

FoodHub Data Analysis

PGP-DSBA _ FoodHub Project

Date 14th September 2023



Contents / Agenda

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

Executive Summary: Conclusion (1/11): Intro

This analysis was carried out on data collected from 1898 online food delivery orders made by 1200 clients in New York via the food aggregator company, FoodHub. Altogether, 14 cuisine types were ordered from 178 restaurants. A total of 9 attributes were recorded and analyzed. The principal attributes under scrutiny are the cuisine type and feedback ratings. In an effort to provide actionable information to FoodHub seeking to ameliorate customer experience and improve business performance, we uncovered the relationship between these two main parameters and the other attributes.

Executive Summary: Conclusion (2/11): Volume of Orders

- 'American' was the most frequently ordered cuisine type (about 580 orders, 30.56%) followed by 'Japanese' (about 480, 25.29%) and 'Italian'(about 280 orders, 14.75%)
- 'Vietnamese' cuisine type was the least frequently ordered (less than 20 orders, 1.05%)
- The number of orders received during the weekend is over twice greater than those received during weekdays. Probably, people tend to use office and school restaurants during weekdays and, therefore, don't need to order food online.
- The top 5 restaurants in terms of the number of orders in descending number of orders are 'Shake Shack' (219 orders, 11.54%), 'The Meatball Shop' (132 orders, 6.95%), 'Blue Ribbon Sushi' (119 orders, 6.27%), 'Blue Ribbon Fried Chicken' (96 orders, 5.06%), and 'Parm' (68 orders, 3.58%)
- The IDs of the top 3 most frequent customers who will, therefore, be benefitting from the discount are 52832 (13 orders, 0.68%), 47440 (10 orders, 0.53%), and 83287 (9 orders, 0.47%)
- All 14 cuisine types were ordered during the weekend

Executive Summary: Conclusion (3/11): Food Preparation Time

- It takes between 20 and 35 minutes to prepare food after an order is placed, and 27 minutes on average
- The orders received are fairly distributed across food preparation times with peaks at 21, 26, 31, and 34 minutes, meaning relatively large numbers of orders registered preparation times around these peaks
- Most orders are prepared between 23 and 31 minutes
- Half of the orders are prepared in less than 27 minutes
- Thai dishes have the greatest variability in preparation time whereas the least variability is registered for Korean dishes
- Italian and Thai dishes generally have the longest preparation time followed by Japanese and Spanish dishes whereas Korean and Vietnamese dishes are the most rapidly prepared dishes

Executive Summary: Conclusion (4/11): Food delivery time

- most orders are delivered in between 20 and 28 minutes
- half of the orders are delivered in less than 25 minutes
- The average delivery time is 24.16 minutes
- A total of 200 (10.54%) orders were delivered in more than 60 minutes from the time the order is placed
- Delivery time during weekdays is significantly higher than during the weekend
- Weekend delivery times tend to be more variable than weekday delivery times
- On average, orders are delivered in 28 minutes on weekdays and 22 minutes on weekends

Executive Summary: Conclusion (5/11): Cost of orders

- most orders cost between 12 and 23 dollars
- half of the orders cost less than 14 dollars
- 555 (29.24%) orders cost above 20 dollars
- Southern dishes have the largest variability in cost whereas korean dishes are the least cost variable
- French dishes tend to be the most expensive meanwhile vietnamese dishes are the cheapest

Executive Summary: Conclusion (6/11): Ratings

- A total of 736 (38.78%) orders are not rated
- Among the orders rated, most had a rating of 5 (about 580, 30.56% of all orders) followed by 4 (about 380, 20.02% of all orders)
- It turns out each restaurant having registered more than 50 ratings also has an average rating of over 4; only 4 (2.25%) restaurants respect these criteria and are, thus, eligible for the promotional offer: 'The Meatball Shop', 'Blue Ribbon Fried Chicken', 'Shake Shack', and 'Blue Ribbon Sushi'

Executive Summary: Conclusion (7/11): Revenue

- The top 14 restaurants in terms of revenue raised between 506 and 3580 dollars each from the orders
- 'Shake Shack' generated the greatest amount of revenue from the orders (3580 dollars) followed by 'The Meatball Shop' (2145 dollars)
- Among the top 14 restaurants, 'Five Guys Burgers and Fries' registered the least revenue (506 dollars) followed by 'Nobu Next Door' (623 dollars)
- The revenue raised by FoodHub is around 6166.3 dollars

Executive Summary: Conclusion (8/11): Food delivery time vs Rating

- Orders with the longest delivery time were generally rated 3 whereas those with the least delivery time were rated 4 on average.

Executive Summary: Conclusion (9/11): Food preparation time vs Rating

- There is hardly any observable relationship between the food preparation time and the rating though orders with the longest preparation time and widest preparation time variability register a rating of 3 on average

Executive Summary: Conclusion (10/11): Cost of dishes vs Rating

- The variability of cost is greatest for orders rated 3
- Among the dishes rated, the most costly dishes are rated 5 whereas the least are rated 3
- The cheapest dishes tend not to be rated

Executive Summary: Conclusion (11/11): Food preparation time vs Food delivery time vs Cost

- The variability of cost is greatest for orders rated 3
- Among the dishes rated, the most costly dishes are rated 5 whereas the least are rated 3
- The cheapest dishes tend not to be rated

Executive Summary: Recommendation (1/2)

- FoodHub should prospect and make promotion offers to restaurants that serve American, Japanese, and Italian dishes since these three cuisine types account for more than half of orders
- Optimal capacity management can be attained by allotting significantly greater logistics allowance (number of delivery cars and personnel and software availability and stability) during the weekend, which registers more than half of orders made during the entire week
- Service delivery agents should be stationed close to restaurants serving Japanese and Italian dishes to reduce total delivery time by compensating for the relatively long preparation time of these popular dishes

Executive Summary: Recommendation (2/2)

- The greater delivery time during weekdays needs to be investigated considering that a lot fewer orders are received than during weekends
- Advertisement of expensive French dishes might result in additional revenue generation
- Customer experience should be carefully gauged when advertising expensive dishes to determine whether the average ratings for this class of dishes remain high as the relative number of orders rises
- Clients tend to be more sensitive to delivery time than preparation time; so FoodHub should think of cost-effective strategies for delivery time reduction



Business Problem Overview

FoodHub, a food aggregator company providing access to several restaurants in New York via a smartphone app is keen on enhancing customer experience and has reached out to its Data Science team for insights that will help attain this goal. As a member of the Data Science team, I have been given the task to carry out the analysis and provide feedback to Management. The task will be accomplished in two main steps

1. Analysis of the Data Set
2. Recommendation of Business Action

Solution Approach (1/2)

A rundown of the approach used in this analysis

1. General exploration of the dataset to determine main features including number of orders registered, clients, restaurants, cuisine types, and total number of attributes under investigation
2. Univariate analysis of key attributes, namely the customer ID, restaurant name, cuisine type, day of the week, food preparation time, food delivery time, cost of orders, and ratings
3. Bivariate analysis uncovering relationships between:
 - i. Cuisine type and day of the week
 - ii. Cuisine type and Food preparation time
 - iii. Cuisine type and Cost
 - iv. Food delivery time and day of the week
 - v. Food preparation time and Rating



Solution Approach (2/2)

3. Bivariate analysis uncovering relationships between:
 - vi. Food delivery time and Rating
 - vii. Restaurant name and rating
 - viii. Restaurant name and Cost
 - ix. Cost and Rating
4. Multivariate analysis among Food Preparation Time, Food delivery time, and Cost
5. Synthesis of Observations
6. Recommendation

Data Overview (1/2): Data Shape, Datatypes, and Data Completeness

```
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)
```

```
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
```

- **1898 records (rows) and 9 attributes (columns)**¹
- **5 numerical variable columns**²:
 - 'order_id', 'customer_id', 'food_preparation_time', and 'delivery_time' have **int datatype**
 - 'cost_of_the_order' has **float datatype**
- **4 columns of object datatype** (categorical variables/str) ²: 'restaurant_name', 'cuisine_type', 'day_of_the_week', and 'rating'; strangely 'rating' has an object datatype
- No missing values³

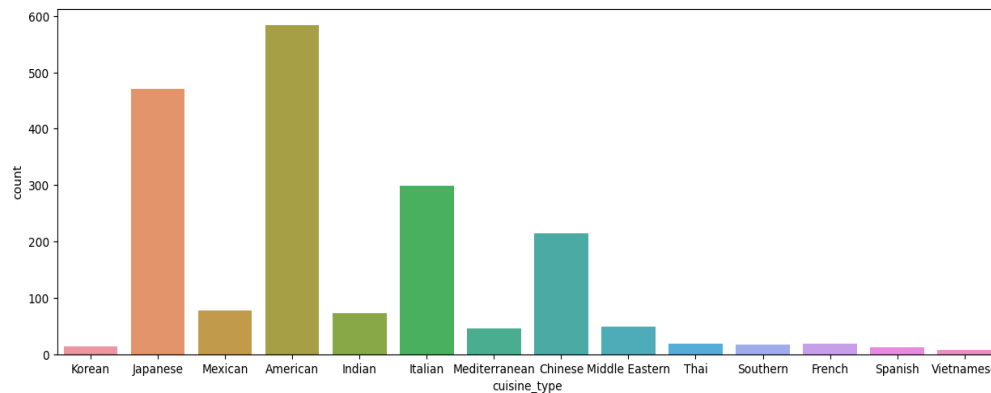
Data Overview (2/2): Statistical Summary

	order_id	customer_id	restaurant_name	cuisine_type	cost_of_the_order	day_of_the_week	rating	food_preparation_time	delivery_time
count	1.898000e+03	1898.000000	1898	1898	1898.000000	1898	1898	1898.000000	1898.000000
unique	NaN	NaN	178	14	NaN	2	4	NaN	NaN
top	NaN	NaN	Shake Shack	American	NaN	Weekend	Not given	NaN	NaN
freq	NaN	NaN	219	584	NaN	1351	736	NaN	NaN
mean	1.477496e+06	171168.478398	NaN	NaN	16.498851	NaN	NaN	27.371970	24.161749
std	5.480497e+02	113698.139743	NaN	NaN	7.483812	NaN	NaN	4.632481	4.972637
min	1.476547e+06	1311.000000	NaN	NaN	4.470000	NaN	NaN	20.000000	15.000000
25%	1.477021e+06	77787.750000	NaN	NaN	12.080000	NaN	NaN	23.000000	20.000000
50%	1.477496e+06	128600.000000	NaN	NaN	14.140000	NaN	NaN	27.000000	25.000000
75%	1.477970e+06	270525.000000	NaN	NaN	22.297500	NaN	NaN	31.000000	28.000000
max	1.478444e+06	405334.000000	NaN	NaN	35.410000	NaN	NaN	35.000000	33.000000

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

- It takes between **20** and **35 minutes** to prepare food after an order is placed⁴
- The average food preparation time is **27 minutes**⁴
- A total of **736 (38.78%)** orders are not rated⁵.

Univariate Analysis (1/6): Order ID, Customer, ID, Restaurant name, and Cuisine type

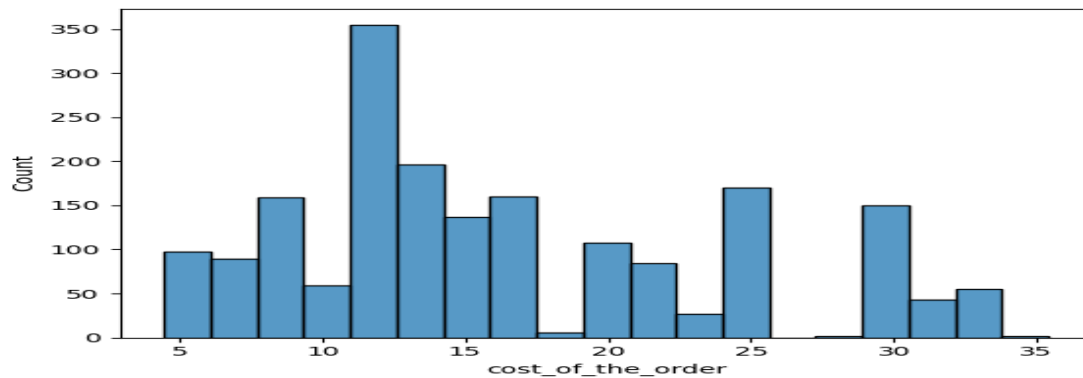


```
Shake Shack      219
The Meatball Shop 132
Blue Ribbon Sushi 119
Blue Ribbon Fried Chicken 96
Parm             68
...
Sushi Choshi    1
Dos Caminos Soho 1
La Follia        1
Philippe Chow    1
'wichcraft       1
Name: restaurant_name, Length: 178, dtype: int64
```

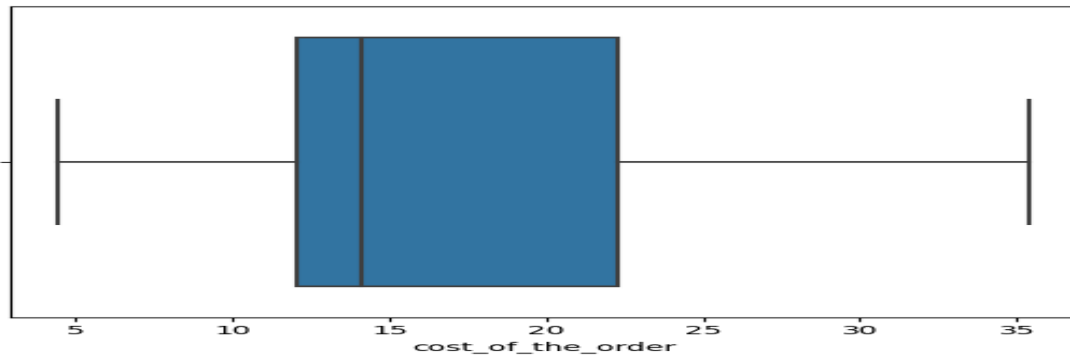
```
52832  13
47440  10
83287   9
Name: customer_id, dtype: int64
```

- 1898 orders were made by 1200 clients from 178 restaurants^{6, 7, 8}
- The IDs of the top 3 most frequent customers who will, therefore, be benefitting from the discount are 52832 (13 orders, 0.68%), 47440 (10 orders, 0.53%), and 83287 (9 orders, 0.47%)⁹
- A total of 14 cuisine types were ordered¹⁰
- 'American' was the most frequently ordered cuisine type (about 580 orders, 30.56%) followed by 'Japanese' (about 480, 25.29%) and 'Italian'(about 280 orders, 14.75%)¹¹
- 'Vietnamese' cuisine type was the least frequently ordered (less than 20 orders, 1.05%)¹¹
- The top 5 restaurants in terms of the number of orders in descending number of orders are 'Shake Shack' (219 orders, 11.54%), 'The Meatball Shop' (132 orders, 6.95%), 'Blue Ribbon Sushi' (119 orders, 6.27%), 'Blue Ribbon Fried Chicken' (96 orders, 5.06%), and 'Parm' (68 orders, 3.58%)¹²

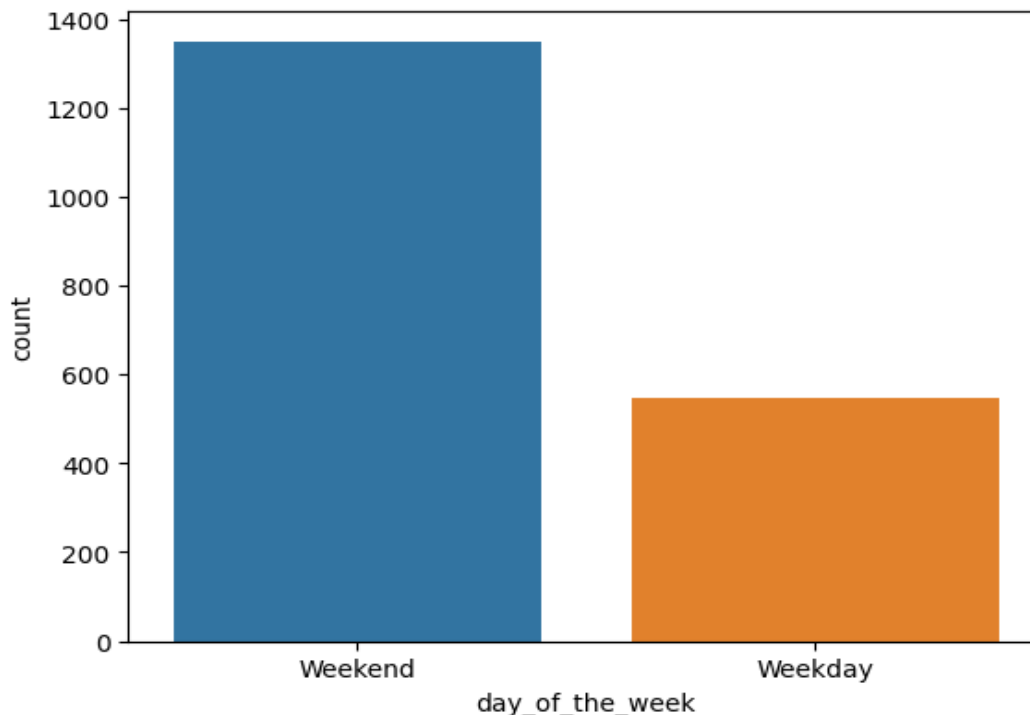
Univariate Analysis (2/6): Cost of order



- 'cost_of_the_order' is **slightly right-skewed**¹³
- **most** orders cost between **12** and **23 dollars**¹³
- **half** of the orders cost less than **14 dollars**¹³
- **555 (29.24%)** orders cost above **20 dollars**¹⁴

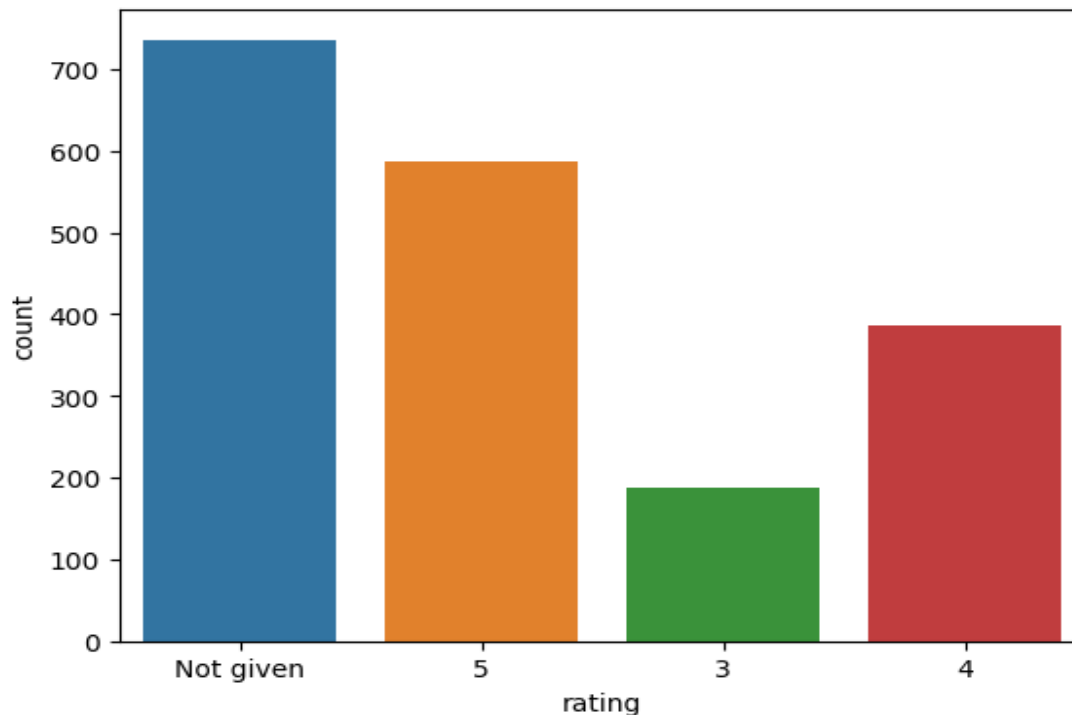


Univariate Analysis (3/6): Day of the week



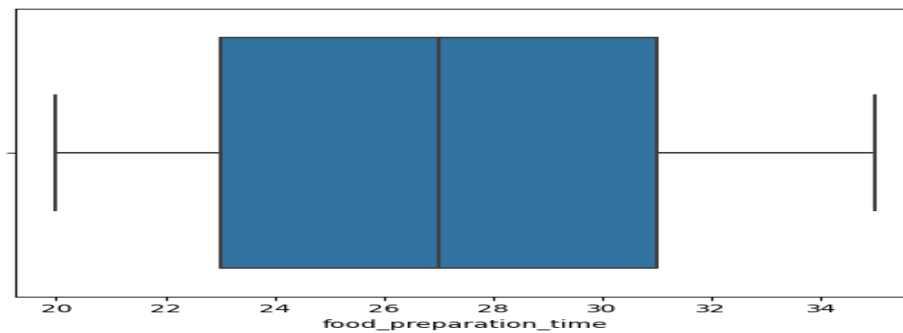
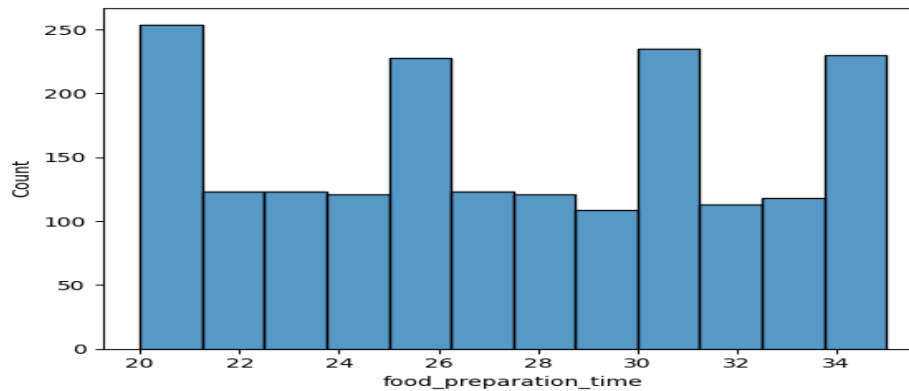
- As expected, 'day_of_the_week' has two values; orders are either placed during **weekdays** or during the **weekend**¹⁵
- The number of orders received during the weekend is **over twice greater** than those received during weekdays. Probably, people tend to use office and school restaurants during weekdays and, therefore, don't need to order food online¹⁶

Univariate Analysis (4/6): Rating



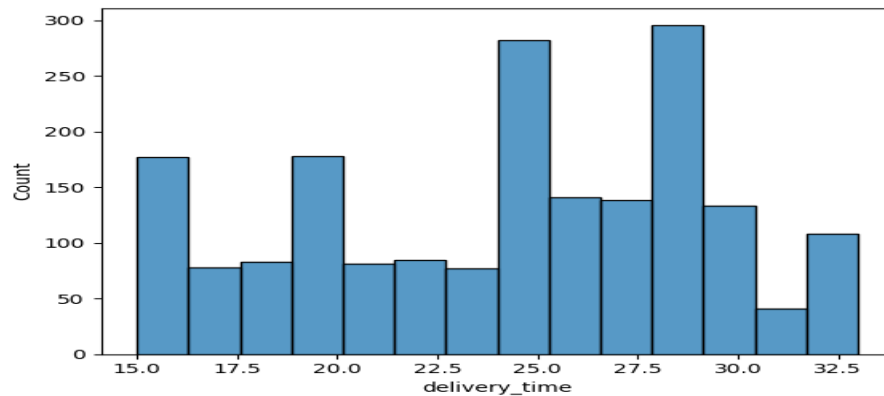
- The only ratings registered are **3**, **4**, and **5**¹⁷
- some orders were not rated (**736**, **38.78%**)¹⁷.
- Among the orders rated, most had a rating of **5** (about **580**, **30.56%** of all orders) followed by **4** (about **380**, **20.02%** of all orders)¹⁸

Univariate Analysis (5/6): Food preparation time

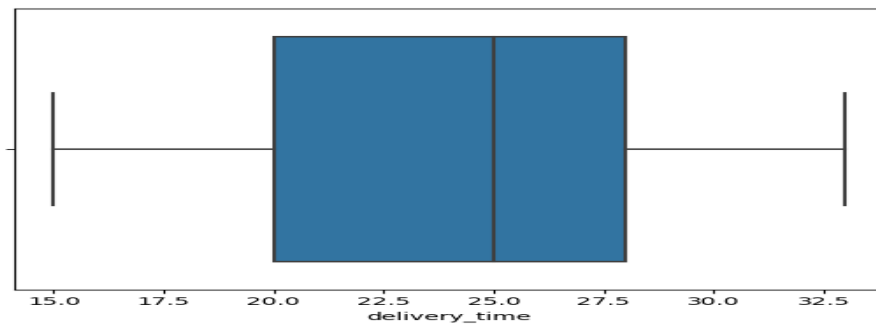


- The orders received are fairly distributed across food preparation times with peaks at **21, 26, 31, and 34** minutes, meaning relatively large numbers of orders registered preparation times around these peaks¹⁹
- Most orders are prepared between **23** and **31 minutes**¹⁹
- Half of the orders are prepared in less than **27 minutes**¹⁹

Univariate Analysis (6/6): Food delivery time



- 'delivery_time' is slightly **left-skewed**²⁰
- most orders are delivered in between **20** and **28 minutes**²⁰
- half of the orders are delivered in less than **25 minutes**²⁰

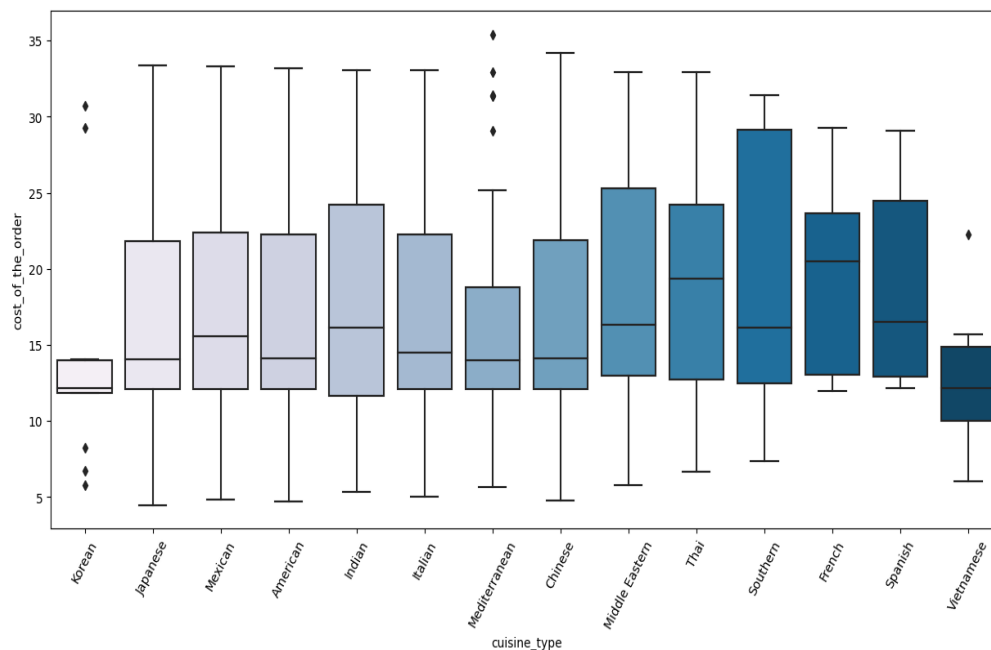


Multivariate Analysis (1/12): Cuisine Type vs Day of the Week

```
array(['Korean', 'Japanese', 'American', 'Italian', 'Mexican',  
      'Mediterranean', 'Chinese', 'Indian', 'Thai', 'Southern', 'French',  
      'Spanish', 'Middle Eastern', 'Vietnamese'], dtype=object)
```

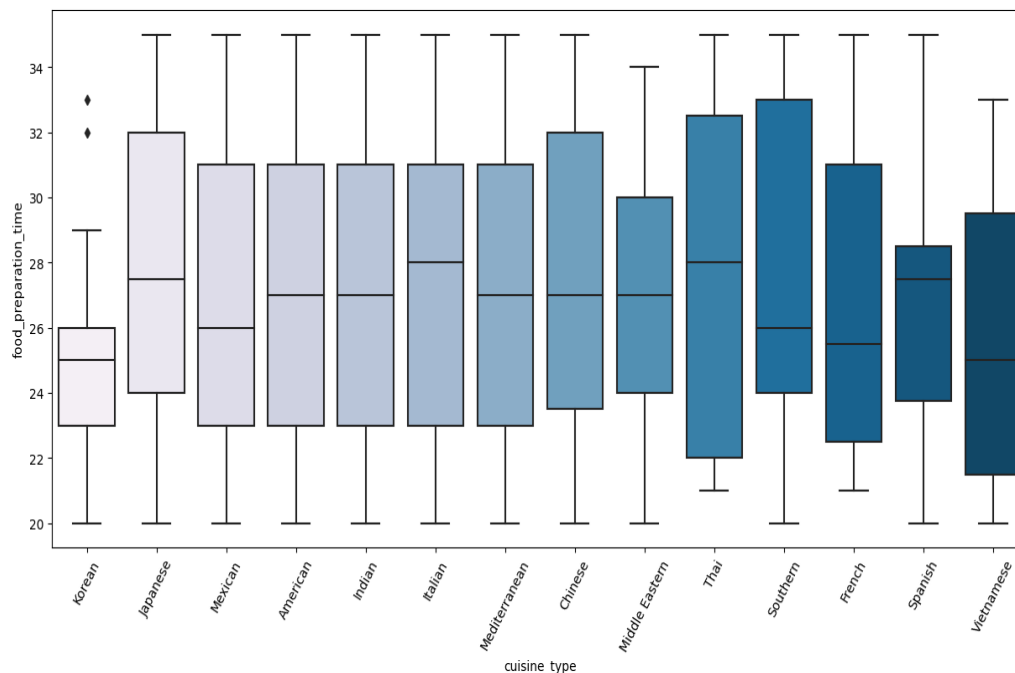
- All **14 cuisine types** were ordered during the weekend²¹

Multivariate Analysis (2/12): Cuisine Type vs Cost²²



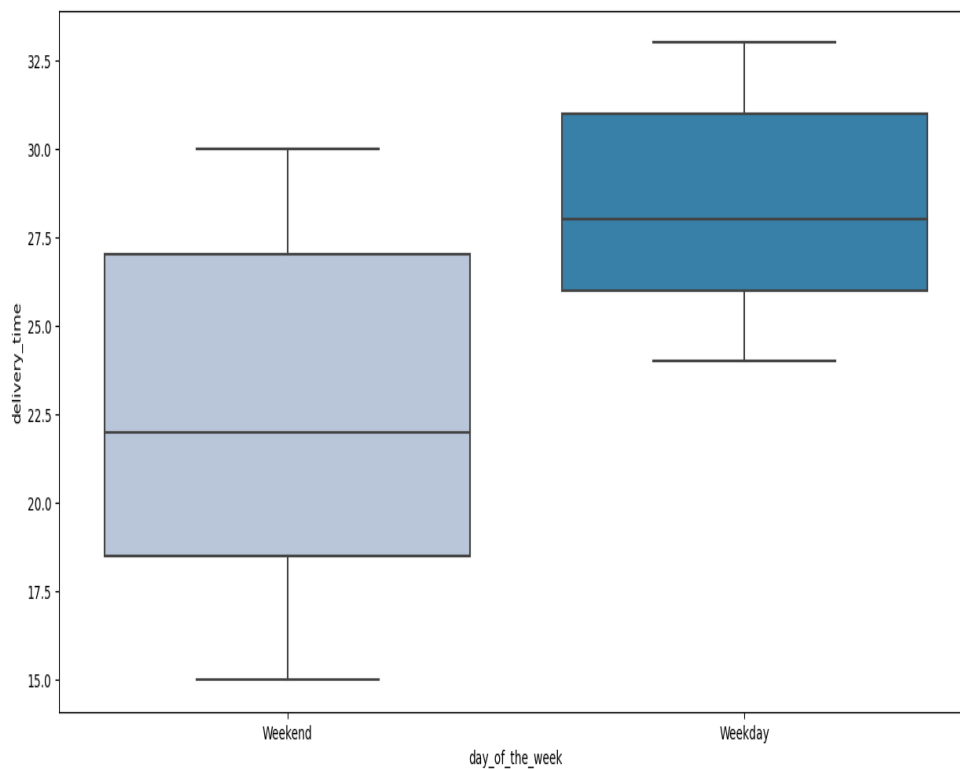
- **Southern dishes** have the largest variability in cost whereas **Korean** dishes are the least cost variable
- **French dishes** tend to be the most expensive meanwhile **Vietnamese** dishes are the cheapest
- **Korean dishes** have outliers in both directions of the distribution
- **Mediterranean** dishes are right-skewed with several upper outliers
- Relatively fewer outliers are registered for **Vietnamese** dishes

Multivariate Analysis (3/12): Cuisine Type vs Preparation Time²³



- **Thai** dishes have the greatest variability in preparation time whereas the least variability is registered for **Korean** dishes
- **Italian** and **Thai** dishes generally have the longest preparation time followed by **Japanese** and **Spanish** dishes whereas **Korean** and **Vietnamese** dishes are the most rapidly prepared dishes
- Some **Korean** dishes took abnormally long to prepare

Multivariate Analysis (4/12): Day of the week vs Delivery Time



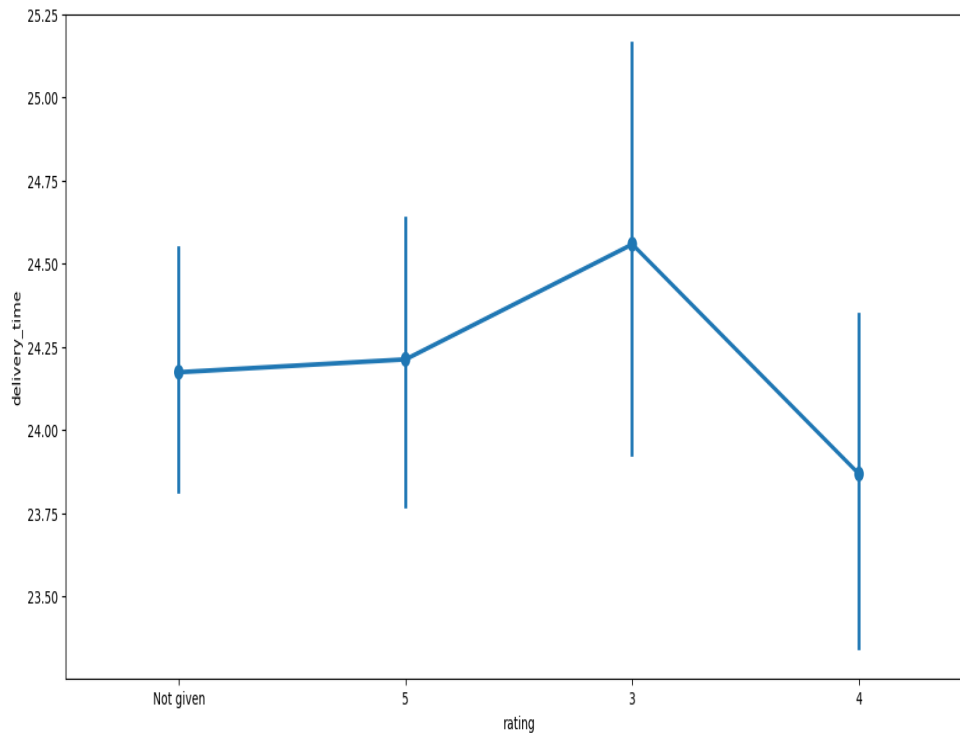
- Delivery time during **weekdays** is significantly higher than during the **weekend**²⁴
- **Weekend** delivery times tend to be more variable than **weekday** delivery times²⁴
- On average, orders are delivered in **28** minutes on **weekdays** and **22** minutes on **weekends**²⁵

Multivariate Analysis (5/12): Restaurant Name vs Cost²⁶

```
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
```

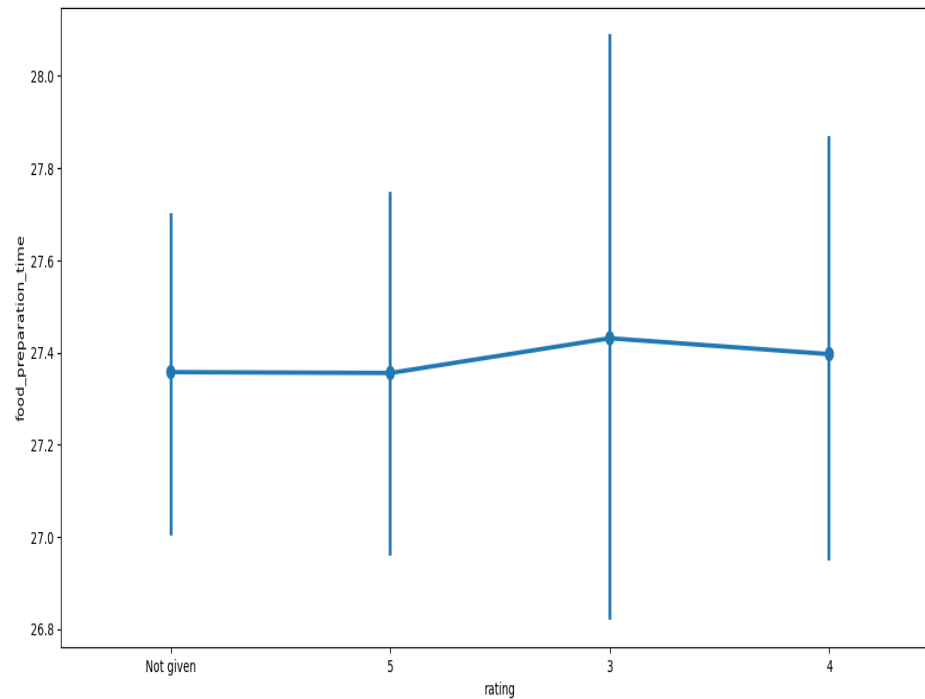
- The top **14** restaurants in terms of revenue raised between **506** and **3580** dollars each from the orders
- **'Shake Shack'** generated the greatest amount of revenue from the orders (**3580** dollars) followed by **'The Meatball Shop'** (**2145** dollars)
- Among the top **14** restaurants, **'Five Guys Burgers and Fries'** registered the least revenue (**506** dollars) followed by **'Nobu Next Door'** (**623** dollars)

Multivariate Analysis (6/12): Rating vs Delivery Time²⁷



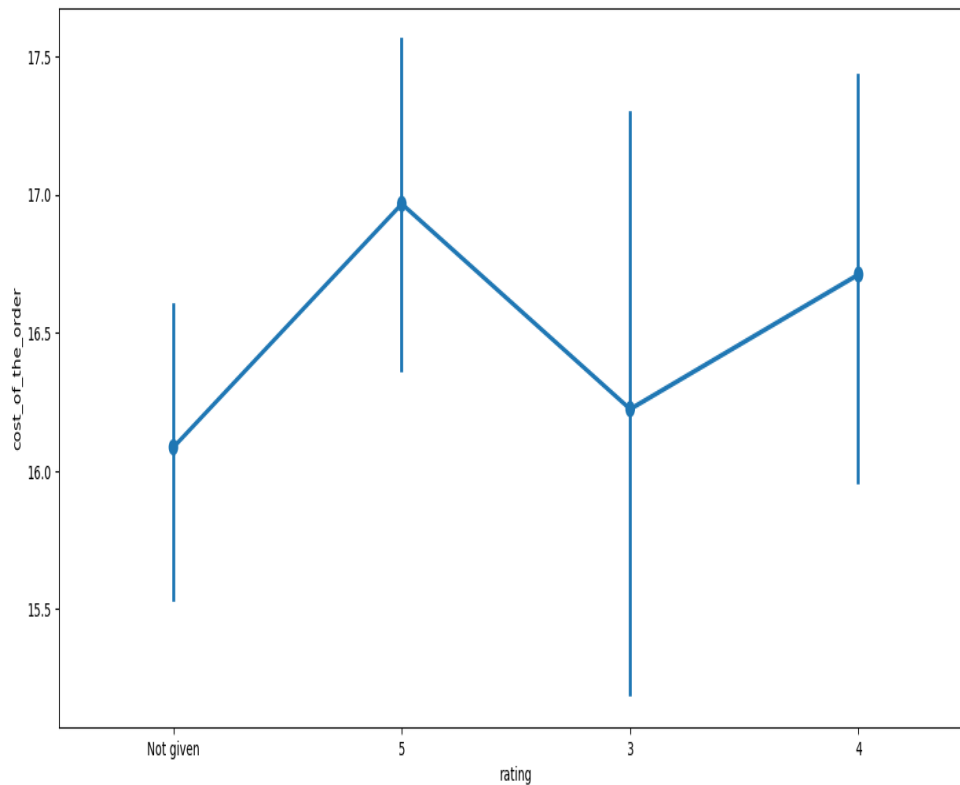
- Orders with the longest delivery time were generally rated **3** whereas those with the least delivery time were rated **4** on average.

Multivariate Analysis (7/12): Rating vs Preparation Time²⁸



- There is hardly any observable relationship between the food preparation time and the rating though orders with the **longest** preparation time and widest preparation time **variability** register a rating of **3** on average

Multivariate Analysis (8/12): Rating vs Cost²⁹



- The **variability** of cost is greatest for orders rated **3**
- Among the dishes rated, the **most** costly dishes are rated **5** whereas the **least** are rated **3**
- The **cheapest** dishes tend not to be rated

Multivariate Analysis (9/12): Cost vs Preparation Time vs Delivery Time³⁰



- There is hardly any correlation among **delivery time**, **food preparation time**, and **cost of the order**
- Correlation between **preparation time** and **cost** and between **preparation time** and **delivery time** is slightly positive
- **Delivery time** and **cost** are slightly negatively correlated

Multivariate Analysis (10/12): Rating vs Restaurant Name³¹

	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

- It turns out each restaurant having registered more than **50 ratings** also has an average rating of over **4**
- Only **4 (2.25%)** restaurants respect these criteria and are, thus, eligible for the promotional offer: '**The Meatball Shop**', '**Blue Ribbon Fried Chicken**', '**Shake Shack**', and '**Blue Ribbon Sushi**'

Multivariate Analysis (11/12): Cost of orders vs FoodHub's Revenue³²

	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 revenue raised by FoodHub is around **6166.3** dollars

Multivariate Analysis (12/12): Total Delivery Time (Delivery Time + Preparation Time)³³

- A total of **200 (10.54%)** orders were delivered in more than **60 minutes** from the time the order is placed

APPENDIX

Code Snippets (1/4)

```
1. df.shape
2. df.info()
3. df.isnull().sum()
4. df.describe(include='all')
5. df['rating'].value_counts()
6. df['order_id'].nunique()
7. df['customer_id'].nunique()
8. df['restaurant_name'].nunique()
9. df['customer_id'].value_counts().head(3)
10. df['cuisine_type'].nunique()
11. plt.figure(figsize = (15,5))
    sns.countplot(data = df, x = 'cuisine_type')
12. df['restaurant_name'].value_counts()
13. sns.histplot(data=df,x='cost_of_the_order')
    plt.show()
    sns.boxplot(data=df,x='cost_of_the_order')
    plt.show()
14. df_greater_than_20 = df[df['cost_of_the_order']>20]
    print('The number of total orders that cost above 20 dollars is:', df_greater_than_20.shape[0])
    percentage = (df_greater_than_20.shape[0] / df.shape[0]) * 100
    print("Percentage of orders above 20 dollars:", round(percentage, 2), '%')
15. df['day_of_the_week'].unique()
16. sns.countplot(data = df, x = 'day_of_the_week')
```


Code Snippets (2/4)

```
17.df['rating'].unique()
18.sns.countplot(data = df, x = 'rating')
19.sns.histplot(data=df,x='food_preparation_time')
    plt.show()
    sns.boxplot(data=df,x='food_preparation_time')
    plt.show()
20.sns.histplot(data=df,x='delivery_time')
    plt.show()
    sns.boxplot(data=df,x='delivery_time')
    plt.show()
21.df_weekend = df[df['day_of_the_week'] == 'Weekend']
    df_weekend['cuisine_type'].unique()
22.plt.figure(figsize=(15,7))
    sns.boxplot(x = "cuisine_type", y = "cost_of_the_order", data = df, palette = 'PuBu')
    plt.xticks(rotation = 60)
    plt.show()
23.plt.figure(figsize=(15,7))
    sns.boxplot(x = "cuisine_type", y = "food_preparation_time", data = df, palette = 'PuBu')
    plt.xticks(rotation = 60)
    plt.show()
```

Code Snippets (3/4)

```
24. plt.figure(figsize=(15,7))
    sns.boxplot(x = "day_of_the_week", y = "delivery_time", data = df, palette = 'PuBu')
    plt.show()
25. print('The mean delivery time on weekdays is around', round(df[df['day_of_the_week'] ==
    'Weekday']['delivery_time'].mean()), 'minutes')
    print('The mean delivery time on weekends is around', round(df[df['day_of_the_week'] ==
    'Weekend']['delivery_time'].mean()), 'minutes')
26. df.groupby(['restaurant_name'])['cost_of_the_order'].sum().sort_values(ascending = False).head(14)
27. plt.figure(figsize=(15, 7))
    sns.pointplot(x = 'rating', y = 'delivery_time', data = df)
    plt.show()
28. plt.figure(figsize=(15, 7))
    sns.pointplot(x = 'rating', y = 'food_preparation_time', data = df)
    plt.show()
29. plt.figure(figsize=(15, 7))
    sns.pointplot(x = 'rating', y = 'cost_of_the_order', data = df)
    plt.show()
30. col_list = ['cost_of_the_order', 'food_preparation_time', 'delivery_time']
    plt.figure(figsize=(15, 7))
    sns.heatmap(df[col_list].corr(), annot=True, vmin=-1, vmax=1, fmt=".2f", cmap="Spectral")
    plt.show()
```

Code Snippets (4/4)

```
31. df_rated = df[df['rating'] != 'Not given'].copy()
    df_rated['rating'] = df_rated['rating'].astype('int')
    df_rating_count = df_rated.groupby(['restaurant_name'])['rating'].count().sort_values(ascending =
False).reset_index()
    df_rating_count.head()
    rest_names = df_rating_count[df_rating_count['rating'] > 50]['restaurant_name']
    df_mean_4 = df_rated[df_rated['restaurant_name'].isin(rest_names)].copy()
    df_mean_4.groupby(['restaurant_name'])['rating'].mean().sort_values(ascending = False).reset_index().dropna()
32. def compute_rev(x):
    if x > 20:
        return x*0.25
    elif x > 5:
        return x*0.15
    else:
        return x*0
    df['Revenue'] = df['cost_of_the_order'].apply(compute_rev)
    df.head()
    total_rev = df['Revenue'].sum()
    print('The net revenue is around', round(total_rev, 2), 'dollars')
33. df['total_time'] = df['food_preparation_time'] + df['delivery_time']
    df_delivery_greater_than_60 = df[df['total_time']>60]
    print('The number of total orders whose delivery time is more than 60 minutes is:',
df_delivery_greater_than_60.shape[0])
    percentage = (df_delivery_greater_than_60.shape[0] / df.shape[0]) * 100
    print("Percentage of orders delivered in more than 60 minutes:", round(percentage, 2), '%')
```



Happy Learning !

