Report: Optimising NYC Taxi Operations

Include your visualisations, analysis, results, insights, and outcomes. Explain your methodology and approach to the tasks. Add your conclusions to the sections.

## Data Preparation

* 1. Loading the dataset
     1. **Sample the data and combine the files --done**

## Data Cleaning

### Fixing Columns

* + 1. **Fix the index -- done**
    2. **Combine the two airport\_fee columns -- done**

### Handling Missing Values

* + 1. **Find the proportion of missing values in each column**

* + 1. **Handling missing values in passenger\_count**

* + 1. **Handle missing values in RatecodeID**
    2. **Impute NaN in congestion\_surcharge**

### Handling Outliers and Standardising Values

* + 1. **Check outliers in payment type, trip distance and tip amount columns**

## Exploratory Data Analysis

### General EDA: Finding Patterns and Trends

* + 1. **Classify variables into categorical and numerical**
    2. **Analyse the distribution of taxi pickups by hours, days of the week, and months**
    3. **Filter out the zero/negative values in fares, distance and tips**
    4. **Analyse the monthly revenue trends**
    5. **Find the proportion of each quarter’s revenue in the yearly revenue**
    6. **Analyse and visualise the relationship between distance and fare amount**
    7. **Analyse the relationship between fare/tips and trips/passengers**
    8. **Analyse the distribution of different payment types**
    9. **Load the taxi zones shapefile and display it**
    10. **Merge the zone data with trips data**
    11. **Find the number of trips for each zone/location ID**
    12. **Add the number of trips for each zone to the zones dataframe**
    13. **Plot a map of the zones showing number of trips**
    14. **Conclude with results**

### Detailed EDA: Insights and Strategies

* + 1. **Identify slow routes by comparing average speeds on different routes**
    2. **Calculate the hourly number of trips and identify the busy hours**
    3. **Scale up the number of trips from above to find the actual number of trips**
    4. **Compare hourly traffic on weekdays and weekends**
    5. **Identify the top 10 zones with high hourly pickups and drops**
    6. **Find the ratio of pickups and dropoffs in each zone**
    7. **Identify the top zones with high traffic during night hours**
    8. **Find the revenue share for nighttime and daytime hours**
    9. **For the different passenger counts, find the average fare per mile per passenger**
    10. **Find the average fare per mile by hours of the day and by days of the week**
    11. **Analyse the average fare per mile for the different vendors**
    12. **Compare the fare rates of different vendors in a distance-tiered fashion**
    13. **Analyse the tip percentages**
    14. **Analyse the trends in passenger count**
    15. **Analyse the variation of passenger counts across zones**
    16. **Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.**

## Conclusions

### Final Insights and Recommendations

* + 1. **Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.**
    2. **Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.**
    3. **Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.**