PORTER

DATA ANALYSIS REPORT

ANIL

INTRODUCTION

Porter, India's largest and most trusted marketplace for intra-city logistics, connects restaurants with customers by facilitating swift and efficient food deliveries. The company works with a wide range of restaurants across different categories, ensuring that meals are delivered fresh and on time. However, predicting delivery times accurately is a crucial challenge. It depends on a variety of factors, including the nature of the order, restaurant characteristics, delivery partner availability, and temporal patterns like time of day or day of the week. An inaccurate estimation can lead to delays, dissatisfied customers, and inefficient resource allocation.

The goal of this project is to perform a detailed analysis of Porter's delivery dataset to understand the factors affecting delivery times and provide recommendations for improving the process. By applying Exploratory Data Analysis (EDA), preprocessing the data, and engineering new features, the project will identify key trends and correlations that influence delivery performance.

Understanding the Dataset

The dataset contains 197,428 entries across 14 columns.

- ***Overview of the dataset:***
  - *There are three types column present in the dataset which are "Numerical", "Categorical" and "Datetime".*
  - *Columns like total_items, subtotal, min_item_price, max_item_price, total_onshift_partners, total_busy_partners, and total_outstanding_orders represent numerical features relevant to delivery and restaurant operations.*
  - *Columns like store_primary_category and order_protocol represent categorical data such as restaurant types and protocols for handling orders.*
  - *created_at and actual_delivery_time represent time-related data for the creation of orders and delivery completion.*
  - *Some columns, like store_primary_category, contain missing values, which will need handling in preprocessing.*
  - *For partner-related columns (total_onshift_partners, total_busy_partners, and total_outstanding_orders), missing values also exist and may require strategies like imputation.*

- ***Statistical Insights:***
  - ***Total_onshift_partners:*** *Mean = 44.81, indicating an average of ~45 on-shift partners per observation. The max value of 171 suggests well-staffed instances, but negative values (e.g., -4) require investigation as they indicate data quality issues.*
  - ***Total_busy_partners:*** *Mean = 41.73, close to the on-shift partner count, implying high utilization during peak times.*
  - ***Total_items:*** *Median = 3, 75th percentile = 4, and max = 411, indicating a typical order size of 3-4 items, with some extremely large orders.*
  - ***Num_distinct_items:*** *Median = 2 and max = 20, showing that while most orders contain only a few distinct items, there are cases of high variety.*
  - ***Subtotal:*** *Mean = 2,682.33 and a max of 27,100, showing the wide variation in order sizes.*
  - ***Outlier Alert****: min_item_price has a minimum of -86, which is clearly an error and requires correction or removal.*

- ***Categorical Data:***
  - *Primary store categories with 74 unique categories with "American" being the most frequent.*
  - *Order protocol to 1 to 7 are in use, with protocol 3 being the most common.*

- ***Datetime Data:***

- *Orders are spread over 180,985 unique timestamps (created_at), indicating active data collection across time.*

- ***Missing Values:***
  - *Missing values in multiple columns. Those are "market_id", "actual_delivery_time", "store_primary_category", "order_protocol" , "total_onshift_partners", "total_busy_partners" and "total_outstanding_orders ".*
  - ***Market_id:** It has **6** unique market and the missing value in the **987 rows** which are very less if we compare it with our whole datasets where all **197428** so we can use the **random method** to impute the missing values.*
  - ***Actual_delivery_time:** It has only **7** rows containing null values in the 'actual_delivery_time' column and also the data is in TimeSeries and Continuous so we can impute it with **ffill or bfill.***
  - ***Store_primary_category:** Our store_primary_category column has **4760** missing rows and the **74** unique we compare it with our whole database i.e, the 197428 rows and 14 columns these missing values are very few so we can impute it with the high occurence of the data i.e, the **mode() method**.*
  - ***Order_protocol:** It has 995 missing values so we can impute with **ffill or bfill**.*
  - ***Total_onshift_partners,total_outstanding_orders & total_busy_partners:** Since our above all three columns have the 16262 rows contain the null values and also these are the numerical columns and rows containing values like*

```
total_onshift_partners   : 172 unique values,
total_busy_partners      : 159 unique values,
total_outstanding_orders : 281 unique values
```

*Observation we can say that several similarities between the columns like each column has **181166** entries, mean vary between the 41-58, standard deviation also vary between 32-52, min vary between -4 to -6, max vary between 154 to 285 and the Quartile range also the same. So, We Can Impute missing values all three columns with same method i.e., **the random method.***

Creating Target Column

- ***Time Taken for Delivery:** Created with help of "Created_at" and "actual_delivery_time" columns. At first observation are minimum time to delivered to any product is **-23 days** and maximum time is **98 days**. So, it is impossible to get time in negative.*
- ***Outlier Detection:** We are removing the outliers with **Quadrantile Range**. After removing outliers, we got 6285 missing values. It is datetime format and best approach to handle missing in datetime is **ffill or bfill.***

Creating New Features

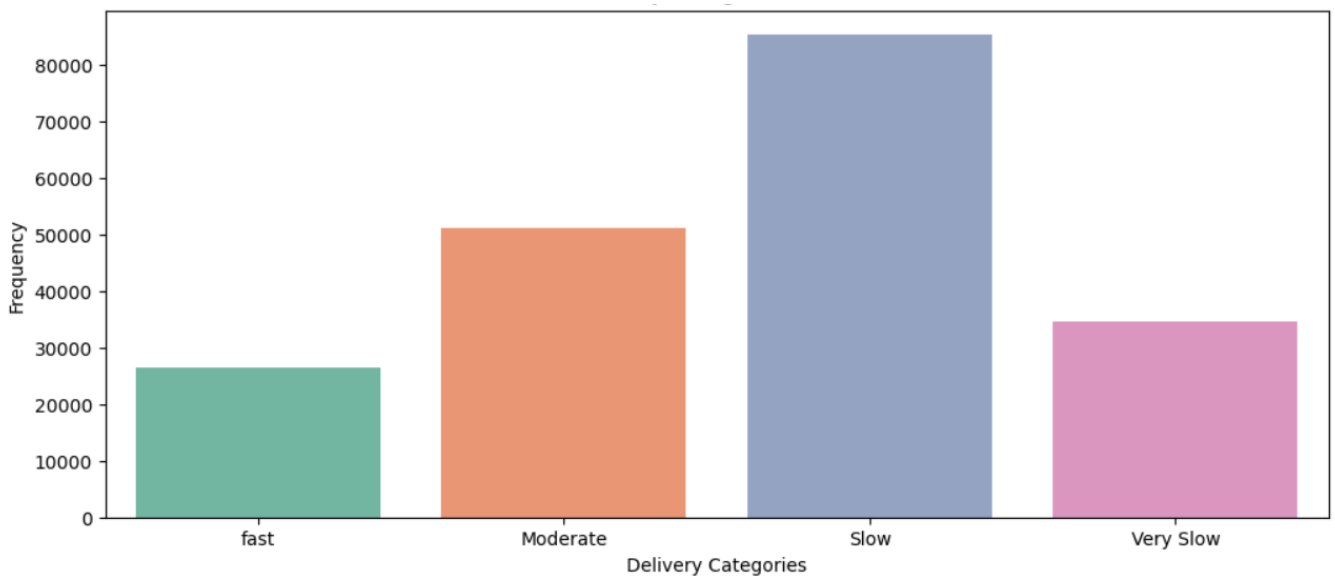
- ***Delivery_categories:*** *It describes how fast and slow are we able to deliver the products. Since we have time format like Minutes, hours and seconds. But we can extract this feature with the help of minutes because maximum, and average delivery time falls within an hour.* ***Since we saw that our maximum order delivered came inside in Slow categories and less order in the fast and moderate.***

Encoding Categorical Columns

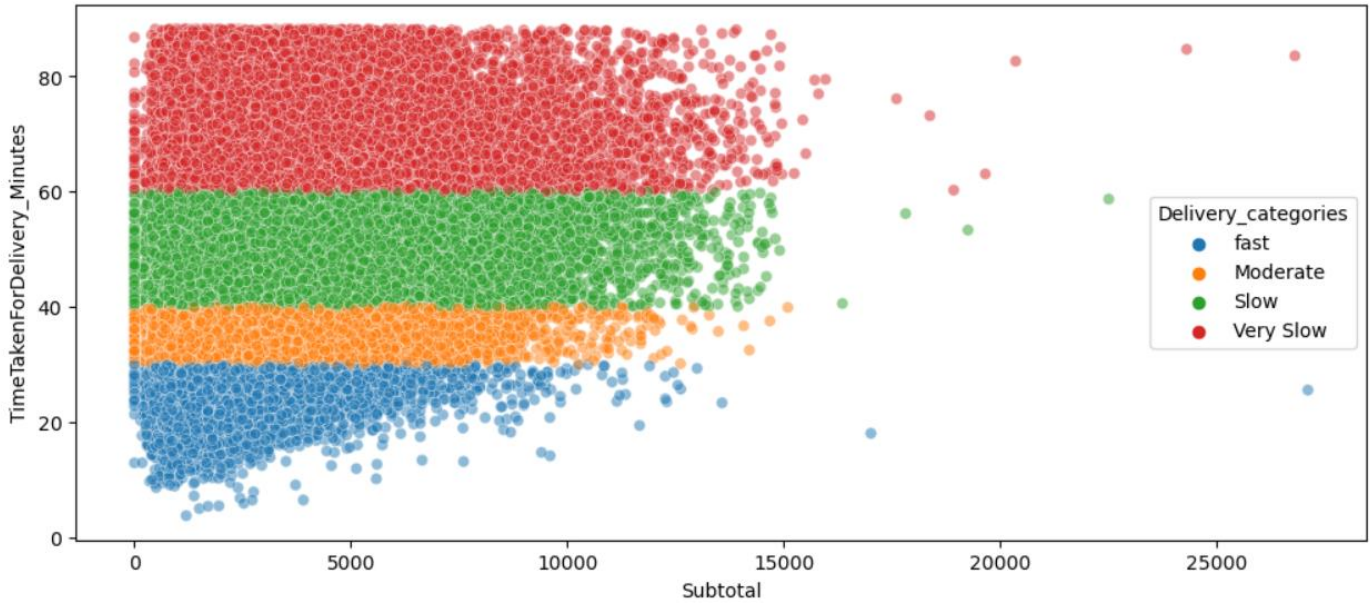
- We encoded categorical columns like "store_id", "store_primary_category". In which it takes 1.3 GB memory usage. Because many columns in Encoded_Categorical_df.
- Pd.get_dummies is used to encoded the categorical columns.

Data Visualization

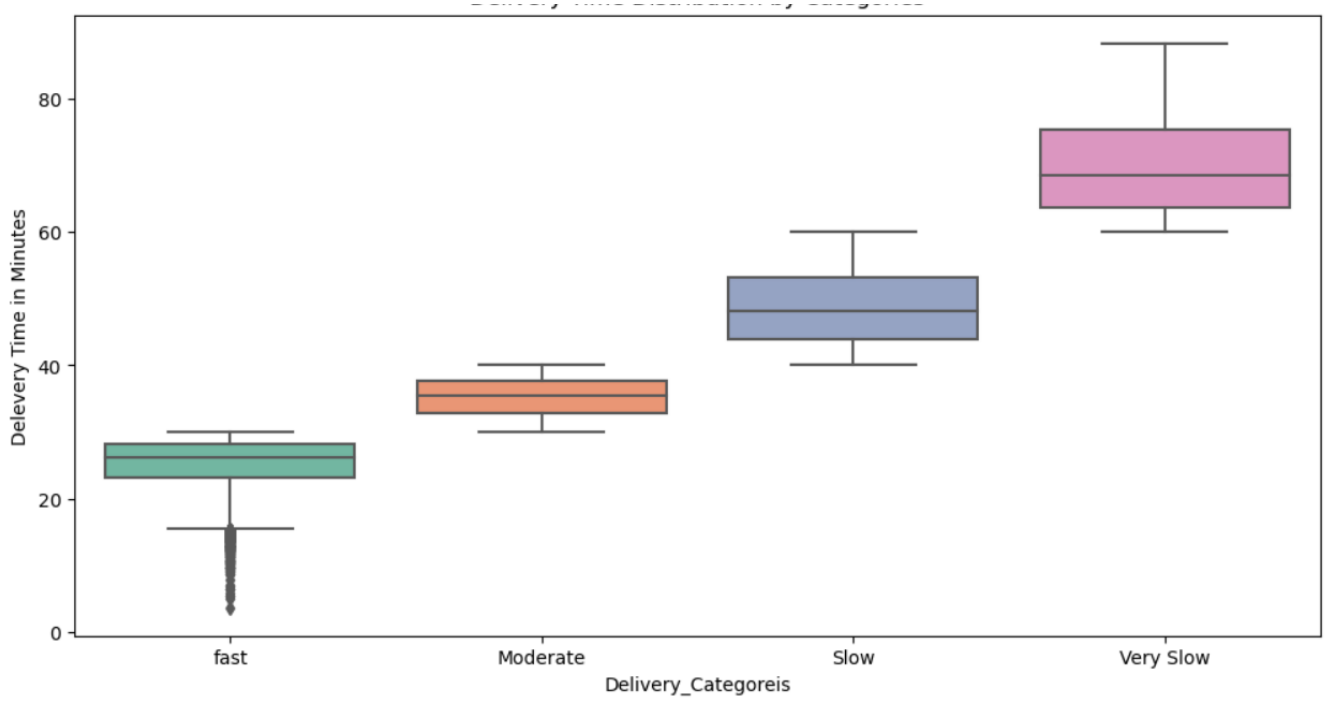
- **Delivery categories Distribution:**



- **Subtotal v/s Time Taken for Delivery in Minutes:**

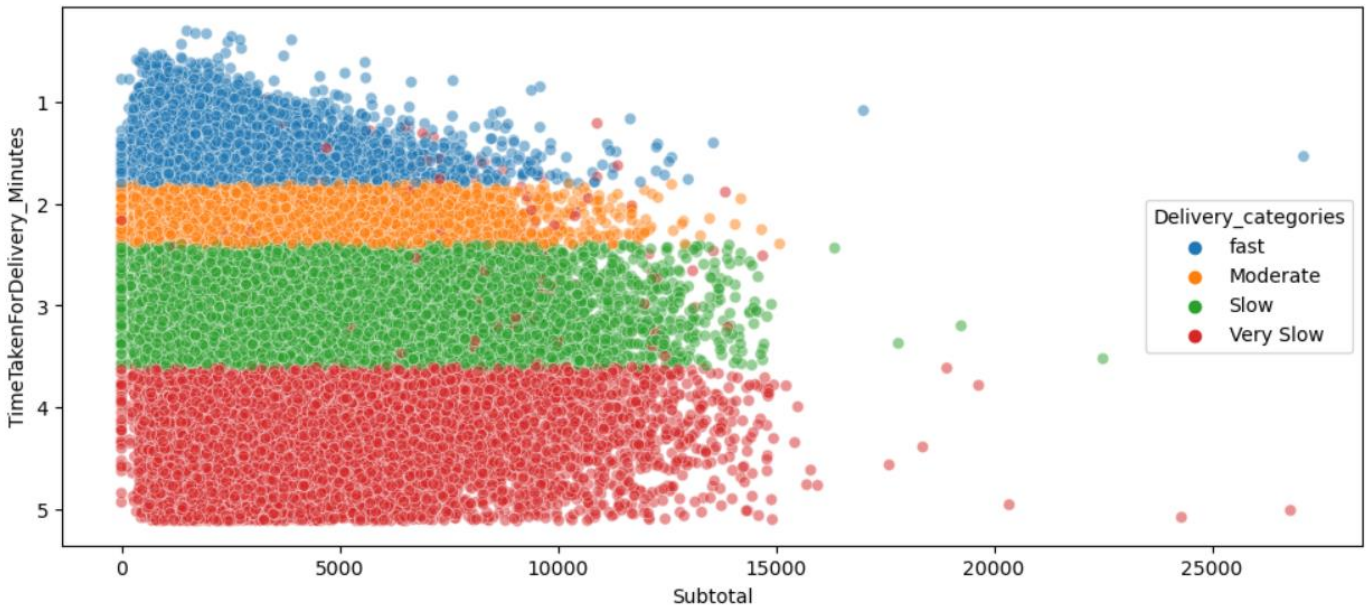


- **Delivery Time Distribution by Categories:**



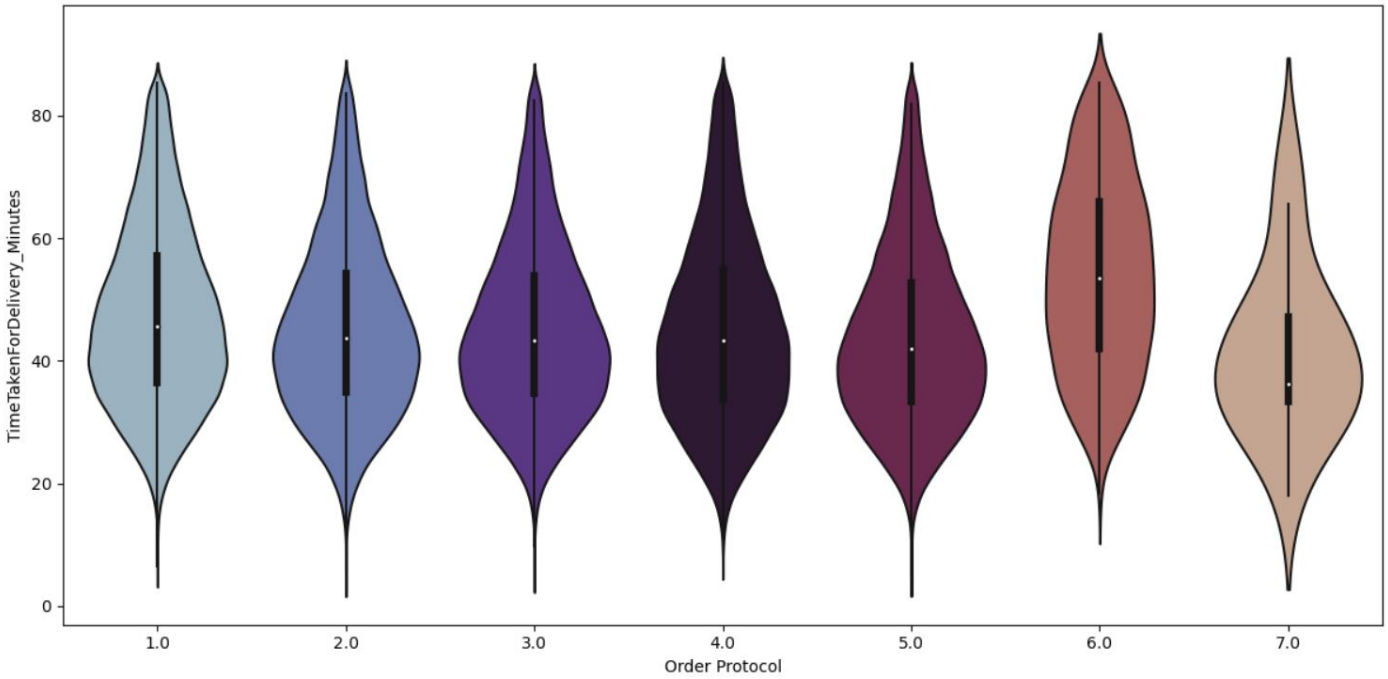
- *After the Analysis we can say that the some of the **outliers** in 'time_taken_for_delivery_minutes' Column and also statical view indicates.*
- Outliers is removed by ***Quadrantile Range*** *method. After imputing outliers we had 1335 missing values which were impute by **ffill or bfill**.*

- **Plotting graph again:**



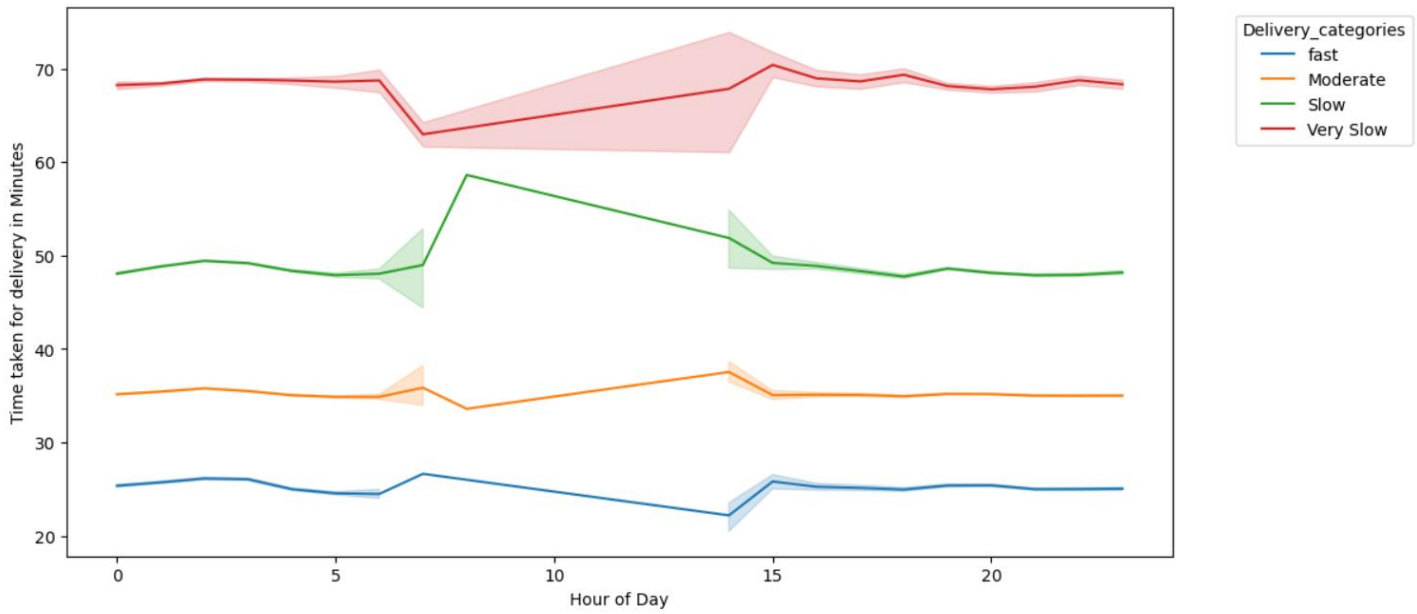
- *After the analyse of "subtotal" and "TimeTakenForDelivery_Minutes" we can say that the these are not strongly corelated bcz generally in correlation both are increasing or decreasing that means the positive corelation and negative corelation. But in our, case it's the constant like "subtotal" amount is 10000 but the "TimeTakenForDelivery_Minutes" is the same as "subtotal" amount is 5000 there no chaninging like positively or negatively.*

- ***Order Protocol and Time Taken for Delivery in Minutes:***



- *In our case maximum order delivered between the 40 to 50 minutes and the order protocol 6.0 shows versatility.*

- ***Delivery time by Hour of Day and Delivery Categories***

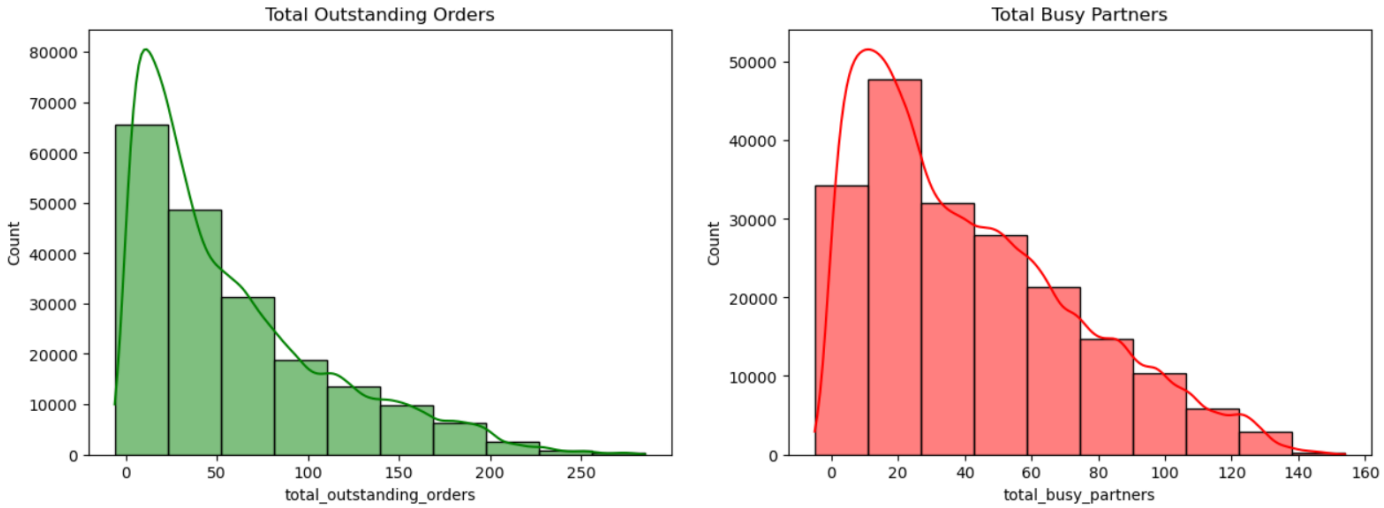


- ***Blue(Fast)*** *most order delivered in fast section within between the 20 and 30 minutes.* ***Yellow(moderate)*** *all the order delivered in moderate section between the 30 and 40 minutes.* ***Green(slow)*** *all the order delivered in slow section between 50 and 60 minutes and the last* ***Red(very slow)*** *section order delivered more than 70 minutes.*

- ***Service/Product related to on 'min_item_price','AvgItemPrice' and 'MaxItemPrice':***

  - *By the observations of "AvergareItemPrice", "MimimumItemPrice and "MaximumItemPrice", we can say that the our all three are similar means in "AvgItemPrice", maximum product/services are in range 0-2000 like the "MinimuItemPrice", and "maxItemPrice". So, our maximum revenue comes from the "MimimumItemPrice" and very less revenue from the "MaxItemPrice" our product/services* ***Budget Friendly and best suited for middle class family.***

- ***How 'total_outstanding_orders' and 'total_busy_partners' are affecting our services:***



  - ***Total Busy Partners:*** *Number of Delivery Partners attending the other tasks. Since Our "TotalBusyPartners" falls within 0-160 and peak is 20 and count more than 50000. After that second most is ~15 and count are ~35000.*
  - ***Total Outstanding Orders:*** *Total number of orders to be fulfilled at the moment. Since our "TotalOutstandingOrders" fall between 0-250 and maximum*

"TotalOutstandingOrders" deliver at ≈30 and count ≈65000 and second most "TotalOutstandingOrders" at ≈ 40 and frequency is ≈ 50000.

Insights

- ***Delivery Speed: -*** *Our Most delivery categories fall inside the SLOW, MODERATE and very less come inside the FAST and VERY SLOW categories. Also, HOUROFDAY at peak time not able to handle the customers. Having NUMBER OF DISINCT ITEMS IN THE ORDER takes more time to deliver the product.*

- ***Total Busy Partners and Total Outstanding Orders: -*** *At the peak performance TOTAL BUSY PARTNERS are less and other time there much more PARTNERS. Also, the TOTAL OUTSTANDING ORDERS comes between 0-40.*

- ***Customer Spending and Item trends: -*** *Our maximum customer is in the minimum categories and services they are using which is Budget Friendly, So, we can say that our service is generally use by middle class family.*

- ***Market and Total Busy Partners: -*** *In market id "2.0" and "4.0" high frequency and the services in the high density these two markets have high demand and rest of all are the in the general but we are not able to full fill the demand because here we delivered maximum order in "moderate" and "slow" Categories.*

- ***Store and Market: -*** *Store are related on the market we observed that the market highly correlated to the market because in the bifurcation of "Delivery_Categories" we got the maximum order are fall in the "slow" categories.*

- ***Order Protocol and Partner Efficiency: -*** *In some protocol we are able to handle the delivery but in some we are not because our onshift partners are less and not full fill the demand.*

- ***Operational Efficiency: -*** *A high number of "total_OutStanding_Orders" are related to delayed deliveries specially in the "TotalBusyPartners" during the peak hour we don't have enough work force to handle the volume of orders. There are Seven "Order_protocol" that's the reason we have to verify orders coming from*

where like through Porter, call to restaurant, pre-booked, third-party, etc. all these thing taking too much time.

Recommendations

- ***Delivery Speed Optimization:*** *- most of our fall in "moderate" and "slow" category and in peak hour we are not able to full fill the demand. We can increase the staff during the peak hour and ensure there are enough delivery partners at peak hours we adjust this to give over time to the delivery partners. Implement incentive program for those delivery who frequently delivered product in the fast Category, this can boost overall speed and keep partners to motivate and redirect their root.*

- ***Flexible workhour: -*** *Offers flexible work hour to the delivery partners, allow them to work at the peak hours handle the work and reduce the totalOutstandingOrders.*

- ***Customer Segmentation and Targeting:*** *- Since Our customer is budget conscious and especially middle class. Introduce the premium services for premium customer those who pay for the superfast delivery orders. Provide some discount in the premium services during the non-peak hours. We can introduce the customer loyalty program for the customer those who are frequently orders and refer their friend's give them the rewards or offer premium services at more discount.*

- ***Market Specific Strategy: -*** *In market 4.0 and market 6.0 high demand and we are not able full fill it also our Delivery Category is slow. Assign more delivery partners and resources during the peak hour's partnership with local retails to take geographic advantages.*

- ***Store Level Improvement: -*** *Stores are highly corelated with the speed of delivery in slow Categories. Regularly audit the high-volume orders and why is it slow that means in preparation time, inventory management, order processing or something for causing the delays. Work with stream line and give the clear instruction and provide training if it necessary, introduce the incentive for who those delivered the high-volume product in fast Category.*

- ***Improving Order protocol Efficiency: -*** *Some order protocols types are more efficient and some are struggling to fulfil their orders due to the lack of partners. Introduce streamline protocols and assign the protocols types like directly pre-booked, and third-party protocols automatically ordered no need to manual verification and*

confirmation. Introduce the AI which assign the assignment of Orders equally or less dependent on a special Store.

- ***Customer Satisfaction Initiatives: -*** *Delay and slower services in some category leading to lower customer satisfaction. Implement a feedback mechanism where customer can give the real time feedback and after use this data, we can identify what is most concern of customers delivery speed or the better communication. Provide real time tracking system and estimated time to delivered the product also inform them if in case the product will deliver late, this all above mentioned things reduce the customer frustration during the peak workhour.*

- ***Long-Term Growth Strategies: -*** *As demand grows particularly high market invest in growing force of delivery partners. We can use predictive model to predict where the demand will increase and according to them, we will prepare. Strengthen relationships with stores and delivery partners provide the analysis and insights of their performance this will help to improve their work and as well as customer behaviour and improve customer satisfaction rate.*


THANK YOU