Exploratory Data Analysis of New York City TLC Dataset

Executive summary Report By Yahya Bhara

Basic Info / Summary of findings

1. There are 22699 entries and 18 columns in total. 8 columns have dtypes: 8 float, 7 have int and 3 have object. There is no missing value.
2. Key metrics used: trip_distance and total_amount
3. Boxplot, histogram and Bar plot used for visualizations.
4. Used matplotlib and seaborn to make visualizations

Key Insights

1. Monthly revenue and daily revenue generally follow same trend as monthly rides and daily rides.
2. Nearly 2/3 of total rides were single occupancy.
3. Out of a total of 16117 total rides ,33 rides have zero occupancy
4. Highest distribution of trips are below 5 miles.
5. The dataset includes trip start and end locations. That is how we are able to calculate the total distance traveled for each trip

Outliers:
There are some outliers up to 35 miles.
There are several trips with trip distance of 0.

Maximizing Value from Data: Actionable Insights

1. Since highest distribution of trips are below 5 miles, a deeper analysis of 0-5 miles trips could be conducted and compared with other segments: 10-15 mile trips and other similar segments.

2. So far we have broken trends into monthly and daily revenues and rides. This data can be further broken down into hourly basis which will help us pinpoint Peak times according to revneue and frequency

3. Pick off and drop off locations can be used to create a heatmap to figure out High Demand areas.

4. Analyze most frequently used routes to figure out which ones are most profitable.

Solution to tackle Outliers and next steps

Outliers:

1. Trips with distances close to 35 miles and trips with 0 miles need to be examined further.
2. Verify whether the outliers are valid. Sometimes these could be data entry errors.
3. Eliminate extreme outliers if they skew the analysis, or mark them for further review.
4. Investigate if 0-mile trips are due to incorrect data or if they reflect specific taxi use cases such as cancelled trips or free rides.

Contact New York City TLC:
Make sure that the sample they provided is an accurate representaion of data as a whole.