# **Introduction**

This comprehensive Report contains a thorough Exploratory Data Analysis of the Rapido Dataset. This EDA offers important insights into customer behavior, operational effectiveness, and engagement patterns across many cities. Given the increasing significance of on-demand transportation services, improving service quality and streamlining operational strategies require an awareness of key performance factors such ride fare distribution, travel duration variance, payment preferences, and driver performance.

Finding trends in ride demand, evaluating the effectiveness of marketing initiatives, comprehending consumer involvement, and pinpointing possible service improvement areas are the objectives of this analysis. The study offers a thorough analysis of the dataset, along with statistical summaries and visuals that draw attention to important patterns, insights, and connections.

The analysis can conclude to assist stakeholders to make well informed decisions to enhance the customer experience, diver engagement and operational effectiveness—all of which will eventually support Rapido's expansion and customer retention objectives.

**Overview of the dataset.**

The dataset contains the details regarding each ride which happened on a particular date. The dataset contains 10 attributes which are describe below.

Customer\_ID: Unique identifier for each customer

Booking\_Date: Date and time of the ride booking

Ride\_Fare:Amount being charged for each ride

Payment\_Method: Payment method options that provided by the company.

Promo\_Code: Discount offered code

Travel\_Time: Time taken to complete each ride

Ride\_Distance: Distanced travelled (Km)

City: City in which the ride occurred

Driver\_ID: Unique Identifier for each driver

Cancellation\_Status: Whether a ride was cancelled or not: Completed: 1 , Cancelled: 0.

# Findings, Insights and Trends discovered through the EDA

The initial step in doing the Exploratory Data Analysis, is loading the dataset to the python framework and identifying the structure and the data type of the variables in the dataset. According to that the Rapido dataset had the structure of the 100000 rows and 10 columns. Based on the function **.info()** in python, the information regarding the dataset was obtained as below.

A screenshot of a computer screen

Description automatically generated

Figure 1: Information on the dataset

Based on the output generated all the attributes contains non-null values except for the Promo\_Code attribute which has about 20% null values. This indicates that the discount has not been applied to all the rides, only for 79935 rides the discount is applied. In the dataset the attributes fall under three main data types as categorical (object), numerical (float64) and integer (int64). Even the Booking\_Date has the data type as categorical (object) which needs to be converted to Date type for the suitability of the analysis. Checking for duplicate rows is other step in preprocessing the data, which this dataset has no duplicates. After the initial preprocessing steps, a summary statistic was done which includes count, mean, min max and standard deviation for the attributes with numeric data type to understand the distribution of those attributes. The below contains the summary statistics of the attributes in the Rapido Dataset.

A screenshot of a computer screen

Description automatically generated

Figure 2: Summary Statistics

The summary statistics can be used to identify missing values, outliers and unexpected ranges. As in the screenshot the Ride Fare range from 20 to 500 and the mean is approximately 260 which has a typical fare charge range. The distribution of the data can be gained from the minimum, maximum and quartiles. The Travel\_Time ranges from 5 to 59 minutes and most of the time falls between the 18 minutes and 46 minutes. The Ride Distance ranges from 1 to 20 km and most of the distances fall between 6 to 15 km and average ride distance is 11 km. Based on the Cancellation\_Status attribute we can say that almost 50% of the rides are being cancelled. This will ensure that we need to investigate more on this issue which will be analyzed in the latter part. After understanding the structure and the distribution of each variable in the dataset, now we can go in depth to the analysis to identify key findings, insights and trends in the dataset.

**Ride Fare Distribution Analysis**

When considering about the Ride Fare Distribution we can see that there’s a uniform distribution or the data is uniformly distributed within the range of 20 to 500. From the graph of Ride Fare Distribution, we can see that there are no peaks, all the ride fare are equally likely to occur. Based on this we can see that there ‘s a pricing model with less variability.

A graph of a number of blue bars

Description automatically generated

Figure 3: Ride Fare Distribution

If we further analyze the fare distribution across time slots a boxplot visualization is used to identify the fare trends across the different time slots. As per the boxplots we can see that there’s a consistent median fare across all the time slots. The fares are distributed evenly across the different time of the day. And there are no outliers in the time slots. So based on this we can say there’s no time slot showing a significantly higher or lower ride fare.

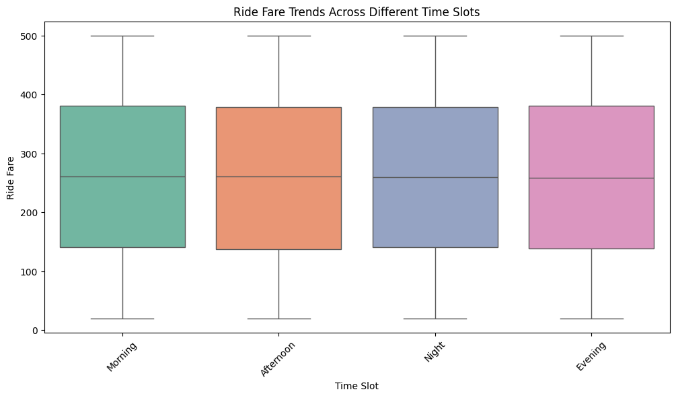


Figure 4: Ride Fare Trends Across Different Time Slots

**Customer Engagement and Preferences.**

According to the graph we can see that bookings had happen throughout the day and each hour has around 4000 to 4500 rides. Based on the graph we can see that there are no particular peak and off peak hours. This illustrates that there's an effective supply and demand management. When evaluating the distribution more thoroughly we can see that there's a slight increase in the bookings within the morning hours from 5 to 9 am and evening from 6 to 7 pm

A graph of different colored bars

Description automatically generated with medium confidence

Figure 5: Most popular time slots for ride bookings

The Figure is used to see the ride trends across the days. In considering all the Days in a week we can see that the bookings are compatible in all the days reaching a booking of 14000 to 15000 per day. There's no difference when it comes to weekends and weekday as the rides are stable in a week. A slight increase in bookings can be seen in Saturday and Sunday. A noticeable drop in bookings cannot be seen in the days. This suggests that people are using the service as a daily shuttle service.

A graph of blue bars

Description automatically generated

Figure 6: Most popular days for ride bookings

**Operational Efficiency and Driver Metrics**

Approximately there's a 31.97 minute of time spent averagely for travel time. This is an illustration that most of the rides are of decent duration which might be influenced by traffic and the route length that might affect the travel time. This might be beneficial for the stakeholders to be used in finding the shortest route to reduce the travel time it takes. When it comes to the cancellation rate there's a 50 % cancellation rate which is a quite high value. This might be due to various reasons which might be occurred due to the riders not available at the point where the customers have booked. Also, users might cancel their ride due to long waiting time and the price might affect the user has well or else the drivers might cancel the ride because of communication issues.

A number on a white background

Description automatically generated

A graph of a number of rides

Description automatically generatedAs per the distribution of the riders per Drivers we can see that the drivers usually have one drive per day which is an anomaly. This illustrates an underutilization of drivers which might be a leading factor for such a cancellation rate.

Figure 7: Distribution of rides per driver

**Effectiveness of the Payment Method**

The below bar chart indicates the number transaction gone for each payment method. Based on the distribution we can say that there's a even distribution across the payment methods. Each of the payment gateways have approximately around 20,000 transactions. The customer do not have a distinct preference of a payment method. The uniform distribution shows that the company has provided with diverse payment options which almost all methods are frequently used. This further illustrates that the company has a flexible payment method where the customer has the option to choose over digital method over the traditional method.

A graph of different colored rectangular shapes

Description automatically generated

Figure 8: Payment method effectiveness

**Promo Code Utilization**

A chart of different colors

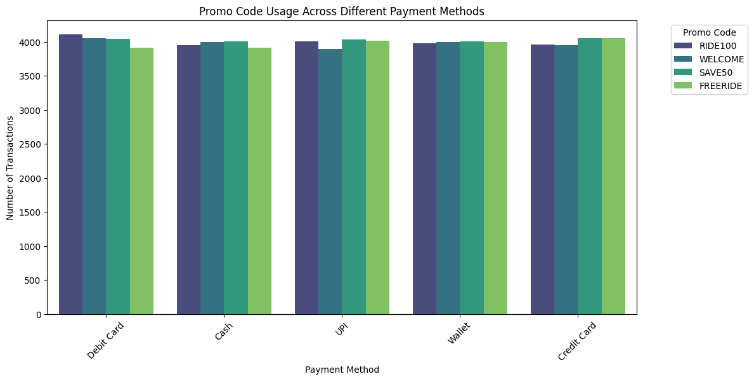
Description automatically generated with medium confidenceWhen it comes to the Promo\_Code utilization, there are four Promo\_Codes which are RIDE100, WELCOME, SAVE50, FREERIDE. All promocodes are evenly distributed across all the payment method. A significant difference is not found in the usage of promo code of customers regardless of how they prefer to pay. Based on the graph we can see that there is equal no. of transaction for each promo code for each payment method. It is not biased around one payment method where the customers are okay of using all promo codes with all methods around. All the Promo\_Codes are being used 20000 times approximately.

Figure 9: Usage of Promo\_Code

**Relationship Between Ride Distance, Fare and Travel Time**

According to the correlation matrix we can see that there's nearly no correlation between ride distance and ride fare ride distance and travel time and ride fare and travel time which are -0.0051, -0.00035 and 0.0024 respectively which are very close to zero and having a nonlinear relationship. This is further proofed by the heatmap. this may be due to flat-rate pricing model also which might be influenced by other factors

A group of blue squares

Description automatically generatedA screenshot of a graph

Description automatically generated

Figure 10: Relationship between Ride Distance, Fare and Travel Time

**Variations in fare rates across cities**

Across all the cities their average fare is consistent which is around 250 to 300 units. There's no difference in the average fare value in different cities. The IQR is almost similar for all the cities which means the fare values falls in the similar range despite of the cities. The fare values range with a minimum of 20 units to maximum of 500 units across all cities which directly says that there's a standardized fare values in all the cities. Outliers are not found which means the distribution of the fare values is evenly distributed and contains a manageable or controllable price mechanism.

A chart of different colored squares

Description automatically generated

Figure 11: Fare Variation by City

**Cancellation rate by cities.**

When considering about the cancellation rate across the cities we can see there's a consistent cancellation rate across all the cities which is having a rate closer to 50 %. This indirectly says that across all the cities the factors that affect the cancellation is similar. The cancellation is extremely high which is half of the booking are being cancelled which needs to be more concerned on as mentioned previously also being proved again with the below chart. In close examination of the visualization, we can see that the highest cancellation rate has been occurred in the city Bengaluru which accounts for approximately to 50.46% cancellation rate.

A graph of different colors

Description automatically generated

Figure 12: Cancellation Rate by city

**Analysis of Cancellation by Payment Method**

A graph of a graph with red and green bars

Description automatically generatedIf we consider which payment method had contributed the most cancellation, we can make a strategic plan to make that payment method familiarized and convenient for the customers to lessen the cancellation rate associated with the payment method. The figure 13 which shows a stacked bar chart illustrates the completed and the cancelled ride distribution occurred under each payment gateway. The key insights that we can gain from this visualization is that in comparison of the completed rides there’s an equal proportion of cancellation rides under each payment method.

Figure 13: Ride Cancellation by Payment Method

**Analysis of Ride Fare, Ride Distance and Cancellation**

To further analyze the factors that affects the cancellation a correlation matrix is used to visualize the relationship between the ride fare, ride distance and cancellation. Based on the below correlation matrix we can see that there’s a weak correlation between these factors which directly tells that the cancellation is not affected by the fare ride, distance which is proved by the correlation.

A blue and red squares with white text

Description automatically generated

Figure 14: Correlation between Ride Fare, Ride Distance and Cancellation

**Cancellation trend over the time in a day by city**

A graph of different colored lines

Description automatically generatedIf we consider the below chart, we can see that in the morning hours between 12AM to 6AM we can see that a very low cancellation rate. This might be due less rides as people travel less in the early mornings. Between the morning hour 6AM to 10AM we can see that there’s an increase in the cancellation in cities Bengaluru and Mumbai. We can say that these hours are rush hours as people need to go work through public transport in India and more demand is created. Although a higher demand is created in these hours since a driver is allocated one drive per day, driver unavailability will occur and creating a longer waiting time leading to a higher probability of cancelling the ride. Apart from that Traffic congestion may lead to another reason for cancelling. Between the Midday from 10AM to 4PM there’s a stability in cancellation rate in most of the city. A moderate demand can be seen in these hours and the availability of the drivers has been improved in these hours. In the evening between 4PM to 8PM which is also a rush hour in most of the cities in India which leads to high cancellation rate as people return from work to home and the drivers might be selective in accepting long distance ride. Between 8PM to 12 AM we can see a mixture of trends which we can see a less cancellation rides in some cities due to low ride requests.

Figure 15: Cancellation over time of the day and city

**Ride Demand by City**

When considering about the ride demand we can see that there's a consistent no.of rides across all cities which has around 16000 rides. This indicates that the service provided by rapido is equal across all the cities which has a well distributed service across all. No city stands significantly for higher ride demand. All the cities have access to platform. This might be due to the standardized service offered by the platform

A graph of different colored bars

Description automatically generated with medium confidence

Figure 16: Demand by city

**Analyzing travel time variation by city**

When it comes to analysis of travel time variation by city, we can see that the median time is between 30 to 35 minutes across all the cities. The average time it takes to travel is consistent across all the cities. The IQR ranges from 20 to 45 minutes. The minimum time taken to travel is close to 10 minutes while the maximum is 60 minutes. This shows that there is a uniformity in Travel experience.

A graph showing different colored squares

Description automatically generated

Figure 17: Boxplot on travel time variation by city

**Customer Segmentation using K- Mean clustering**

For the business to understand the behavior if the customer a customer segmentation is performed using the K- mean clustering. The reason behind the use of K- mean clustering is that it effectively handles big data and is versatile which works for both numerical and categorical parameters and have clear boundaries which is the assigning of customers to a one cluster. As the initial step in applying K-Mean clustering first step involves feature selection which is selection of parameters/features that will affect the customer behavior. So the features that were chosen are the Ride\_Fare, Ride\_Distance, Travel\_Time and Cancellation\_Status. The next step involves preparing the data which involves in identifying NaN values which have been already performed in the initial stage itself since there are not any missing values in the selected features the normalization is done to bring the features to same scale. The next step involves is selecting the correct no of clusters or selecting the optimal no. of clusters for this the Elbow method is used where plotting the sum of squared distances within the clusters with the no of clusters. Then identification of the elbow point, which is considered as the K, the no. of clusters. According to the Figure 18 we can see that the elbow point is at 3 which means that the division of the dataset to 3 clusters will be effective in segmenting the customers.

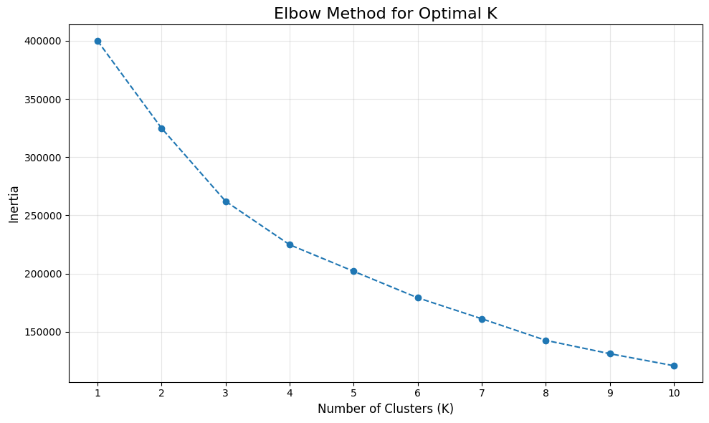


Figure 18: Elbow Method for the optimal K

After the selection of the 3 clusters as the optimal cluster next is the application K-Means Clustering with the chosen cluster size and visualized the result using the scatter plot which is shown below in the Figure 19.



Figure 19: K-Mean Clustering

Based on the scatter plot we can derive the following key insights:

The clusters are labeled as 0, 1,2 which are segmented on Ride Fare in the y-axis and the ride distance on the x axis. The Green cluster labeled as 0 shows customers with long distance rides with a middle level ride fare. The Orange cluster labeled as 1 shows customers with short to moderate distance rides with low level ride fare. The Blue cluster labeled as 2 includes the customers with ride fares that is high irrespective of the ride distance. Based on this we can say that the cluster 2 composite customers who use premium services as their rides are longer. Cluster 1 customers are more conscious of their cost who are more likely to travel shorter distance for a more affordable price. Cluster 0 composite with customers who are regular customers and casual which prefer a moderate spending habit.

# Recommendation

Based on the EDA there are few recommendations for the Rapido management to consider.

Introduction of a dynamic pricing model which needs to be adjusted according to the demand peak hours and traffic jam. This will enable the company to bring a more competitive price for customers with off-peak times.

Introduction of the fare based on time and distance. Based on the correlation matrix that we build we saw that there's no correlation between the time and distance and fare. By introducing a fare based on time and distance will improve customer conception for value for money.

Offering discounts by targeting less busy time and for shortest route options, also can offer a personalized discount based on the history of that customer. This will improve customer engagement and customer loyalty.

Also to enhance operational efficiency should introduce an effective method to reduce the cancellation rate which is high across all cities as we observed. To do this we can allocate more rides for drivers per day by providing incentives for drivers so that they are enthusiased to get more rides on a day. Also, they can utilize the data to allocate drivers for forecasted demand. This will make the customers waiting time to be eliminated as the most optimum driver is allocated.

Optimize the route with AI. Can use AI driven optimization routes to find the shortest route to reduce the travel time specially in peak hours.

Implementing a method to give customer feedback after every ride. This real time feedback can be used to enhance the customer satisfaction. And giving the option to rate the driver regarding the ride. This will enable the riders to be more motivated. Implement a loyalty program so that the customers get encouragement to rebooking. Awarding them with discounts and free rides after a certain number of bookings. Should provide transparency in fare calculation, how the fare is calculated based on distance and time. Should have a flexible payment option.