Exploratory Data Analysis

Python Foundations: Foodhub

June 23, 2023

Vaibhav Pradhan

Contents / Agenda

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

Executive Summary

Foodhub is a food aggregator company that has collected and stored data on orders made by registered customers in their online portal and is looking to understand demand for different restaurants to help enhance customer experience. The data was analyzed to gain valuable insights to understand customer preferences and behavior around the following

Restaurant demand

- The top five restaurants with the most orders (Shake Shack, the Meatball Shop, Blue Ribbon Sushi, Blue Ribbon Fried Chicken, and Parm) account for 33.4% of all 1898 orders
- Foodhub can offer promotions for these restaurants on their app/portal which has the potential to drive demand for the restaurants and increased revenue for FoodHub.
- Repeat customers can also be further incentivized to generate demand

Popular Cuisines

- There are 14 unique cuisines across 178 restaurants and “American” cuisine is the most popular (584 orders)
- Top 5 popular cuisine in order of popularity are: American, Japanese, Italian, Chinese and Mexican
- These restaurants could be encouraged to offer promo offers to customers further increasing demand and Foodhub revenue

Customer ratings

- 736 out of 1898 orderswere not rated.
- Only neutral (3 star) and positive (4 and 5 star) reviews are given. 50.6% of reviews given were 5-star reviews, 33.2% were 4-star reviews, and 16.1% were 3 star reviews.
- It is possible that customers with negative experience did not rate their orders
- In order to gain a better understanding of negative customer experiences, and to gain a holistic view of customer experiences with FoodHub for each restaurant, FoodHub should incentivize customers to provide ratings for Foodhub

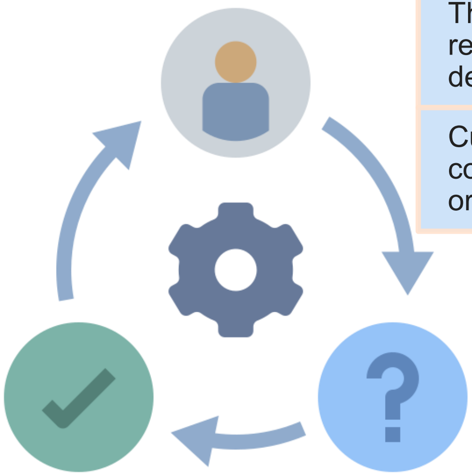
Ordering Trends on different days and delivery times

- Delivery times are longer on weekdays (28 minutes on average) than on weekends (22 minutes on average)
- The spread of delivery times on weekend is much higher than the spread of delivery times of weekdays. This indicates weekend delivery times have wider variance
- In order to improve customer delivery times on weekdays, FoodHub could evaluate their delivery staffing during weekday rush hour and recruit more people as needed. They could also review if the traffic routing on their app is optimized to avoid areas with heavy traffic on weekdays..

Typical customer food order cost

- The minimum money spent on an order was $4.47, and the most expensive order was $35.41 with an average cost of $16.5 and a mean cost of $14.14. The median value is lower than the mean value for cost_of_the_order and hence data is skewed to the right and tend to be on the higher side than the median value. Percentage of orders above 20 dollars 29.24 %
- Korean and Vietnamese food orders tend to be lower side and have a lower spread on the cost
- Day of the week does not have any impact on cost of the order
- Data analysis did not indicate that cost of the order was contributing to customer demand

Business Problem Overview and Solution Approach



Business Context

FoodHub is a food aggregator company that offers a smartphone app that allows customers to access multiple restaurants and help place online orders.

The app facilitates the entire order and delivery process, from the restaurant receiving the order to a delivery person picking it up and delivering it to the customer's location.

Customers can rate their orders, and FoodHub earns revenue by collecting a fixed margin from the restaurants for each delivery order.

Problem Overview

Foodhub has collected and stored data on orders made by registered customers in their online portal and is looking to understand demand for different restaurants to help enhance customer experience

Foodhub data can be analyzed to gain valuable insights to understand customer preferences and behavior around the following

- Restaurant demand
- Popular Cuisines
- Customer ratings
- Order delivery times
- Ordering trends on different days
- Typical customer food order cost

Solution Approach

Exploratory Data Analysis approach is to be leveraged to analyze and summarize main characteristics and uncover hidden insights from the data.

Observations from the analysis can help to enhance customer experience, improve operational efficiency, and make informed business decisions to offer promotions that can drive demand

Data Analysis includes univariate and multivariate analysis of the dataset to review summary statistics and visualization to identify data patterns, relationship between data elements and data anomalies

Data Overview

The data contains details on the orders that were submitted by Foodhub customers on the online portal and includes restaurant name, cuisine, customer ratings and delivery related information. Table below outlines the data dictionary

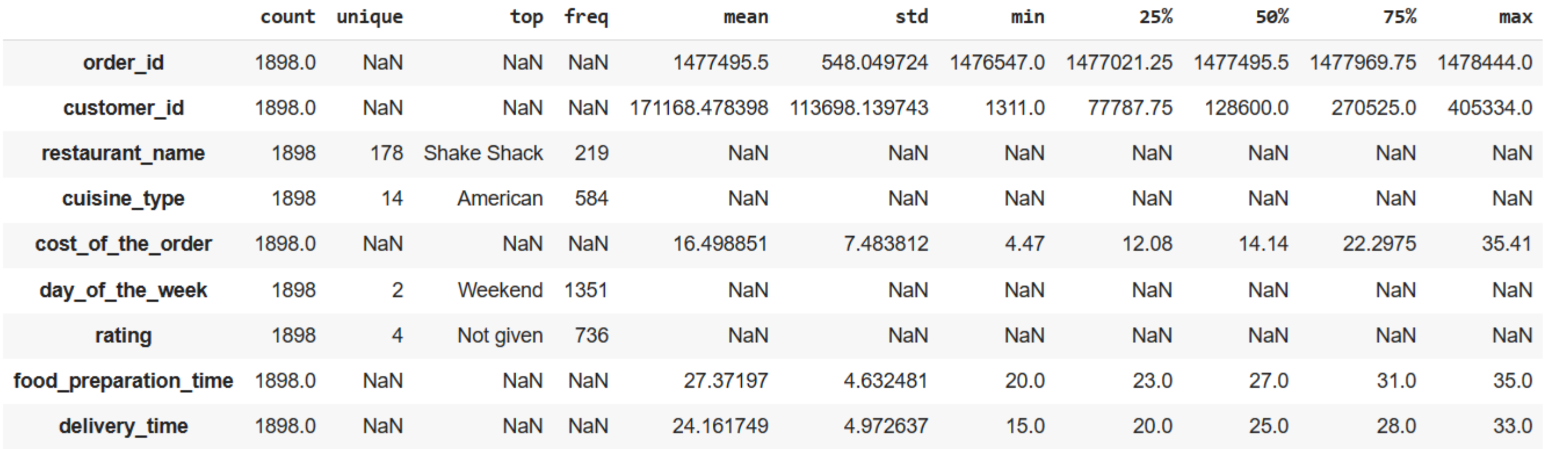
Variable	Description	Datatype	# of missing values
order_id	Unique ID of the order	Integer (int64)	0
customer_id	ID of the customer who ordered the food	Integer (int64)	0
restaurant_name	Name of the restaurant	String (Object)	0
cuisine_type	Cuisine ordered by the customer	String (Object)	0
cost_of_the_order	Cost of the order	Float (float64)	0
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)	String (Object)	0
rating	Rating given by the customer out of 5	String (Object)	0
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.	Integer (int64)	0
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	Integer (int64)	0

Note:

- The dataset provided included a total of 1898 rows (customer orders)
- A total of 9 data columns / variables are included
- All variables have data values available. However, about 736 orders have the customer rating variable value as “Not Given” which even though available is not very helpful

Data Overview

The statistical summary of the dataset for all data columns/variables is represented below.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
order_id	1898.0	NaN	NaN	NaN	1477495.5	548.049724	1476547.0	1477021.25	1477495.5	1477969.75	1478444.0
customer_id	1898.0	NaN	NaN	NaN	171168.478398	113698.139743	1311.0	77787.75	128600.0	270525.0	405334.0
restaurant_name	1898	178	Shake Shack	219	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cuisine_type	1898	14	American	584	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cost_of_the_order	1898.0	NaN	NaN	NaN	16.498851	7.483812	4.47	12.08	14.14	22.2975	35.41
day_of_the_week	1898	2	Weekend	1351	NaN	NaN	NaN	NaN	NaN	NaN	NaN
rating	1898	4	Not given	736	NaN	NaN	NaN	NaN	NaN	NaN	NaN
food_preparation_time	1898.0	NaN	NaN	NaN	27.37197	4.632481	20.0	23.0	27.0	31.0	35.0
delivery_time	1898.0	NaN	NaN	NaN	24.161749	4.972637	15.0	20.0	25.0	28.0	33.0

Note:

- Although order_id and customer_id are Integer, the summary statistics are not very relevant
- The dataset includes data for a total of 178 restaurants with most orders (219) from “Shake Shack” restaurant
- There are 14 unique cuisines across these 178 restaurants and “American” cuisine is the most popular (584 orders)
- Order amount varied from a minimum of $4.47 to a maximum of $35.41 with an average order amount of ~ $16.50
- The dataset includes primarily “Weekend” order with 1351 out of 1898 orders over the Weekend
- 736 out of 1898 customer orders do not have a customer defined rating
- Food Preparation Time varied from a minimum of 20 mins to a maximum of 35 mins with an average of ~27.37 mins
- Delivery Time varied from a minimum of 15 mins to a maximum of 33 mins with an average of ~24.16 mins

Univariate Analysis – order_id, customer_id, restaurant_name

Top 5 restaurants in terms of the number of orders received

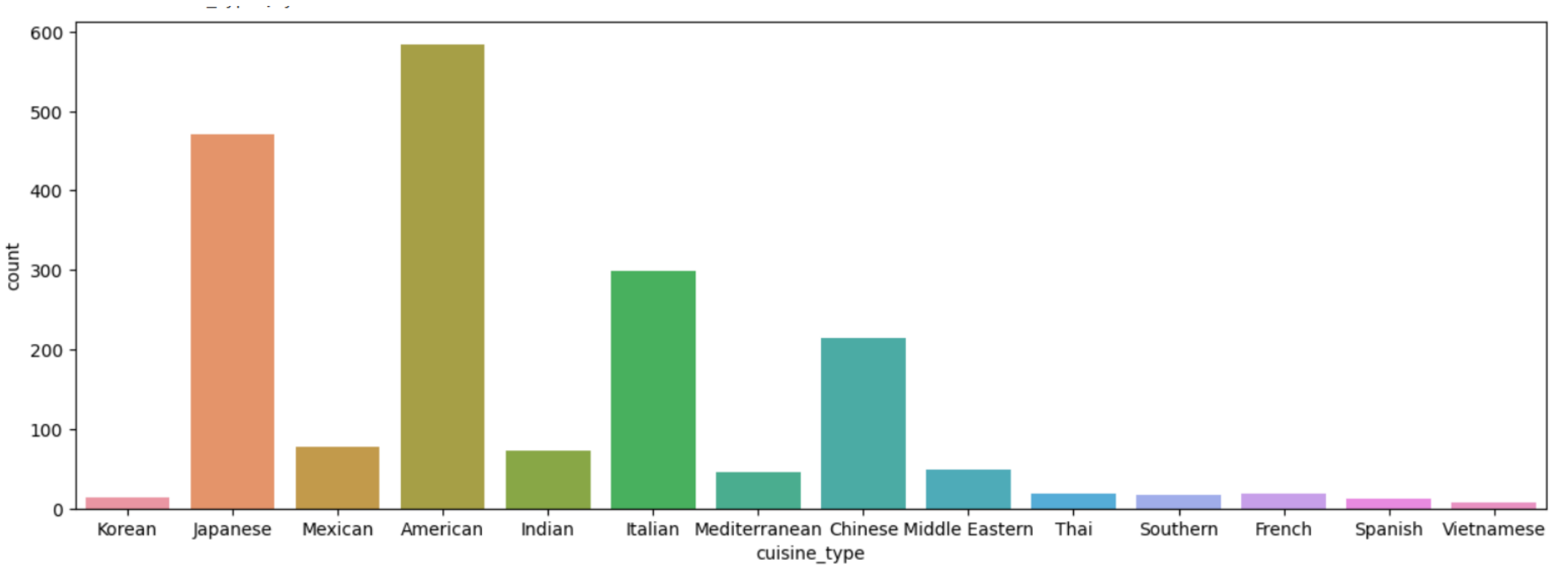
restaurant_name	# of orders
Shake Shack	219
The Meatball Shop	132
Blue Ribbon Sushi	119
Blue Ribbon Fried Chicken	96
Parm	68

Top 3 most frequent customers that can be offered 20% discount vouchers

Customer_id	# of orders
52832	13
47440	10
83287	9

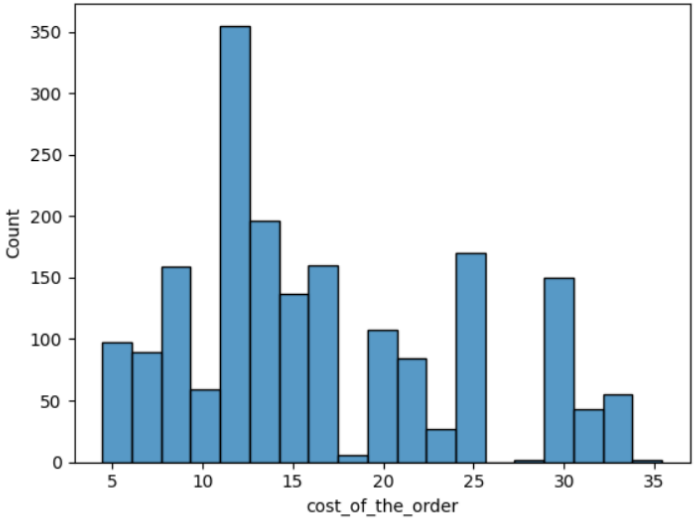
- The dataset includes 1898 unique customer orders. Minimum order_id is 1476547 and max order_id is 1478444
- The dataset includes orders placed by 1200 unique customers. Minimum customer_id is 1311 and maximum customer_id is 405334
- order_id and customer_id are Integer and represent unique identifiers for order and the customer, hence the summary statistics are not very relevant for these fields
- The dataset includes order data for a total of 178 restaurants with most orders (219) from “Shake Shack” restaurant
- Top 5 restaurants on the left account for ~33% of total orders
- Many customers have repeat orders and the table on the left indicates top 3 repeat customers

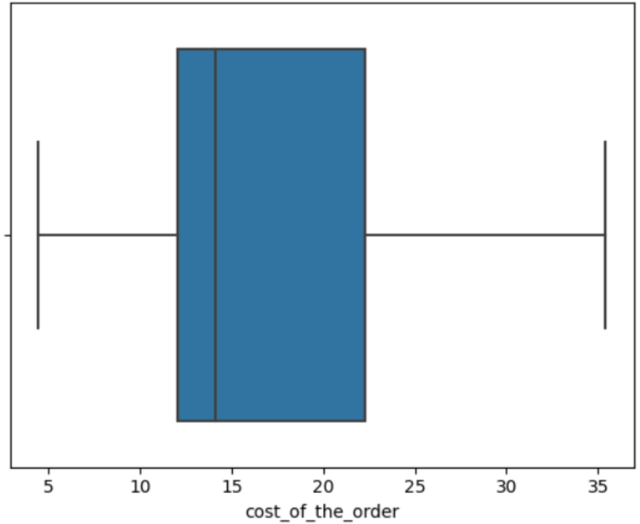
Univariate Analysis – cuisine_type



- There are 14 unique cuisines across 178 restaurants and “American” cuisine is the most popular (584 orders)
- Top 5 popular cuisine in order of popularity are: American, Japanese, Italian, Chinese and Mexican

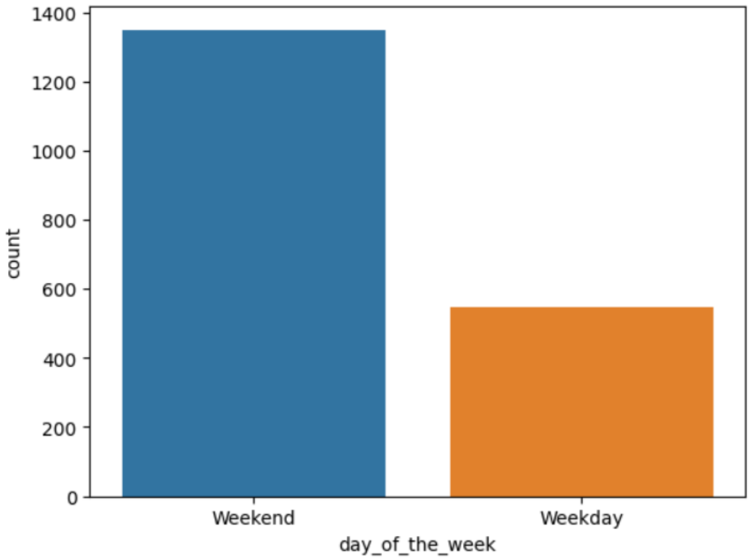
Univariate Analysis – cost_of_the_order





- The number of total orders that cost above 20 dollars is: 555
- Percentage of orders above 20 dollars: 29.24 %
- The minimum money spent on an order was $4.47, and the most expensive order was $35.41 with an average cost of $16.5 and a mean cost of $14.14
- The above is also evident from the histogram which depicts a longer tail of data to the right
- The boxplot also indicates that the median value is lower than the mean value for cost_of_the_order and hence data is skewed to the right
- There are no outliers on either minimum or maximum side of the data

Univariate Analysis – day_of_the_week

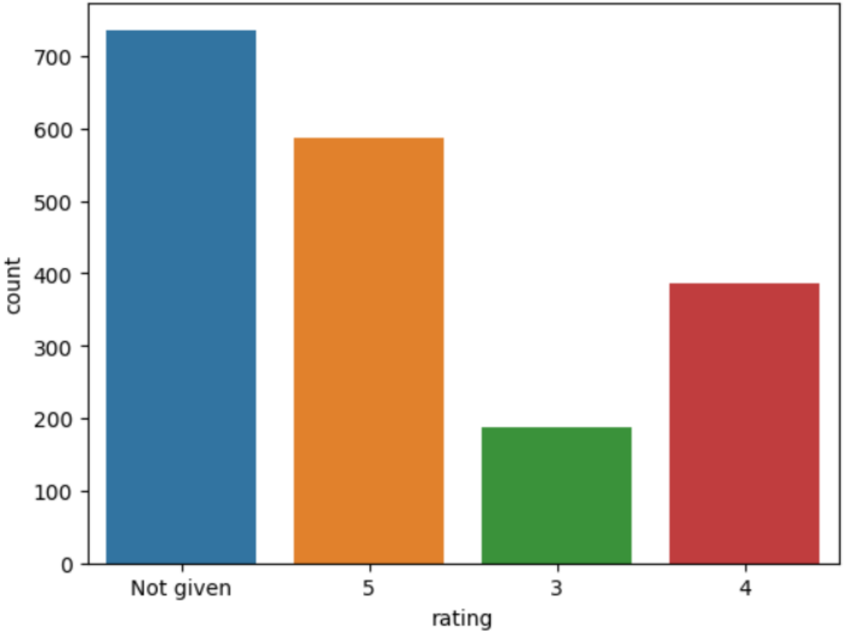


- The dataset includes primarily “Weekend” order with 1351 out of 1898 (~ 71%) orders over the Weekend
- Most popular cuisines over the weekend are depicted on the right which is aligned to the overall cuisine popularity

Most popular cuisine on weekends

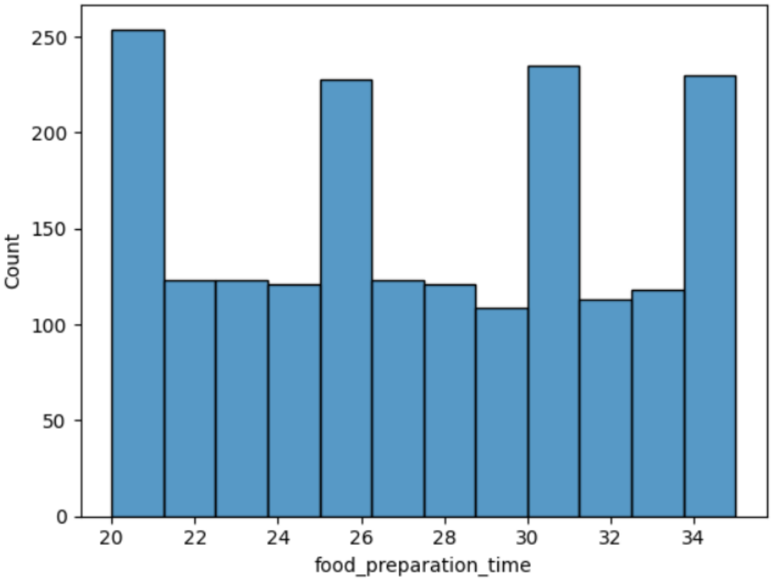
Cuisine	# of orders
American	415
Japanese	335
Italian	207
Chinese	163
Mexican	53
Indian	49
Mediterranean	32
Middle Eastern	32
Thai	15
French	13
Korean	11
Southern	11
Spanish	11
Vietnamese	4

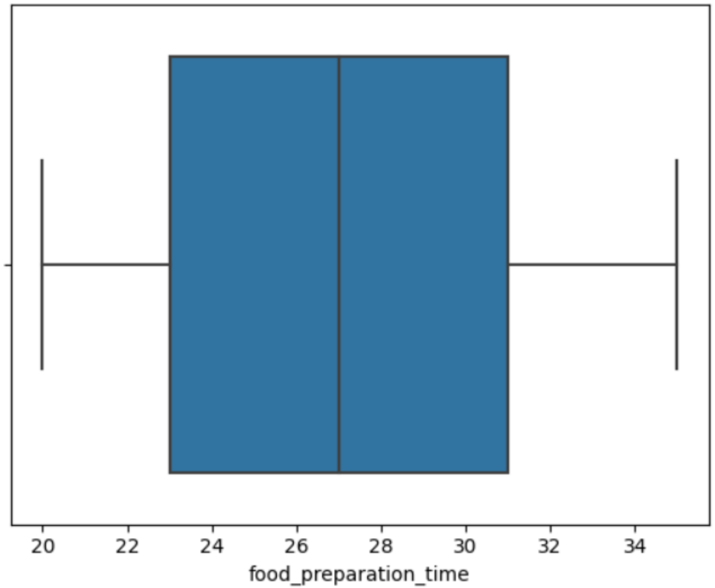
Univariate Analysis – rating



- 736 out of 1898 customer orders (~ 38% orders) do not have a customer defined rating
- 588 out of 1898 customer orders (~ 31% orders) have highest rating
- 386 orders received 4 stars
- 188 orders received 3 stars
- No order received 1 or 2 stars

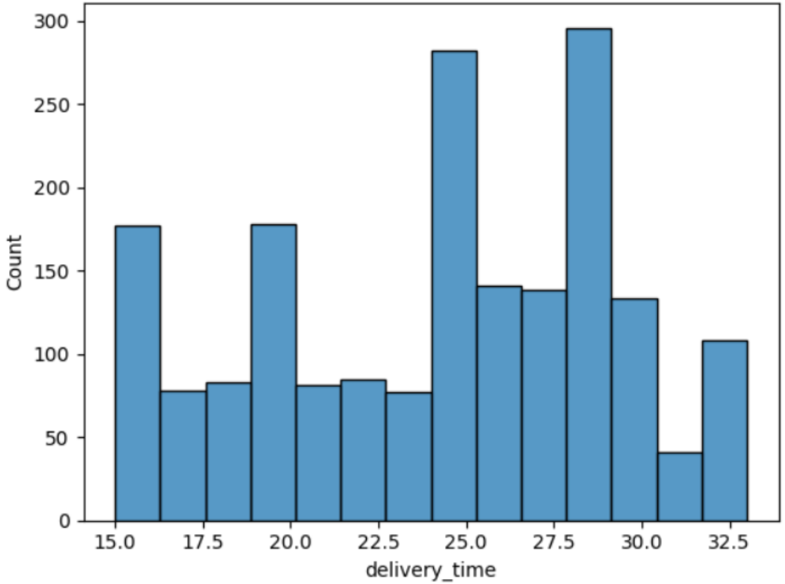
Univariate Analysis – food_preparation_time

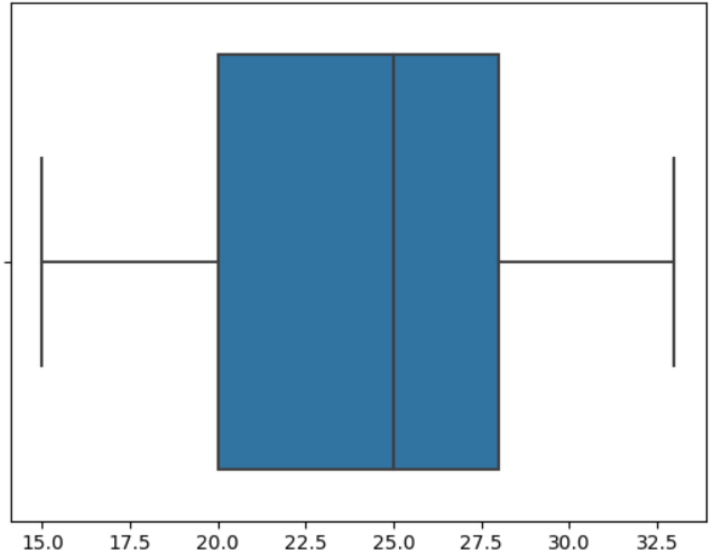




- Food Preparation time has a very slight righ skew but is mostly evenly distributed with the mean time = 27.37 and median time = 27
- Most food preparation takes between 23 – 31 minutes
- There are no outliers

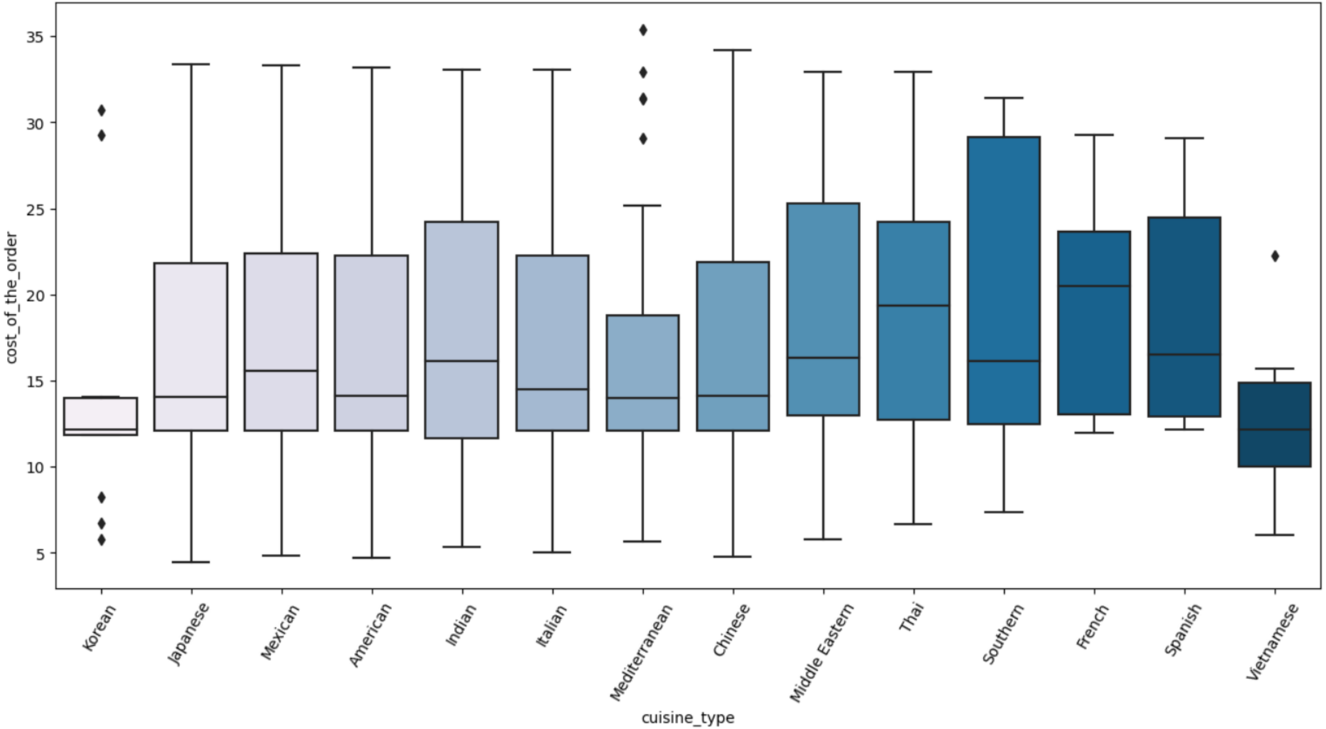
Univariate Analysis – delivery_time





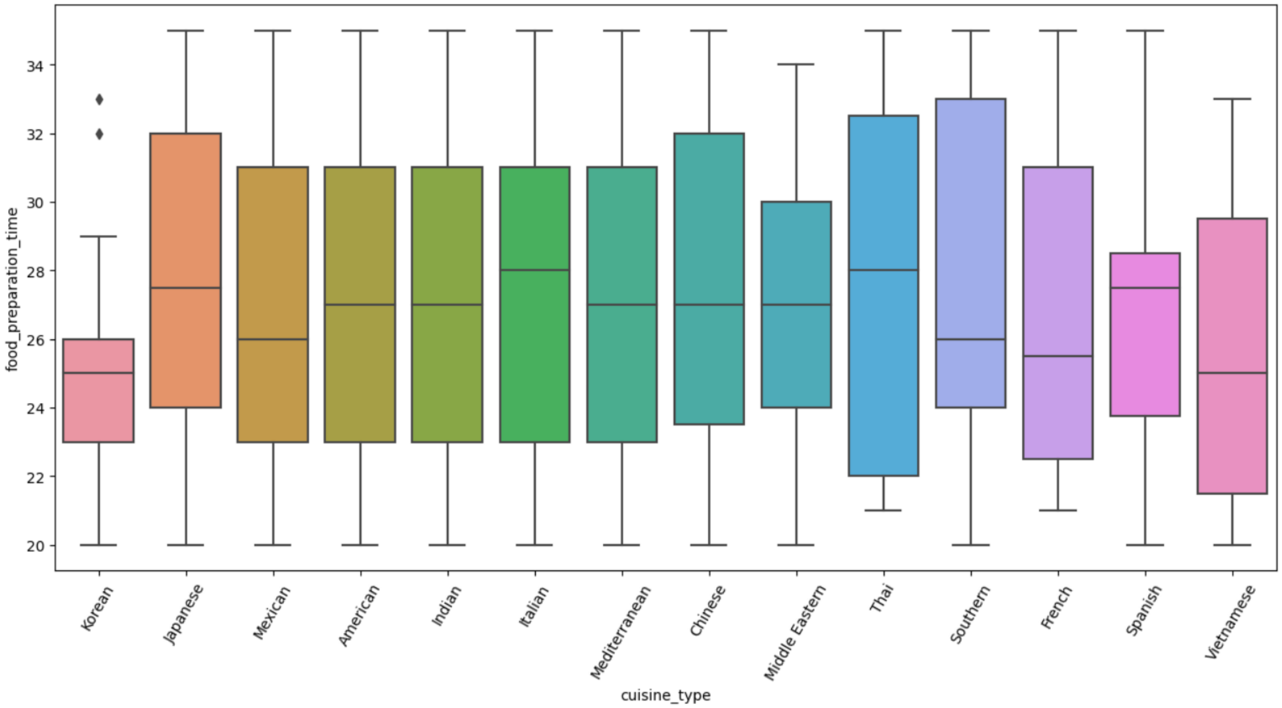
- Delivery time has a slight left skew with a mean time = 24.16 and median time = 25
- min delivery time = 15 and max time = 33
- The middle 50% of delivery times are within the 20 – 28 mins
- There are no outliers

Multivariate Analysis – cuisine v/s cost_of_the_order



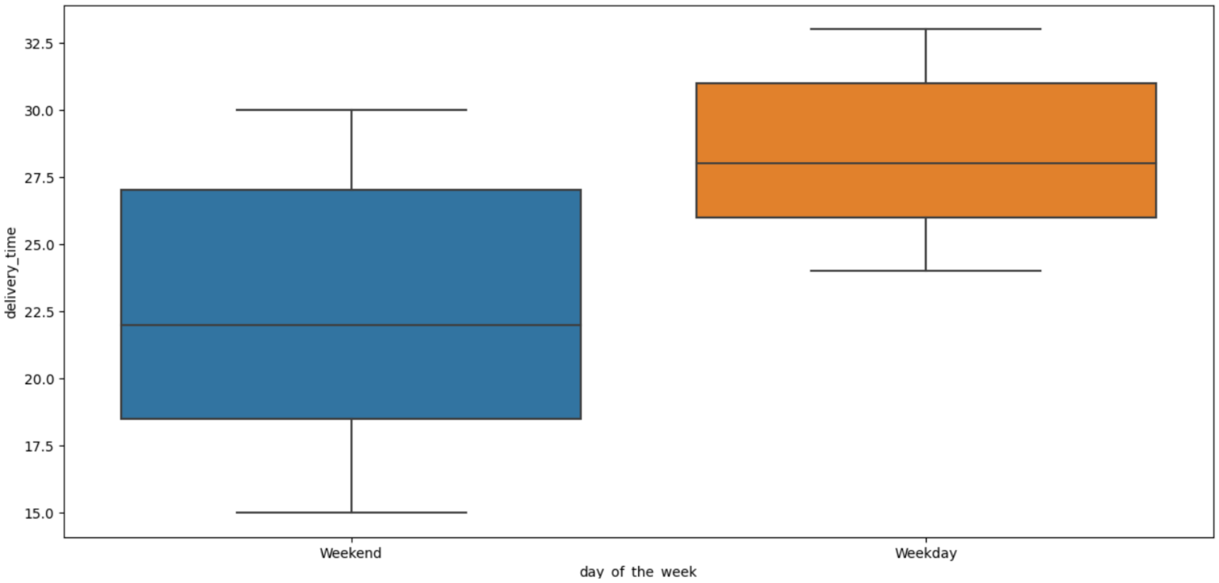
- cost_of_the_order for most cuisines is right skewed with the exception of Thai and French cusines, which indicates the average the cost of the order is higher than the mean cost. This means that most of these cuisines do have many high cost orders
- There are outliers for Korean, Mediterranean and Vietnamese cuisine
- Median cost of Vietnamese cuisine is lowest whereas French cuisine has highest median cost
- Chinese cuisine has a wide range of cost whereas Korean cuisine is less variant with the exception of some outliers

Multivariate Analysis – cuisine v/s total_preparation_time



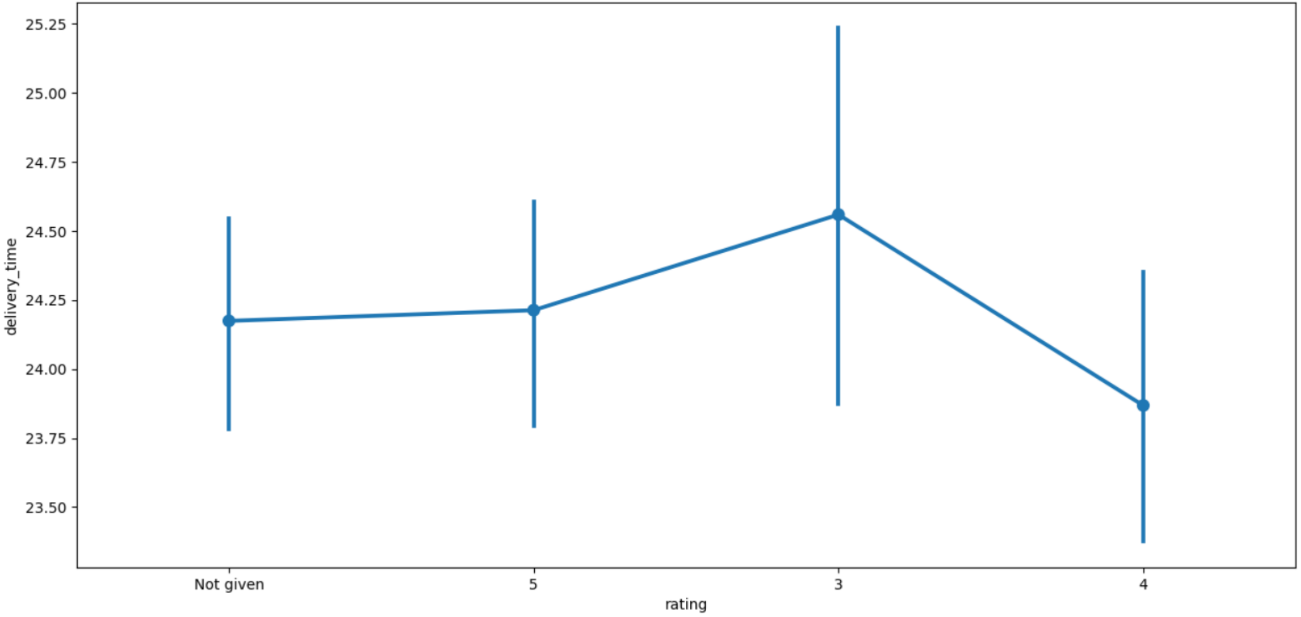
- Korean and Vietnamese cuisines have the lowest median preparation times of 25 mins
- Korean cuisine has the lowest mean
- Italian cuisine has the highest median preparation time of 28 mins
- Min food preparation time across all cuisines except for Thai and French is 20 mins. Thai and French cuisine the minimum time is 21 minutes
- Southern cuisine followed by Thai cuisine have the most spread out preparation times

Multivariate Analysis – day_of_the_week v/s delivery_time



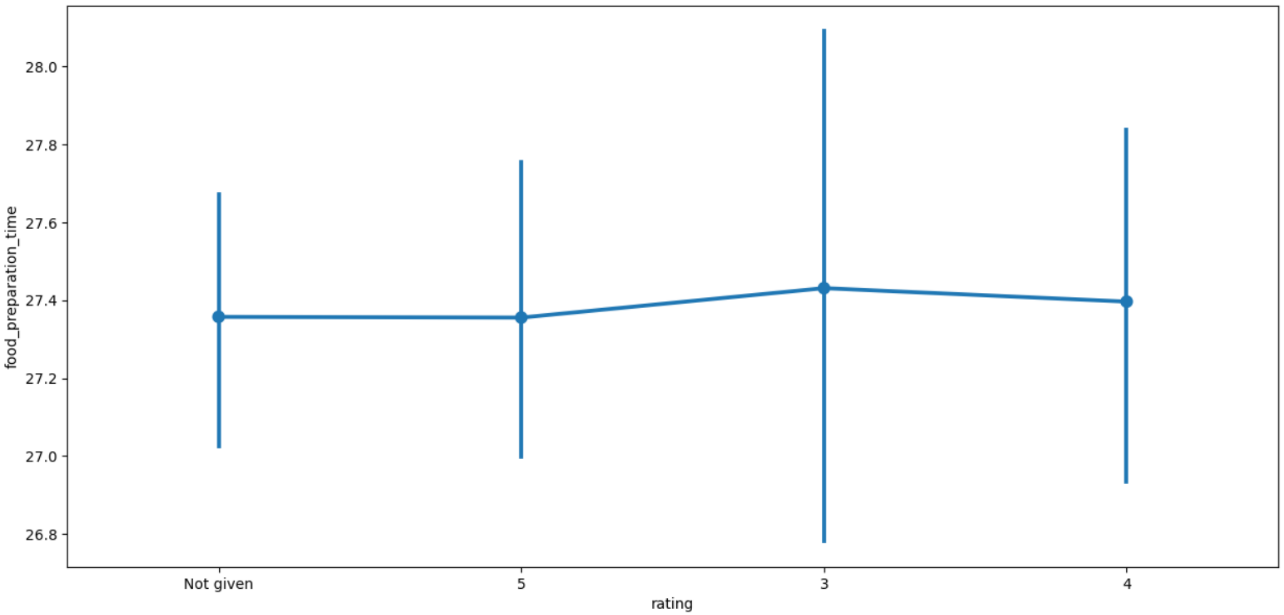
- The minimum, maximum, and median of weekday delivery times is higher than the same measures for the weekend.
- The spread of delivery times on weekdays is much smaller than the spread of delivery times of weekends.
- The data indicates much higher number of weekend orders than weekday hence more data points may need to be explored to understand the cause for longer delivery times on weekdays. It could be due to less delivery staff on weekdays, increased traffic during delivery due to rush hours etc

Multivariate Analysis – rating v/s delivery_time



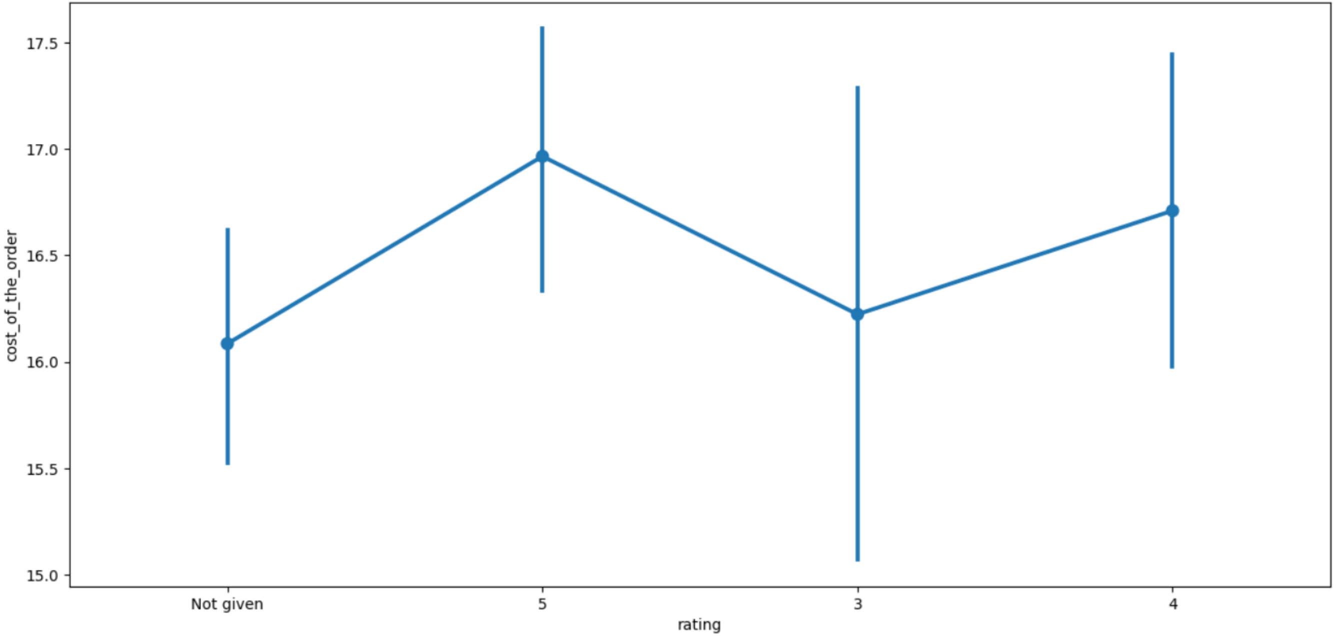
- Higher delivery times have a rating of 3
- Delivery time for orders rated 5 are higher than orders rated 4
- There does not seem to be any relationship between rating and delivery_time

Multivariate Analysis – rating v/s food_preparation_time



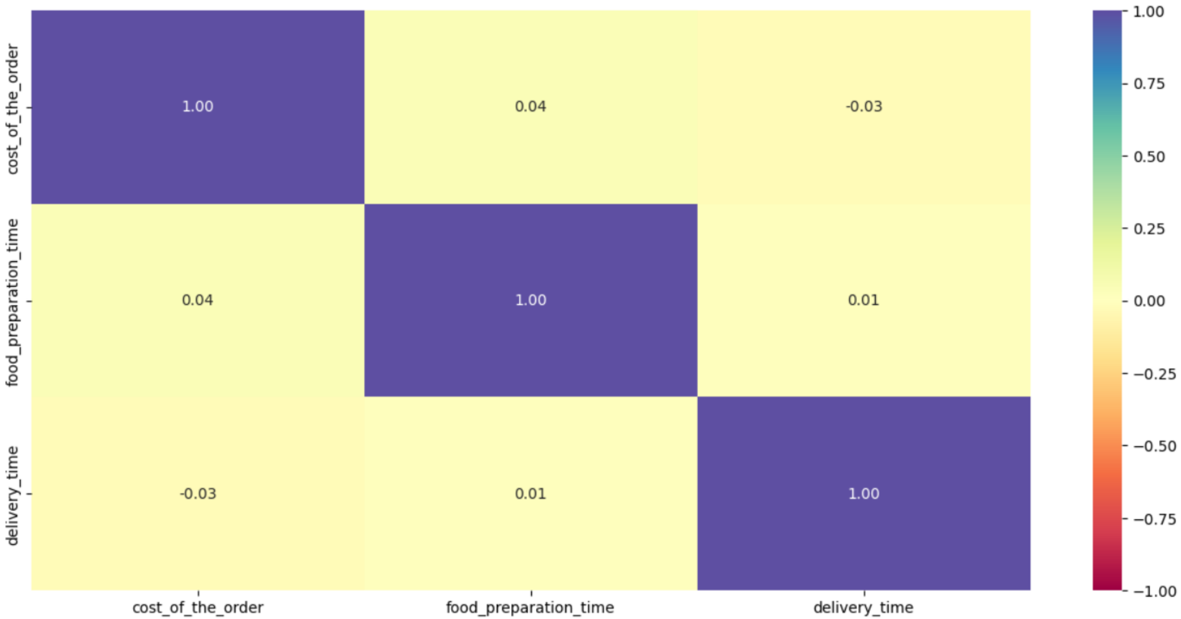
- Food preparation time has a slight negative correlation with rating
- As the food preparation time decreases the rating is higher

Multivariate Analysis – rating v/s cost_of_the_order



- Cost of order seems to have a slight positive correlation with rating

Multivariate Analysis – Correlation among variables



- The chart on the left depicts the correlation heatmap between the following variables
    - Cost_of_the_order
    - Food_preparation_time
    - Delivery_time
- Based on the chart it is very evident that there is extremely low between the above three variables
- The cost of the order has a slightly positive correlation with prep time, indicating that as the time to prepare food increases, the cost of the of the order increases slightly.
- The cost of the order has a slightly negative correlation with delivery time, indicating that as cost of the order increases, the delivery time decreases slightly

Multivariate Analysis – Top Restaurants by Revenue

restaurant_name	Revenue
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	834.5
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

- The table on the left depicts top 14 restaurants by revenue
- Shake Shack has the highest revenue amongst all the 178 restaurants that had the orders
- Foodhub can also negotiate revenue based commisions with restaurants e.g. higher commisions with top restaurants to increase revenue or promote lowere revenue restaurants to increase their sales and gain more revenue share

Multivariate Analysis – Restaurant Promo Offer

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.

The restaurants fulfilling the criteria to get the promotional offer were arrived by the following steps

- Identify all the restaurant orders that were rated and aggregate the data by the restaurant names to get a count of ratings for each restaurant. Sort the result in descending order of the number of ratings to get below
- From the subset of data above, filter out all restaurants that have less than 50 ratings
- Using the above dataset, calculate the average rating and sort by the average rating.
- All the restaurants with an average rating greater than 4 are eligible for promotional offer

restaurant_name	Count of rating
Shake Shack	133
The Meatball Shop	84
Blue Ribbon Sushi	73
Blue Ribbon Fried Chicken	64
RedFarm Broadway	41

These restaurants have more than 50 ratings

This restaurant was not further included in analysis as it has less than 50 ratings

List of restaurants eligible for promotional offer based on the criteria

restaurant_name	Average rating
The Meatball Shop	4.511905
Blue Ribbon Fried Chicken	4.328125
Shake Shack	4.278195
Blue Ribbon Sushi	4.219178

Multivariate Analysis

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. The net revenue generated by the company across all orders is calculated in below steps

- Calculate revenue for each order based on the cost_of_the_order applying appropriate percentage of commission
- Aggregate revenue across all the orders

Foodhub Net Revenue

$6166.30

The company wants to analyze the total time required to deliver the food. The percentage of orders that take more than 60 minutes to get delivered from the time the order is placed is calculated as below

- Calculate total time for each order based food_preparation_time and delivery_time
- Select the number of orders where total time is greater than 60 mins and calculate the percentage

# of Orders with Total Time > 60 mins

200

% of Orders with Total Time > 60 mins

10.54 %

The company wants to analyze the delivery time of the orders on weekdays and weekends. The mean delivery time varies during weekdays and weekends and is indicated on the right

Average Delivery Time on Weekdays

28 mins

Average Delivery Time on Weekends

22 mins

Great Learning

Happy Learning !