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.

1. Data Preparation

1.1. Loading the dataset

- First I have imported all the essential libraries that are used for analysis
- I have read one file to see the data structure for the columns

1.1.1. Sample the data and combine the files

- Loaded and combined 12 separate `.parquet` files containing monthly NYC taxi trip records.
- I sampled the data using a 0.5% sampling fraction (n = 42) from a total of around 18 million records. This gave us a smaller dataset of about 300,000 records. I saved this sampled data into a CSV file and worked with it as `df_sampled` throughout the analysis.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

- After sampling and saving the data, I reloaded the CSV and noticed that the index column from the original DataFrame got saved too. So, we reset the index and dropped the old one to keep things clean.

- Then I dropped few columns that are not needed for analysis

2.1.2. Combine the two airport_fee columns

- I have noticed that there are two airport columns with Airport_fee and ariport_fee in the dataset, maybe because of how the data was merged. Combined them into one by checking if either had a value, and then dropped the extra column.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

Checked how many missing values are there in each column by calculating the proportion of NaNs.

2.2.2. Handling missing values in passenger_count

Rows with missing or zero `passenger_count` were dropped to ensure accuracy in analysis.

2.2.3. Handle missing values in RatecodeID

Missing `RatecodeID` values were filled with the most frequent rate code in the dataset.

2.2.4. Impute NaN in congestion_surcharge

congestion_surcharge was dropped in the data cleaning process assuming it is not need for the analysis

2.3. Handling Outliers and Standardising Values

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

- Verified if there are any unexpected or invalid values in the `payment_type` column that don't match known payment types.
- Identified extremely high or zero values in `trip_distance` that could be errors or unusual cases, and removed them for better analysis.
- Spotted unusually high or negative values in `tip_amount`, and filtered them out to keep the analysis realistic.

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

3.1.1. Classify variables into categorical and numerical

Categorical Variables

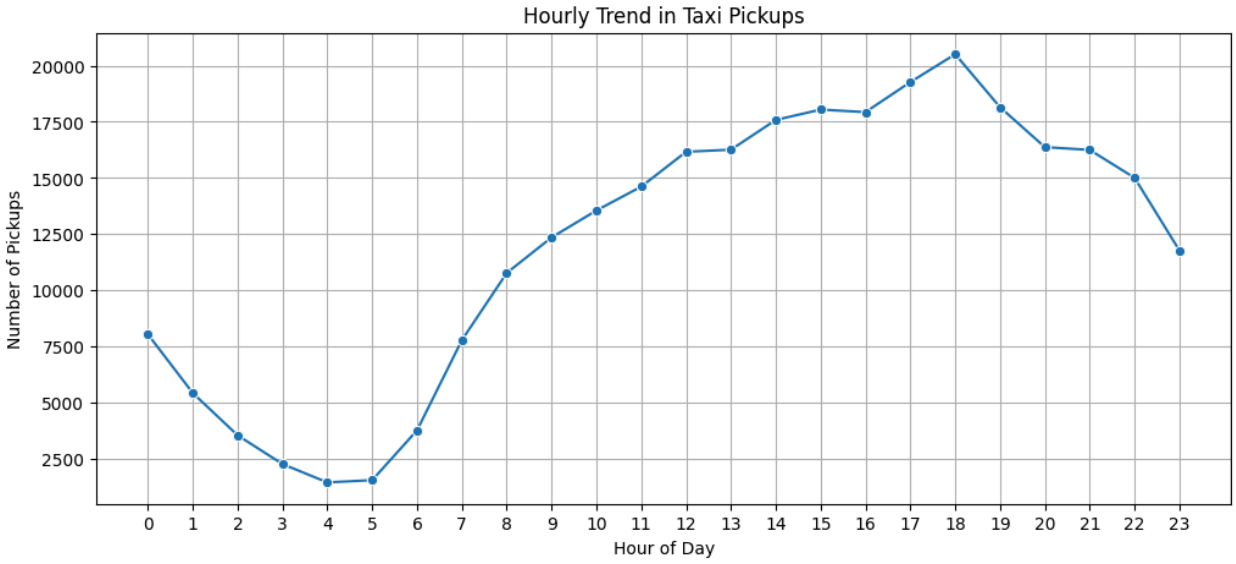
- `VendorID`
- `RatecodeID`
- `PULocationID`
- `DOLocationID`
- `payment_type`
- `Pickup_hour`

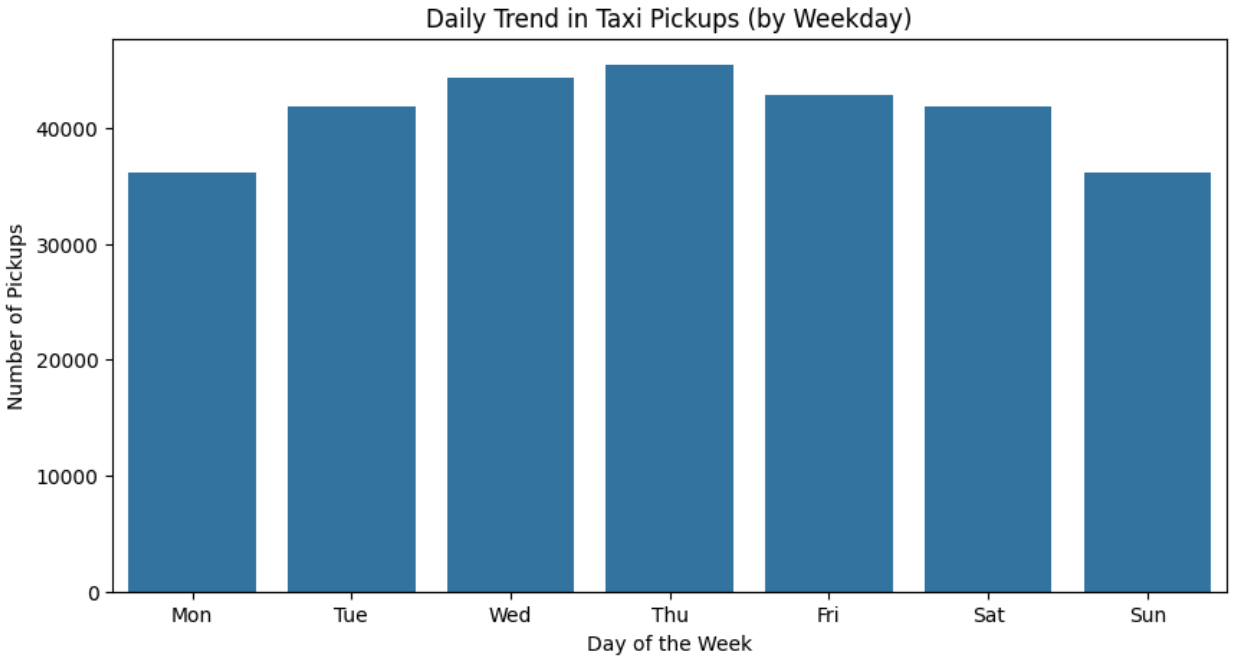
Numerical Variables

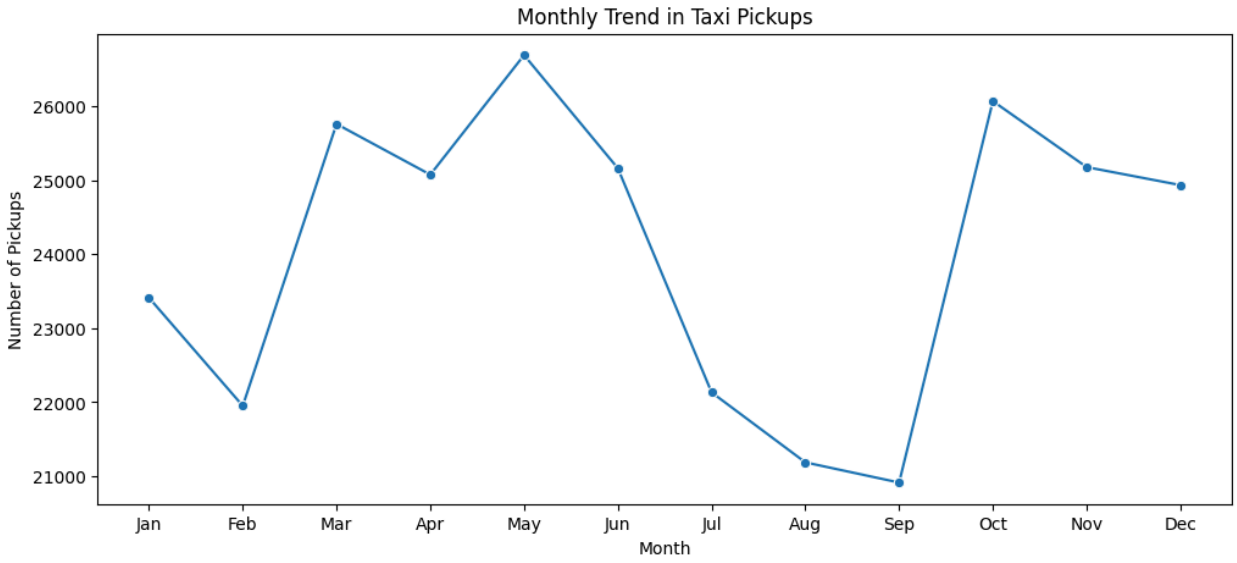
- `passenger_count`
- `trip_distance`
- `trip_duration`

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

- Most taxi pickups happen around 6 PM, and the lowest around 4 AM. Evening hours need more cabs than early mornings.
- Thursdays see the highest number of trips. Cab supply can be increased mid-week to match demand.
- May had the most trips, likely due to better weather or tourism. Operations can scale during peak months.







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

- Removed trips with zero or negative fare, distance, or tips to keep the data clean and reliable.

3.1.4. Analyse the monthly revenue trends

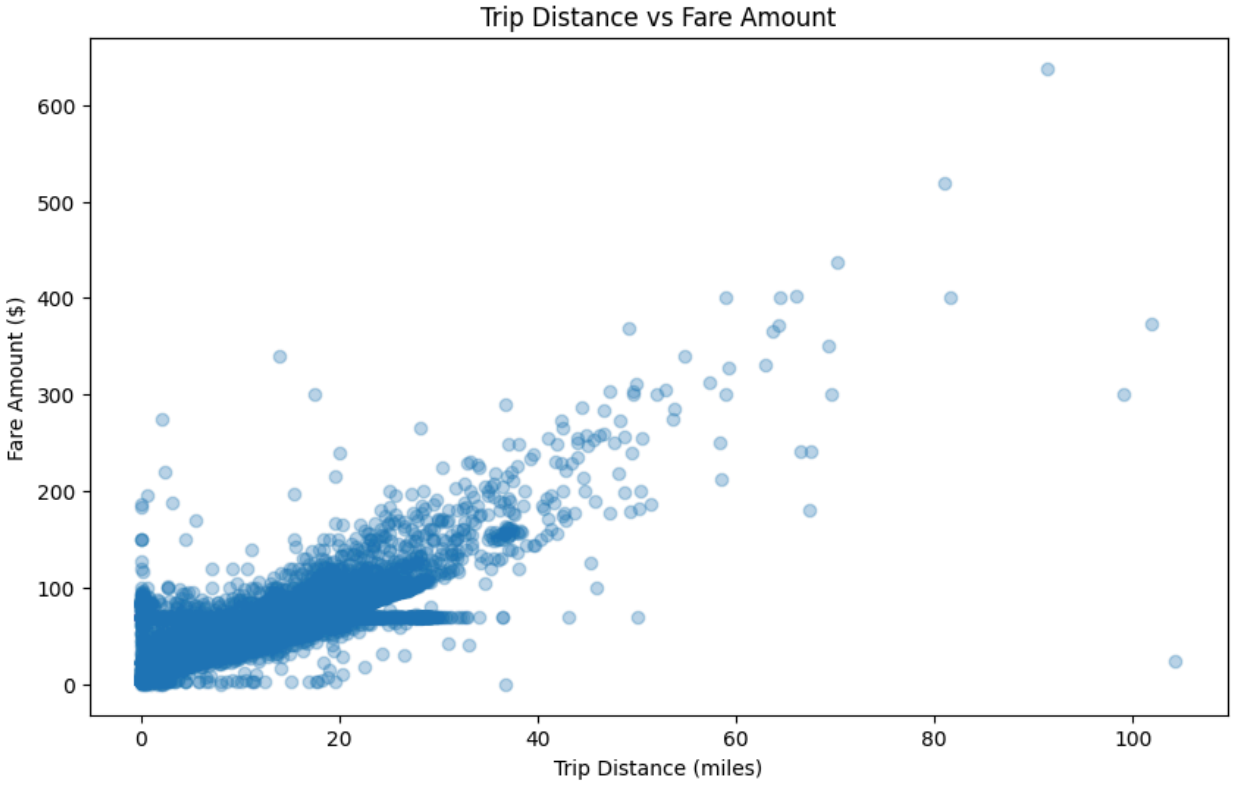
- Revenue was highest in May, showing strong demand that month. Plotted a line chart to visualize month-wise revenue.

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

- Calculated how much each quarter contributed to the yearly revenue. Q2 had the highest share.

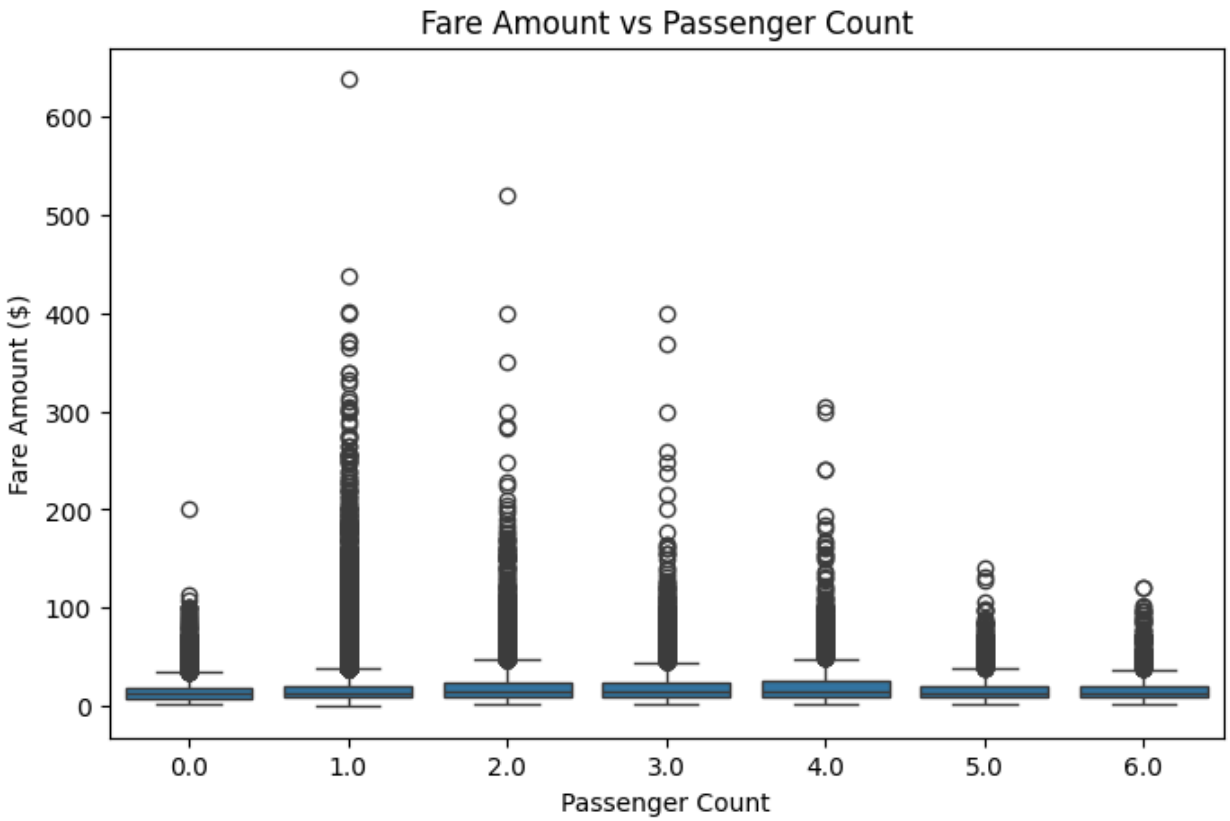
3.1.6. Analyse and visualise the relationship between distance and fare amount

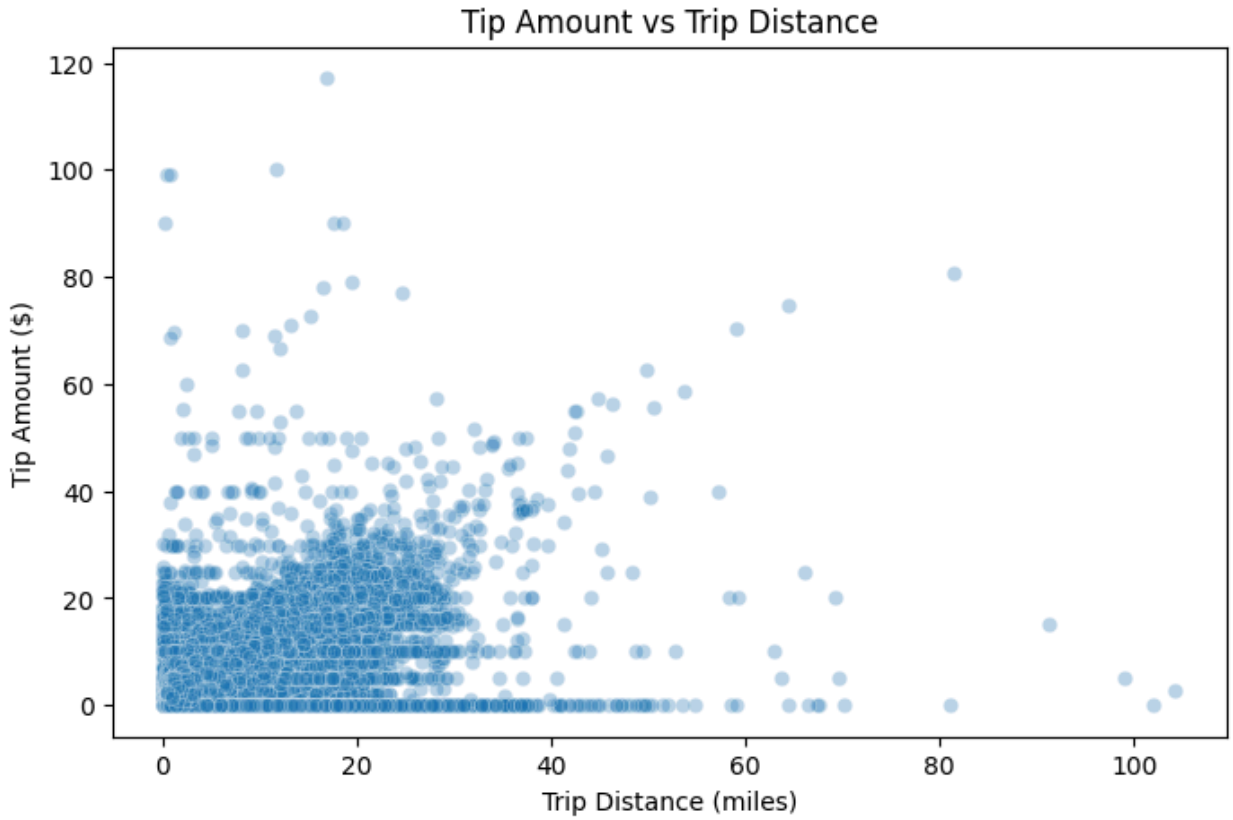
- Found a strong positive relation—more distance usually means higher fare. Used a scatter plot for visual clarity.
- Correlation between trip distance and fare amount: 0.945



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

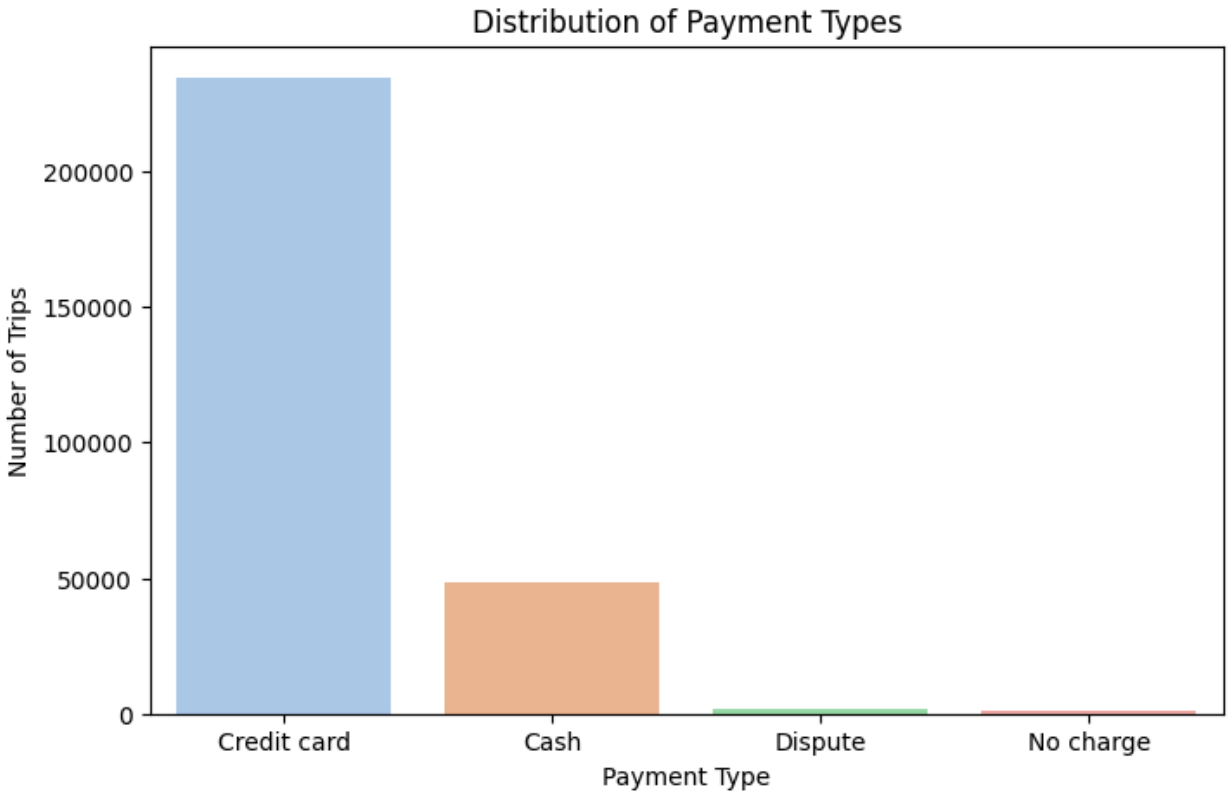
- Higher fare and tips were generally seen with longer trips and more passengers. Analyzed this using group means.
- Correlation between fare_amount and trip_duration: 0.275
- Correlation between fare_amount and passenger_count: 0.042


Fare Amount vs Passenger Count
Fare Amount ($)
0
100
200
300
400
500
600
0.0
1.0
2.0
3.0
4.0
5.0
6.0
Passenger Count


Tip Amount vs Trip Distance
Tip Amount ($)
0
20
40
60
80
100
120
0
20
40
60
80
100
Trip Distance (miles)

3.1.8. Analyse the distribution of different payment types

- Majority of trips were paid using credit cards. Created a pie chart to show the share of each payment method.



3.1.9. Load the taxi zones shapefile and display it

- Loaded the NYC zones shapefile using GeoPandas to map locations. Helped in location-based analysis.

3.1.10. Merge the zone data with trips data

- Merged trip data with zone names using location IDs. This allowed us to analyze demand by pickup and dropoff zones.

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

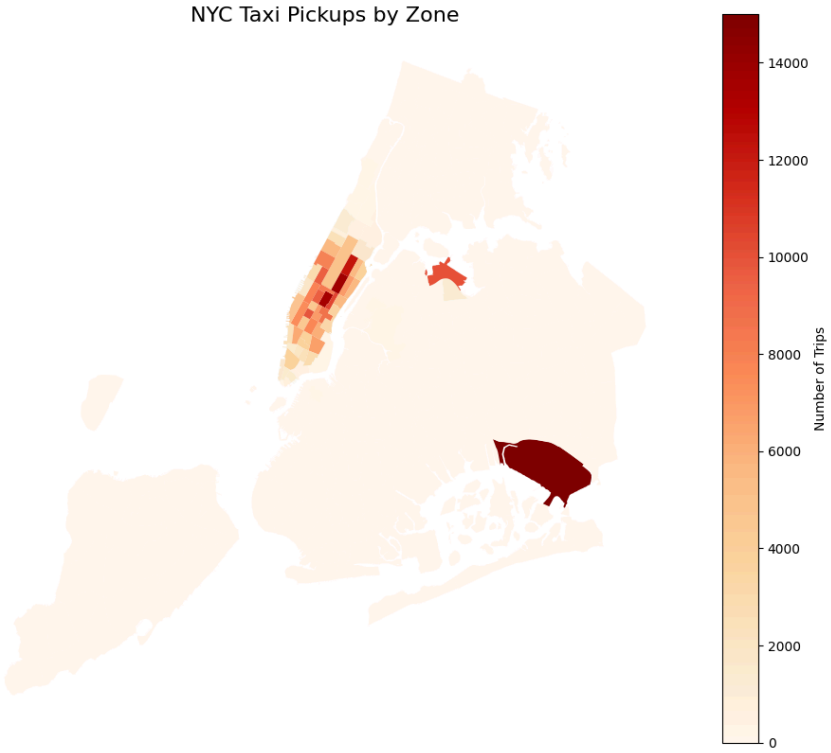
- Counted how many trips started from each pickup zone using groupby on `PULocationID`.

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

- Merged the trip counts into the taxi zones data, so we can see the demand per area.

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

- Plotted a choropleth map where darker zones had more trips. It clearly showed busy and low-demand areas.



- Plotted a choropleth map where darker zones had more trips. It clearly showed busy and low-demand areas.

3.1.14. Conclude with results

- The map helped identify which areas had the most taxi activity. These zones can be useful to focus on for better cab distribution.

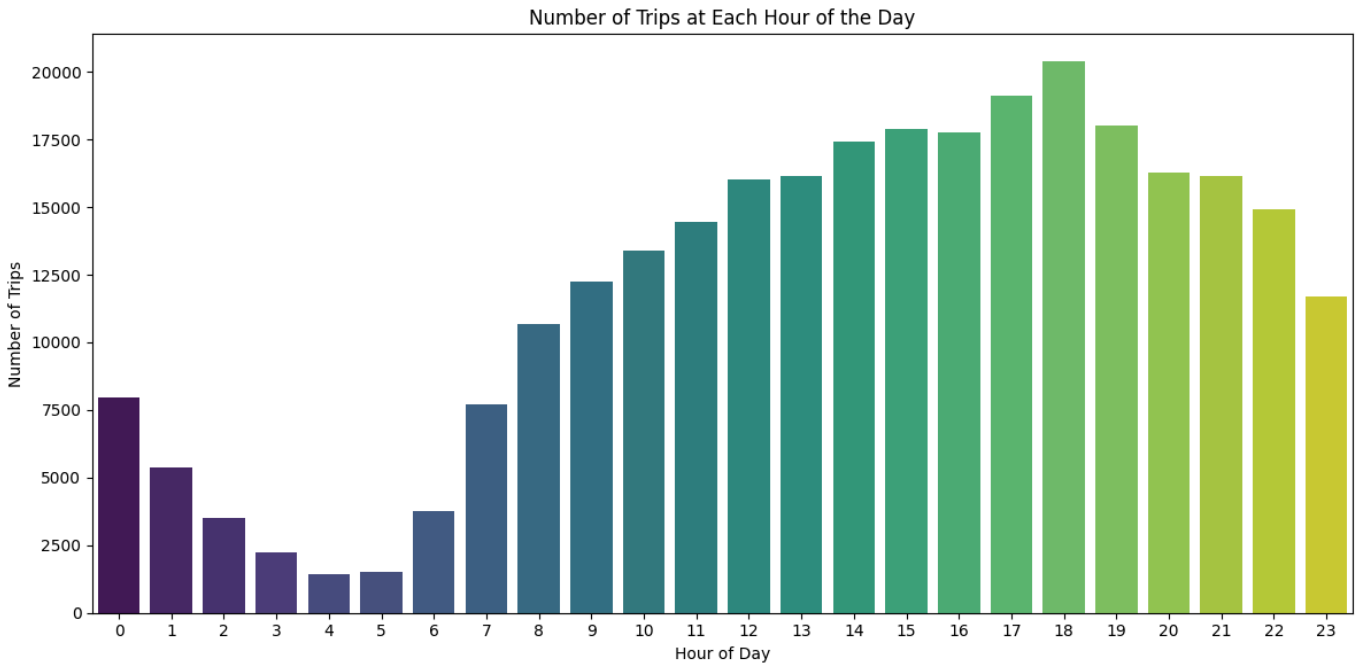
3.2. Detailed EDA: Insights and Strategies

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

- Calculated speed as distance divided by trip duration. Found certain zones with consistently lower speeds, indicating slower or congested routes.

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

- Grouped trips by pickup hour. Found that 6 PM (18:00) had the highest number of trips, and 4 AM had the least.

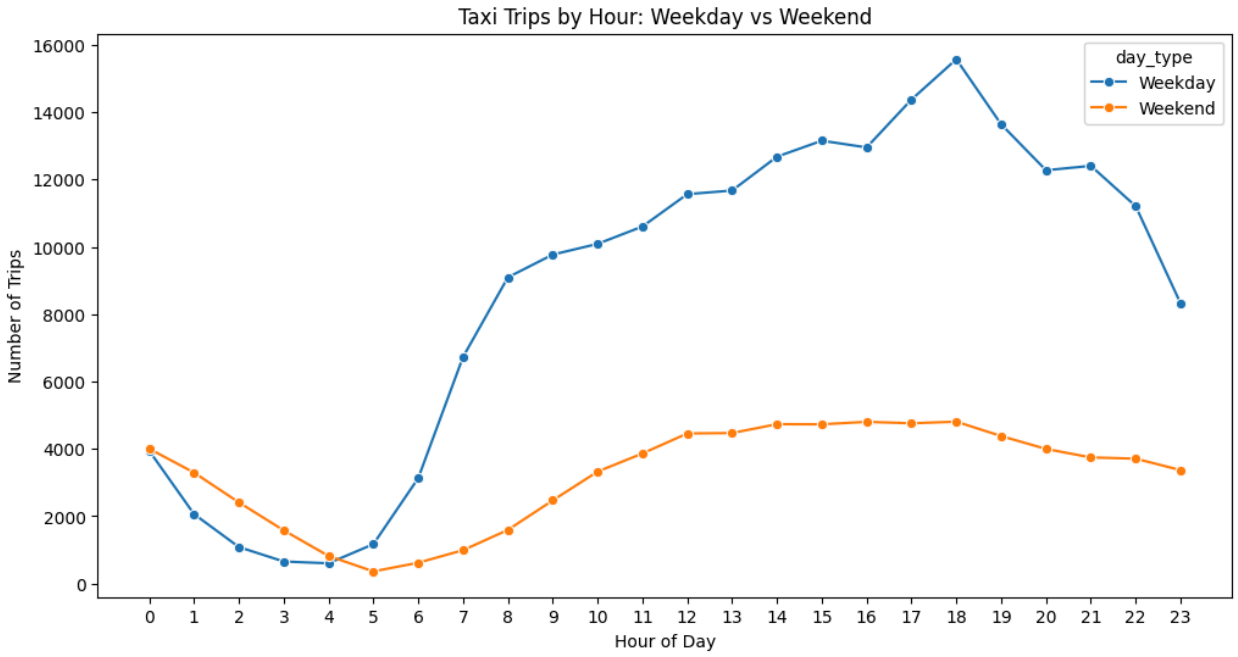


3.2.3. Scale up the number of trips from above to find the actual number of trips

- Used sample ratio to estimate actual trip counts. Helped understand real-world demand in each hour.

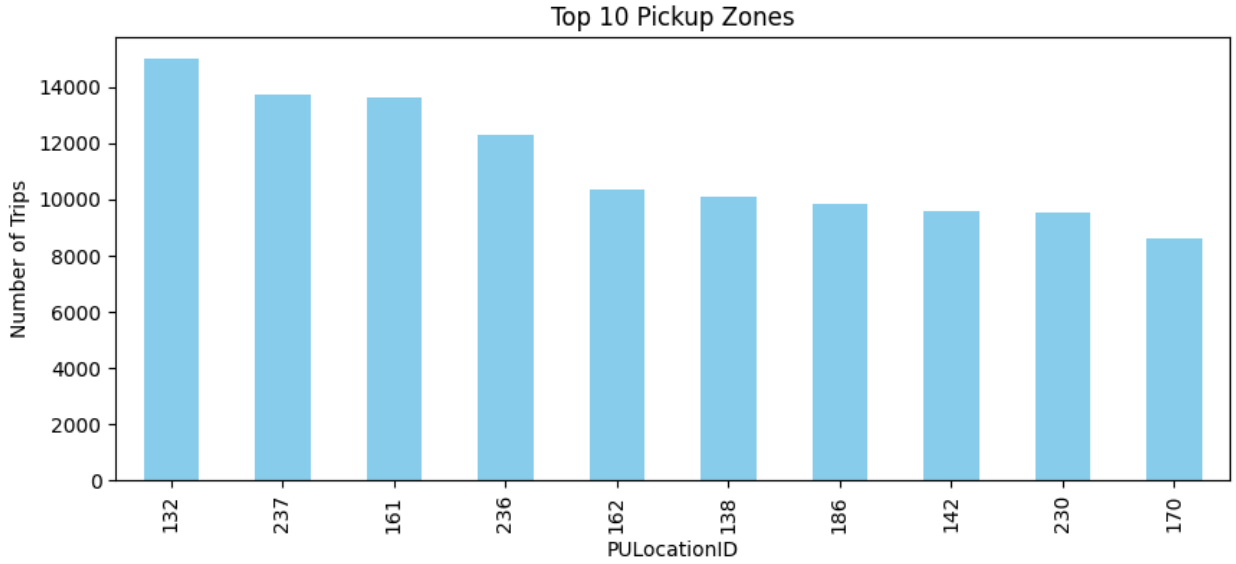
3.2.4. Compare hourly traffic on weekdays and weekends

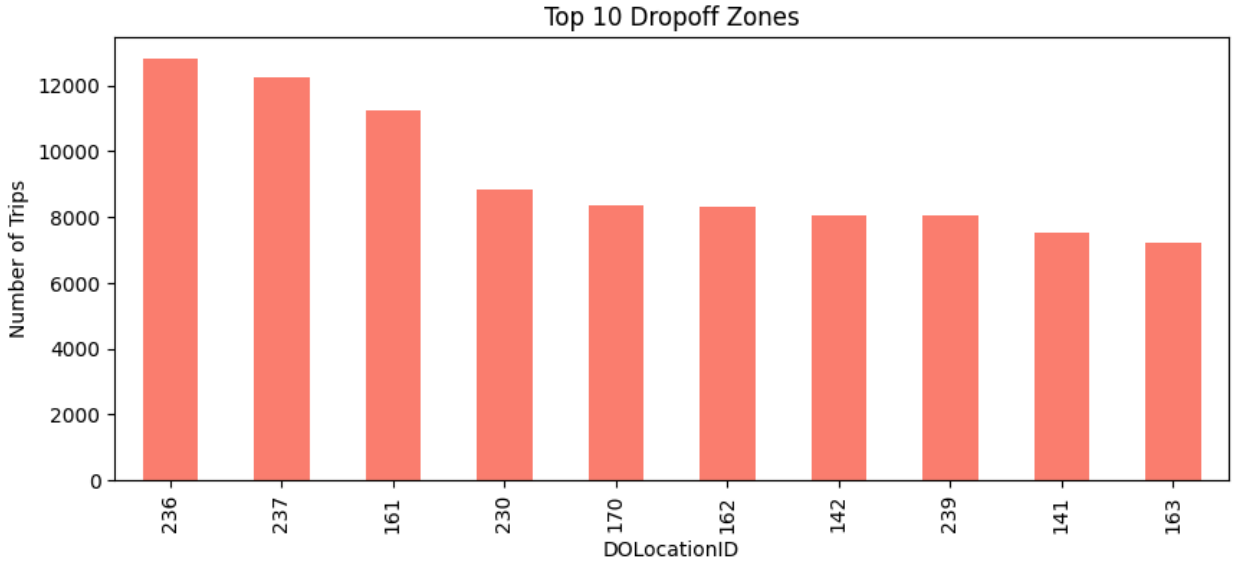
- Split the data by weekdays and weekends. Weekday evenings and weekend afternoons had higher traffic.



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

- Found the busiest zones by counting pickups and dropoffs by hour with PULocationID





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

- Calculated the pickup-to-dropoff ratio for each zone. Some areas had more pickups (starting points), others more dropoffs (destinations).

3.2.7. Identify the top zones with high traffic during night hours

- Filtered trips between 11 PM to 5 AM. Identified zones like nightlife areas and airports with more traffic during these hours.

3.2.8. Find the revenue share for nighttime and daytime hours

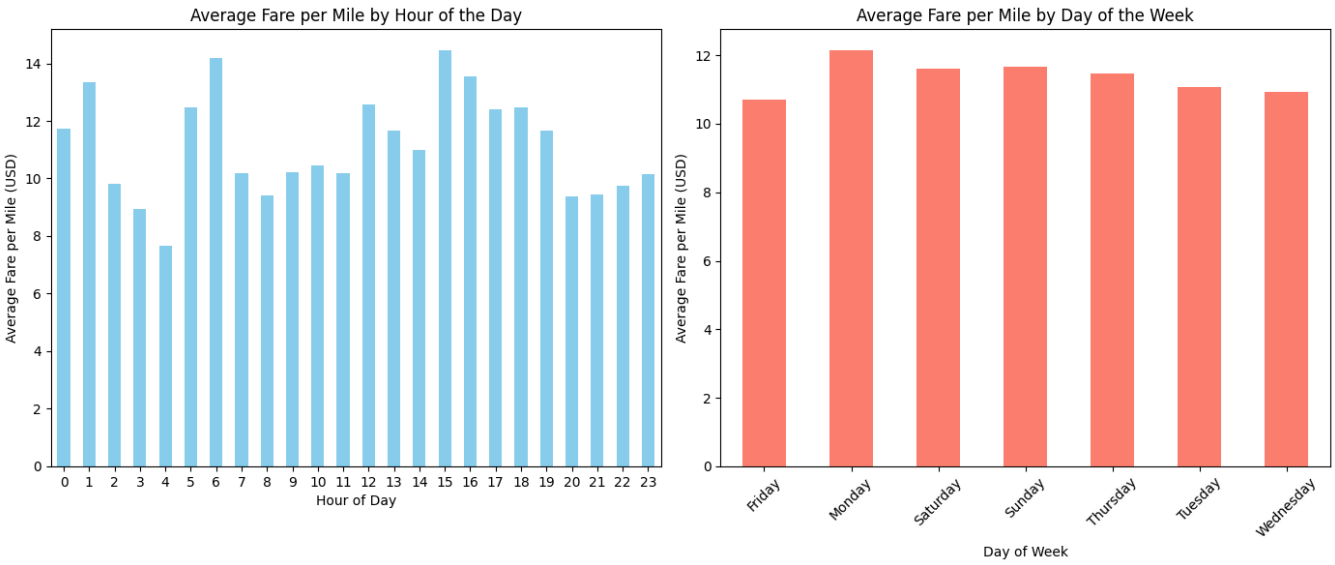
- Separated day (5 AM–11 PM) and night (11 PM–5 AM) trips. Most revenue came from daytime, but night trips still made a decent share.

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

- Divided fare per mile by passenger count. Saw that more passengers lowered the fare per mile per person, making group travel cheaper.

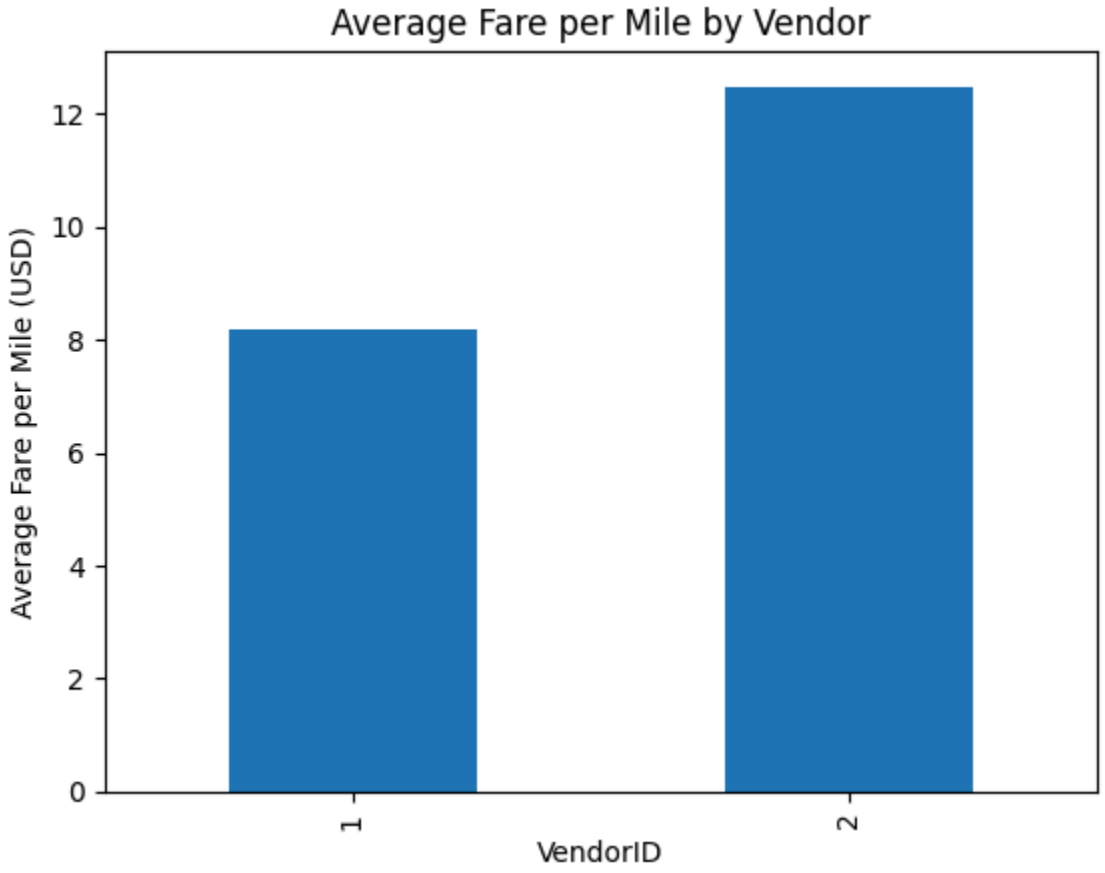
3.2.10. Find the average fare per mile by hours of the day and by days of the week

- Calculated fare per mile for each trip and grouped them by hour and weekday. Early mornings and late-night hours had higher average fare per mile, while mid-day was more stable.



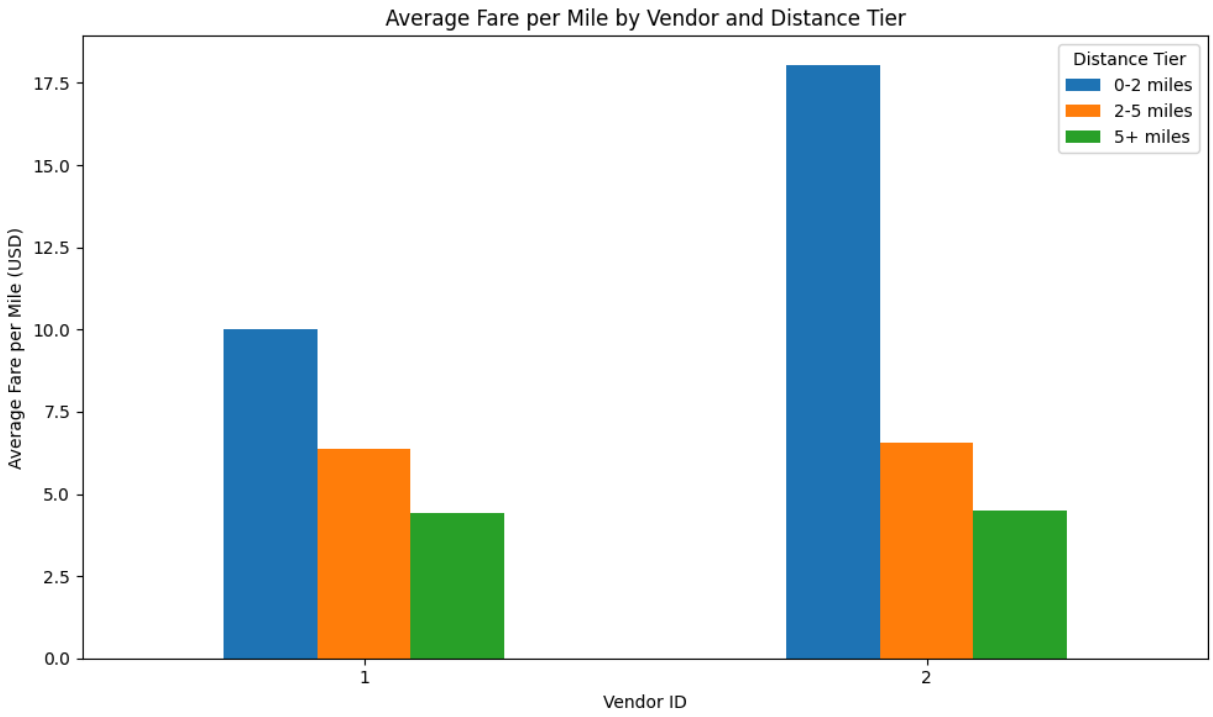
3.2.11. Analyse the average fare per mile for the different vendors

- Compared Vendor 1 and Vendor 2. Vendor 1 had slightly lower average fare per mile compared to Vendor 2.



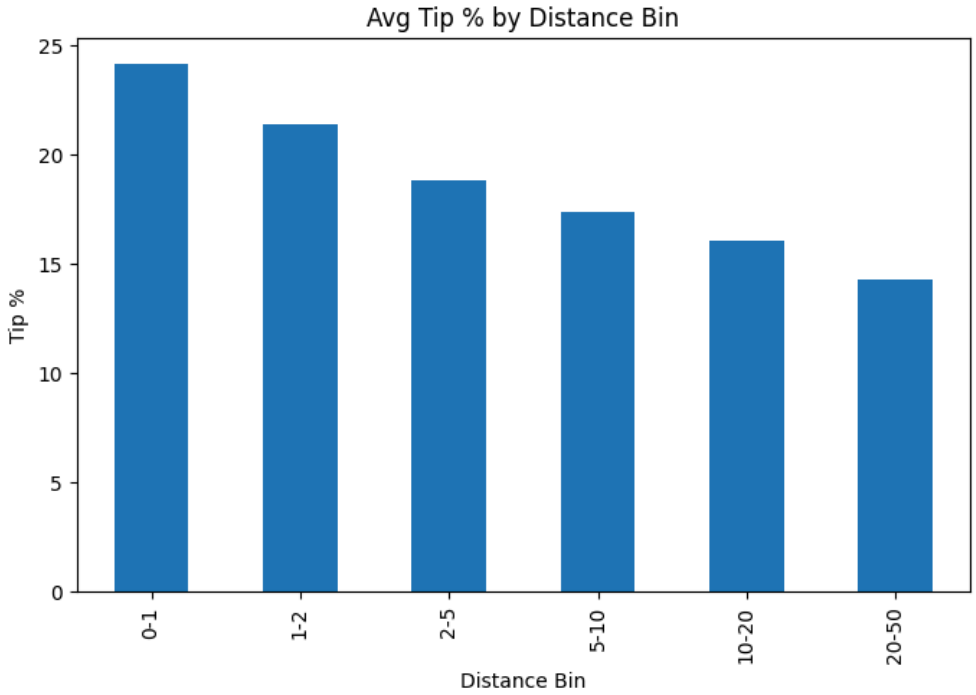
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

- Split trips into short, medium, and long distances. Vendor 1 was more cost-effective for short trips, while Vendor 2 was costlier across all tiers.



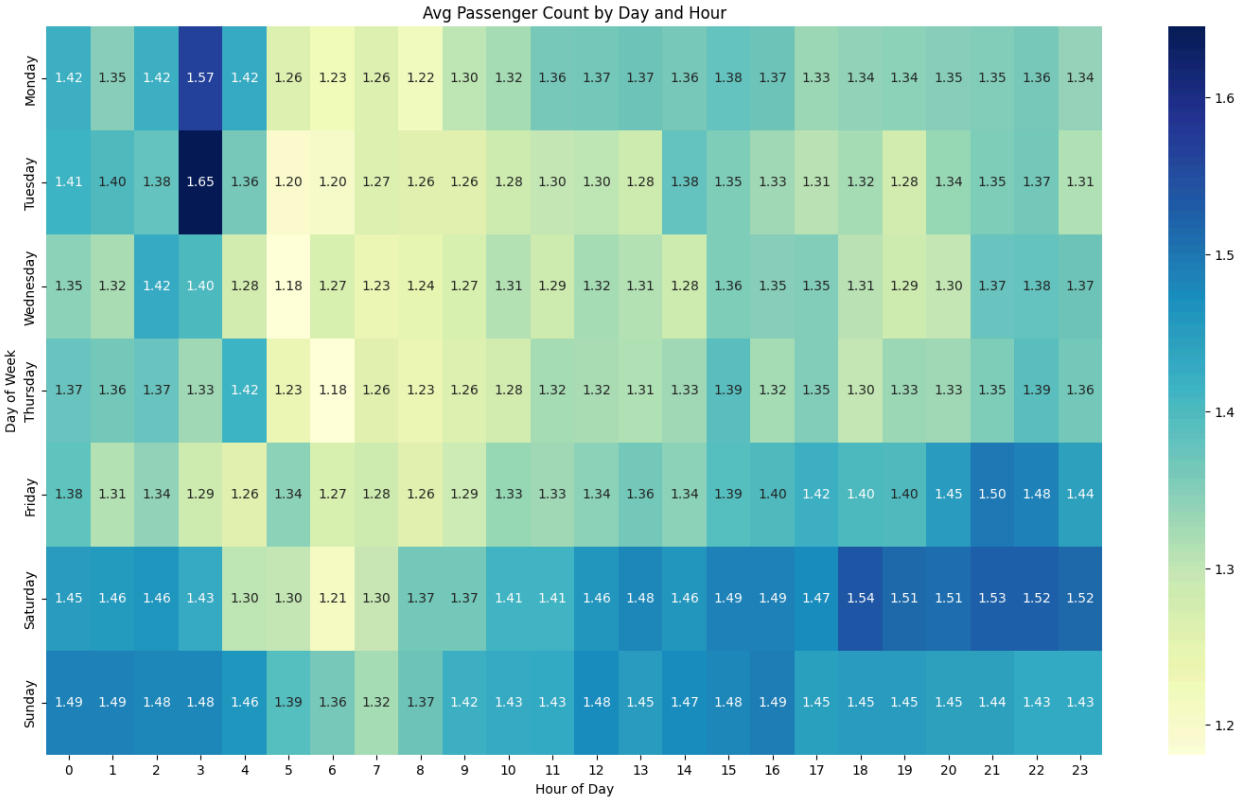
3.2.13. Analyse the tip percentages

- Calculated tip as a percentage of fare. Most trips had low tip percentages, with very few trips crossing 20%. Shorter trips with fewer passengers often had lower tips.



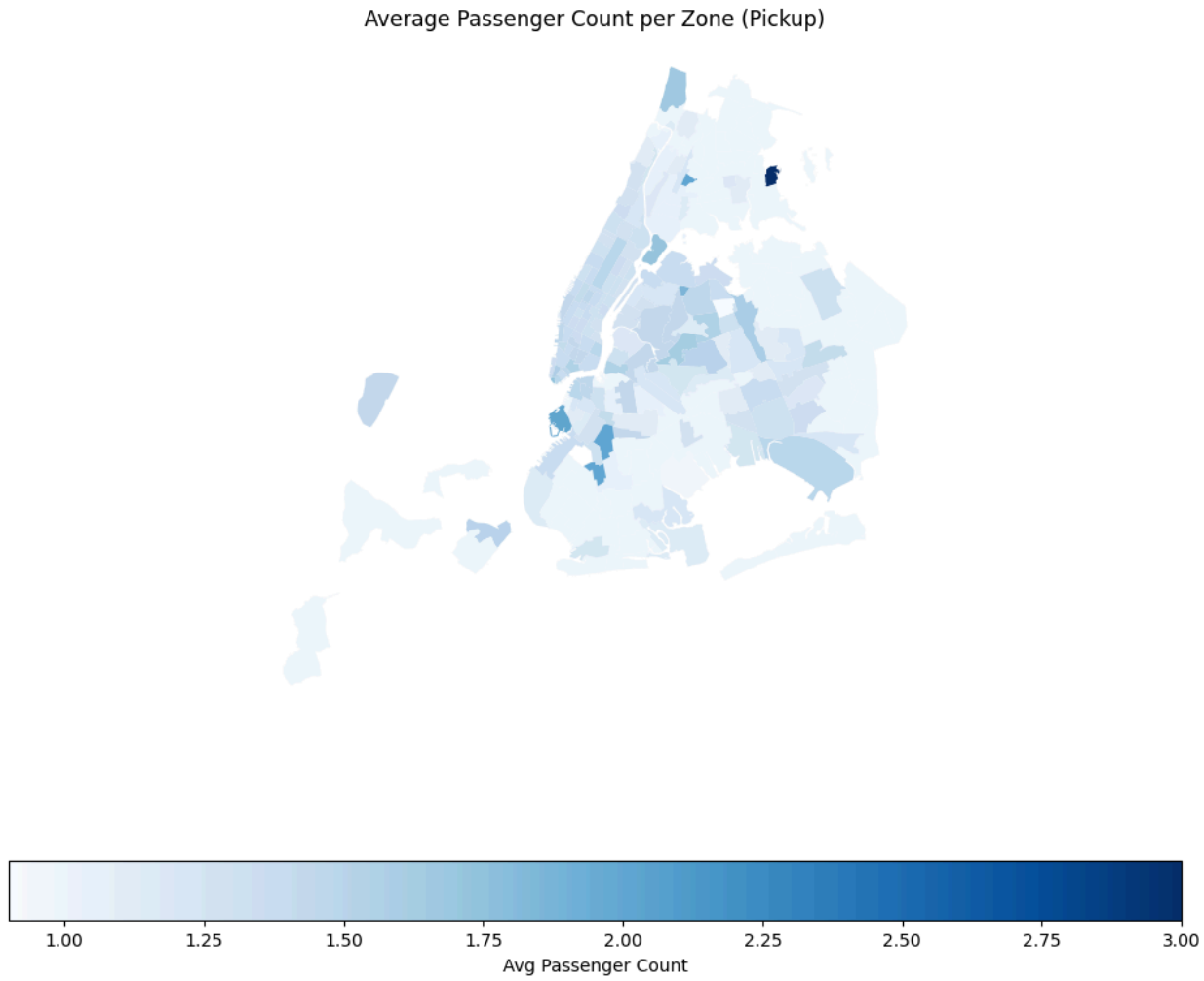
3.2.14. Analyse the trends in passenger count

- Most trips had 1 or 2 passengers. Higher passenger counts were rare, and passenger count remained fairly consistent throughout the day.



3.2.15. Analyse the variation of passenger counts across zones

- Mapped average passenger count per zone. Some zones, especially airport zones and downtown areas, had slightly higher averages due to shared rides or group travel.



3.2.16. Analyse the pickup/dropoff zones or times when extra charges are applied more frequently.

- I had removed the surcharge column during data cleaning process

4. Conclusions

4.1. Final Insights and Recommendations

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

- Place more cabs in zones with high demand like Midtown during peak hours.
- Early morning (7–9 AM) and evening (5–7 PM) need more cabs on busy routes.
- Use fewer cabs in low-demand zones to avoid idle time and fuel waste.
- At night, keep cabs ready in areas like Downtown and around nightlife or restaurant areas.
- Weekends showed different patterns, so plan cab movement accordingly.
- Reduce waiting time for passengers by adjusting cab availability based on trip counts by hour and day.
- Improve routing by checking average speeds — avoid slow routes during traffic hours.

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.

- Keep cabs ready in residential areas during mornings (people go to work or airport).
- Position cabs near commercial and office areas during evenings (for return trips).
- Add more cabs near shopping areas and entertainment zones on weekends.
- Reduce cabs in zones with fewer pickups and drops unless there's an event.

- Zones with high pickups but low dropoffs can become dispatching hubs.

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

- Slight price increases during peak hours when demand is high.
- Offer discounts for long-distance rides to attract more passengers and stay competitive.
- Keep base fare and fare-per-mile reasonable to avoid losing customers to other vendors.
- Tips are low overall — maybe provide better driver incentives or loyalty rewards.
- Track fare per mile and optimize it based on vendor comparisons.
- Nighttime trips make up a good share of revenue, so offer attractive night fares with some safety features.