upGrad

International Institute of Information Technology Bangalore

LIVERPOOL JOHN MOORES UNIVERSITY

Master of Science in Machine Learning & AI

Report: Optimising NYC Taxi Operations

D. Jerard Ashwin

1. Data Preparation

1.1. Loading the dataset

1.1.1. Sample the data and combine the files

Imported libraries and suppressed warnings.

Checked library versions for compatibility.

Loaded and verified a sample Parquet file.

Sampled 0.7% data from each Parquet file and combined them.

Handled file errors using a try-except block.

Saved the final dataset as data_New.csv.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

Loaded the cleaned dataset from data_New.csv.

Displayed the first five rows using head() for a quick check.

Verified data types and missing values using info().

Identified and removed unnecessary columns containing "index" or "unnamed".

Reset the index to maintain a clean, continuous sequence.

2.1.2. Combine the two airport_fee columns

Combined the duplicate airport_fee columns by summing them and handling missing values with .fillna(0).

Created a new column Combined_Airport_Fee to store the combined result.

Removed the original airport_fee and Airport_fee columns using df.drop().

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

Identified missing values in the dataset using df.isnull().sum().

Columns passenger_count, RatecodeID, store_and_fwd_flag, and congestion_surcharge each have **9038 missing values**.

No missing values in other columns, including the combined Combined_Airport_Fee

2.2.2. Handling missing values in passenger_count

Displayed rows where **passenger_count** is missing.

Imputed missing values in **passenger_count** with its **median** value.

Identified and **removed rows** where **passenger_count** was **zero**.

Final distribution of **passenger_count** updated, with zero values eliminated.

2.2.3 Handle missing values in RatecodeID

Display rows where RatecodeID is missing.

Impute missing values in RatecodeID with its median value.

Identify and remove rows where RatecodeID was zero (if applicable).

Final distribution of RatecodeID updated, with missing values handled.

2.2.4 Impute NaN in congestion_surcharge

Fill the missing values in congestion_surcharge with its median value.

After imputation, check the final distribution of congestion_surcharge to ensure there are no missing values.

2.3. Handling Outliers and Standardising Values

2.3.1. Check outliers in payment type, trip distance and tip amount columns

Payment Type: Identify any invalid or unexpected values in the payment_type column. Check for rows where payment_type is missing or has values not defined in the data dictionary.

Trip Distance: Use a box plot to check for any outliers in the trip_distance column. Remove any entries with trip distances greater than 250 miles, as they may be erroneous.

Tip Amount: Check for outliers in the tip_amount column by visualizing it with a box plot. This will help identify any unusually high or low tip amounts that are outliers.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

Categorical Variables:

VendorID

RatecodeID

PULocationID

DOLocationID

payment_type

Numerical Variables:

passenger_count

trip_distance

pickup_hour

trip_duration

fare_amount

extra

mta_tax

tip_amount

tolls_amount

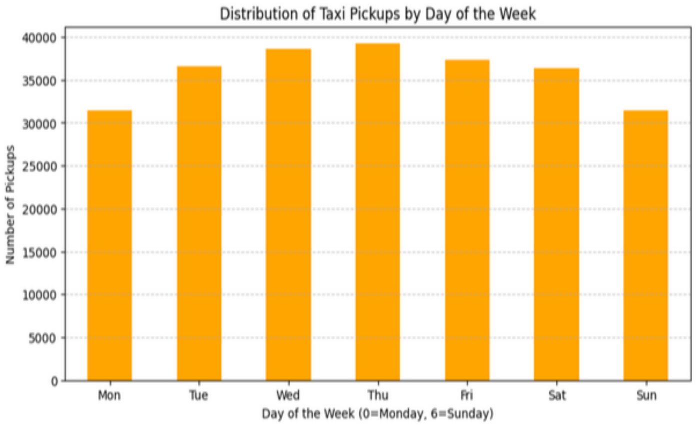
improvement_surcharge

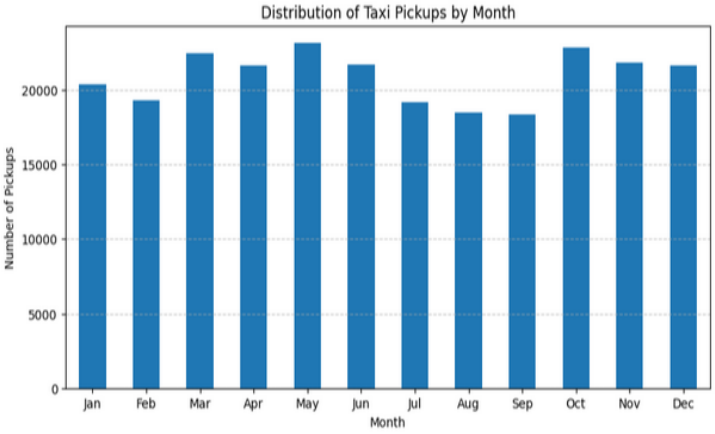
total_amount

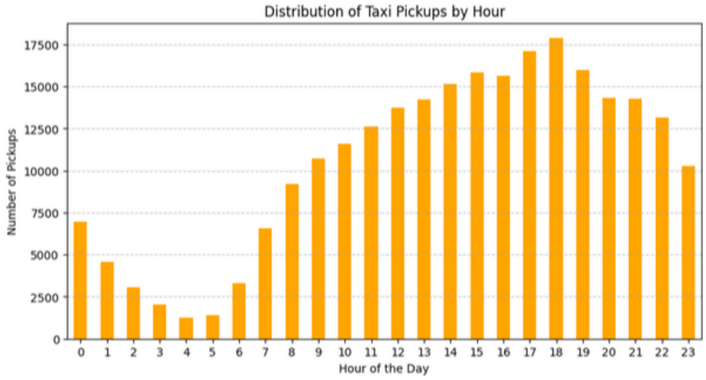
congestion_surcharge

airport_fee

3.1.2. Analyse the distribution of taxi pickups by hours, days of the week, and months







3.1.3. Filter out the zero/negative values in fares, distance and tips

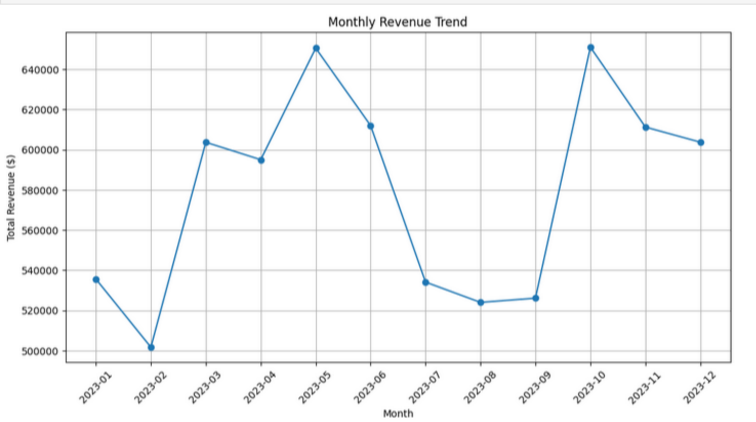
Fare Amount: You checked for rows where the fare_amount was zero or negative and removed those rows from the data.

Total Amount: You did the same for the total_amount, removing rows where the value was zero or negative.

Trip Distance: You removed rows where the trip_distance was zero or negative.

3.1.4. Analyse the monthly revenue trends

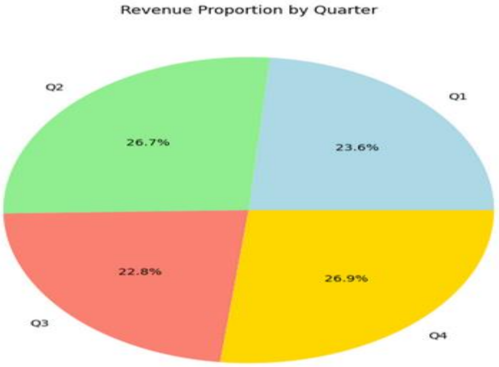
Highest Revenue: The revenue peaked in **April (2023-04)** and **November (2023-11)**, with both months showing the highest total revenue.
Lowest Revenue: The revenue was lowest in **January (2023-01)** and **July (2023-07)**, indicating lower earnings during these months.
There's a **significant rise in April** and another in **November**, with steady periods in between. The trend suggests potential seasonal influences or other factors that cause revenue to fluctuate throughout the year.



3.1.5. Find the proportion of each quarter's revenue in the yearly revenue

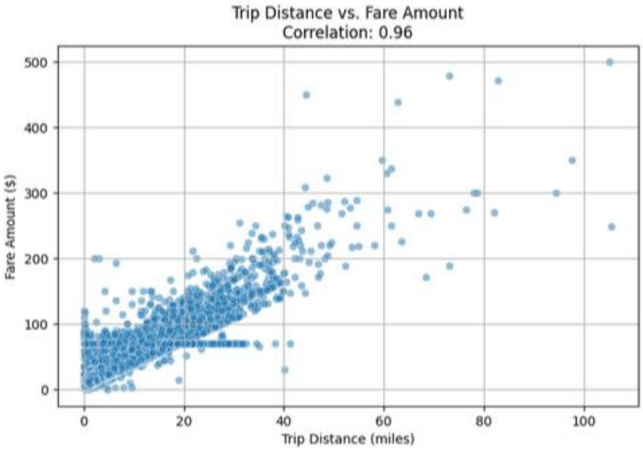
Proportion of Revenue by Quarter:

- **Q1 (January - March)**: 23.6%
- **Q2 (April - June)**: 26.7%
- **Q3 (July - September)**: 22.8%
- **Q4 (October - December)**: 26.9%



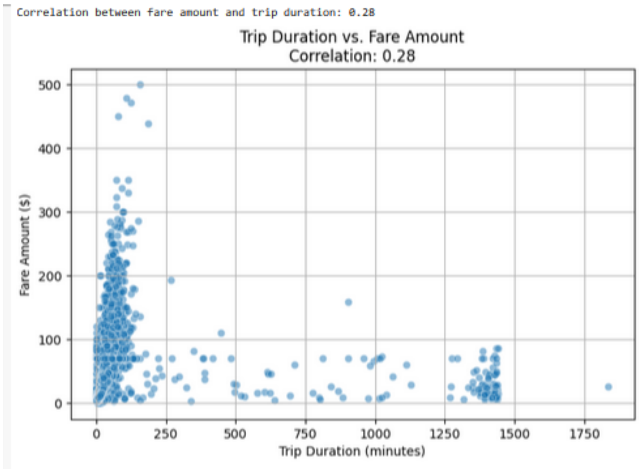
3.1.6. Analyse and visualise the relationship between distance and fare amount

The chart shows the relationship where trip distances from 0 to 100 miles correspond to fare amounts ranging from $0 to $500. As the distance increases, the fare increases significantly, confirming the direct relationship between the two.



3.1.7. Analyse the relationship between fare/tips and trips/passengers

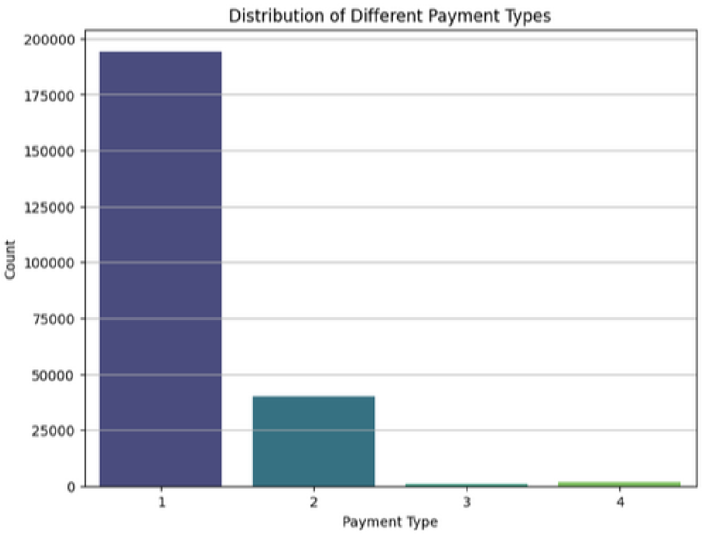
The scatter plot shows a weak correlation (0.28) between Trip Duration and Fare Amount. This suggests that trip duration does not strongly affect the fare amount. The data is spread out, and there are some outliers with very high durations but not necessarily high fares.



```
Correlation between fare amount and trip duration: 0.28
```

3.1.8. Analyse the distribution of different payment types

Payment Type 1 is the most widely used, with a significantly higher count than other payment types. Payment Types 2, 3, and 4 are rarely used.



3.1.9. Load the taxi zones shapefile and display it

loaded the taxi zones shapefile using geopandas, which contains information about each zone like its ID, name, area, location, borough, and its geographical shape. This data is now ready for analysis or visualization on a map.

3.1.10. Merge the zone data with trips data

I have merged the Zone data and trips data

3.1.11. Find the number of trips for each zone/location ID

Grouped the data by **zone** and counted how many trips occurred in each zone. This gives the number of trips for each location, such as 237 trips for "Alphabet City" and 80 trips for "Astoria."

3.1.12. Add the number of trips for each zone to the zones dataframe

Aggregate the trip counts from the taxi trip data by pickup_zone or dropoff_zone.

Merge the aggregated data with the zones dataframe.

3.1.13. Plot a map of the zones showing number of trips

Plot a map showing the number of trips in each zone, you merged the trip counts with the zone geometries and then used the .plot() function to create a map. The map shades each zone based on the number of trips, with the number of trips displayed in a color legend. This visualizes the distribution of trips across the different zones.

3.1.14. Conclude with results

Trip Trends: Most trips happen in the late afternoon and weekdays, with fewer trips at night and on weekends.

Revenue: Revenue is highest in April and November, with Q2 and Q4 contributing the most to yearly revenue.

Correlations: Trip distance has a strong impact on fare, but trip duration doesn’t affect fare much.

Payment Methods: Most people pay using **Payment Type 1**, with fewer using other methods.

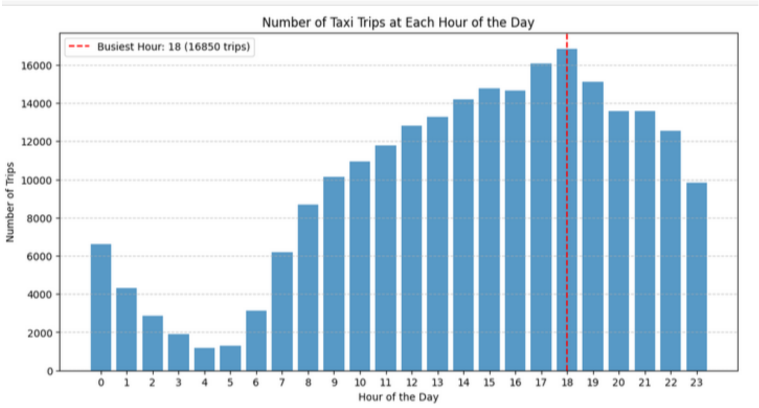
Zone Activity: Some zones, like **Yorkville West**, have significantly more trips, showing higher demand in those areas.

3.2. Detailed EDA: Insights and Strategies

3.2.1. Identify slow routes by comparing average speeds on different routes

3.2.2. Calculate the hourly number of trips and identify the busy hours

calculated the number of trips for each hour by extracting the hour from the pickup time. Then, you counted how many trips happened each hour and found that **18:00** (6 PM) was the busiest hour with the most trips. The graph shows this trend with a red line marking the busiest hour.



3.2.3. Scale up the number of trips from above to find the actual number of trips

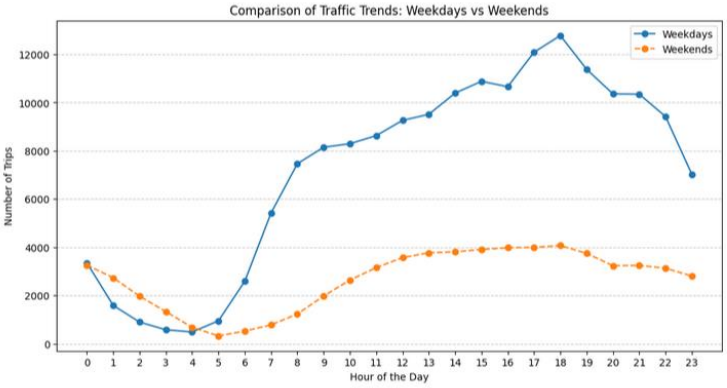
3.2.4. Compare hourly traffic on weekdays and weekends

Compared **weekdays** and **weekends** by counting trips for each hour of the day.

Weekdays: Trips are higher in the late afternoon (around 5-7 PM).

Weekends: Trips are more consistent throughout the day, with no clear peak.

The analysis shows weekday traffic has a clear evening peak, while weekend traffic is more even.



3.2.5. Identify the top 10 zones with high hourly pickups and drops

Pickup-to-dropoff ratio, the **top 10 zones** with the highest pickup-to-dropoff ratios are:

LocationID 59 - Highest ratio with a very large difference between pickups and dropoffs.

LocationID 67 - Significant ratio of pickups compared to dropoffs.

LocationID 124 - A high pickup-to-dropoff ratio indicating high demand.

LocationID 138 - Strong pickup demand with fewer dropoffs.

LocationID 103 - Another area with high pickups relative to dropoffs.

LocationID 86 - More pickups compared to dropoffs.

LocationID 101 - High demand for pickups in this area.

LocationID 106 - Large ratio, indicating a high number of pickups.

LocationID 134 - Similar high pickup ratio.

LocationID 154 - Good ratio indicating higher pickups.

Bottom 10 Zones (Lowest Pickup-to-Dropoff Ratio):

These zones have a very low or almost zero pickup-to-dropoff ratio, meaning taxis are more often used for **dropoffs** than pickups.

LocationID 183 - Very low ratio indicating mostly dropoffs.

LocationID 184 - Similar low ratio.

LocationID 185 - Fewer pickups, more dropoffs.

LocationID 186 - High dropoff count compared to pickups.

LocationID 187 - Most trips are dropoffs in this zone.

3.2.6. Find the ratio of pickups and dropoffs in each zone

Location 2977 – Highest ratio of pickups to dropoffs (13,839 pickups and 1 dropoff).

Location 3184 – Another zone with a high pickup-to-dropoff ratio.

Location 3120 – High pickups compared to dropoffs.

Location 3093 – High pickup ratio.

Location 3101 – Another zone with a significant pickup-to-dropoff ratio.

(These zones show high demand for pickups compared to dropoffs.)

Bottom 10 Zones (Lowest Pickup-to-Dropoff Ratio):

Location 480 – Almost equal pickups and dropoffs.

Location 1859 – Similar low ratio, more balanced pickups and dropoffs.

Location 5642 – More dropoffs than pickups.

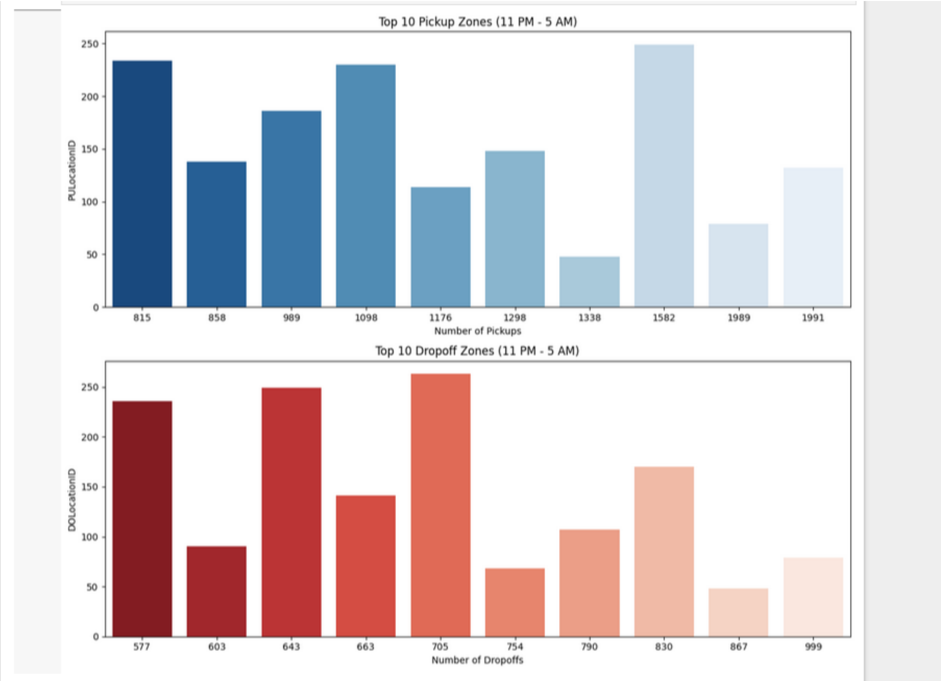
Location 5643 – Dropoffs higher than pickups.

Location 5461 – Few pickups compared to dropoffs.

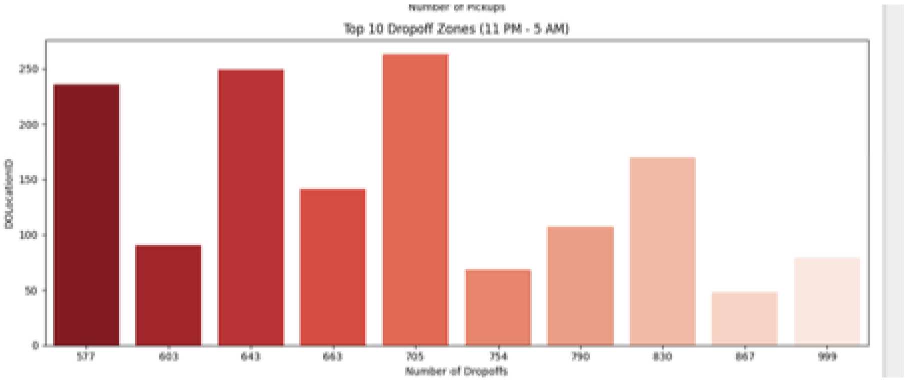
3.2.7 Identify the top zones with high traffic during night hours

Pickup zones like **815**, **858**, and **1582** are busy with many people starting their trips during the night.

Dropoff zones like **577**, **705**, and **643** are areas where many passengers are being dropped off late at night.



3.2.8 Find the revenue share for nighttime and daytime hours



3.2.9 For the different passenger counts, find the average fare per mile per passenger

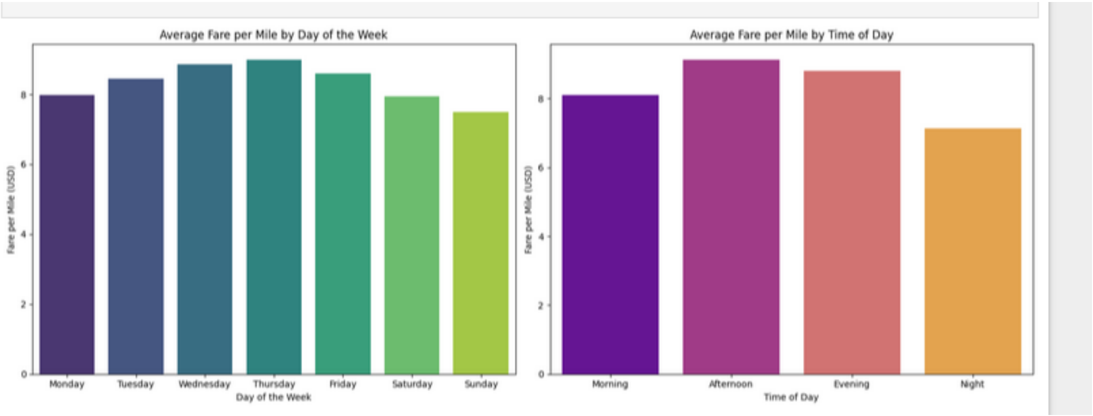
Passenger Count 2 has the highest average fare per mile per passenger, indicating that trips with 2 passengers tend to have the highest fare per mile.

Passenger Counts 3 to 6 have lower average fares, with the fare decreasing as the passenger count increases.

3.2.10 Find the average fare per mile by hours of the day and by days of the week

Days **of the Week**: **Monday** and **Tuesday** tend to have higher average fares, possibly due to increased demand at the beginning of the week.

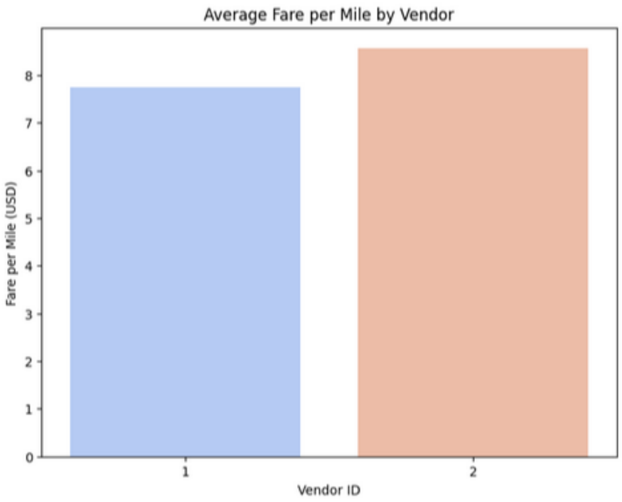
Time of Day: **Morning** fares are highest, possibly due to rush hour or higher demand for early trips, while **night** fares are the lowest, indicating less demand at night.



3.2.11 Analyse the average fare per mile for the different vendors

Vendor 1 has a lower average fare per mile, around **$6**.

Vendor 2 has a higher average fare per mile, around **$8**.



3.2.12 Compare the fare rates of different vendors in a distance-tiered fashion

Short Distance (0-2 miles):

Vendor 2 has a higher **average fare per mile** compared to **Vendor 1** for short trips. Vendor 2’s fares are significantly more expensive for these short distances.

Medium Distance (2-5 miles):

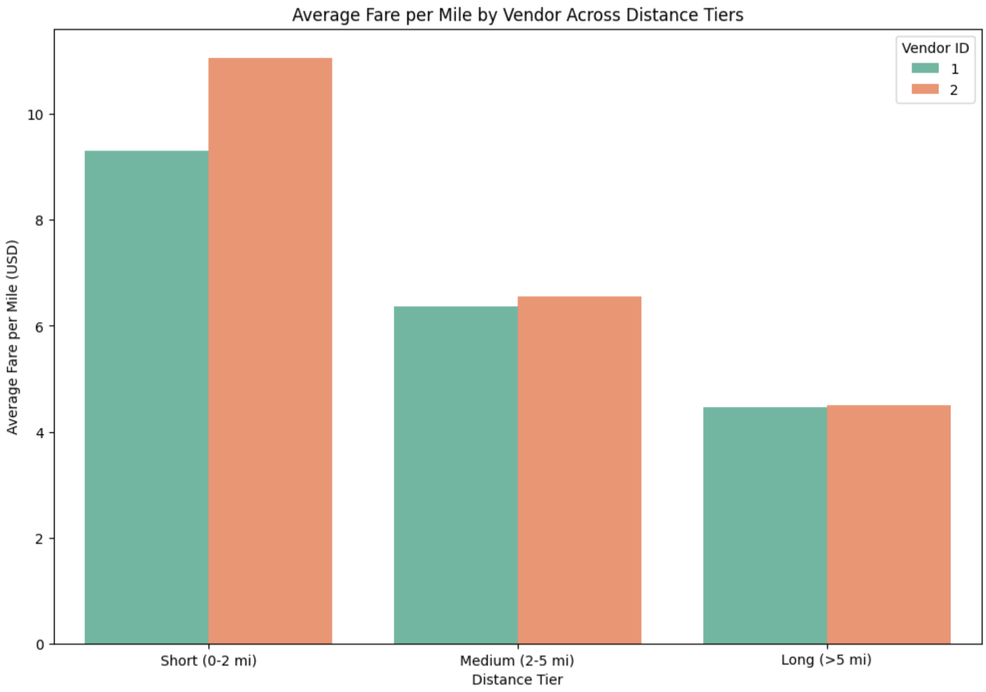
The fare difference between **Vendor 1** and **Vendor 2** is smaller for **medium distances**, with both vendors charging similar rates.

Long Distance (>5 miles):

Vendor 2 again has a slightly **higher fare per mile** than **Vendor 1** for longer trips, but the difference is less pronounced compared to short distances.

Key Insight:

Vendor 2 tends to have higher fares across **short** and **long-distance trips**, while **Vendor 1** has more consistent rates across all distance tiers, charging slightly less for both short and long distances.



3.2.13 Analyse the tip percentages

Tip Percentage by Distance Tier:

Short Trips (0-2 miles) have the highest **average tip percentage** at around **20%**.

Medium Trips (2-5 miles) have a slightly lower tip percentage, and **Long Trips (>5 miles)** have the lowest tip percentage.

Key Insight: Passengers tend to tip more on **short trips** compared to **longer trips**, possibly due to shorter travel times or higher perceived service quality.

Tip Percentage by Passenger Group:

The **average tip percentage** is quite **consistent** across all passenger groups, with only **slightly higher tips** for **1 passenger** and **2 passengers**.

The **tip percentage** does not vary much when there are more passengers in the group (3-4 or 5+ passengers).

Key Insight: The **number of passengers** does not seem to significantly impact the **tip percentage**, with relatively similar averages across the groups.

Tip Percentage by Pickup Hour:

Tip percentages are highest in the **late evening (around 8-9 PM)**, reaching over **21%**.

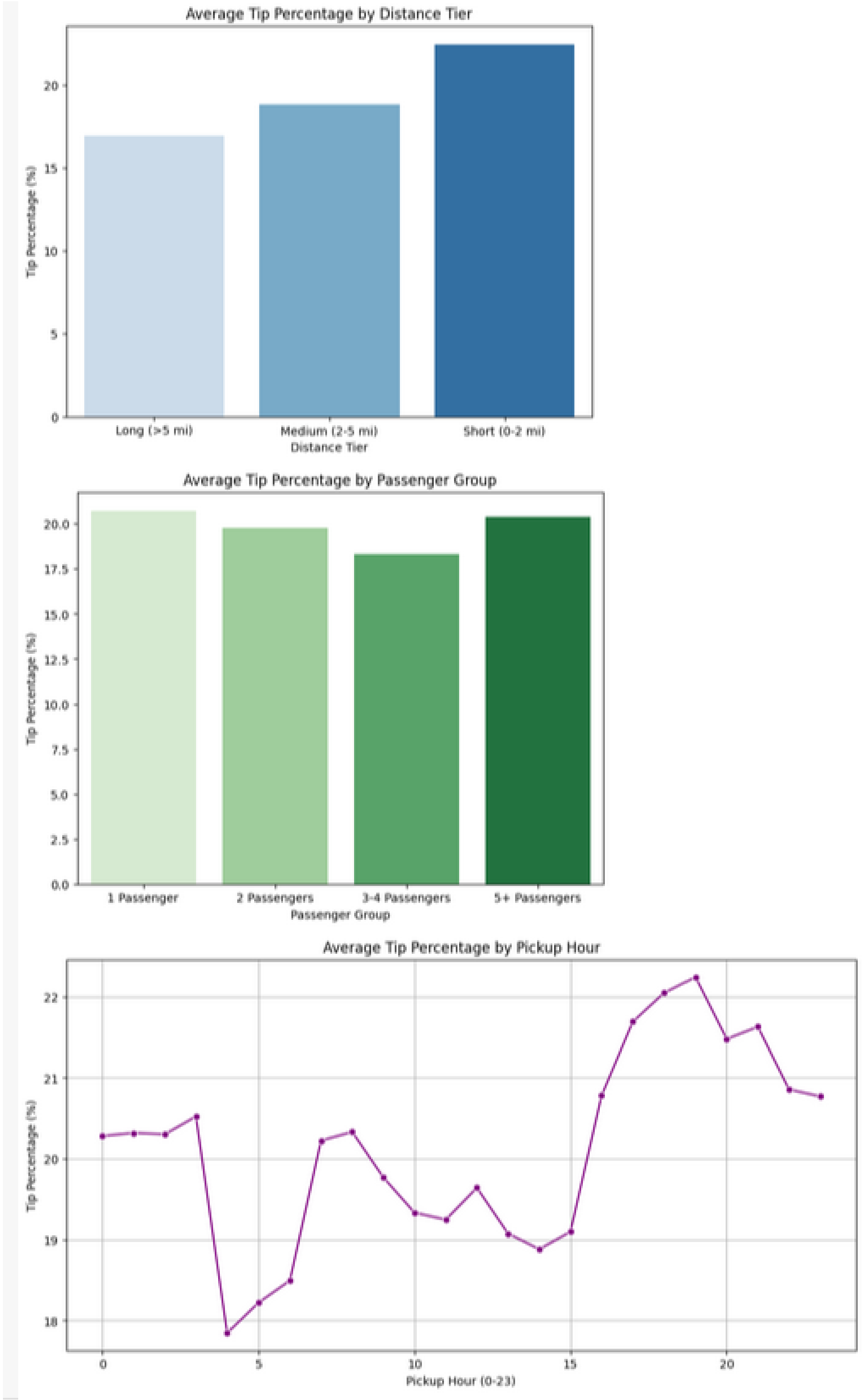
The **lowest tip percentages** are observed around **early morning hours** (5 AM) and some **afternoon hours**.

Key Insight: **Higher tips** are given during **peak evening hours**, possibly due to factors like busier periods or better service expectations, while **morning** and **afternoon** trips have lower tip percentages.

Conclusion:

Short trips tend to get the highest tips, with **passenger count** not significantly affecting the tip percentage.

Evening hours, especially around **8-9 PM**, see the highest tip percentages



3.2.14. Analyse the trends in passenger count

Weekdays (Monday to Friday):

Passenger counts are generally higher in the **late afternoon and early evening** (around **5 PM to 7 PM**).

The **highest passenger counts** are seen on **Friday and Saturday evenings**.

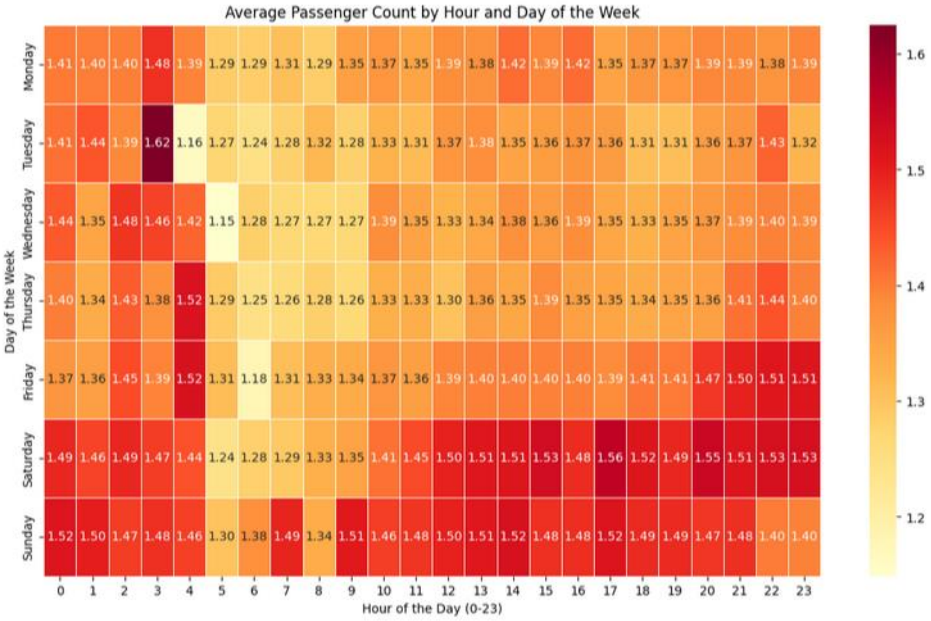
There are **lower counts** during early morning hours (**12 AM to 6 AM**), especially on weekdays.

Weekend (Saturday and Sunday):

Saturday shows higher passenger counts throughout the day, particularly in the **afternoon** and **evening**.

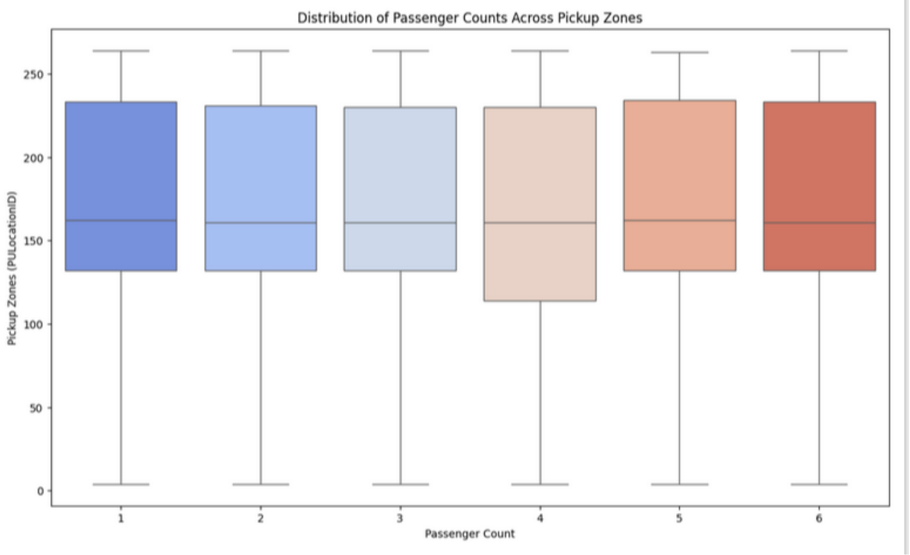
Sunday has slightly higher counts compared to weekdays, with a noticeable peak in the **evening (around 7 PM)**.

The overall trend on **Sunday** shows a **higher passenger count** than most weekdays, particularly during evening hours.



3.2.15. Analyse the variation of passenger counts across zones

The boxplot shows that the passenger count is fairly similar across all zones, whether there are 1 to 6 passengers. Zones with more passengers (3-6) tend to have slightly higher average counts, but the variation is not very large. There are a few outliers, but they are not significant. In simple terms, the number of passengers is pretty consistent across different zones.



4 Conclusions

4.1Final Insights and Recommendations

4.1.1. Recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies.

Dynamic Dispatching:

Prioritize high-demand zones during peak hours like **morning rush**, **evening**, and **night surge**.
Use **real-time monitoring** to adjust vehicle allocation based on demand.

Geofencing & Supply Management:

Set up **geofences** around busy areas such as **airports** and **business hubs**.
Move **idle vehicles** to **under-served regions** to maintain coverage.

Route Optimization:

Use **algorithms** to find the **shortest and most efficient routes**.
Minimize **empty return trips** by **smart fleet positioning**.

Driver Incentives:

Offer **bonuses** for working in **low-demand areas** or during **off-peak hours**.
Reward **long-distance trips** to ensure broader coverage across regions.

4.1.2 Suggestions on strategically positioning cabs across different zones to make best use of insights uncovered by analysing trip trends across time, days and months.

Peak Hour Hotspots:

Morning (7 AM - 10 AM): Position cabs in **residential zones** to serve people commuting to work.

Evening (5 PM - 8 PM): Focus on **business districts** for **return trips** as people head home after work.

Night Demand Zones:

11 PM - 3 AM: Place cabs near **entertainment hubs**, **airports**, and **transport stations** to capture **late-night travellers**.

Weekend Strategy:

Increase **cab availability** in **malls**, **tourist spots**, and **event venues** during weekends, especially in the **afternoon and evening**.

Low-Demand Redistribution:

Move **idle cabs** from **low-traffic areas** to zones with **emerging trends**, such as newly developed areas or places hosting seasonal events.

Seasonal Adjustment:

Adjust **cab supply** based on monthly trends. Increase **availability near holiday destinations** during vacation seasons and around **business hubs** during work periods.

Zone-Specific Supply Balancing:

Ensure a **balanced distribution** of cabs by analysing **pickup/drop-off imbalances**. Position more cabs in areas with **high outbound demand** to meet customer needs.

These strategies help optimize cab availability and efficiently distribute resources to meet demand at different times, places, and seasons.

4.1.2. Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates with other vendors.

Dynamic Surge Pricing:

Increase fares during peak hours and in **high-demand zones** to manage demand.

Distance-Based Tiers:

Lower fares for **short trips**, maintain **standard rates** for **medium trips**, and offer **discounts** for **long trips** to incentivize longer journeys.

Passenger-Based Rates:

Apply **higher rates** for **1-2 passengers** and offer **discounts** for **3+ passengers** to encourage shared rides.

Night & Off-Peak Discounts:
Offer **lower fares** during **off-peak hours (10 PM - 6 AM)** to boost demand when there are fewer riders.

Loyalty Programs:

Provide **discounts** for **frequent riders** and introduce **subscription plans** for regular commuters to retain customers.

Competitor Matching:

Adjust your fares based on **competitor pricing** to ensure you stay competitive in the market. These strategies help maximize both **profitability** and **customer satisfaction** by offering flexible pricing based on time, distance, demand, and competition.