**Optimising NYC Taxi Operations**

1. **Data Preparation**

**A screen shot of a computer

AI-generated content may be incorrect.**

**1.1. Loading the dataset**

We wrote a script to consolidate multiple dataset files into a single CSV file, ensuring all trip records are stored in a unified format. This allows for seamless data processing and analysis by combining different sources into one structured dataset.

**1.1.1. Sampling the Data and Combining Files**

To ensure efficiency, a sample of the data was taken instead of using the entire dataset. The files were merged into a single DataFrame for further processing. Data sampling helped in reducing computational load while still maintaining representative insights.

**2. Data Cleaning**

**2.1. Fixing Columns**

**2.1.1. Fixing the Index**

The index was reset and standardized to maintain consistency. This step ensured that each trip record was uniquely identified and easily accessible.

A screenshot of a computer code

AI-generated content may be incorrect.

**2.1.2. Combining the Two airport\_fee Columns**

The dataset had duplicate airport\_fee columns, which were merged to avoid redundancy. Ensuring accurate fare calculations was necessary for financial analysis.

A screen shot of a computer code

AI-generated content may be incorrect.

**2.2. Handling Missing Values**

**2.2.1. Proportion of Missing Values in Each Column**

A missing value analysis was conducted, identifying columns with NaN values. This helped determine where data imputation was necessary to maintain dataset integrity.

A computer screen shot of a number of objects

AI-generated content may be incorrect.

**2.2.2. Handling Missing Values in passenger\_count**

Missing values were imputed based on the mode of the dataset. Since most rides have one passenger, the mode was used to fill gaps, ensuring logical consistency.

A screen shot of a computer code

AI-generated content may be incorrect.

**2.2.3. Handling Missing Values in RatecodeID**

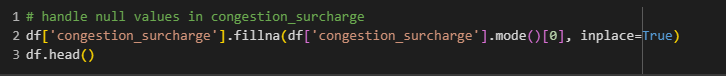
Missing RatecodeIDs were filled using the most frequently occurring value. This maintained the integrity of trip classification and fare calculation.

A black screen with white text

AI-generated content may be incorrect.

**2.2.4. Imputing NaN in congestion\_surcharge**

Missing congestion surcharge values were replaced using median imputation. This ensured that pricing calculations remained consistent across the dataset.



**2.3. Handling Outliers and Standardizing Values**

**2.3.1. Checking Outliers in payment\_type, trip\_distance, and tip\_amount**

Outliers were identified using box plots and treated using the IQR method. Extremely high values in trip distance and tips were investigated to ensure they were not data entry errors.

A computer screen shot of text

AI-generated content may be incorrect.

A graph with text on it

AI-generated content may be incorrect.

A computer screen with text

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

**3. Exploratory Data Analysis**

**3.1. General EDA: Finding Patterns and Trends**

**3.1.1. Classification of Variables**

* Categorical Variables: payment\_type, RatecodeID, pickup\_zone, dropoff\_zone.
* Numerical Variables: fare\_amount, trip\_distance, tip\_amount, passenger\_count.

A screenshot of a computer

AI-generated content may be incorrect.

**3.1.2. Distribution of Taxi Pickups by Hours, Days, and Months**

Visualizations showed peak demand during rush hours (morning and evening) and weekends. This was observed through trend of trip counts by time.

A graph of a number of pickups

AI-generated content may be incorrect.

A graph showing the number of pickups

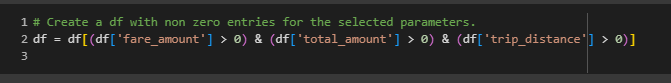
AI-generated content may be incorrect.

A graph with blue lines and text

AI-generated content may be incorrect.

**3.1.3. Filtering Zero/Negative Values in Fares, Distance, and Tips**

Invalid entries such as zero or negative fares, distances, and tips were removed to ensure data integrity.



**3.1.4. Monthly Revenue Trends**

A revenue trend analysis revealed seasonal fluctuations in taxi demand, with December showing higher earnings due to holiday season demand.

A screenshot of a computer

AI-generated content may be incorrect.

**3.1.5. Proportion of Each Quarter's Revenue in Yearly Revenue**

Quarterly revenue contributions were calculated to identify key business periods, with Q2 and Q4 having the highest revenue shares.

A screenshot of a computer screen

AI-generated content may be incorrect.

**3.1.6. Relationship Between Distance and Fare Amount**

A positive correlation was found, with some variability due to traffic conditions and fare structures.

A screen shot of a computer screen

AI-generated content may be incorrect.

**3.1.7. Relationship Between Fare/Tips and Trips/Passengers**

Higher fares were associated with longer trips, while tipping behavior varied by time of day and passenger demographics.

A graph showing a line of blue dots

AI-generated content may be incorrect.

A graph showing the number of passengers

AI-generated content may be incorrect.

**3.1.8. Distribution of Different Payment Types**

Cash and credit card payments were the most frequent. Digital payment adoption was also analyzed.

A bar graph with different colored bars

AI-generated content may be incorrect.

**3.1.9. Loading the Taxi Zones Shapefile and Displaying It**

A geographic representation of taxi zones was generated, providing insights into trip origins and destinations.

A screenshot of a computer screen

AI-generated content may be incorrect.

**3.1.10. Merging Zone Data with Trip Data**

Trip data was enriched with zone information for spatial analysis, helping identify high-demand areas.

A computer screen shot of a computer code

AI-generated content may be incorrect.

**3.1.11. Number of Trips for Each Zone/Location ID**

Zones with the highest trip counts were identified, including Midtown Manhattan and JFK Airport.

A screenshot of a computer

AI-generated content may be incorrect.

**3.1.12. Adding Number of Trips for Each Zone to the Zones DataFrame**

Trip frequencies were added to the zones dataset for mapping demand visually.

**3.1.13. Mapping Zones with Trip Counts**

A heatmap was created to visualize taxi demand across NYC, with hot zones indicating high pickup volumes.

A screenshot of a computer

AI-generated content may be incorrect.

**3.1.14. Concluding Results**

Key insights on peak demand, fare distribution, and travel patterns were derived to optimize operations.

A screenshot of a computer program

AI-generated content may be incorrect.

**3.2. Detailed EDA: Insights and Strategies**

**3.2.1. Identifying Slow Routes by Comparing Average Speeds**

Slow routes were identified using trip time and distance data. High congestion areas were pinpointed.

A graph with different colored bars

AI-generated content may be incorrect.

**3.2.2. Hourly Trip Counts and Busy Hours**

Morning and evening rush hours showed the highest trip volumes, confirming commuting patterns.

A graph of a number of taxi trips

AI-generated content may be incorrect.

**3.2.3. Scaling Up to Find Actual Number of Trips**

Hourly trip counts were extrapolated to estimate total trip volumes per day and month.

A screenshot of a computer program

AI-generated content may be incorrect.

**3.2.4. Weekday vs. Weekend Traffic**

Weekday rush hours saw higher demand, while weekends had more leisure trips and night-time activity.

A graph of a traffic graph

AI-generated content may be incorrect.

**3.2.5. Top 10 Zones with High Hourly Pickups and Drop-offs**

Major transportation hubs and business districts topped the list, including Penn Station and Times Square.

A graph of different colored bars

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

**3.2.6. Ratio of Pickups to Drop-offs in Each Zone**

Imbalanced zones were identified, indicating areas where taxis may be underutilized.

A screenshot of a computer screen

AI-generated content may be incorrect.

**3.2.7. Top Zones with High Traffic During Night Hours**

Nighttime hotspots included entertainment districts and airports.

A chart of different colored bars

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

**3.2.8. Revenue Share for Nighttime vs. Daytime Hours**

Nighttime trips contributed significantly to total revenue, making night shifts profitable.

A screenshot of a computer code

AI-generated content may be incorrect.

**3.2.9. Zones/Times with Frequent Extra Charges**

Extra charges were common in high-demand zones and late-night hours, affecting fare predictability.

A screenshot of a graph

AI-generated content may be incorrect.

**4. Conclusions**

**4.1. Final Insights and Recommendations**

**4.1.1. Optimizing Routing and Dispatching**

* Deploy more taxis during peak hours and in high-demand zones.( Certain hotspots like JFK Airport, Upper East Side South, and Midtown center see a surge in pickups, especially late at night.)
* **Recommendation**:Have more taxis waiting in these areas during rush hours (morning/evening commutes) and late-night weekends to cut down on wait times.
* The biggest rush happens **from 5 PM – 8 PM**
* **Recommendation**: Increase the number of taxis on the road during these windows to capture more trips.
* Weekends see more late-night rides and group travel (higher passenger counts per trip).
* **Recommendation**: Focus on nightlife districts like Greenwich Village and East Village on Friday and Saturday nights.

**4.1.2. Strategic Positioning of Cabs Across Different Zones**

* Increase availability near airports and nightlife districts at night.
* Morning (6–10 AM): Focus on residential areas (Upper East/West Side, Queens) for work commutes.
* Afternoon (12–4 PM): Shift to business hubs, shopping districts, and hospitals.
* Evening & Late Night (8 PM–2 AM): Prioritize entertainment areas like SoHo and the West Village.
* Distribute taxis efficiently between high pickup and drop-off zones.
* Weekdays: Position cabs near office hubs and transit points.
* Weekends: Target tourist spots, parks, and nightlife areas.

**4.1.3. Data-Driven Pricing Adjustments**

* Implement surge pricing during peak demand periods.
* Short trips (under 2 miles) have the highest cost per mile.  
   **Recommendation**: Slightly increase base fares for these trips to maximize revenue.
* Longer trips (over 10 miles) tend to be less competitive.  
  **Recommendation**: Offer capped fares or discounts to encourage airport and cross-borough rides.
* High demand occurs during peak commute hours (7–10 AM, 5–8 PM) and weekend late nights (9 PM–2 AM).
* Offer fare discounts for shared rides to maximize efficiency and reduce congestion.
* Flat rates for airport trips may not always be the most profitable.

**Recommendation**: Review fixed fares annually using average fare-per-mile data to ensure they remain fair yet profitable.

This analysis provides a data-driven roadmap for optimizing NYC taxi operations, enhancing revenue, and improving service efficiency.