EDA Optimizing NYC Taxis

Name: Prem

Assignment: NYC Taxi Data Analysis

Introduction

This report provides an in-depth analysis of NYC taxi trips using Exploratory Data Analysis (EDA). The goal is to identify key trends, operational inefficiencies, and propose data-driven strategies for optimizing pricing, dispatching, and cab positioning to maximize efficiency and revenue.

Assumptions

* The dataset is assumed to be a representative sample of NYC taxi trips.

* Passenger count is assumed to be accurate and does not account for multiple fares in shared rides.

* Fare and tip amounts are correctly recorded without major data discrepancies.

* Traffic conditions affecting trip duration are inferred from trip durations rather than real-time congestion data.

* Outliers such as extremely high fares or trip distances are assumed to be valid unless proven otherwise.

Handling Null Values & Outliers

1. Null Values

During data preprocessing, missing values were detected in:

* passenger_count, RatecodeID, and trip_distance.

* These were either filled with median values (for numerical features) or removed if they significantly impacted analysis.

2. Outlier Handling

To ensure data consistency and accuracy, the following outlier removal steps were applied:

Passenger Count

* Removed entries where passenger_count > 6, as standard taxis have a maximum seating capacity of 6.

Trip Distance & Fare Amount

* Removed trips where:
  * trip_distance == 0 and fare_amount == 0 but pickup and dropoff zones were different.
  * These records were likely incorrect or canceled trips.
* Capped trip_distance at 250 miles to remove extreme outliers.

Tip Amount

* Capped tip_amount at $90 (99th percentile) to reduce skewed values caused by extreme tipping cases.

Trip Duration

* Capped trip_duration at the 99th percentile to remove excessively long trips that may be anomalies.
* Set a lower limit of 1 minute to remove invalid 0-minute trips unless proven valid.

These steps ensure that the dataset is clean, reliable, and ready for further analysis.

Exploratory Data Analysis (EDA)

The analysis includes visualizations of demand patterns, pricing structures, and trip behaviors. Key insights from different plots are summarized below:

1. Average Passenger Count by Hour & Day

* Peak demand: 12 AM - 2 AM and 8 PM - 11 PM (nightlife and evening commuters).
* Lowest demand: 4 AM - 7 AM (early morning hours).
* Highest passenger count: Saturdays (leisure travel).
* Lowest passenger count: Wednesdays (mid-week drop in demand).

2. Heatmap of Pickups by Hour and Location

* Consistent demand in business districts during morning & evening rush hours.

* Nightlife areas (bars, clubs, entertainment zones) peak at night.

* Residential areas show low demand during work hours and higher demand in the evenings.

* Transit hubs (airports, train stations) have steady demand throughout the day.

3. Monthly Trends in Taxi Demand

* High demand between January - May (likely due to increased business travel and tourism).

* Lower demand from June - December, possibly due to seasonality and holiday periods.

* Adjusting fleet size seasonally could help reduce idle time and improve efficiency.

Results & Key Findings

1. Trip fare is strongly correlated with trip distance, but some long trips are underpriced.

2. Waiting time and congestion are not well-accounted for in fares, leading to revenue loss in high-traffic areas.

3. Short-distance trips have inconsistent pricing, leading to inefficiencies in fare structure.

4. Tip amounts increase with trip distance, indicating that passengers are willing to pay more for longer trips.

5. Peak hour pricing is effective, but mid-day demand is low, suggesting that promotional fares could help increase ridership.

Recommendations

1. Optimizing Routing & Dispatching

* Increase cab supply during peak hours (12 AM - 2 AM, 8 PM - 11 PM) in high-demand areas.

* Reduce fleet during low-demand hours (4 AM - 7 AM) to cut idle time & fuel costs.

* Dynamically allocate cabs based on hourly demand trends to reduce downtime and improve coverage.

2. Strategic Cab Positioning

* Business districts should have high availability during morning and evening rush hours.

* Entertainment & nightlife zones (bars, clubs, restaurants) require more taxis at night.

* Seasonal adjustments: Increase cabs during January - May when demand is higher.

3. Data-Driven Pricing Strategy

* Increase fares for long-distance trips that are currently underpriced.

* Introduce congestion-based pricing to account for waiting time in traffic.

* Offer mid-day discounts (8 AM - 3 PM) to boost demand and utilize taxis efficiently.

* Apply moderate surge pricing in nightlife areas to balance demand.

Conclusion

By leveraging data-driven insights, NYC taxi operations can be optimized for efficiency and profitability.

* Better fleet distribution, dynamic dispatching, and strategic pricing adjustments can significantly enhance revenue.

* Implementing these recommendations will ensure higher service availability, reduced operational inefficiencies, and improved customer experience.