SMDM PROJECT REPORT

DSBA

Contents-

List of Figure-

Problem Statement-

The food aggregator company has stored the data of the different orders made by the registered customers in their online portal. They want to analyse the data to get a fair idea about the demand of different restaurants which will help them in enhancing their customer experience. Suppose you are hired as a Data Scientist in this company and the Data Science team has shared some of the key questions that need to be answered. Perform the data analysis to find answers to these questions that will help the company to improve the business.

Data Description-

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

Data Dictionary-

- order_id: Unique ID of the order
- customer_id: ID of the customer who ordered the food
- restaurant_name: Name of the restaurant
- cuisine_type: Cuisine ordered by the customer
- cost: Cost of the order
- day_of_the_week: Indicates whether the order is placed on a weekday or weekend(The weekday is from Monday to Friday and the Weekend is Saturday and Sunday)
- rating: Rating given by the customer out of 5
- food_preparation_time: Time (in minutes) taken by the restaurant to prepare the food. This is calculated by taking the difference between the timestamps of the restaurant's order confirmation and the delivery person's pick-up confirmation.
- delivery_time: Time (in minutes) taken by the delivery person to deliver the food package. This is calculated by taking the difference between the timestamps of the delivery person's pick-up confirmation and drop-off information .

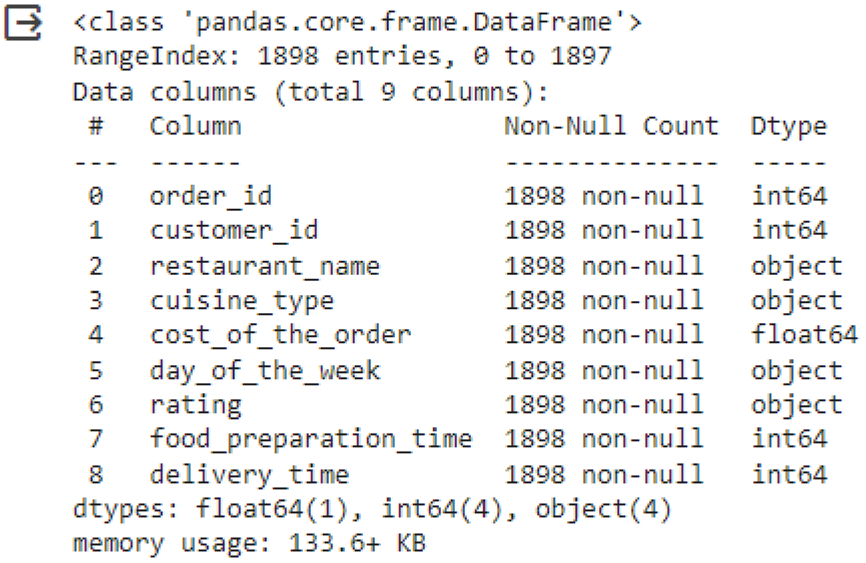
Data Overview-

Importing required libraries for data Manipulation.
Importing libraries for data visualisation.

1. How many rows and columns are present in the dataset?

In the dataset **(1898) rows and (19) columns** are present.

2. What are the datatypes of the different columns in the dataset?

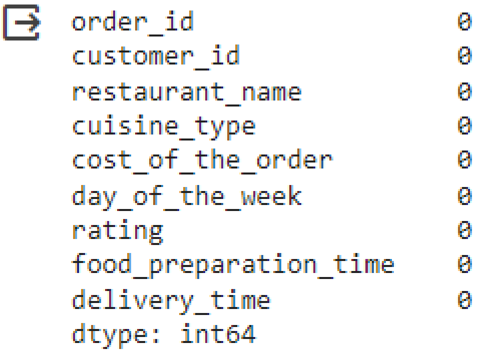


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1898 entries, 0 to 1897
Data columns (total 9 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   order_id               1898 non-null   int64  
 1   customer_id            1898 non-null   int64  
 2   restaurant_name        1898 non-null   object 
 3   cuisine_type           1898 non-null   object 
 4   cost_of_the_order      1898 non-null   float64
 5   day_of_the_week        1898 non-null   object 
 6   rating                 1898 non-null   object 
 7   food_preparation_time  1898 non-null   int64  
 8   delivery_time          1898 non-null   int64  
dtypes: float64(1), int64(4), object(4)
memory usage: 133.6+ KB
```

The dataset has **1898 instances and 19 attributes**. 1 float type, 4 integer type and 8 object type.

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

- There are **no null values** in any of the columns which is evident from the below result.

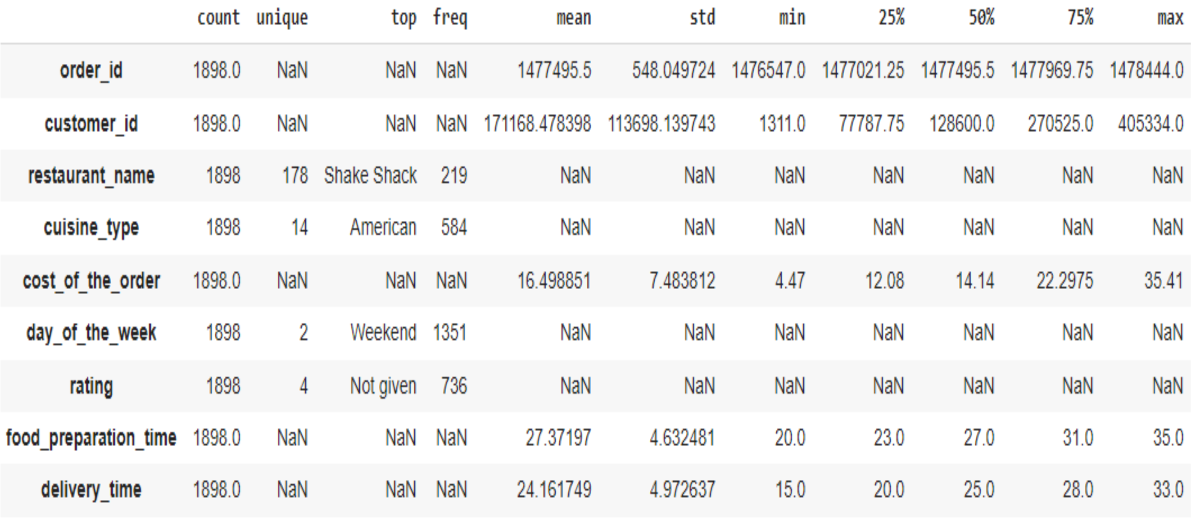


```
order_id                 0
customer_id              0
restaurant_name          0
cuisine_type             0
cost_of_the_order        0
day_of_the_week          0
rating                   0
food_preparation_time    0
delivery_time            0
dtype: int64
```

- Data set has **9 variables** ‘order_id’, ‘customer_id’, ‘restaurant_name’, ‘cuisine_type’, ‘cost_of_the_order’, ‘day_of_the_week’, ‘rating’, ‘food_preparation_time’, ‘delivery_time’.

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?

Statistical summary helps to know the various aspects like max,min values of all the columns,their mean and standard deviation.Their values at 25%,50% and 75% can also be determined from here. The Descriptive summary is shown 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

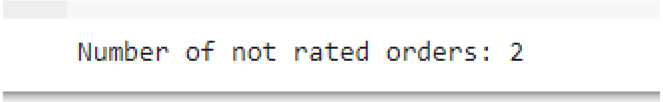
a) The minimum time it takes for food_preparation_time is 20.0

b) The average time it takes for food_preparation_time is 27.37197

c) The maximum time it takes for food_peparation_time is 35.0

5. How many orders are not rated?

Let’s check the total number of order in the given data are not rated-



```
Number of not rated orders: 2
```

So the output of the total number of **orders not rated is 2.**

Univariate Analysis-

6.Explore all the variables and provide observations on their distributions.(Generally, histogram,boxplot, countplot,etc.are used for Univariate exploration).

6.1. Order ID -

Number of unique **order ID 1898**.

6.2. Customer ID -

Number of unique **customer ID 1200.**

6.3. Restaurant Name -

Number of unique **restaurant names 178.**

6.4. Cuisine type -

Number of **unique cuisine type 14.**

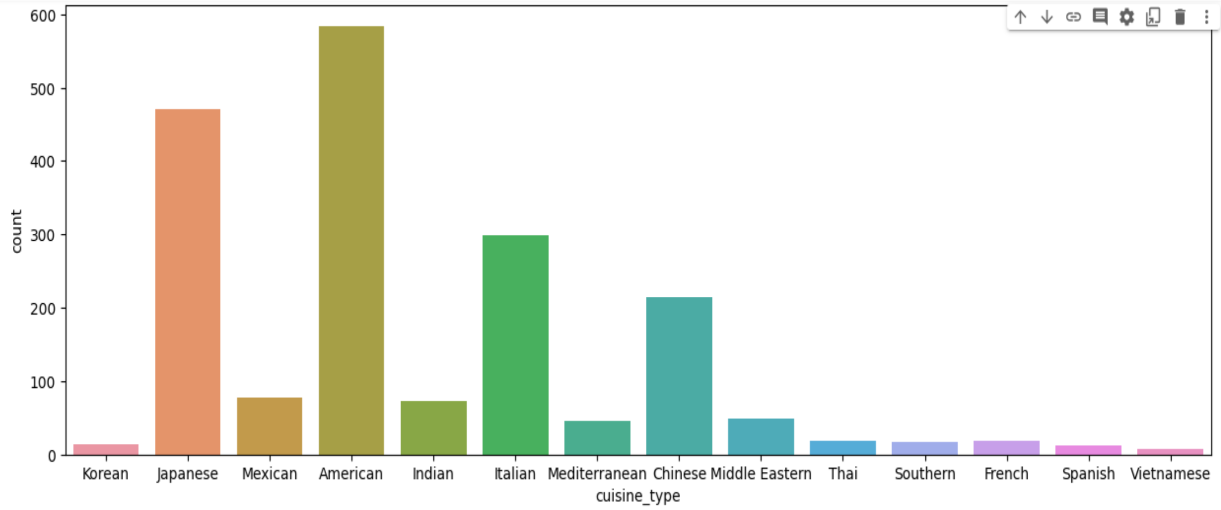


Fig.1 Countplot

From the above categorical plot below points can be inferred:

a) There are 14 unique cuisine_types in the dataset.
b) The most popular cuisine in the dataset is American cuisine_type.
c) The least cuisine in the dataset is Vietnamese cuisine_type.

6.5. Cost of the order-

Cost of the order is shown below in the fig2.1 and fig2.2

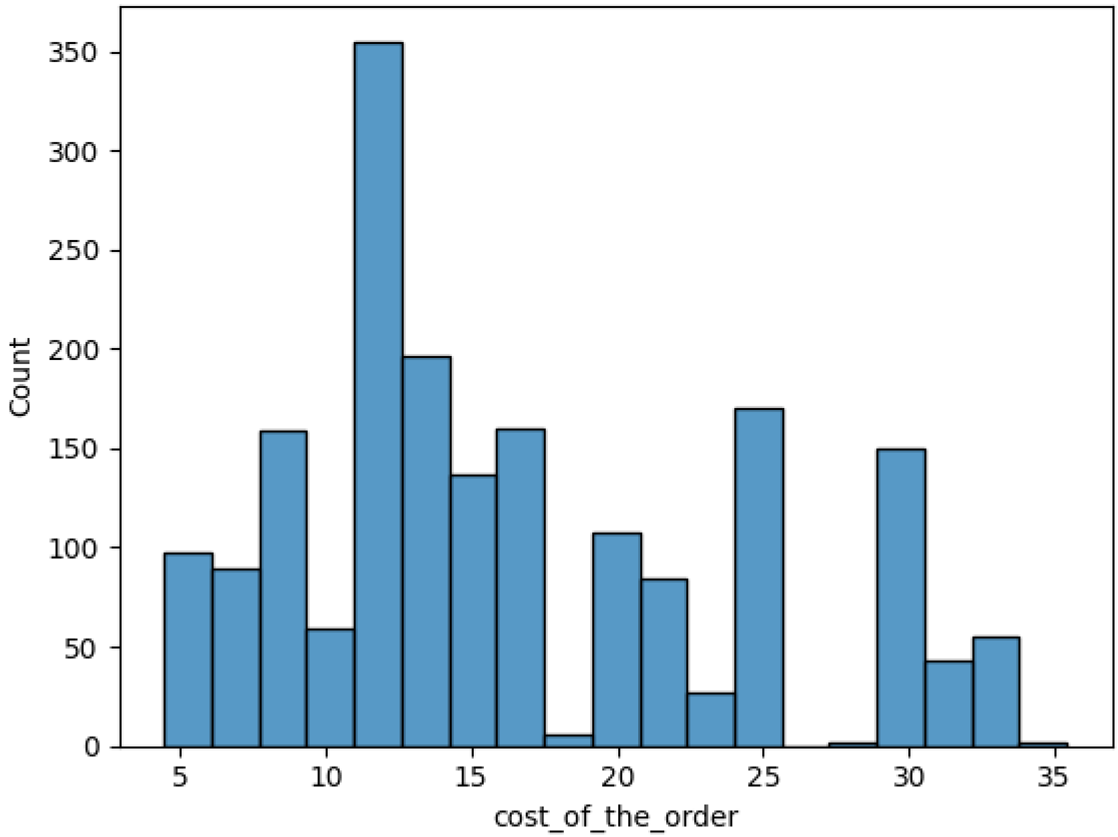


Fig.2.1Cost of the order vs Count Histogram

- From the above histogram for the continuous variables across all the regions it is evident that all the Cost of the order does not behave similarly across all the regions.
- There are a **total of 18 bins.** In the X-axis is Cost_of_the order and Y-axis is Count.
- As we see in the above graph, the **maximum cost of the order is about 35** something and the total count of the order is about 350.

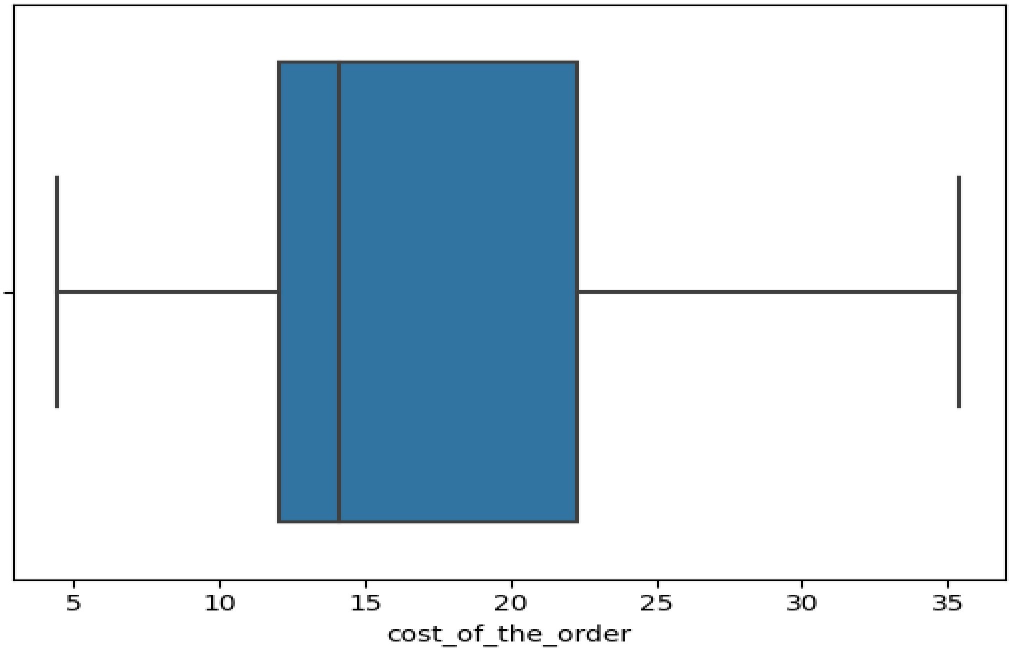


Fig.2.2 Cost of the order Boxplot

- In a box plot, when the median is closer to the left of the box and the whisker is shorter on the left end of the box, we say that the distribution is **positively skewed(skewed right).**
- Similarly, when the median is closer to the right of the box and the whisker is shorter on the right end of the box, we say that the distribution is **negatively skewed(skewed left).**
- As we see in the above boxplot the Q3 minus one and half time IQR is around 5 cost of the order and the Q3 plus one and half time IQR is somewhere around 35 cost of the order.
- So here the median of the cost of the order is somewhere around 14 or something like that.
- The Third quartile is somewhere around 23 cost of the order and the First quartile is somewhere around 13 cost of the order.

6.6. Day of the week-

Unique values for the day of the week column is 2.

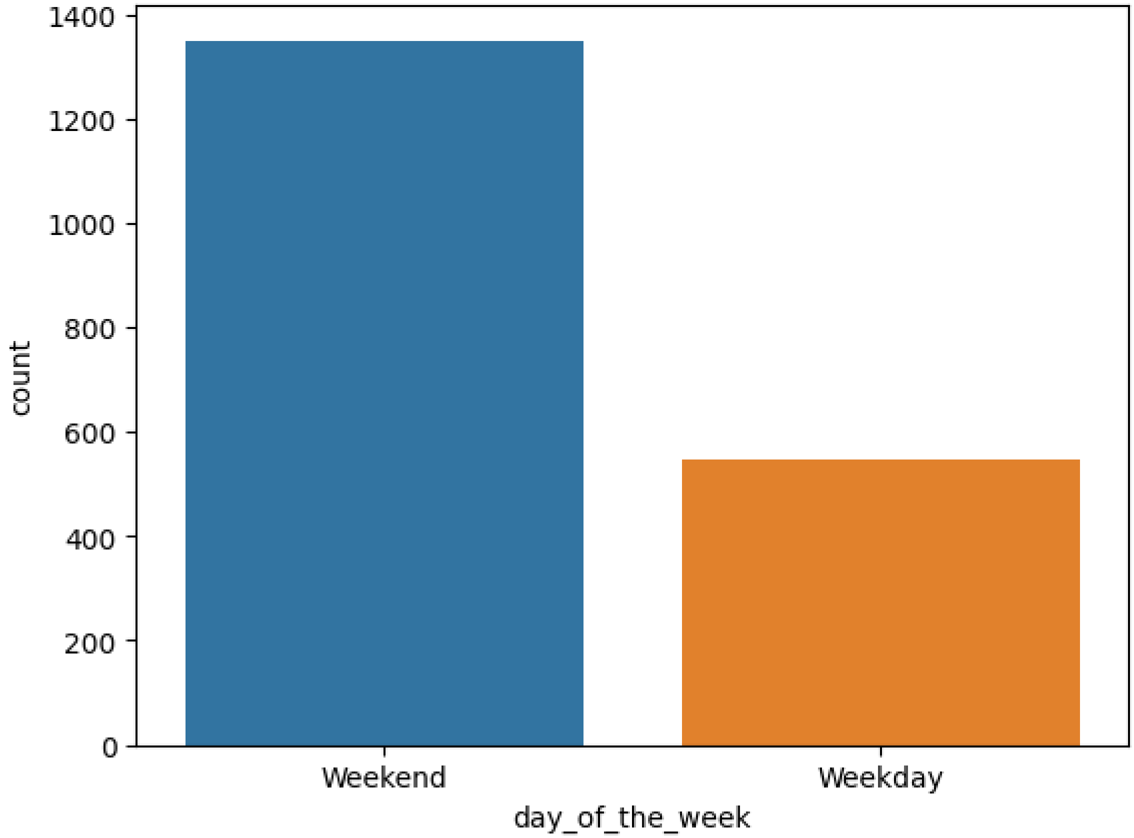


Fig.3 Day_of_the_week vs Count Bar Graph

We will calculate the count of order based on the day_of_the_week. From the above plot, we can see that the order with **Weekend day_of_the_week has the highest count** and the order of **Weekday day_of_the_week is lowest count.**

6.7. Rating-

As shown in the below fig.4 rating vs count countplot-

- There are a total of 4 rating columns in the dataset **'Not given', '5', '3', and '4'.**
- The highest rating count for 'Not given' is somewhere around 700 or something like that.
- The lowest rating count for '3' is somewhere around 190 or something like that.
- The rating count for '5' is somewhere around 590 or something like that.
- The rating count for '4' is somewhere around 480 or something like that.

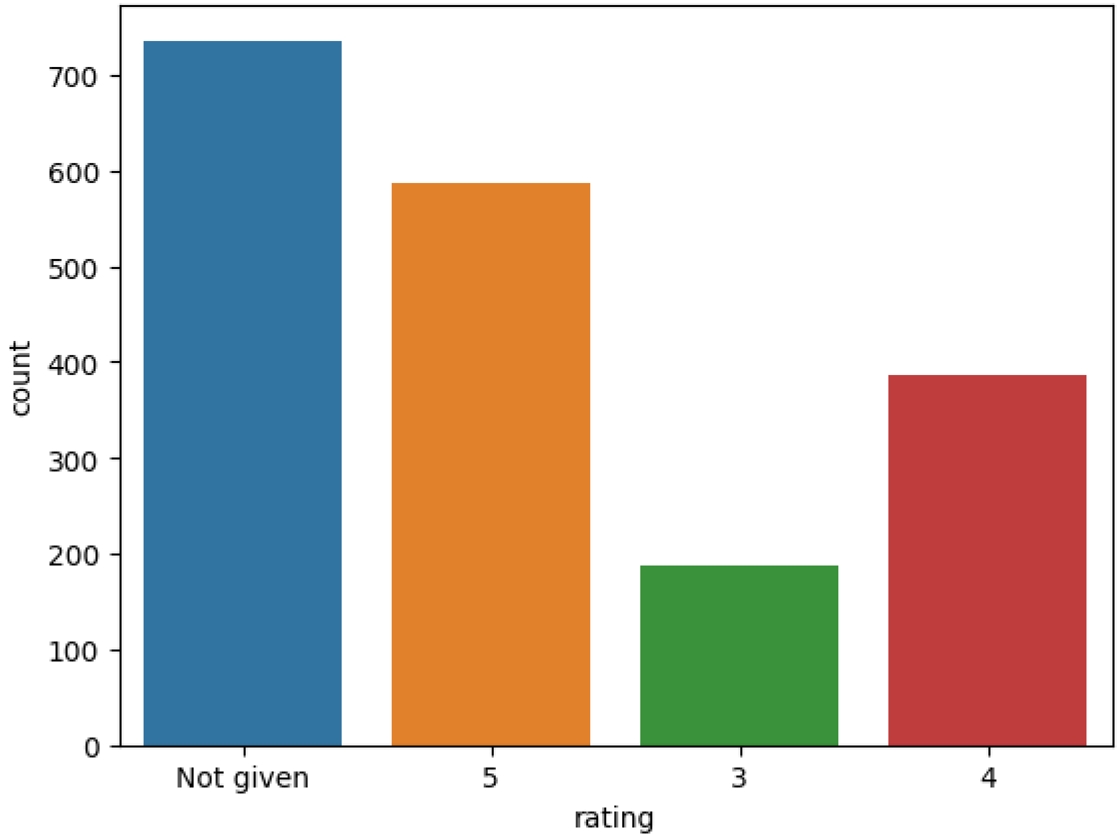


Fig.4. Rating vs Count Bar Graph

6.8. Food preparation time-

- From the below fig.5.1 food_preparation_time vs count histogram for the continuous variable across all regions it is evident that all the count order of food preparation time does not behave similarly across all the regions as shown in the below graph.
- In the X-axis is the food_preparation_time and Y-axis is count.
- As we see in the graph below, the count of the maximum food preparation time is about 250 or something like that.
- The maximum food preparation time is about 34 minutes or something like that.

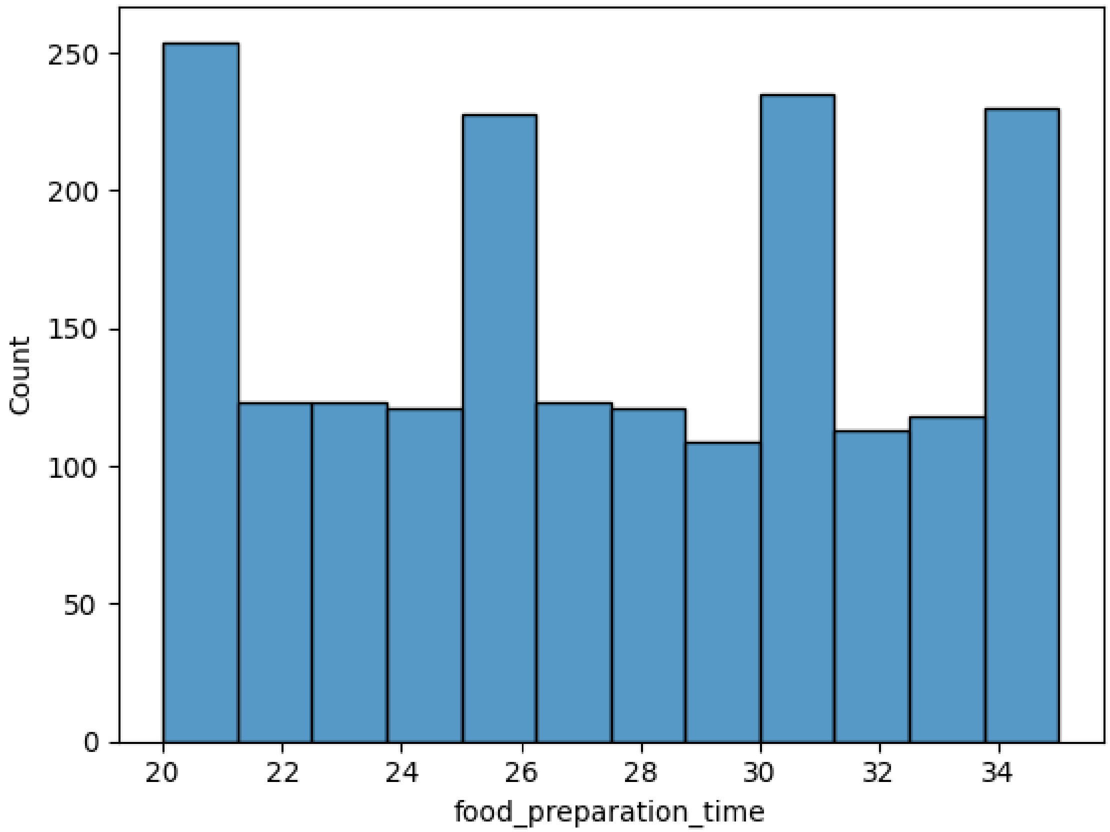


Fig.5.1 Food_preparation_time vs Count Histogram

- From the below fig.5.2 food_preparation_time boxplot , the Q3 minus one and half time IQR **is around 20 food preparation time** and the Q3 plus one and half time IQR is somewhere around **35 food preparation time** or something like that.
- So here the **median of the food preparation time** is somewhere around **27 or something like that**.
- The **Third Quartile is somewhere around 31** food preparation time and the **first quartile is somewhere around 23** food preparation time or something like that.

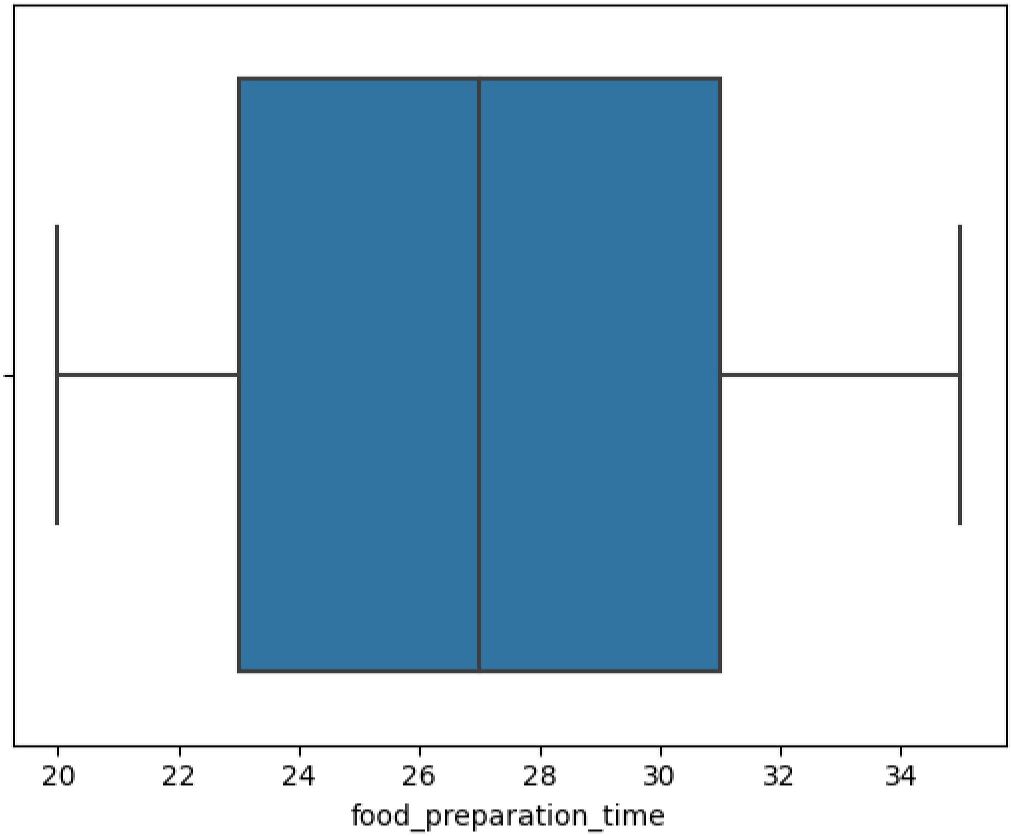


Fig.5.2 Food_preparation_time Boxplot

6.9. Delivery time-

- From the below fig.6.1 delivery time vs count histogram for the continuous variables across all the regions it is evident that all the delivery time does not behave similarly across all the regions.
- There are a total **14 bins displayed on the graph** . In the X-axis is delivery_time and the Y-axis is count.
- As we see in the below graph, the **maximum delivery time is about 32.5** or something like that and the **maximum count for the delivery time is about 300** or something like that.
- Similarly, **the minimum delivery time is about 15.0** or something like that and the **minimum count for the delivery time is about 48** or something like that.

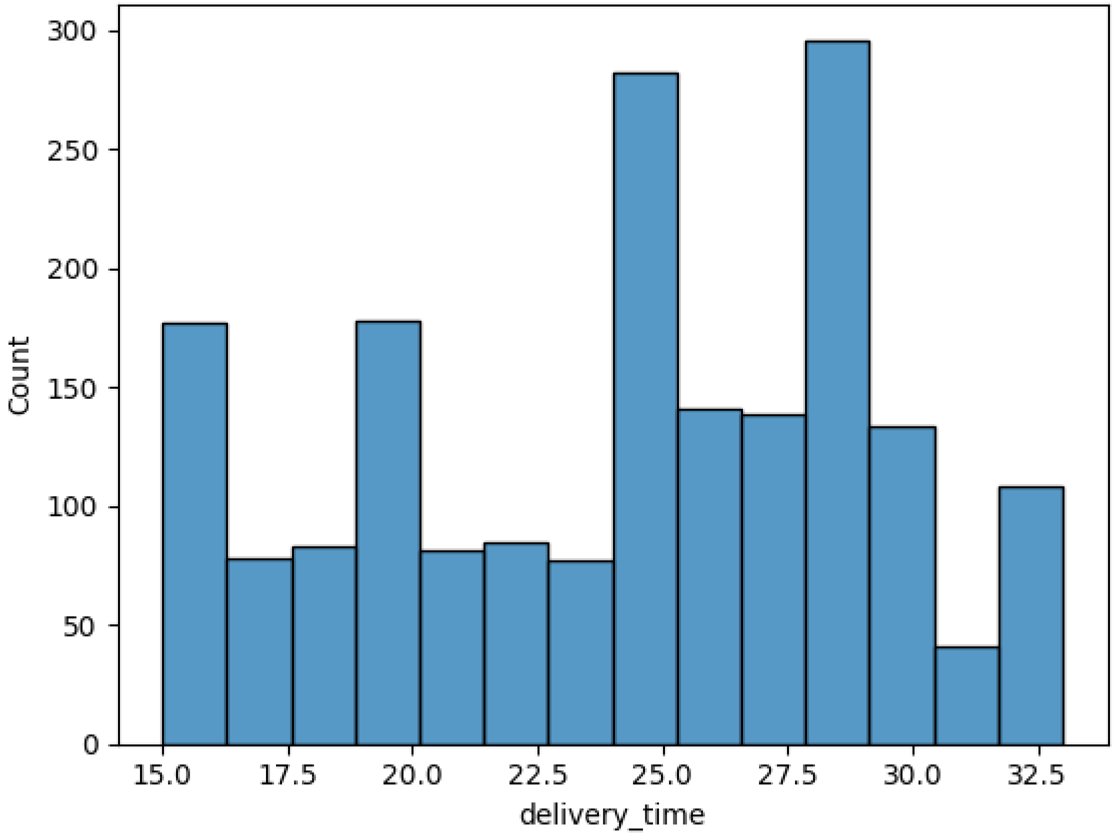


Fig.6.1 Delivery_time vs Count histogram

- From the below fig.6.2 delivery time boxplot ,as we know when the median is closer to the left of the box and the whisker is shorter on the left end of the box, we say that the distribution is positively skewed.
- Similarly, when the median is closer to the right of the box and the whisker is shorter on the right end of the box, we say that the distribution is negatively skewed.
- As we see in the below box plot **the Q3 minus one and half IQR is around 15.0** delivery time and **the Q3 plus one and half IQR is somewhere around 32.5** or something like that.
- So here the **median of the delivery_time is somewhere around 25.0** or something like that.
- The **third quartile is somewhere around 27.5** delivery time or something like that and the **first quartile is somewhere around 20.0** or something like that.

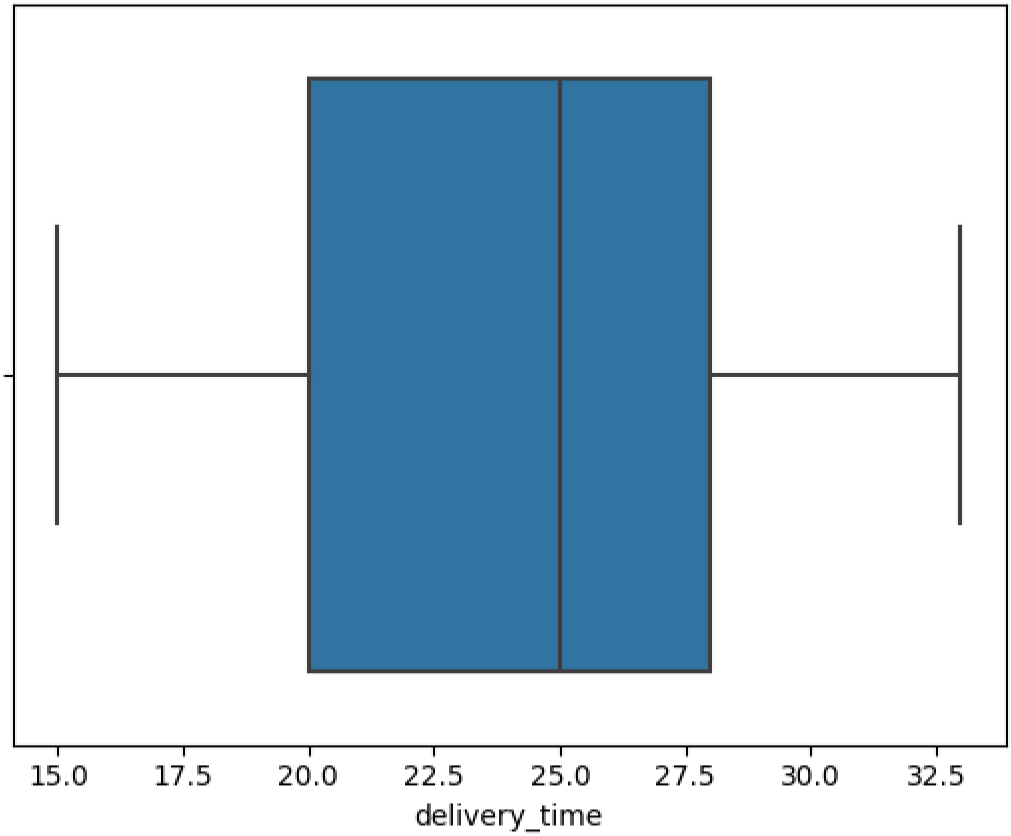
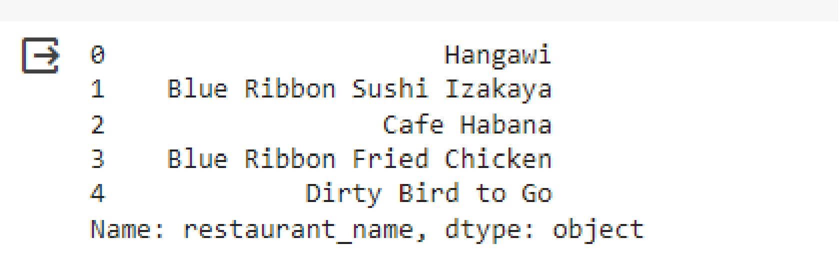


Fig.6.2 Delivery_time Boxplot

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

- The top 5 restaurants in the terms of the number of orders received is-



```
0                      Hangawi
1    Blue Ribbon Sushi Izakaya
2                  Cafe Habana
3    Blue Ribbon Fried Chicken
4              Dirty Bird to Go
Name: restaurant_name, dtype: object
```

- Dataset has the top 5 restaurants in terms of the number of orders received: **'Hangawi', 'Blue Ribbon Suzi Izakaya', 'Cafe Habana', 'Blue Ribbon Fried Chicken', 'Dirty Bird to Go'.** Data Type is Object.

8. Which is the most popular cuisine on weekends?

As shown in the below graph the most popular cuisine type on weekends is **American** and the unique values for the cuisine type on weekends is **14**.

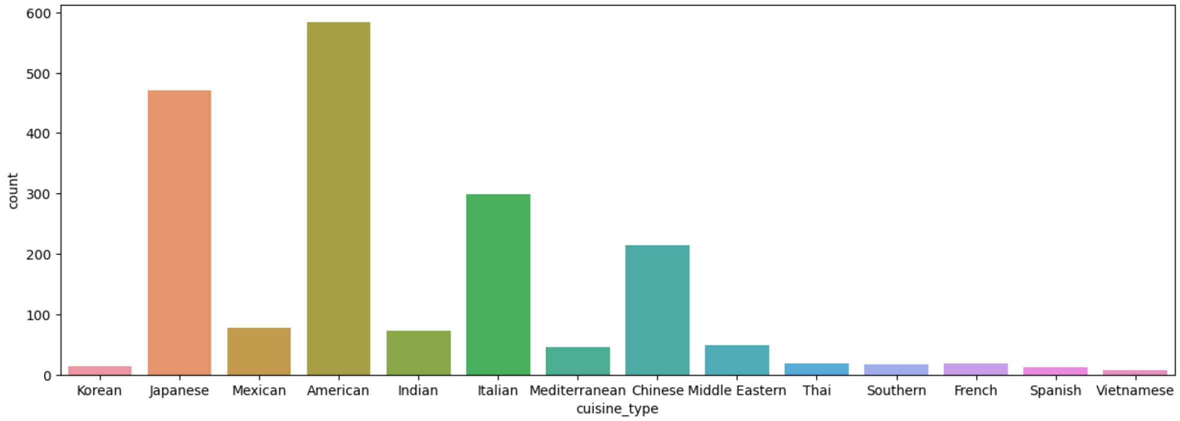
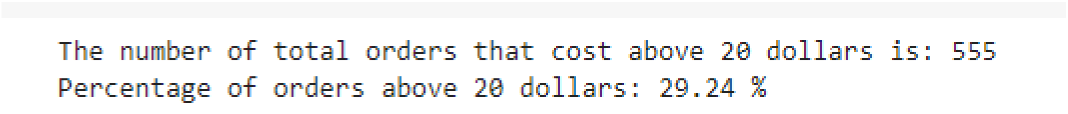


Fig.7 Cuisine_type vs Count Barplot

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

To get a percentage that costs above 20 dollars, Firstly we write an appropriate column name to get the orders above $20. After that we calculate the number of total orders where the cost is above $20 and then we calculate the percentage of such orders in the dataset. So we get the number of total orders that cost above 20 dollars and the percentage of orders above 20 dollars.

So we get the following output-

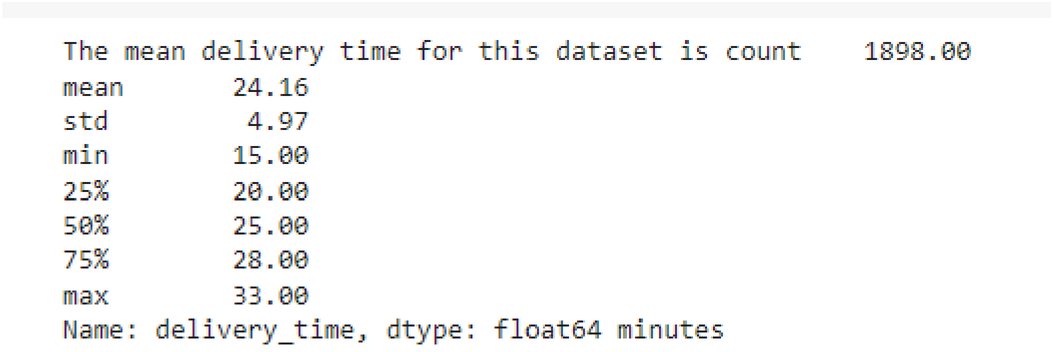


```
The number of total orders that cost above 20 dollars is: 555
Percentage of orders above 20 dollars: 29.24 %
```

10. What is the mean order delivery time?

To get the mean delivery time we write an appropriate function to obtain the mean delivery time.

So we get the following output-

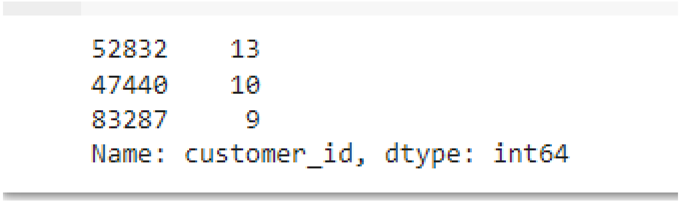


```
The mean delivery time for this dataset is count    1898.00
mean        24.16
std          4.97
min         15.00
25%         20.00
50%         25.00
75%         28.00
max         33.00
Name: delivery_time, dtype: float64 minutes
```

From the above descriptive statistics we can see that the mean delivery time for this dataset is **count 1898.00** and the **mean order delivery time is 24.16**

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.

To get the counts of each customer_id we need to print the IDs and order counts of top 3 customers we get the following output-



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

So the top 3 customer_id is: **52832, 47440, 83287**

Number of orders is: **13,10,9**

Data type is : **integer**

Multivariate Analysis -

12. Perform a multivariate analysis to explore the relationship between the important variables in the dataset.(It is a good idea to explore the relations between numerical variables as well as relations between numerical and categorical variables).

12.1. Cuisine vs Cost of the order -

From the below fig.8 Cuisine_type vs Cost_of_the_order boxplot we can see that-

- In the dataset the outliers present only three boxplot regions : **'Korean', 'Mediterranean', and 'Vietnamese'.**
- The top cuisine type in the dataset is **'American'.**
- The least cuisine type in the dataset is **'Vietnamese', and 'Korean'.**
- The maximum cost of the order for cuisine type is somewhere around **35.0 or something like that.**

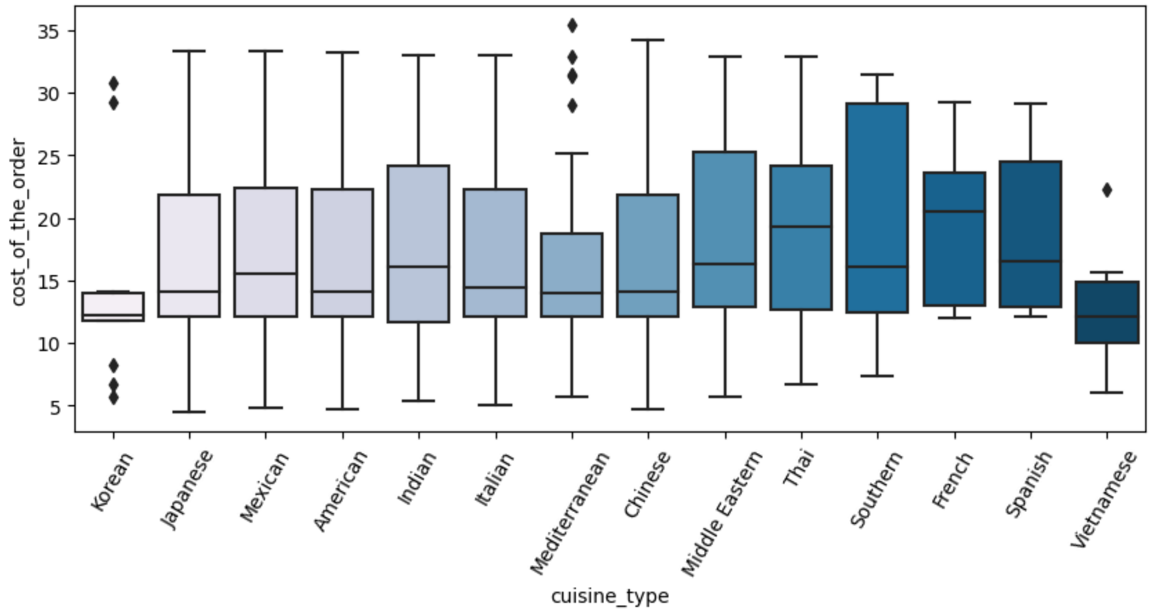


Fig.8. Cuisine_type vs Cost_of_the_order Boxplot

12.2. Cuisine vs Food preparation time -

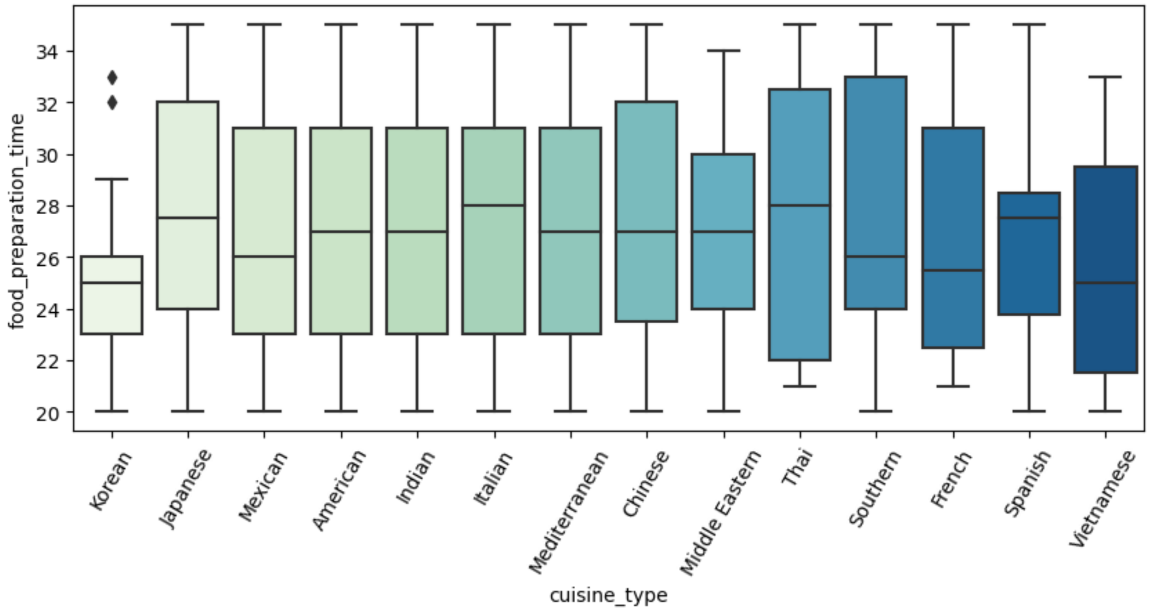


Fig.9 Cuisine_type Vs Food_preparation_time Boxplot

From the above fig.9 Cuisine_type vs Food_preparation _time boxplot we can see that-

- The maximum food_preparation_time is **somewhere around 35** or something like that.

- The mean of the food_preparation_time is **somewhere around 27.3** or something like that.
- Also we can see that 75% food_preparation_time is somewhere around **31.0 or something like that**.
- There are **14 unique cuisine_type** shows in the graph above.

12.3. Day of the week vs Delivery time -

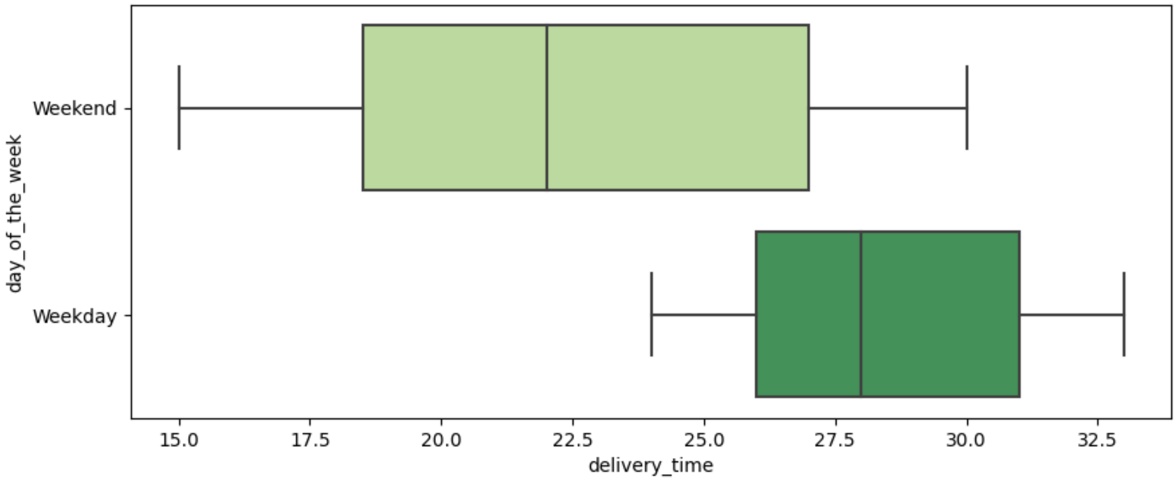


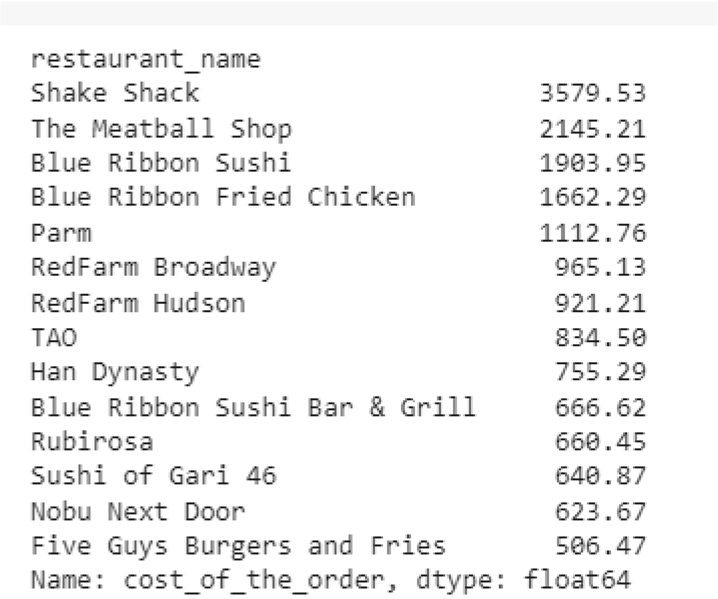
Fig.10 Delivery_time Vs Day_of_the_week Boxplot

From the above fig.10 delivery_time vs day_of_the_week boxplot we can see that-

- The delivery_time of Weekday is maximum **somewhere around 33.0** or something like that .
- The delivery_time of the **Weekend is less compared to Weekday delivery_time.**
- The median for the Weekday delivery_time is **somewhere around 28.0** or something like that.
- The median for the Weekend delivery_time is **somewhere around 21.0** or something like that.

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

The observations on the revenue generated by the restaurants is-



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

We can see the **highest revenue generated by the restaurant is 'Shake Shack' which is 3579.53** and the **least revenue generated by the restaurant is 'Five Guys Burgers and Fries' which is 506.47** . Data Type is Float.

12.5. Rating Vs delivery time -

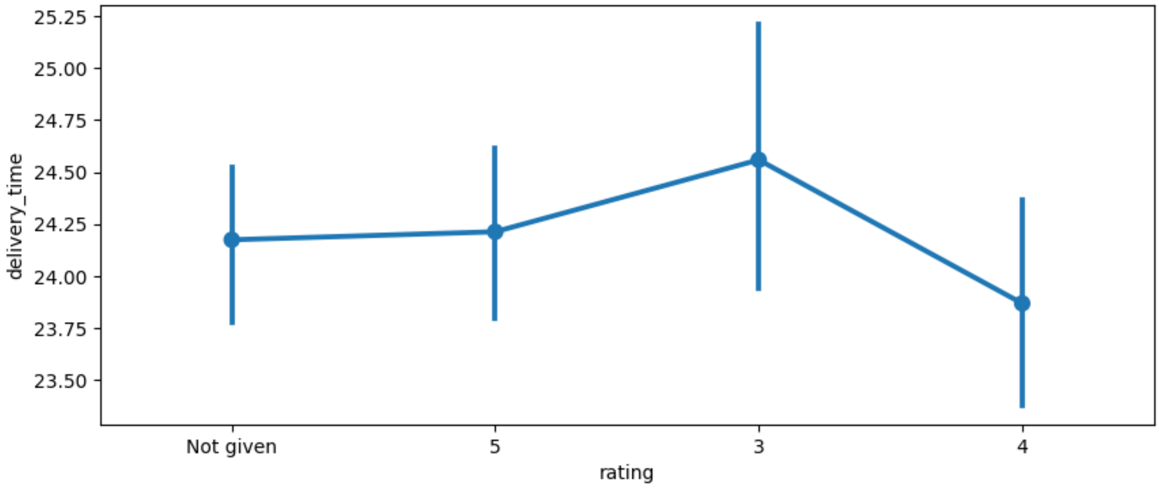


Fig.11 Rating vs Delivery_time pointplot

- As we can see in the above fig.11 in the rating 3 delivery_time with a **24.50 confidence interval somewhere between 23.95 and 25.25.**
- Similarly, in the rating 4 delivery_time with a **23.85 confidence interval somewhere between 23.35 and 24.25 .**
- Similarly for rating 'Not given' and rating '5' we can see the relationship rating vs delivery_time.

12.6. Rating Vs Food preparation time -

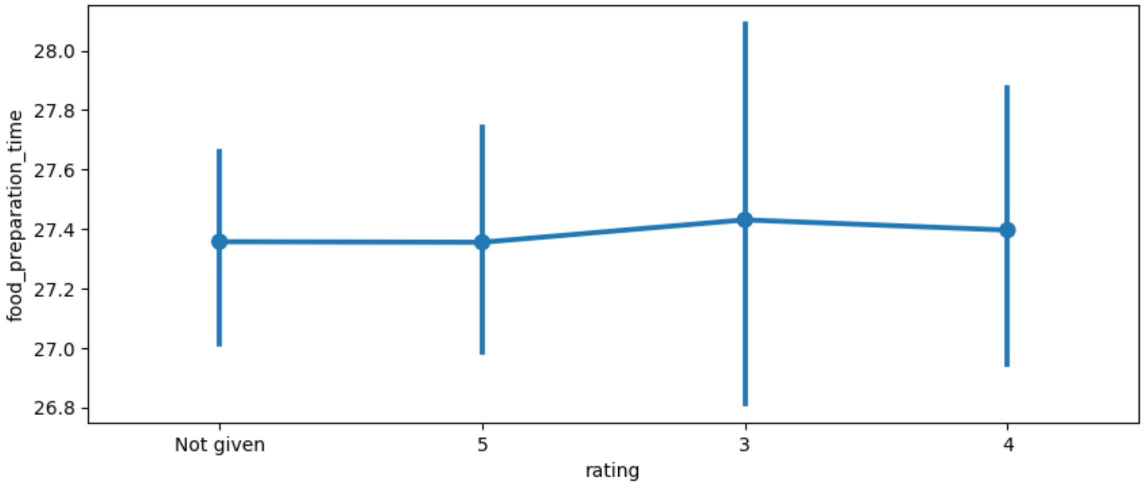


Fig.12 Rating vs Food_preparation_time pointplot

- As we can see in the above fig.12 in the rating ‘3’ food_preparation_time **with a 27.4 confidence interval somewhere between 26.8 and 28.5** or something like that.
- Similarly, in the rating ‘4’ food_preparation_time **with a 27.3 confidence interval somewhere between 26.9 and 27.8** or something like that.
- Similarly, for rating ‘Not given’ and rating ‘3’ we can see the relationship rating vs food_preparation_time.

12.7. Rating Vs Cost of the order -

- As we can see in the below fig.13 in the rating ‘5’ the cost_of_the_order **with a 17.0 confidence interval somewhere between 16.4 and 17.5** or something like that.
- Similarly, in the rating ‘3’ cost_of_the_order **with a 16.4 confidence interval somewhere between 15.0 and 17.4** or something like that.
- Similarly, for rating ‘Not given’ and rating ‘5’ we can see the relationship rating vs cost_of_the_order.

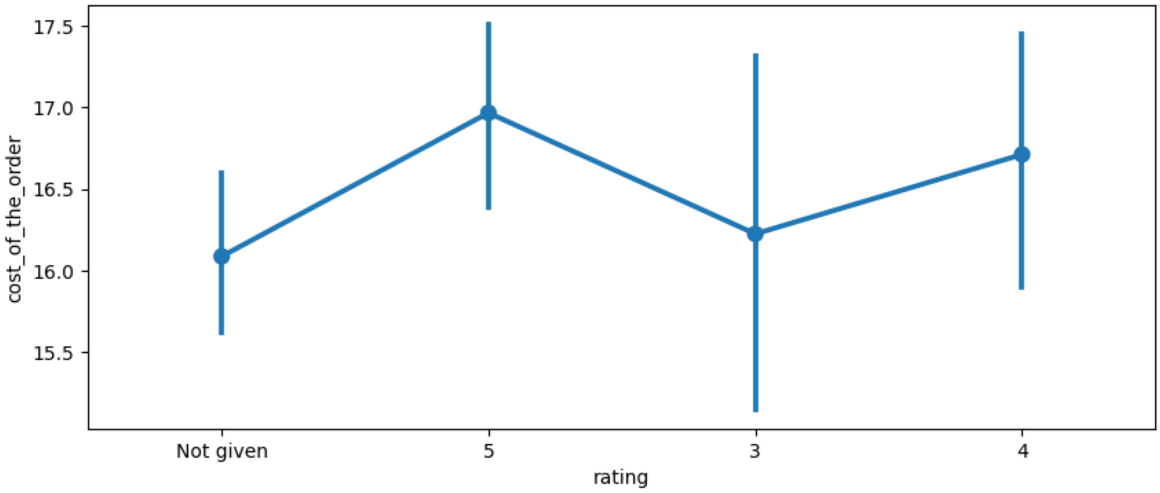


Fig.13 Rating vs Cost_of_the_order Pointplot

12.8. Correlation among the variables -

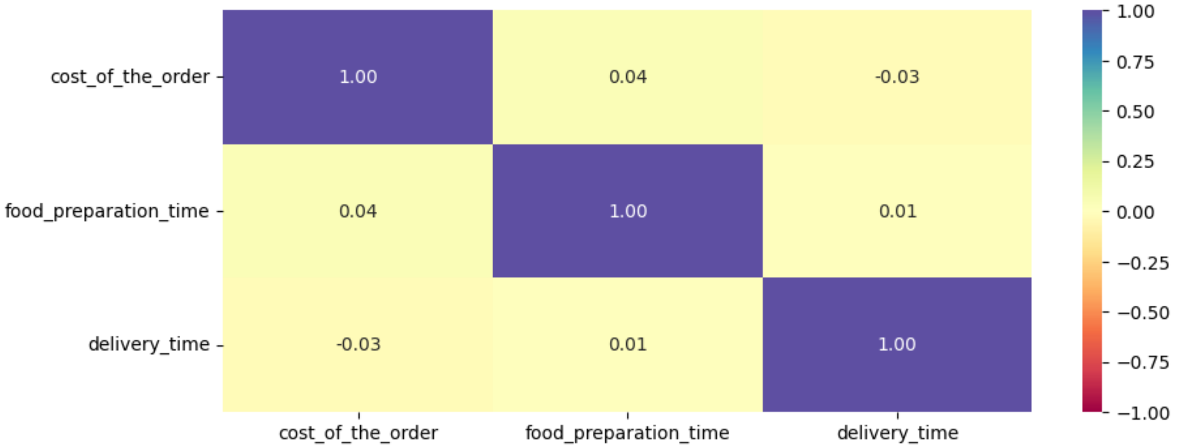


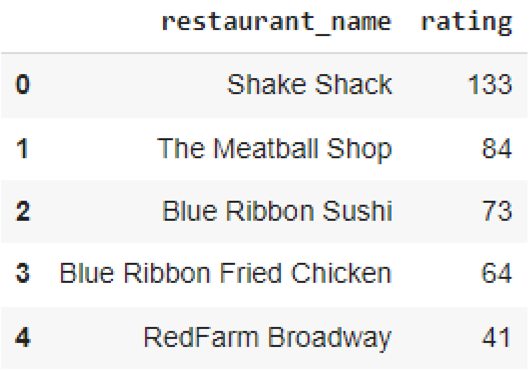
Fig. 14 Correlation Plot

From the correlation plot, we can see that

- The correlation between cost_of_the_order and food_preparation_time is **0.04.**
- The correlation between Cost_of_the_order and delivery_time is **-0.03.**
- The correlation between food_preparation_time and delivery_time is **0.01.**

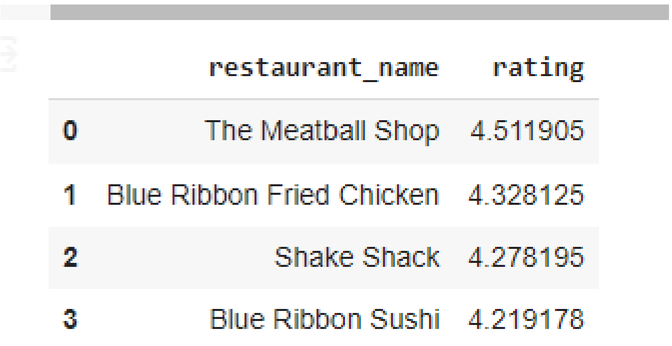
13. The company wants to provide a promotional offer in the advertisement of the restaurants. The condition to get the offer is that the restaurants must have a rating count of more than 50 and the average rating should be greater than 4. Find the restaurants fulfilling the criteria to get the promotional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

Firstly, filter the rated restaurants and convert the rating column from object to integer. After that create a dataframe that contains the restaurant names with their rating counts. We get output restaurant name and rating counts-



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

To get the restaurant names that have rating counts more than 50. Again filter to get the data of restaurants that have rating counts of more than 50 and the average rating should be greater than 4. Finally we got the output to find the mean rating.

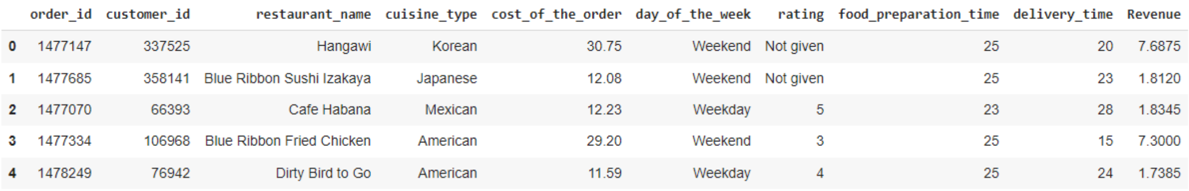


	restaurant_name	rating
0	The Meatball Shop	4.511905
1	Blue Ribbon Fried Chicken	4.328125
2	Shake Shack	4.278195
3	Blue Ribbon Sushi	4.219178

As we can see, **the top rated restaurant_name is 'The Meatball Shop'.**

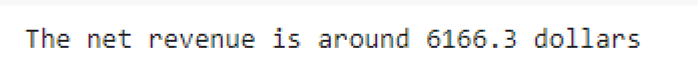
14. The company charges the restaurant 25% on the order having cost greater than 20 dollars and 15% on the order having cost greater than 5 dollars. Find the net revenue generated by the company across all orders.

We can see in the below output, Firstly we write an appropriate column name to compute the revenue. The **highest revenue was 7.6875 day_of_the_week** Weekend and the **lowest revenue was 1.7385 day_of_the_week** Weekday.



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

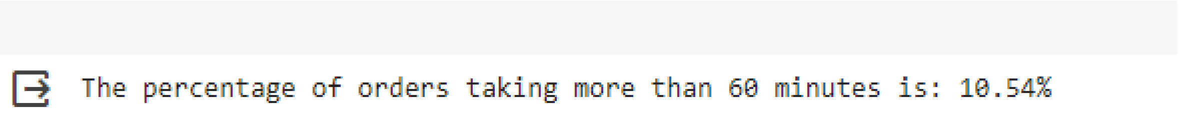
The net revenue generated by **the company across all orders is around 6166.3$**.



```
The net revenue is around 6166.3 dollars
```

15. The company wants to analyse 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).

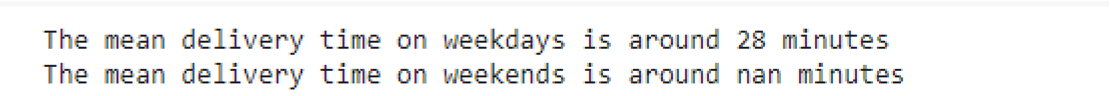
Firstly, we calculate the total delivery time and add a new column to get the data frame to store the total delivery time. After this we calculate the percentage of orders that take more than 60 minutes. So we get the **percentage of orders taking more than 60 minutes is 10.54%.**



```
The percentage of orders taking more than 60 minutes is: 10.54%
```

16. The company wants to analyse the delivery time of the orders on weekdays and weekends . How does the mean delivery time vary during weekdays and weekends?

For each group(Weekdays and Weekend), we calculate the mean delivery time. Firstly we separate the data into two groups: Weekdays and Weekends. We can use day of the week and delivery time information to determine which category each order falls into.



```
The mean delivery time on weekdays is around 28 minutes
The mean delivery time on weekends is around nan minutes
```

So we can see that in the above output the mean delivery time on weekdays is around **28 minutes** and the mean delivery time on weekends is around **NAN minutes.**

17. What are your conclusions from the analysis? What recommendations would you like to share to help improve the business?(You can use cuisine type feedback rating to drive your business recommendations).

Conclusion and Recommendations-

- We can see that FoodHub’s online food delivery service **has the potential for growth and success.**
- The most **popular cuisine_type is American**, which is on the weekend.
- There are a **total four rated orders** and **two not rated orders** . The company enhances its operations, maintains food quality and meets the evolving needs of its customers in the dynamic restaurant industry of America.
- **Analyse customer feedback and rating to identify issues** with restaurants delivering the overall services.

THE END…