Report: Optimizing NYC Taxi Operations

This report presents an exploratory data analysis (EDA) of 2023 yellow taxi trip data in New York City. The goal is to uncover insights that could help optimize taxi operations, improve service efficiency, maximize revenue, and enhance passenger experience.

1. Data Preparation

1.1 Loading the Dataset

The dataset consists of twelve Parquet files, one for each month of 2023. Due to the large size of the dataset, a sampling strategy was employed to reduce computational load while preserving data representativeness.

1.1.1 Sampling the Data and Combining the Files

A 5% random sample of trips was taken for each hour of every day across all months. This approach ensures a balanced representation of temporal patterns. The sampled data from all months was then combined into a single Parquet file for further analysis.

2. Data Cleaning

2.1 Fixing Columns

2.1.1 Fix the Index

The index of the combined DataFrame was reset to ensure sequential numbering. Unnecessary columns ('VendorID', 'store_and_fwd_flag') were dropped.

2.1.2 Combine the two airport_fee columns

Two columns representing airport fees were identified and combined into a single column ('airport_fee') to avoid redundancy.

2.2 Handling Missing Values

2.2.1 Find the Proportion of Missing Values in Each Column

The proportion of missing values in each column was calculated. 'Passenger_count', 'RatecodeID', and 'congestion_surcharge' had missing values.

2.2.2 Handling missing values in passenger_count

Missing values in 'passenger_count' were imputed using the median value. Zeroes in 'passenger_count' were also addressed.

2.2.3 Handle missing values in RatecodeID

Missing values in 'RatecodeID' were imputed using the mode (most frequent value).

2.2.4 Impute NaN in congestion_surcharge

Missing values in 'congestion_surcharge' were imputed with 0, assuming no congestion surcharge was applied in those cases.

2.3 Handling Outliers and Standardizing Values
2.3.1 Check outliers in payment type, trip distance, and tip amount columns
Outliers were identified and handled based on logical reasoning:

Trips with near-zero distance and high fare amounts were removed.
Trips with zero distance and fare but different pickup and drop-off zones were removed.
Trips with distances exceeding 250 miles were removed.
Trips with invalid payment types (0) were removed.
3. Exploratory Data Analysis
3.1 General EDA: Finding Patterns and Trends
3.1.1 Classify variables into categorical and numerical
Variables were categorized as follows: Categorical: VendorID, RatecodeID, PULocationID, DOLocationID, payment_type, pickup_hour, pickup_day, pickup_month Numerical: tpep_pickup_datetime, tpep_dropoff_datetime, passenger_count, trip_distance, trip_duration, and all monetary parameters (fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee).

3.1.2 Analyze the distribution of taxi pickups by hours, days of the week, and months
Visualizations (bar plots) were created to show hourly, daily, and monthly pickup trends. Peak hours, days, and months were identified.

3.1.3 Filter out the zero/negative values in fares, distance, and tips
Trips with zero or negative values for fare, distance, and tips were filtered out for specific analyses where these values were irrelevant.

3.1.4 Analyze the monthly revenue trends
Monthly revenue trends were visualized using a line plot, showing fluctuations and overall patterns.

3.1.5 Find the proportion of each quarter’s revenue in the yearly revenue
The proportion of revenue generated in each quarter was calculated and presented.

3.1.6 Analyze and visualize the relationship between distance and fare amount
A scatter plot was used to visualize the relationship between trip distance and fare amount, revealing a positive correlation. The correlation value was also calculated.

3.1.7 Analyze the relationship between fare/tips and trips/passengers
Relationships between fare/tips and trip duration/passenger count were visualized using scatter plots and box plots. Correlation values were calculated.

3.1.8 Analyze the distribution of different payment types

A count plot was used to show the distribution of different payment types. The proportions of each payment type were also calculated.

3.1.9 Load the taxi zones shapefile and display it
The taxi zones shapefile was loaded using GeoPandas and displayed on a map.

3.1.10 Merge the zone data with trips data
The zone data (from the shapefile) was merged with the trip data using location IDs.

3.1.11 Find the number of trips for each zone/location ID
Trips were grouped by location ID to calculate the total number of trips for each zone.

3.1.12 Add the number of trips for each zone to the zones DataFrame
The trip counts were added to the zones GeoDataFrame.

3.1.13 Plot a map of the zones showing the number of trips
A choropleth map was created to visualize the number of trips per zone, using color-coding to represent trip density.

3.1.14 Conclude with results
General EDA Findings:

Busiest Hours: Evenings and late nights (especially 6 PM to 9 PM) saw the highest pickup activity.
Busiest Days: Weekdays, particularly Tuesday to Friday, had higher trip counts compared to weekends.
Busiest Months: February and March had a slightly higher number of pickups compared to other months.
Revenue Trends: Monthly revenue exhibited fluctuations, but there was a general upward trend over the year.
Quarterly Revenue: Revenue was relatively distributed across quarters.
Fare-Distance Relationship: A strong positive correlation exists between trip distance and fare amount.
Fare-Trip Duration/Passenger Count Relationship: Fare amount also showed a moderate correlation with trip duration and passenger count.
Tip-Distance Relationship: A moderate positive correlation was observed between tip amount and trip distance.
Busiest Zones: Zones in and around Manhattan (e.g., Midtown, Upper East Side) had the highest pickup and drop-off activity.
3.2 Detailed EDA: Insights and Strategies
3.2.1 Identify slow routes by comparing average speeds on different routes
Average speeds on routes were calculated by dividing distance by trip duration. The slowest routes during different hours were identified.

3.2.2 Calculate the hourly number of trips and identify the busy hours
The number of trips per hour was calculated and visualized. The busiest hour and the corresponding number of trips were identified.

3.2.3 Scale up the number of trips from above to find the actual number of trips
The actual number of trips was estimated by scaling up the sampled trip counts using the sampling fraction (0.05).

3.2.4 Compare traffic trends for the weekdays and weekends
Hourly traffic patterns were compared between weekdays and weekends using line plots, showing distinct variations in demand.

Detailed EDA Insights and Strategies:

Operational Efficiency:

Slow Routes: Identification of slow routes can inform route optimization and traffic management strategies, potentially reducing travel times and improving efficiency.
Busy Hours: Understanding busy hours allows for better allocation of resources, such as deploying more taxis during peak demand periods.
Weekday/Weekend Patterns: Different traffic patterns on weekdays and weekends suggest the need for tailored operational strategies for each day type.
Zone-wise Traffic: Identifying zones with high pickup and drop-off traffic during specific times helps with strategic positioning of taxis.
Pricing Strategy:

Fare per Mile per Passenger: Analysis of fare per mile per passenger for different passenger counts can inform pricing models to ensure fairness and revenue optimization.
Revenue Share by Time: Understanding revenue distribution across nighttime and daytime hours can inform decisions regarding surge pricing or promotional offers.
Customer Experience:

Wait Times: Analyzing pickup and drop-off trends in busy zones can help to minimize passenger wait times and improve overall satisfaction.
Route Optimization: Avoiding slow routes or areas prone to traffic congestion contributes to a smoother and more efficient ride experience for passengers.
Overall Conclusions:

This analysis provides valuable insights into various aspects of NYC taxi operations. The findings can be used to inform strategic decision-making regarding operational efficiency, pricing strategies, and customer experience. By implementing data-driven solutions based on these insights, taxi companies can optimize their services, increase profitability, and enhance passenger satisfaction.