Executive Summary – NYC Yellow Taxi EDA 2023

Project Overview

This project conducts a comprehensive Exploratory Data Analysis (EDA) of New York City Yellow Taxi trip records for the year 2023. The objective is to identify demand patterns, fare structures, passenger behavior, and operational inefficiencies, enabling data-driven recommendations for fleet optimization, pricing strategies, and service improvements. The dataset, provided by the NYC Taxi & Limousine Commission (TLC), contains detailed trip-level information including timestamps, locations, passenger counts, trip distances, fare components, and payment types.

Data Sources & Scope

- Dataset: NYC Yellow Taxi Trip Records (Jan 2023 – Dec 2023) - Format: Parquet files (one file per month) - Source: NYC TLC Trip Record Data - Key Attributes: Pickup/drop-off timestamps & locations, passenger count, trip distance, fare breakdown, surcharges, and payment type.

Methodology

1. Data Sampling: - Loaded monthly Parquet files and merged into a unified dataset. - Applied 5% hourly stratified sampling to maintain representativeness while reducing data size. 2. Data Cleaning: - Merged duplicate columns (airport_fee / Airport_fee). - Imputed missing values using median or probability distribution sampling (for passenger_count). - Removed unrealistic or erroneous records (negative values, >6 passengers, extreme fares/distances). 3. Exploratory Data Analysis: - Time-based demand trends (hourly, daily, monthly) - Geospatial analysis with taxi zone shapefile - Fare structure and tipping behavior analysis - Passenger trends and operational efficiency insights

Key Findings

- Peak Hours: 5 PM – 7 PM weekdays; steady leisure demand on weekends. - Busiest Zones: Midtown Manhattan, JFK Airport, LaGuardia Airport. - Seasonality: Q2 and Q4 revenue peaks align with tourism and holidays. - Fare Structure: Strong correlation between distance and fare; short trips have higher cost per mile. - Vendor Pricing: Vendor 2 charges higher fares per mile than Vendor 1. - Tipping: Evening trips receive higher tips. - Passenger Trends: 1–2 passengers most common; weekend trips more likely to have groups.

Recommendations

1. Fleet Optimization: - Deploy more taxis in Midtown, airports, and high-demand zones during peak hours. - Adjust fleet positioning for weekday vs. weekend patterns. 2. Dynamic Pricing: - Implement time-based and location-based surge pricing. - Offer off-peak discounts to boost demand. 3. Route Efficiency: - Use real-time traffic data to avoid slow routes. - Provide drivers with alternative routing options. 4. Passenger Experience: - Promote ride-sharing in high group-travel zones. - Target high-tipping times and locations. 5. Operational Balance: - Reduce empty trips from airports via targeted dispatching. - Use geo-fencing to balance fleet distribution.

Impact Potential

If implemented, these data-driven strategies can: - Reduce idle time and empty trips - Increase driver earnings - Improve passenger satisfaction - Optimize fleet efficiency - Strengthen competitive

position against ride-hailing services