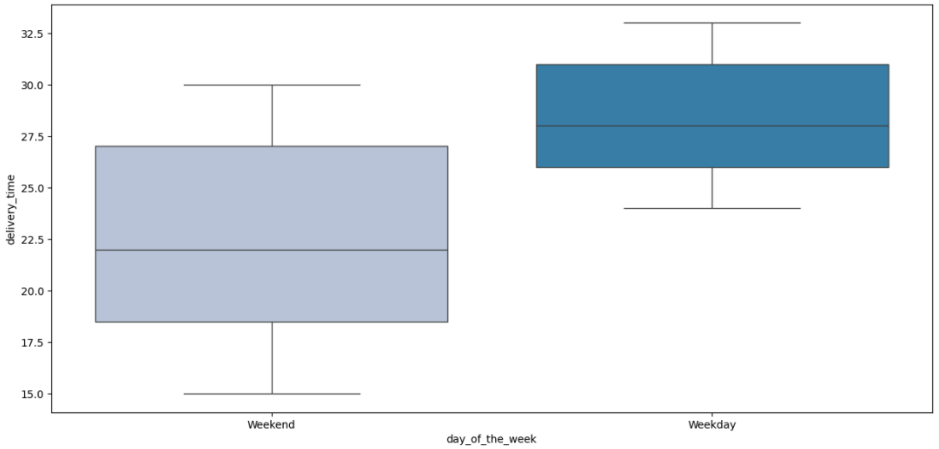
Delivery times are generally faster on weekends than on weekdays, likely due to lower traffic and fewer potential delays on weekends. This pattern suggests that weekday traffic may be a factor in extending delivery times, and focusing on weekday delivery efficiency could help improve overall customer satisfaction.

Similar IQR on Weekdays and Weekends:

- The **IQR** (the range containing the middle 50% of delivery times) for both weekdays and weekends is fairly compact and similar in size. This suggests that the delivery times are relatively stable within their own range, regardless of whether it's a weekday or weekend.
- The compact IQR for both days implies that while the median delivery time is higher on weekdays, the consistency within the middle 50% of times is comparable across both types of days.

Interpretation of Consistent IQR:

- Since the IQR is similar for both weekdays and weekends, it indicates that most deliveries are consistently clustered around their respective median times, without large deviations.
- However, while the IQR is compact for both, the slight difference in median delivery times means that weekday deliveries tend to be longer even within this consistent middle range.

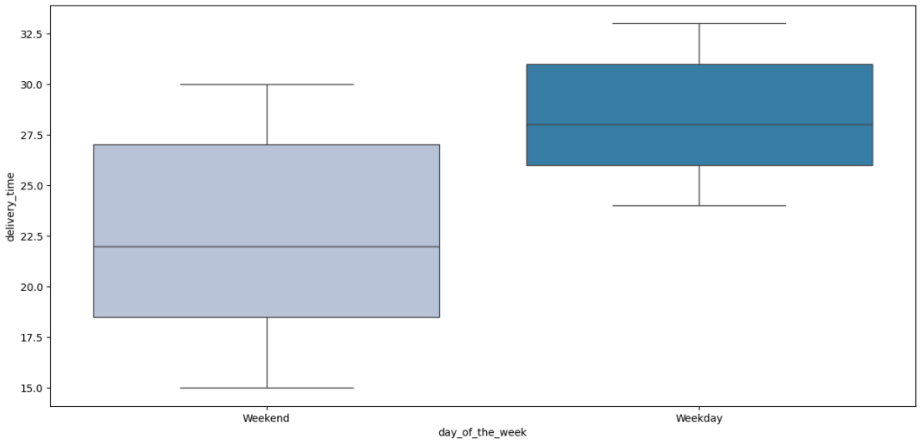


Observation of Multivariate Analysis of Key Relationships

Day of the Week vs. Delivery Time (2/2):

Variability Outside the IQR:

- The primary difference in variability arises outside the IQR. Weekdays show a narrower range overall, extending up to about 32.5 minutes, and weekends have a lower maximum at 30 minutes. This higher maximum on weekdays could indicate occasional delays, which may be less frequent on weekends.



The similar length of the whiskers for both weekdays and weekends suggests that delivery times are generally consistent for the majority of orders, regardless of the day of the week. The main difference lies in the median and the overall range.

- **Weekday deliveries** have a higher median and and maximum delivery time, indicating that while most deliveries are stable, the delivery time is higher on weekdays.
- **Weekend deliveries** have a lower median, suggesting that deliveries are faster and less likely to experience delays on weekends.

Observation of Multivariate Analysis of Key Relationships

Observations on the revenue generated by the restaurants

Total revenue = 31.314.82 dollars

- **Shake Shack** stands out as the top revenue-generating restaurant, with a total order cost of **3579.53**. This indicates that it is either highly popular or customers tend to place higher-value orders at this restaurant. We've seen in earlier plots that **Shake Shack** is most popular with highest numbers of order.
- **The Meatball Shop** and **Blue Ribbon Sushi** follow in second and third place, with revenues of **2145.21** and **1903.95**, respectively. There is a significant gap between Shake Shack and these two, suggesting a much higher demand or average order cost for Shake Shack compared to the others.
- **Blue Ribbon Fried Chicken** and **Parm** also have high revenues, although significantly lower than the top three. These restaurants, along with **RedFarm Broadway** and **RedFarm Hudson**, have revenues around 1000 or slightly below, positioning them as moderately popular in terms of revenue.

Overall, it appears that the top restaurants account for a large portion of the total revenue in the dataset, while the lower-ranked restaurants generate significantly less. This could indicate that a few highly popular restaurants are responsible for a majority of the revenue. This piece of information can help understanding which restaurants are the most profitable and potentially worth promoting further in the app or through other means.

restaurant_name	cost_of_the_order
Shake Shack	3579.530
The Meatball Shop	2145.210
Blue Ribbon Sushi	1903.950
Blue Ribbon Fried Chicken	1662.290
Parm	1112.760
RedFarm Broadway	965.130
RedFarm Hudson	921.210
TAO	834.500
Han Dynasty	755.290
Blue Ribbon Sushi Bar & Grill	666.620
Rubirosa	660.450
Sushi of Gari 46	640.870
Nobu Next Door	623.670
Five Guys Burgers and Fries	506.470

Observation of Multivariate Analysis of Key Relationships

Rating vs. Delivery Time:

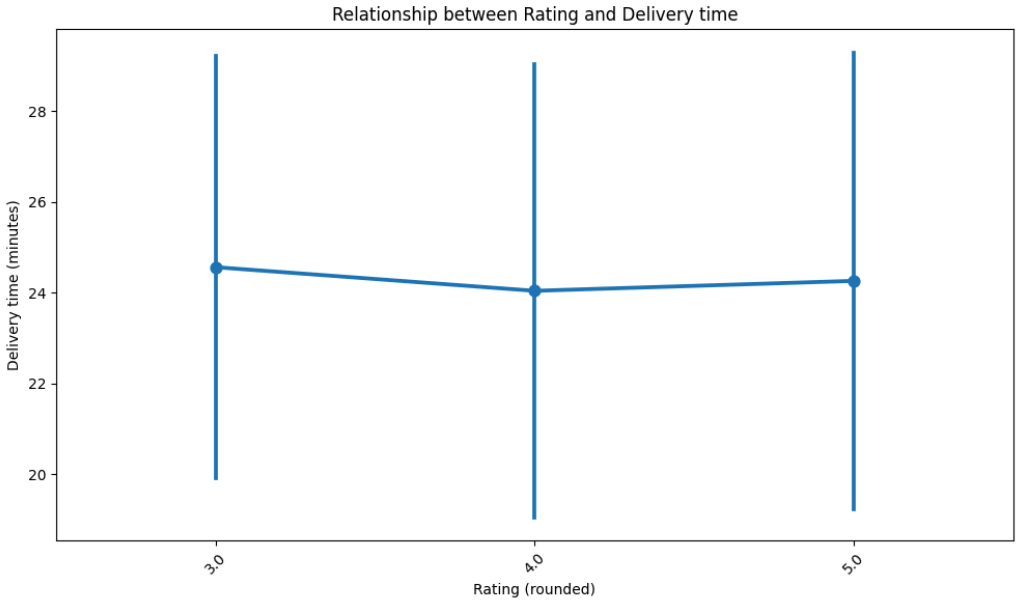
This plot shows the relationship between customer ratings and delivery time in minutes.

- The line appears relatively flat, suggesting that there is no significant difference in average delivery times across the different rating levels. This indicates that delivery time, on average, may not be a major factor influencing customer ratings in this dataset.

- The vertical lines represent the variability (possibly the standard deviation or confidence interval) in delivery times for each rating level. All rating levels show a relatively high variability in delivery times, especially at the extremes. This could mean that some customers experience faster deliveries while others face delays, regardless of the rating they give.

- Since the average delivery times are close across all ratings, it suggests that factors other than delivery time might play a more significant role in determining customer satisfaction (as reflected in ratings). To improve ratings, the business might need to focus on factors like food quality, order accuracy, or communication rather than solely on reducing delivery time.

This plot suggests that delivery time consistency might not be directly linked to customer ratings, as even high ratings show similar delivery times to lower ratings.



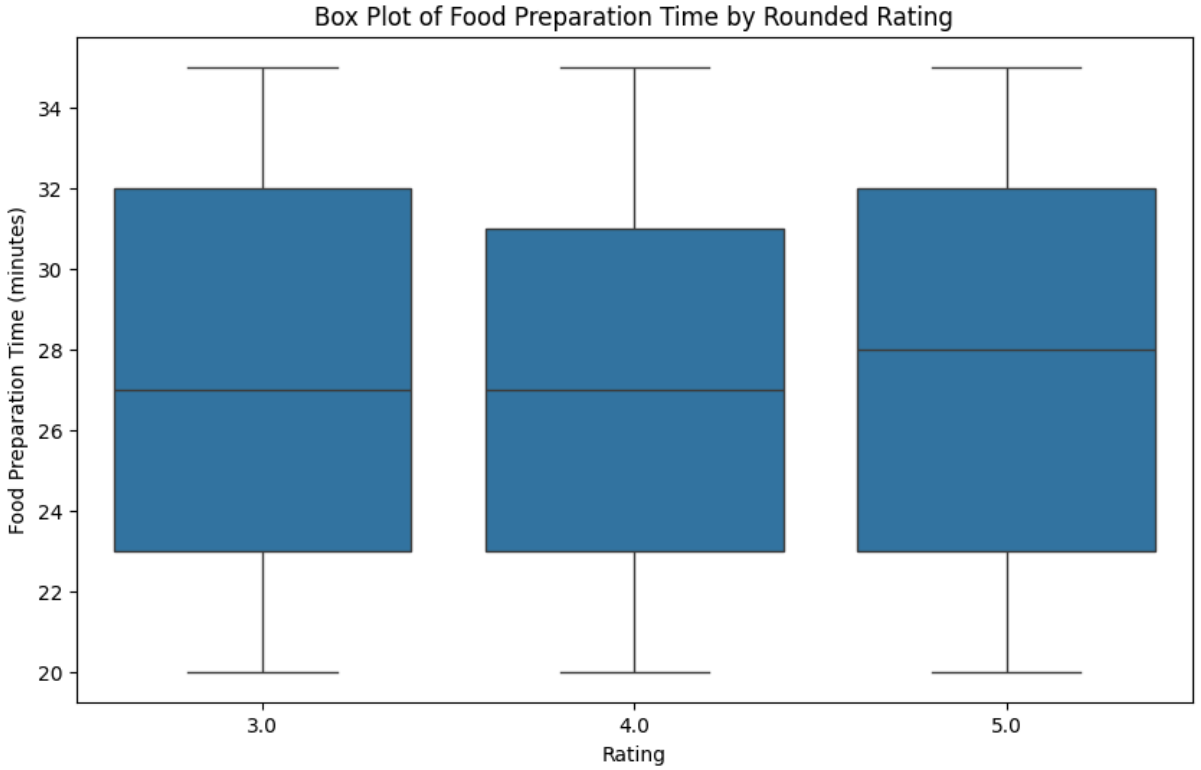
Observation of Multivariate Analysis of Key Relationships

Rating vs Food preparation time

This box plot shows the relationship between customer ratings and food preparation time.

Overall, the food preparation time remains relatively consistent across all ratings. There's no significant upward or downward trend that suggests a strong correlation between preparation time and ratings.

This plot implies that factors other than preparation time likely play a more significant role in determining customer ratings.

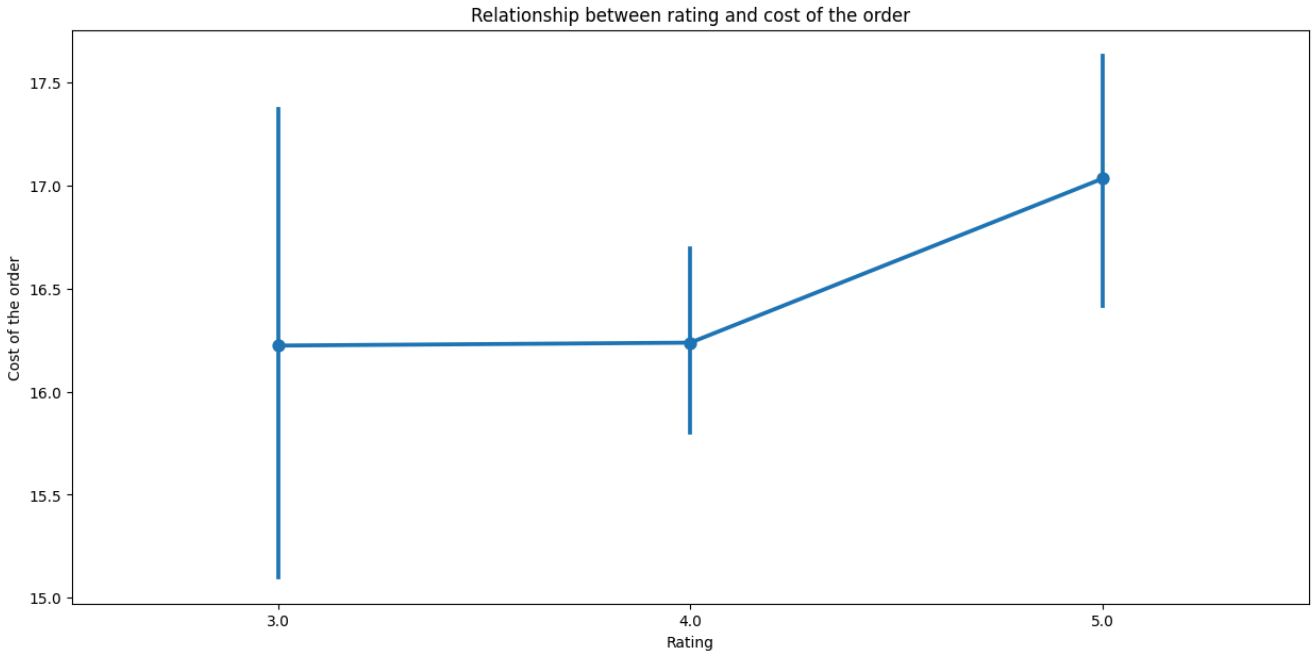


Observation of Multivariate Analysis of Key Relationships

Rating vs Cost of the Order

This plot show the relationship between rating and cost of the order

The key takeaway is that higher-rated orders tend to have higher and more consistent costs, while the lower ratings such as rating 3, have greater variability. This pattern suggests that inconsistent or varied order costs might contribute to a mid-range rating (3), while consistent, possibly higher-value orders are associated with higher ratings (4 and 5).



Observation of Multivariate Analysis of Key Relationships

Heatmap (Rating vs. Delivery Time vs Cost of the Order vs Food preparation time)

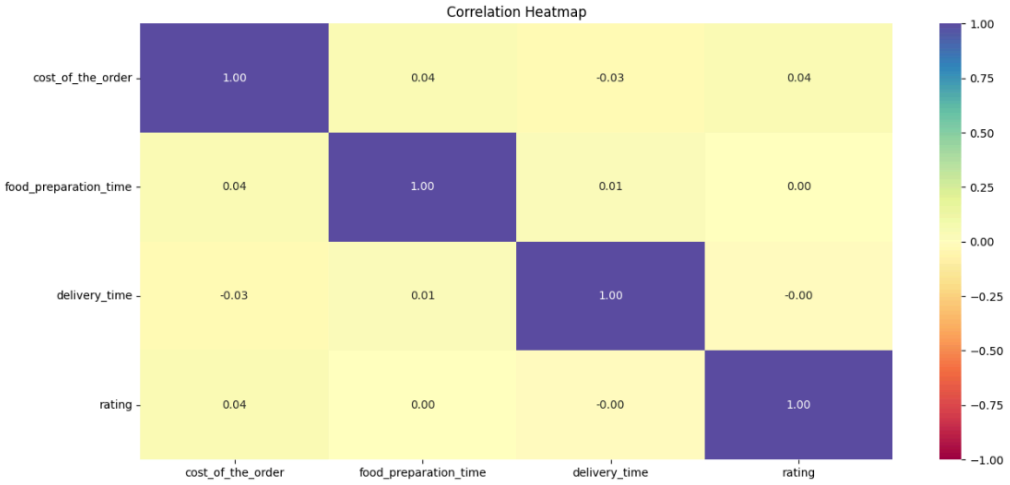
This heatmap shows the correlation matrix among several variables.

All correlations between the variables are close to zero, indicating very weak or no linear relationship among them. This suggests that changes in one variable do not predict changes in another:

- `cost_of_the_order` and `rating` have a very slight positive correlation (0.04), but it's minimal, suggesting that the cost does not significantly impact the rating.
- `food_preparation_time` and `rating` show virtually no correlation (0.00), suggesting that variations in food preparation time have no observable impact on ratings.
- `delivery_time` and `rating` have a near-zero negative correlation (-0.00), indicating that delivery time also does not significantly affect ratings.

Weak Correlations Among Operational Variables:

- `cost_of_the_order` with `food_preparation_time` and `delivery_time` shows very low correlations (0.04 and -0.03, respectively), suggesting that neither the preparation time nor delivery time is closely associated with order cost.
- `food_preparation_time` and `delivery_time` have a minimal positive correlation (0.01), which indicates that these two operational times are largely independent.

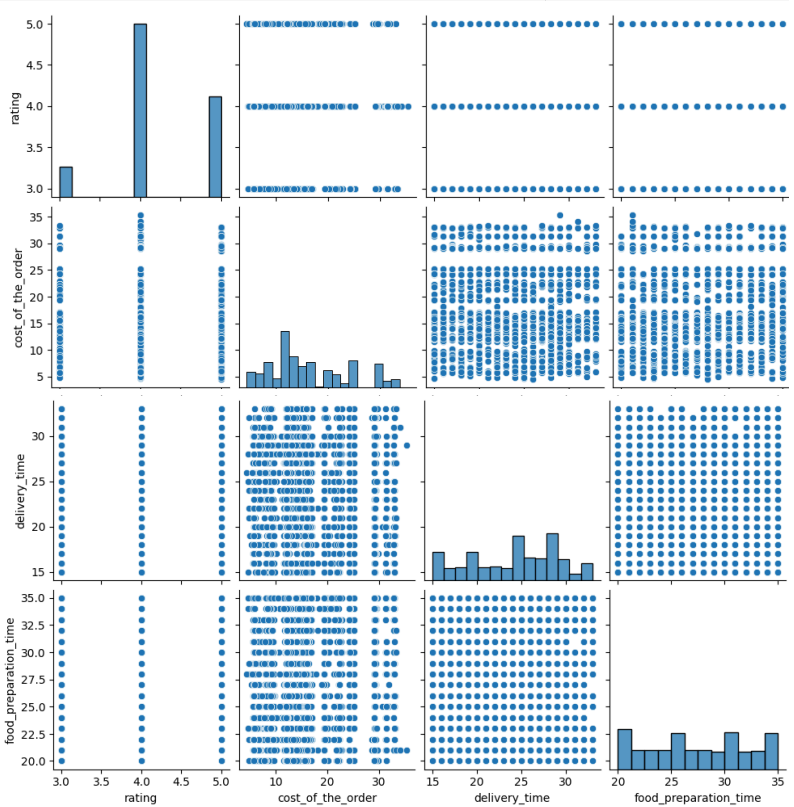


This heatmap indicates that there are no significant linear relationships among `cost_of_the_order`, `food_preparation_time`, `delivery_time`, and `rating`. Each variable seems to operate independently, with minimal influence on the others, especially in terms of customer ratings. This lack of correlation suggests that factors like cost, preparation time, and delivery time are not directly driving customer satisfaction, as measured by ratings, in a linear way.

Observation of Multivariate Analysis of Key Relationships

Pairplot

This pair plot shows the correlation matrix among several variables: `cost_of_the_order`, `food_preparation_time`, `delivery_time`, and `rating`.



The key takeaway from this pairplot is that there are weak or no strong correlations among the variables.

- Ratings do not show a clear relationship with order cost, delivery time, or preparation time. This suggests that customers' satisfaction levels, as measured by ratings, are likely influenced by factors not included in this dataset, maybe for examples service quality, food quality, or other external factors.
- There is no apparent relationship between the cost of an order and the time taken for delivery or preparation, indicating that higher-priced orders don't necessarily take longer or shorter to prepare and deliver.
- Each variable seems to vary independently, meaning they do not significantly impact each other. This independence suggests that operational factors like preparation and delivery times are managed separately from order cost and are not likely to be directly impacting customer satisfaction.

Multivariate Analysis of Key Relationships

Answers to Business Questions

Restaurants Eligible for Promotional Offer

- **Criteria**: Rating count > 50 and average rating > 4.
- **Eligible Restaurants**: in total there are 7 restaurants that meet the criteria

	restaurant_name	rating
0	The Meatball Shop	4.326
1	Blue Ribbon Fried Chicken	4.219
2	RedFarm Broadway	4.169
3	Shake Shack	4.169
4	Blue Ribbon Sushi	4.134
5	RedFarm Hudson	4.109
6	Parm	4.074

Net Revenue Generated by the Company

- **Revenue Calculation**:
  - For orders costing over $20: 25% commission rate.
  - For orders costing between $5 and $20: 15% commission rate.
- **Total Revenue**: The restaurants generated **31.314.82 dollars in total.** After calculation, Foodhub's revenue is **6166.3** dollars.

High-value orders contribute significantly to revenue, highlighting the importance of targeting this segment.

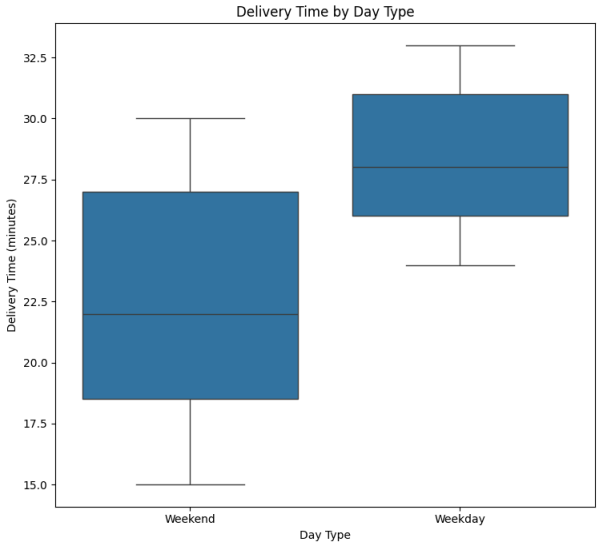
Multivariate Analysis of Key Relationships

Percentage of Orders Taking More Than 60 Minutes

- **Criteria**: Total time = food preparation time + delivery time.
- **Orders Taking >60 Minutes**: 10.54% of all orders take more than 60 minutes.
  A noticeable portion of orders have a total time more than 60 minutes.

Mean Delivery Time Variation Between Weekdays and Weekends

- **Weekday Delivery Time**: Average of **28 minutes**.
- **Weekend Delivery Time**: Average of **22 minutes**.
  Weekday deliveries are generally slower, possibly due to higher traffic.



APPENDIX

`order_id`: Unique identifier for each order.
`customer_id`: ID of the customer who placed the order.
`restaurant_name`: Name of the restaurant that fulfilled the order.
`cuisine_type`: Type of cuisine ordered by the customer.
`cost_of_the_order`: Monetary cost of the order.
`day_of_the_week`: Indicates whether the order was placed on a weekday (Monday to Friday) or weekend (Saturday and Sunday).
`rating`: Customer rating of the order, given out of 5.
`food_preparation_time`: Time (in minutes) taken by the restaurant to prepare the food. Calculated as 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. Calculated as the difference between the timestamps of the delivery person's pick-up confirmation and drop-off confirmation.

Great Learning

Happy Learning !