Exploring Delivery Time Data

Understanding the Key Factors Influencing Delivery Time

Introduction

Delivery speed is a critical factor in evaluating both customer happiness and operational efficiency in the fast-paced world of food delivery services. In addition to guaranteeing that consumers receive their goods hot and fresh, prompt delivery encourages repeat business and customer loyalty. However, delivery delays can result in irate clients, bad reviews, and lost business. Therefore, for food delivery businesses looking to optimize operations, cut expenses, and enhance customer satisfaction, knowing the elements that affect delivery time is crucial.

The primary objective of this research is to examine a dataset that documents many facets of the food delivery process and examine the ways in which variables like distance, vehicle type, and geographic location affect delivery time. We want to use this exploratory data analysis (EDA) to find hidden patterns and insights that may help direct enhancements in route design, vehicle allocation, and delivery logistics.

Objective: This project's main goal is to examine the delivery dataset and look at the ways that different factors affect delivery time. In particular, we want to

Recognize the Connection Between Delivery Time and Distance: We postulate that longer delivery times might arise from greater distances between the restaurant and the delivery site. We seek to quantify this relationship by determining the geographical distance using latitude and longitude data.

Evaluate How Vehicle Type Affects Delivery Time: Delivery times may be directly impacted by the speed and capability of various vehicle types, such as motorcycles, scooters, and electric scooters. To comprehend the function of transportation efficiency in the delivery process, we will look at how delivery times vary among vehicle kinds.

Examine the Function of Order Type and Delivery Person Characteristics: It may take different lengths of time to make and deliver different kinds of food orders (such as meals, snacks, and drinks). The efficiency of the delivery procedure may also be impacted by variables like the delivery person's rating and age.

We anticipate that these studies will yield useful information that will enable food delivery companies to streamline their processes. This might entail suggestions on how to choose the most effective cars for particular kinds of orders, figure out the most effective delivery routes, and improve delivery time prediction.

Why Delivery Time Matters

Delivery time has a direct impact on customer happiness and profitability in the cutthroat food delivery sector. On-time delivery is essential for clients. Customers are more inclined to stick with services that continuously meet or beyond their delivery expectations, according to research. Conversely, delays may result in unfavorable client experiences, unfavorable evaluations, and a decline in brand loyalty.

Businesses may operate more effectively if they are aware of the variables that influence delivery time. In addition to increasing customer satisfaction, cutting delivery times enables companies to make more deliveries in a given amount of time, which boosts profitability. Businesses may improve their logistics, better allocate resources, and cut expenses by understanding which operational factors—like vehicle type or geographic location—have the biggest effects on delivery time.

Moreover, predicting delivery times more accurately can improve customer expectations, reduce anxiety around delivery status, and help businesses manage resources like drivers and vehicles more efficiently. In summary, efficient management of delivery times is a key differentiator in the food delivery market, offering both operational and customer satisfaction benefits.

Data Overview

Several facets of the delivery process are captured in the project's dataset, which includes comprehensive records of delivery transactions. The dataset offers comprehensive information about the delivery time, order details, and geographic areas involved in each transaction, with a total of 45,593 records and 11 columns.

Columns in the Dataset:

1. **ID**: A unique identifier for each delivery record.
2. **Delivery_person_ID**: The unique identifier for the delivery person, which helps track individual performance.
3. **Delivery_person_Age**: The age of the delivery person, which could be a factor influencing the delivery time based on experience or physical capability.
4. **Delivery_person_Ratings**: Rating of the delivery person, usually given by customers. This could reflect the person's efficiency or reliability.
5. **Restaurant_latitude**: The latitude coordinate of the restaurant where the order originates.
6. **Restaurant_longitude**: The longitude coordinate of the restaurant.
7. **Delivery_location_latitude**: The latitude of the delivery destination (the customer's location).
8. **Delivery_location_longitude**: The longitude of the delivery destination.
9. **Type_of_order**: This indicates the type of food ordered (e.g., **Snack**, **Buffet**, **Meal**, **Drinks**). Different types of orders might have different preparation and delivery requirements.
10. **Type_of_vehicle**: The vehicle used for delivery, which could include **motorcycle**, **scooter**, or **electric scooter**. The type of vehicle likely affects the speed and efficiency of the delivery.
11. **Time_taken(min)**: The time, in minutes, it took to complete the delivery, which is the dependent variable we aim to analyze.

The dataset offers a thorough understanding of the food delivery process by concentrating on three important characteristics: delivery time, distance, vehicle type, and geographic coordinates. Better operational decisions, including labor management, vehicle allocation, and delivery route optimization, will be possible with the capacity to examine how these factors interact and affect delivery time.

Additionally, using the customer's latitude and longitude data, the Haversine formula was utilized to determine the distance between the restaurant and their delivery location. We will be able to directly evaluate how geographic distance affects delivery time thanks to this new functionality, which may also provide further insights into delivery operations.

Data Preprocesing

1. Handling Missing Data

It is not uncommon for real-world datasets to contain missing data. Numerous factors, including mistakes made during data collection, inaccurate input, or unrecorded information, may cause it. We came across a number of missing or zero values during this project, especially in the geographic coordinates for some records. To preserve the dataset's integrity, these missing values were handled carefully.

Imputation or Removal:

- Missing Geographic Coordinates: In certain records, the restaurant or delivery location's latitude or longitude (with values like 0 or NaN) was absent. We decided to eliminate these rows completely because delivery distance calculations depend on geographic location. This guarantees the accuracy and significance of our distance computations using the Haversine formula. Inconsistencies that might distort the results are also removed by removing rows with missing coordinates.

- Delivery Time & Other Columns: We looked for missing or inconsistent values (such as negative time values or zeroes where they shouldn't be) in the Time_taken(min) and other numerical columns. Records that lacked the Time_taken(min) were not included in the analysis. This guarantees that the data we use for our analysis is accurate and comprehensive.

By removing these records, we ensured the dataset’s accuracy and reliability. While imputation is a common approach for missing values, in this case, removing the rows with missing or zeroed-out values for geographic data was more appropriate due to the significance of those features in our analysis.

2. Data Types

Data can be categorized into **numerical** and **categorical** types. Understanding these data types is important because they dictate the types of analysis and preprocessing techniques we can apply to each column.

Numerical Data:

- **Delivery_person_Age**: This column represents the age of the delivery person. As an integer, it is a **numerical** data type. We used this feature to explore if age impacts delivery time or efficiency.
- **Delivery_person_Ratings**: The ratings for each delivery person are **numerical** (with decimal values). This column reflects customer feedback and can be used to investigate the potential relationship between delivery time and the performance of delivery personnel.
- **Time_taken(min)**: This is the key target variable for our analysis and represents the time taken for each delivery in minutes. This is **numerical** data, and we will explore how it is influenced by other features in the dataset.
- **Distance (Calculated)**: The distance between the restaurant and delivery location was calculated using the **Haversine formula** and is a **numerical** feature. This distance is crucial for understanding how geographic factors affect delivery time.

Categorical Data:

- **Delivery_person_ID**: This is a **categorical** feature identifying the delivery person. It's a nominal variable used to group and track individual delivery performance.
- **Type_of_order**: This represents the category of food ordered, such as **Snack**, **Buffet**, **Meal**, or **Drinks**. As a **categorical** feature, it helps us understand how different types of orders may have varying delivery times or processing requirements.
- **Type_of_vehicle**: This column categorizes the vehicle used for delivery (e.g., **motorcycle**, **scooter**, or **electric scooter**). We will analyze whether certain vehicles are associated with faster delivery times based on their speed and efficiency.
- **Restaurant_latitude** and **Restaurant_longitude**: While these are **numerical** by nature, they represent **geographical data** and are often treated as **coordinates**. This allows us to calculate the **distance** to the delivery location.

Categorizing the data correctly is vital for the subsequent analysis. Numerical data is often used in correlation analyses, regressions, and distance calculations, while categorical data helps us segment the dataset into different groups, enabling more targeted insights.

3. Feature Engineering

Feature engineering is an essential step in improving the performance of machine learning models and gaining deeper insights from data. In this project, one of the most crucial features that was engineered is the **calculation of the distance** between the restaurant and delivery location using the **Haversine formula**.

Haversine Formula for Distance Calculation:

The **Haversine formula** is a mathematical method used to calculate the shortest distance between two points on the Earth's surface given their **latitude** and **longitude**.
The formula is particularly useful when working with geographical data and allows us to quantify the spatial relationship between the restaurant and the delivery location.
The Haversine formula is defined as:

$$a = \sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1)\cdot\cos(\text{lat}_2)\cdot\sin^2\left(\frac{\Delta\text{lon}}{2}\right)$$

$$c = 2\cdot\text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R\cdot c$$

Where:

- lat_1,lon_1 are the coordinates of the first point (restaurant).
- lat_2,lon_2 are the coordinates of the second point (delivery location).
- R is the Earth's radius (mean radius = 6,371 km).
- dis the distance between the two points in kilometers.

By applying this formula to the latitude and longitude of both the **restaurant** and the **delivery location**, we were able to create a new feature, **Distance** (in kilometers), which is a crucial factor influencing the delivery time. This feature allows us to analyze how delivery time is affected by the physical distance between the two locations.

4. Data Cleaning

Data cleaning is an essential part of ensuring that the dataset is consistent, reliable, and free from errors. In this project, several steps were taken to clean and preprocess the data:

1. **Removing Duplicates**: Duplicate records can skew the results of our analysis. We identified and removed any duplicate rows to ensure the dataset contains only unique records.

2. **Handling Inconsistent or Erroneous Data**: Some entries had erroneous or impossible values. For example, there were instances where the **latitude** and **longitude** values were zero, indicating missing or invalid geographic data. These records were removed to prevent the introduction of inaccuracies into our analysis.

3. **Standardizing Data Formats**: We ensured that all columns were in the correct data format. For instance, columns such as **Delivery_person_Age** and **Delivery_person_Ratings** were converted to the appropriate **numeric** format, and **Type_of_order** and **Type_of_vehicle** were ensured to be **categorical**.

4. **Consistency Checks for Delivery Time**: Some records had negative or zero delivery times, which are not realistic in the context of this dataset. These were identified and removed to maintain the integrity of the analysis.

5. **Handling Missing Values**: As previously mentioned, rows with missing geographic data (e.g., NaN or 0 values for coordinates) were removed. This ensured that the distance calculation and subsequent analysis of delivery time were based on valid information.

By conducting thorough data cleaning and preprocessing, we ensured that the dataset was ready for accurate analysis and meaningful insights.

Visualizations

- **Histogram (Distribution of Delivery Time)**

  Purpose: Visualizes the frequency distribution of delivery times across the dataset. **Insights**: Helps identify the central tendency (e.g., mode), skewness (positive/negative), and whether the delivery times are concentrated around a specific range.
- **Boxplot of Delivery Time**

  Purpose: Shows the summary statistics of delivery time, including median, IQR (Interquartile Range), and outliers.

  Insights: Helps identify whether the data has outliers and the spread of delivery times. If there are significant outliers, this could indicate exceptional cases or issues.
- **Density Plot of Delivery Person Age**

  Purpose: Displays the distribution of delivery person ages in a smooth curve. **Insights**: Helps understand the concentration of delivery people in specific age groups. Peaks in the plot show the most common ages.
- **Violin Plot of Delivery Time by Type of Order**

  Purpose: Combines boxplot and density plot to show the distribution of delivery time by different types of orders.

  Insights: Visualizes the spread, median, and skewness of delivery time for each order type. Helps understand how delivery times vary across different order categories.
- **Line Plot of Time Taken and Delivery Person Age**

  Purpose: Shows the trend of delivery time relative to the delivery person's age.

  Insights: Helps identify any patterns or relationships between delivery time and the age of the delivery person.
- **Pie Chart: Proportion of Types of Vehicles**

  Purpose: Displays the percentage distribution of delivery vehicles in the dataset.

  Insights: Allows you to quickly understand the most common vehicle types used for deliveries (e.g., motorcycle, scooter, etc.).
- **Pie Chart: Proportion of Types of Orders**

  Purpose: Shows the proportion of different order types in the dataset (e.g., snacks, drinks, meals).

  Insights: Provides an understanding of which types of orders are most frequent and how they affect delivery time.

- **Barplot: Average Time by Vehicle Type**

  Purpose: Visualizes the average delivery time for different types of vehicles.

  Insights: Helps compare delivery times based on vehicle type and identify which vehicle type tends to take more or less time.

- **Countplot: Count of Deliveries by Vehicle Type**

  Purpose: Displays the frequency of deliveries based on vehicle type.

  Insights: Helps identify which vehicle types are used the most and can help explain any delivery time differences (e.g., more motorcycles could imply faster deliveries).

- **Histogram: Distribution of Delivery Distance**

  Purpose: Displays the distribution of delivery distances.

  Insights: Helps in understanding the spread of delivery distances and how far deliveries are generally occurring.

- **Density Plot of Delivery Person Ratings**

  Purpose: Visualizes the distribution of delivery person ratings.

  Insights: Helps assess whether most delivery people have high ratings (positive skew) or lower ratings.

- **Boxplot of Delivery Person Ratings**

  Purpose: Shows the distribution and spread of ratings for delivery people.

  Insights: Helps identify any outliers or particularly low/high ratings for the delivery personnel.

- **Scatter Plot: Time Taken vs Delivery Person Age**

  Purpose: Shows the relationship between delivery time and the delivery person’s age.

  Insights: Identifies if there’s any linear/non-linear relationship between the two variables.

- **Scatter Plot: Time Taken vs Delivery Person Rating**

  Purpose: Investigates the relationship between delivery time and the delivery person’s rating.

  Insights: Helps understand if higher-rated delivery persons tend to have faster or slower deliveries.

- **Jointplot on Distance and Time Taken**

  Purpose: Shows the relationship between delivery distance and time taken using scatter and density plots.

  Insights: Helps determine if longer distances lead to longer delivery times.

- **Line Plot: Time Taken vs Delivery Person Rating**

  Purpose: Shows how time taken varies with the delivery person's rating.

  Insights: This could help identify whether higher ratings correlate with faster delivery times.
- **Scatterplot: Geospatial Plot - Restaurant Latitude vs Delivery Location Latitude**

  Purpose: Visualizes the geographical relationship between the restaurant and the delivery location (latitude).

  Insights: Helps identify if deliveries tend to occur in specific geographic areas or if there are clusters of delivery locations.
- **Scatterplot: Geospatial Plot - Restaurant Longitude vs Delivery Location Longitude**

  Purpose: Similar to the latitude plot but focuses on longitude.

  Insights: Helps with understanding the distribution of deliveries in terms of longitude.
- **Geospatial Heatmap: Density of Deliveries**

  Purpose: Visualizes the density of deliveries in different geographical locations using a heatmap.

  Insights: Helps identify regions with high delivery frequency or clusters of delivery points.
- **3D Scatter Plot: Time Taken vs. Age vs. Distance**

  Purpose: Shows the relationship between **Time Taken**, **Delivery Person Age**, and **Distance**.

  Insights: This complex visualization helps in understanding how multiple factors interact and affect delivery times.
- **Violin Plot: Delivery Time by Vehicle Type and Location**

  Purpose: Combines a violin plot with delivery time by both vehicle type and geographic location.

  Insights: Helps understand the combined effect of vehicle type and location on delivery time.
- **Box Plot: Delivery Time by Region**

  Purpose: Displays how delivery times vary across different regions (e.g., by city or state).

  Insights: Helps identify if certain regions have consistently faster or slower delivery times.
- **Clustered Bar Plot: Average Time by Order and Vehicle**

  Purpose: Shows average delivery time across different order types and vehicle types.

  Insights: Helps compare how vehicle types and order types interact to influence delivery time.

- **Stacked Bar Plot: Distance by Order Type and Vehicle**

  Purpose: Visualizes the breakdown of delivery distances by order type and vehicle type.

  Insights: Helps you understand how far deliveries travel based on the type of order and vehicle.
- **Pairplot: 'Delivery_person_Age', 'Delivery_person_Ratings', 'Distance', 'Time_taken(min)'**

  Purpose: Provides a pairwise comparison of multiple variables, helping to identify relationships between delivery time, age, ratings, and distance.

  Insights: Great for spotting correlations or patterns between multiple variables.
- **Pairplot: `sns.pairplot(df, hue='Type_of_vehicle', palette='colorblind')**

  Purpose: Compares multiple variables, with the color hue representing the vehicle type.

  Insights: Helps explore how different vehicle types impact delivery time, distance, and other features.
- **Heatmap: Delivery Person Age, Delivery Person Ratings, Distance, Time Taken**

  Purpose: Displays correlations between different continuous variables.

  Insights: Helps visualize which variables are highly correlated, aiding in feature selection for modeling.
- **Bubble Plot: Time Taken vs. Delivery Person Age with Ratings**

  Purpose: Visualizes the relationship between **Time Taken**, **Delivery Person Age**, and **Delivery Person Ratings** with bubble size representing the number of deliveries.

  Insights: Shows how the delivery time and ratings are affected by age, with a bubble size indicating frequency.

Statistics Analytics

- **T-Test:** Used to compare the means of two variables (Time Taken and Distance) to see if they are significantly different.
- **ANOVA:** Used to compare the means of multiple variables (e.g., Delivery Person Ratings, Time Taken, and Distance) to see if there are significant differences. One-Way ANOVA compares three or more groups, while Two-Way ANOVA evaluates the effects of two categorical variables and their interaction on a continuous outcome.
- **Chi-Square Test:** Used to examine the relationship between two categorical variables (e.g., Distance Category and Time Taken).
- **Grouping and Averaging:** Helps identify patterns by calculating the average delivery time for different combinations of variables (e.g., Type of Order and Vehicle).
- **Two-Way ANOVA:** Investigates the effects of two categorical variables on a continuous variable and their interaction.

These statistical analyses help understand relationships between different variables in the dataset, and provide evidence for making data-driven decisions, improving delivery times, and optimizing business processes.

Key Findings

Strong Correlation Between Delivery Time and Distance

A key finding from the dataset is that distance has a strong correlation with delivery time. This suggests that, as expected, longer distances take more time for delivery. This relationship can be seen through various statistical tests and visualizations, like the scatterplot between Distance and Time Taken.

The Chi-Square Test and ANOVA results also support the idea that distance category (Low, Medium, High) significantly influences delivery time, further emphasizing the importance of distance as a key determinant of delivery efficiency.

Vehicle Type Significantly Affects Delivery Time

The type of vehicle used for delivery plays a significant role in determining the delivery time. Faster vehicles (e.g., scooters or motorcycles) tend to reduce delivery time compared to slower ones (e.g., bicycles or vans). This conclusion is supported by visualizations like the Boxplot of Delivery Time by Type of Vehicle, which shows that delivery times for vehicles like Scooters and Motorcycles are lower than for Vans.

The Two-Way ANOVA analysis shows that the interaction between vehicle type and order type significantly affects the delivery time, highlighting the importance of choosing the right vehicle for different types of deliveries.

Delivery Person Ratings Have a Significant Impact on Delivery Time

Delivery person ratings (how well the delivery person is rated by customers) have a statistically significant impact on delivery time. This is particularly interesting because it suggests that better-rated delivery persons might be more efficient or experienced, thus reducing the time taken for deliveries.

The ANOVA and Chi-Square tests provide evidence that higher-rated delivery persons tend to have better performance in terms of delivery time. Ratings can also be used as a metric to optimize delivery processes by selecting high-performing delivery persons for more time-sensitive orders.

The Importance of Optimizing Routes, Improving Ratings, and Choosing the Right Vehicle Type

Given the strong correlations between distance, vehicle type, and delivery time, optimizing the delivery route is crucial. Long-distance deliveries can be made more efficient by selecting faster vehicles and optimizing the path using route optimization algorithms.

Improving delivery person ratings is another important factor. Investing in training and customer satisfaction can potentially improve delivery times by enhancing the efficiency of high-rated delivery persons.

The choice of vehicle is also essential. Businesses should consider using faster vehicles (e.g., scooters or motorcycles) for short-distance deliveries, while more spacious vehicles (e.g., vans) may be better suited for larger orders or longer distances.

Conclusion

This analysis of the delivery dataset has provided valuable insights into the key factors influencing delivery time. Through statistical testing, visualizations, and feature engineering, we were able to uncover several critical relationships that can drive business decisions aimed at improving operational efficiency and customer satisfaction.

1. Distance as a Primary Factor: The strong correlation between delivery time and distance reaffirms the intuitive notion that longer distances require more time for deliveries. This suggests that optimizing delivery routes, and potentially using different vehicle types for varying distances, is crucial to improving overall delivery efficiency.

2. Vehicle Type Matters: The analysis demonstrated that the type of vehicle used significantly affects delivery times. Faster vehicles such as scooters and motorcycles reduce delivery time, making them ideal for short-distance or time-sensitive deliveries. This highlights the importance of having a varied fleet and strategically selecting vehicles based on the specific requirements of each delivery.

3. Impact of Delivery Person Ratings: Delivery person performance, as reflected in their ratings, plays a substantial role in determining delivery time. Well-rated delivery persons tend to perform more efficiently, potentially due to experience, professionalism, or better customer interaction. Businesses should consider focusing on improving delivery person ratings as part of their strategy for reducing delivery times.

4. Statistical Evidence Supports Operational Strategies: The statistical tests (ANOVA, T-tests, and Chi-Square) and visualizations have provided solid evidence that businesses can optimize delivery operations by focusing on factors such as distance, vehicle type, and delivery person ratings. These factors are not only correlated with delivery time but also significantly influence it.