Data Preparation

There are total twelve files. The data first needs to be combined into one file. So before this

We will import the required libraries into the code. The main library to fetch the data files from folder is glob. So we iterate the files and combine the data files until the all data files are combined.

So coming to sample we are using sample method with .008 fraction so that the data records lie between 250000 and 300000

So after this we export this data to system to use it again in future cases

Data Cleaning

So we havent drop any columns we we are doing a ny pre processing for model building where target and depended variables are not needed. So we left out the columns like that

On analsyis we found two airport_fee columns so we combined it to use it likely wise

Using combine_first method

Fixing Negative Values:

So we seen theres are some numercial attrutes contating negative values. So to solve this either we can do is impute those with absolute values or make it 0 or drop those columns

So for this analysis I have taken or imputed neagtive values with 0 using mask method

Hadeling Missing Values:

So to do this I used .isna() method to find the null values and to get the proportion of those attributes I used .mean()

So there are cloumns like passenger_count, RatecodeID, congestion_surcharge etc

So I filled the NaN values wrt mode values of that columns, Coming to this point we can use mean, median and mode. So the columns here are more likely or pavaroble to take mode.

And I seen some outliers in attributes like ratecodeID which contatin 99, which moslty resonateds to values 6 because the count of 6 is less

Handeling Outlers:

So to find outlies first we need to describe the tables to see the the sats of each column

Upon this we can see theres an vendorID with max value 6 which is not should be

Also theres a passenger_count of 8 which is not true because most of the

case the max go to 6 need to analyse on this

Also Maybe need to check with trip_distance column values

So se taken only the valid values of there attributes like

For passenger_count - df = df[df['passenger_count'] <= 6]

For VendorID - df = df[df['VendorID'].isin([1,2])]

Also we removed the records wher the fareamount is more and 300 and tripdistaince is 0

Same like tripdistance is 0, farte amount is 0 where pickup and drop location ID is different

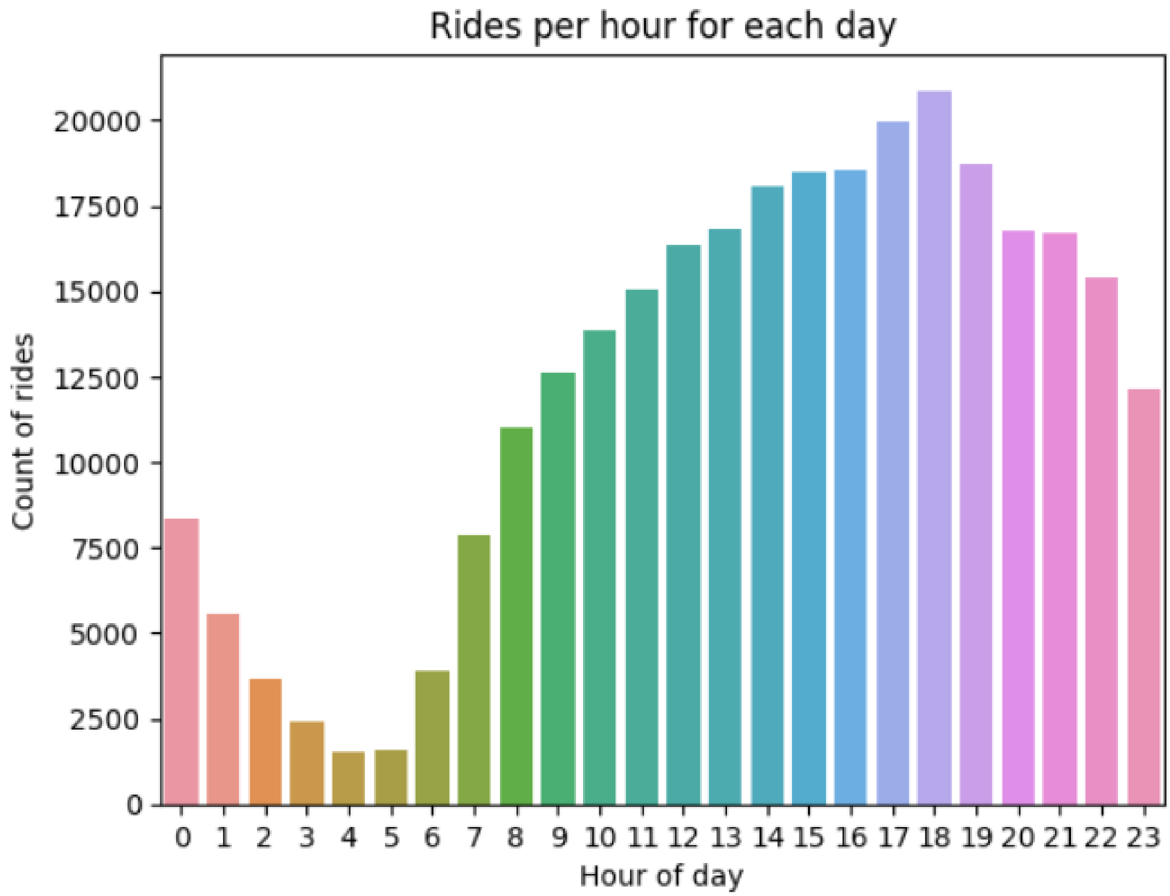
Also filtered out the tripdistance > 250

So I seen theres a payemnt type 0 which is in valid we I removed those records where payment_type is 0

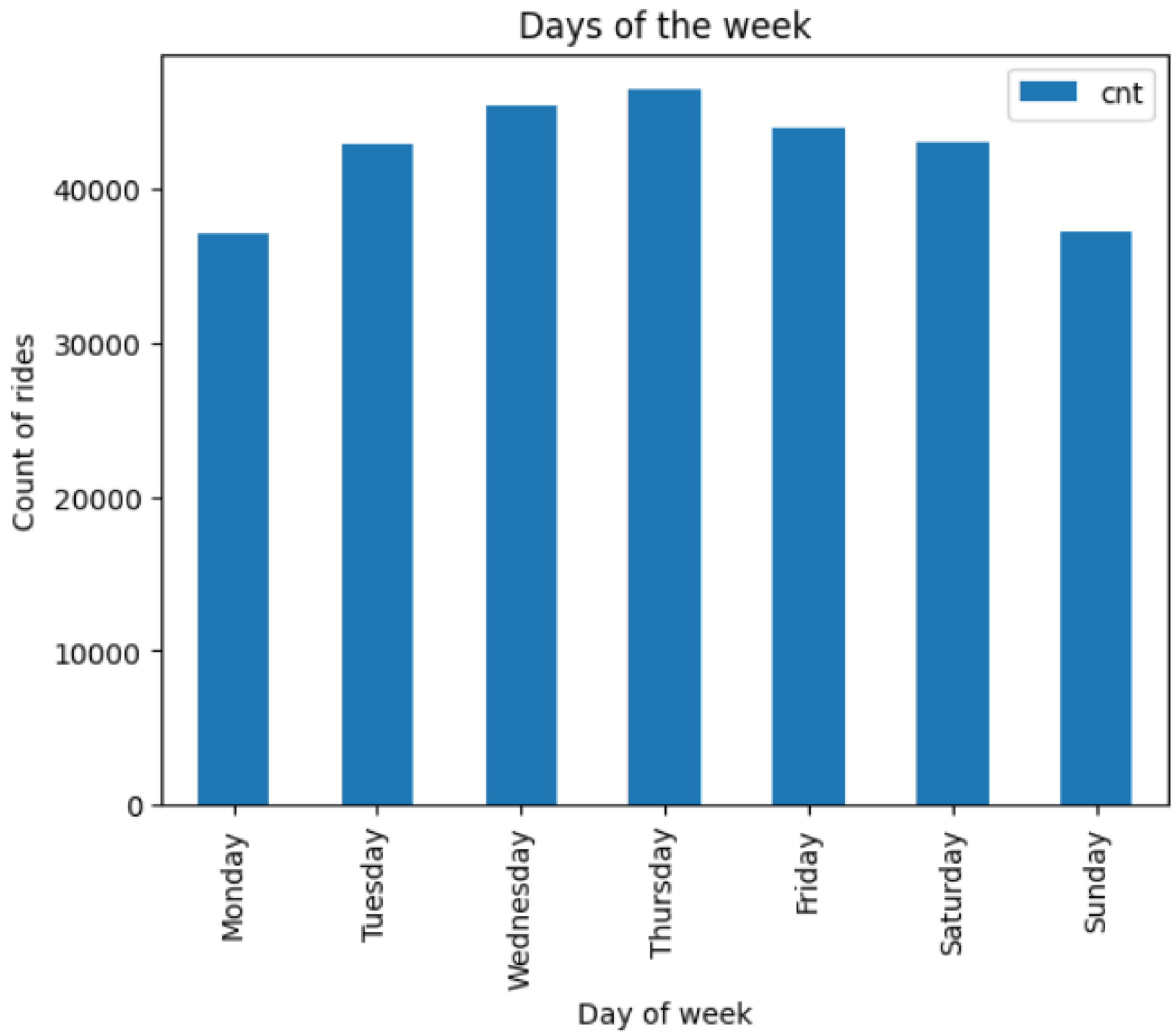
Exploratory Data Analysis:

Got some patterns and trens on analysis

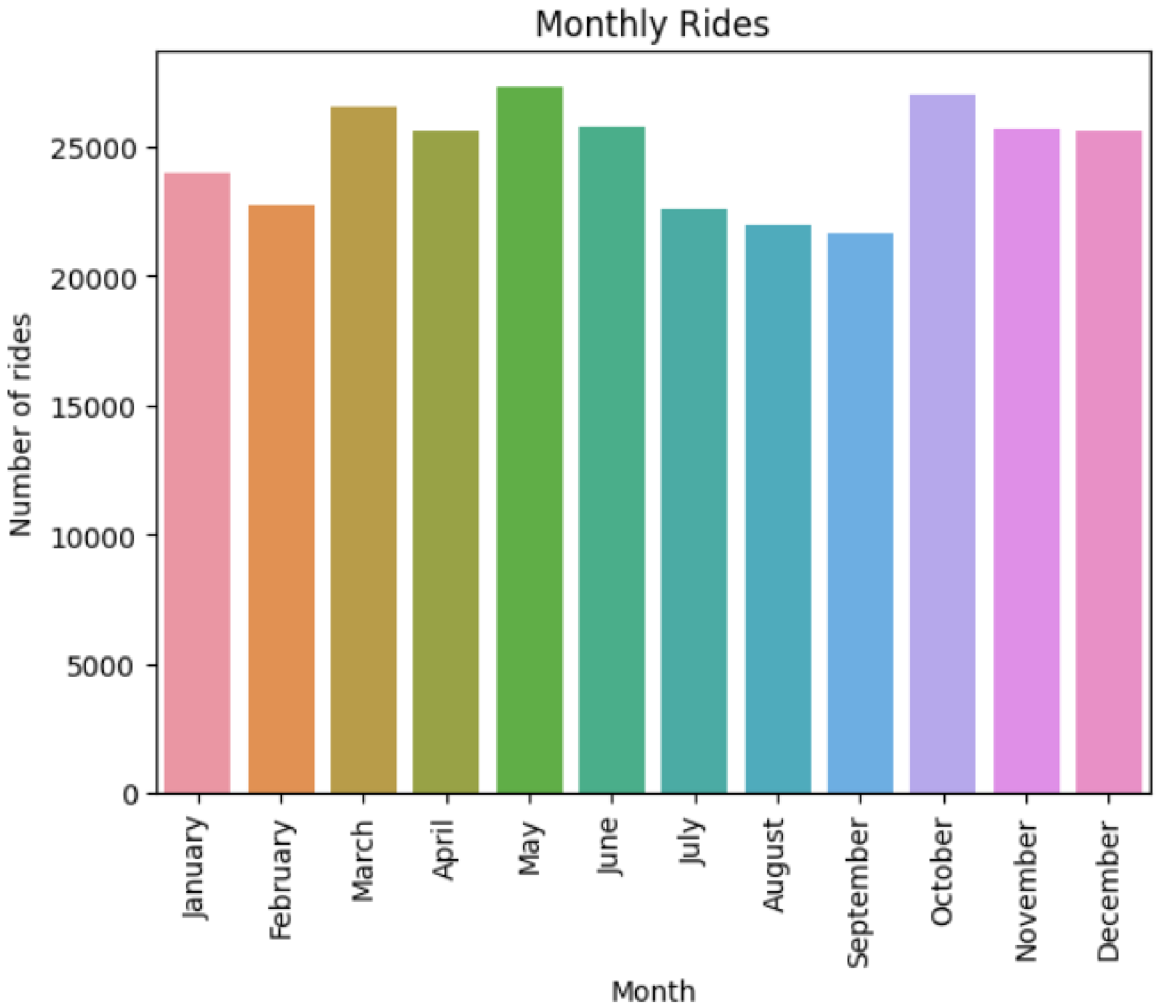
1. **Rides per hour for each day**



2. Days of the week



2. **Monthly Rides**



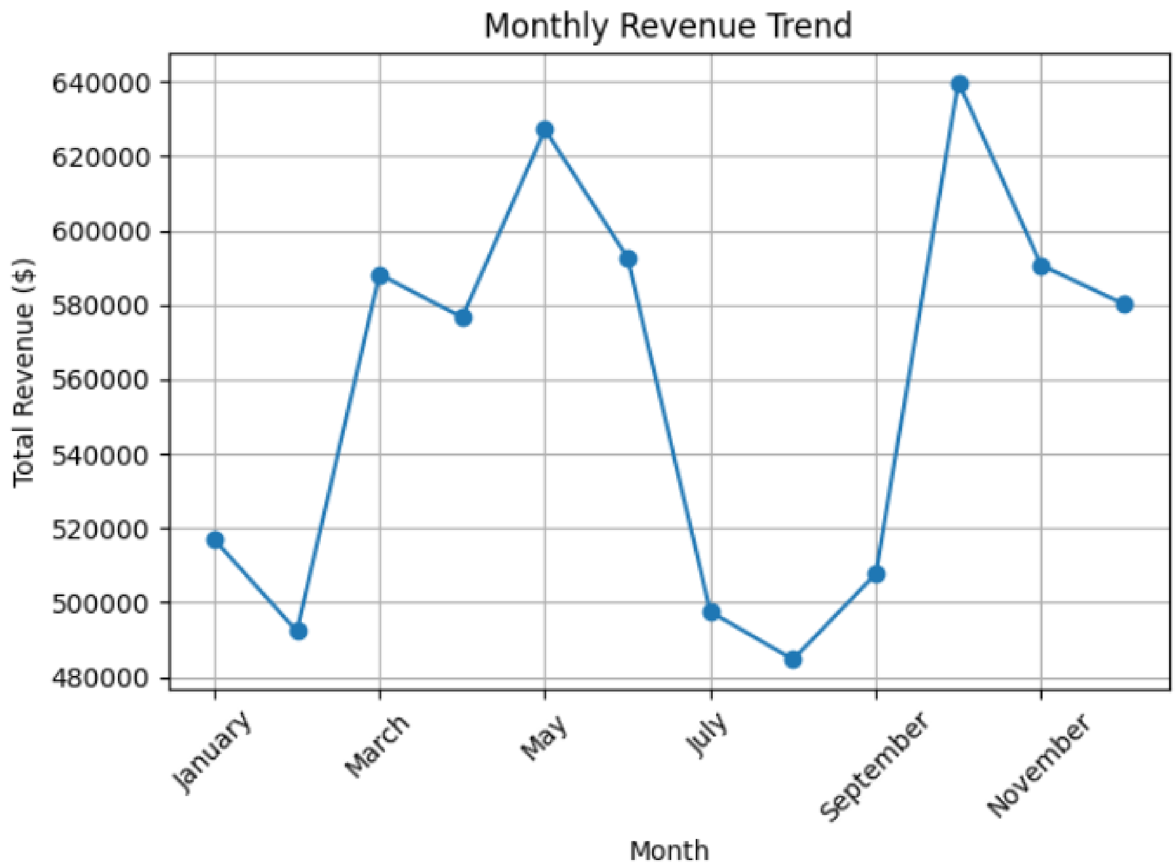
Finanacial Analysis:

So some monetory atrtibutes coantin zero and negative values

So we filterd out tose data and get into new data frame caleed filter_df

Upcomming Trends:

Monthy Revenue Trend



Trip Distance Vs Fare Amount


Trip Distance vs Fare Amount
Fare Amount ($)
Trip Distance (miles)
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100

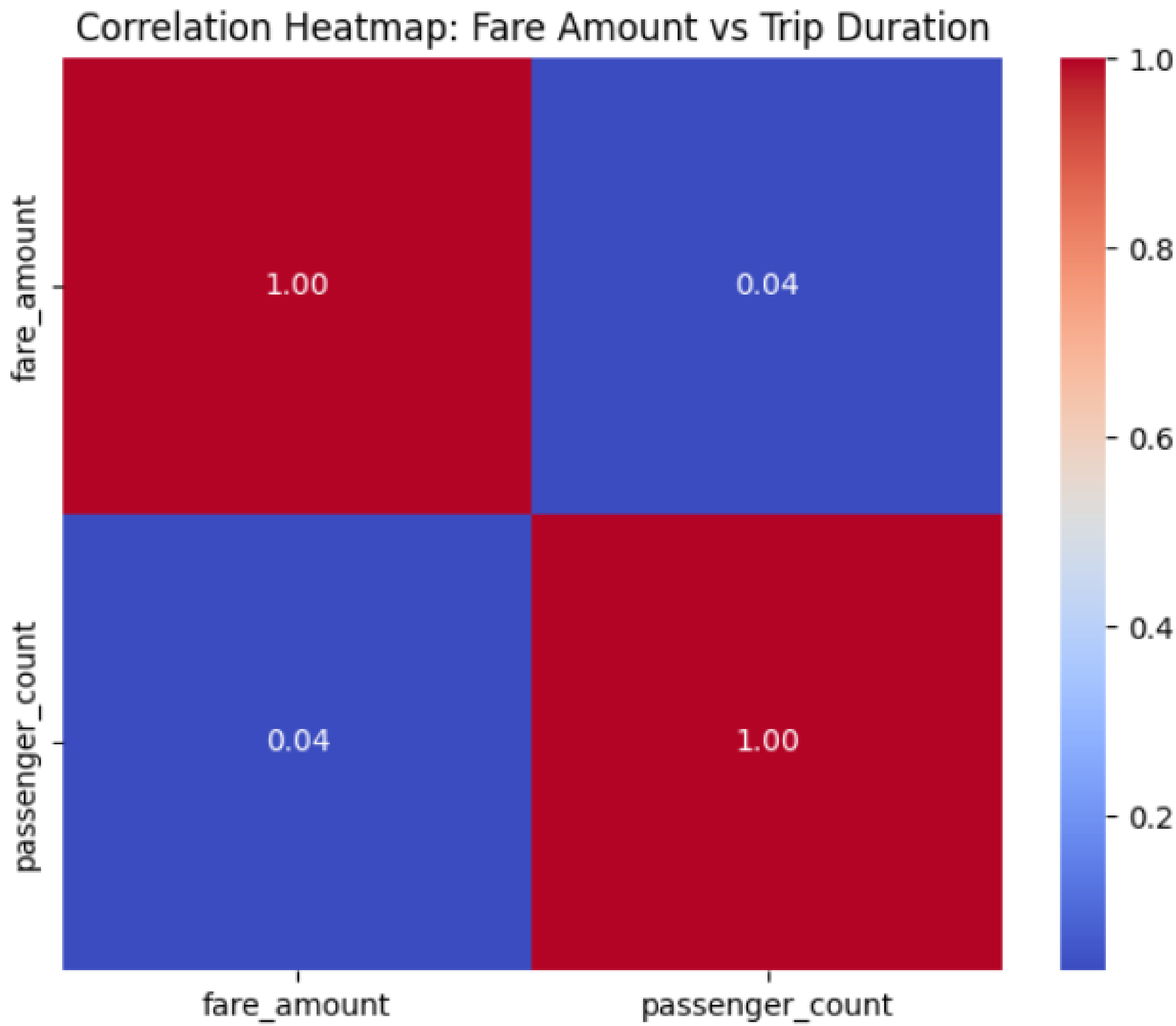
Trip Distance vs Fare Amount
Fare Amount ($)
Trip Distance (miles)
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100


Trip Distance vs Fare Amount
Fare Amount ($)
Trip Distance (miles)
0
50
100
150
200
250
300
350
400
0
20
40
60
80
100

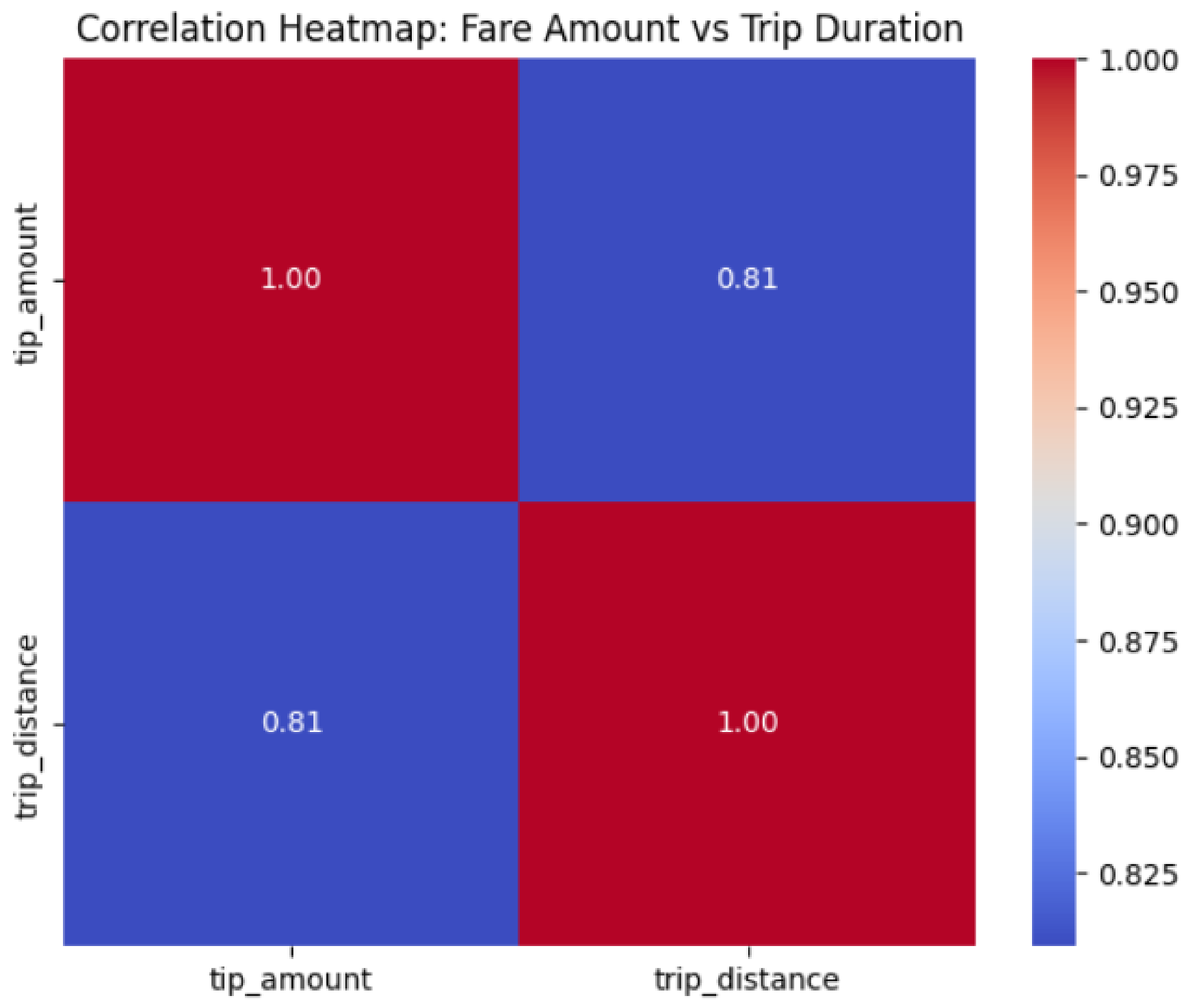
a

Theres a postive correaltion of ~0.94 between fare_amount and trip_distance

Got some heat maps to get on correlation betweeen attributes:



Correlation Heatmap: Fare Amount vs Trip Duration



Done some analysis of payment_types and seen that most payment for taxis are done with credit card

Geographical Analysis:

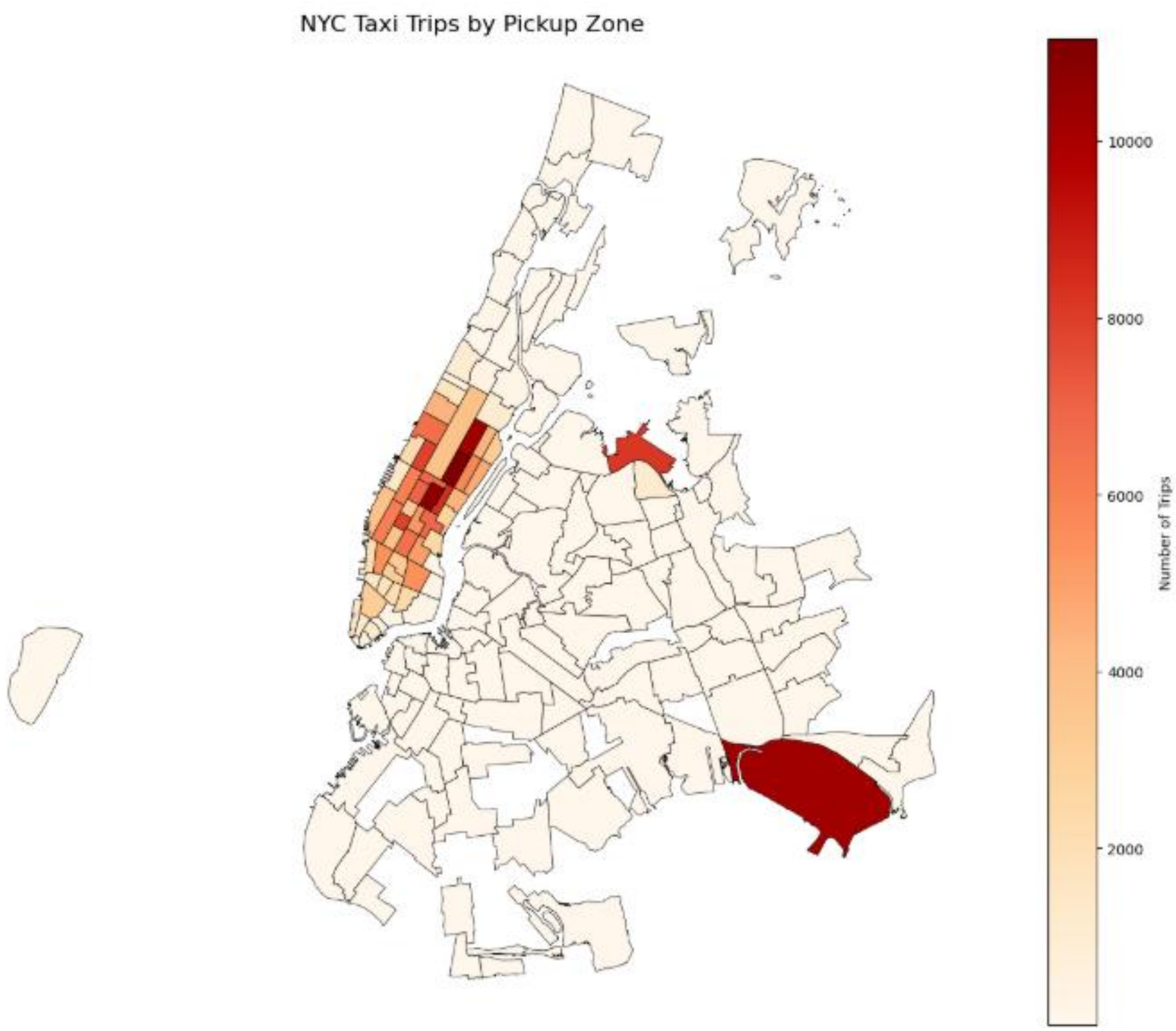
Usisng geopandas library to plot the map of zones data file

Merging zones data file with filter_df on locationID and PULocationID

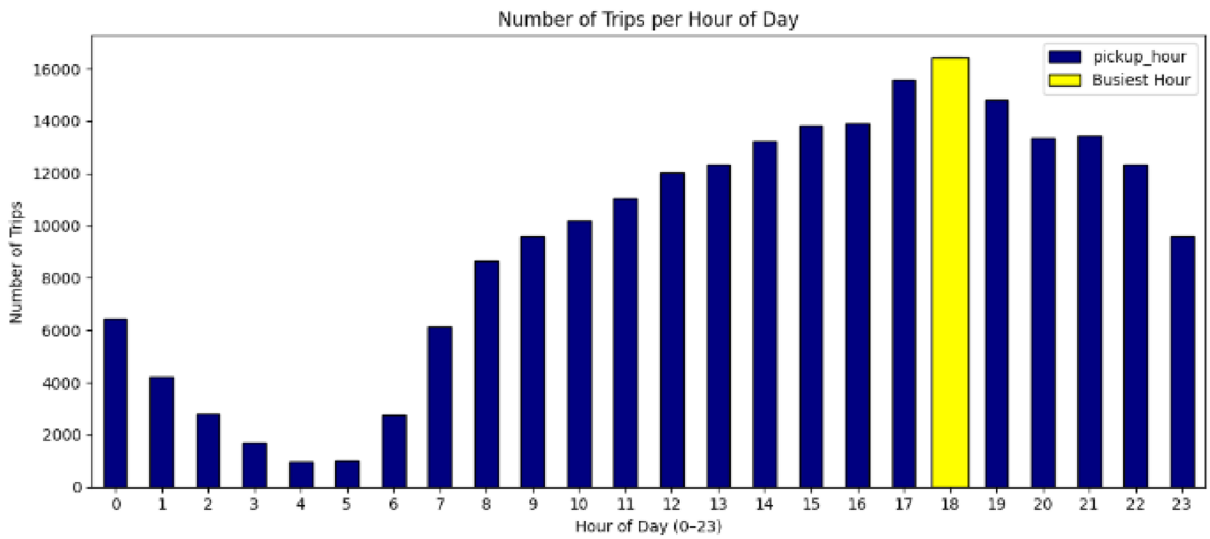
Grouping data to find the total number of trips per locations ID

	LocationID	trip_count
143	237.0	11150
97	161.0	10694
75	132.0	10235
142	236.0	10220
98	162.0	8336

Ploting a color coded map showing zone sise trips



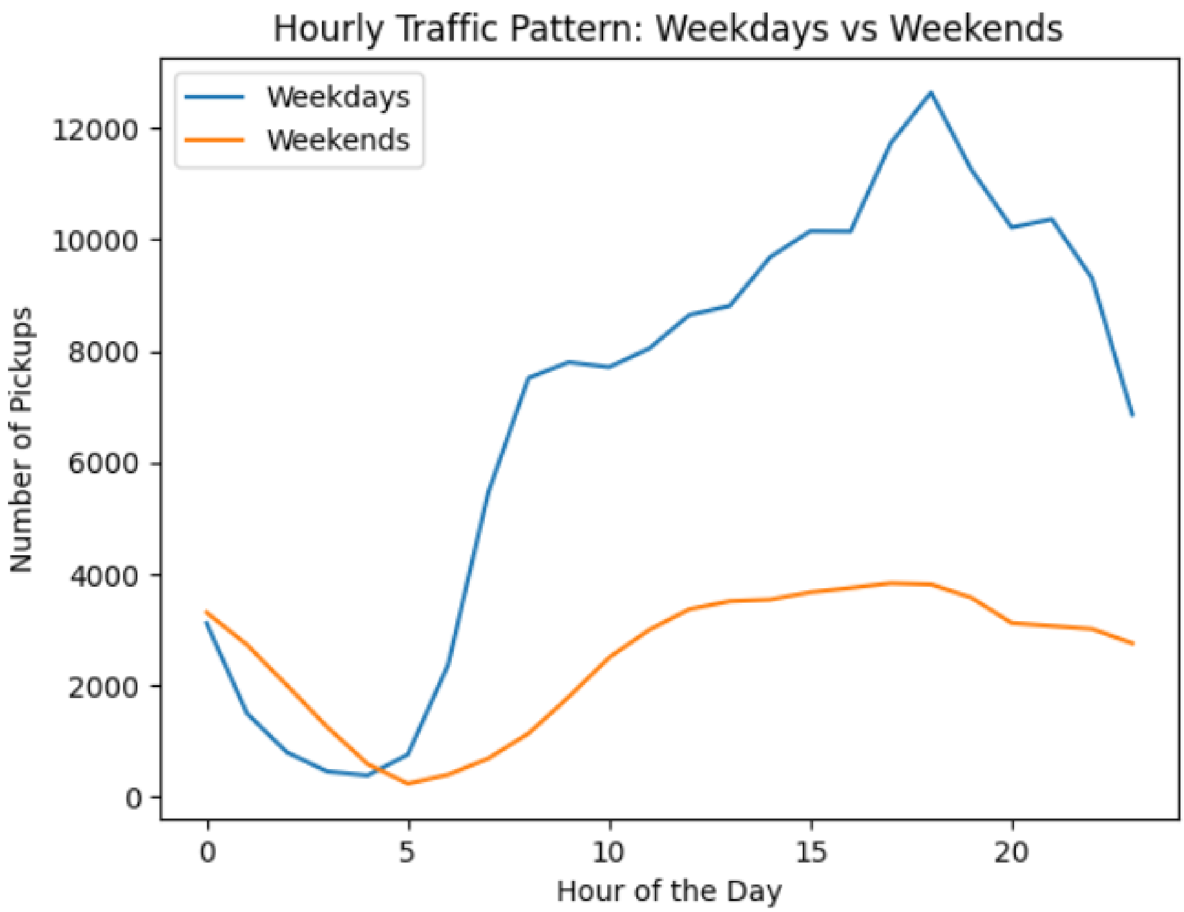
Calculated the number of trips per hour and from that hilighting the busiest hour



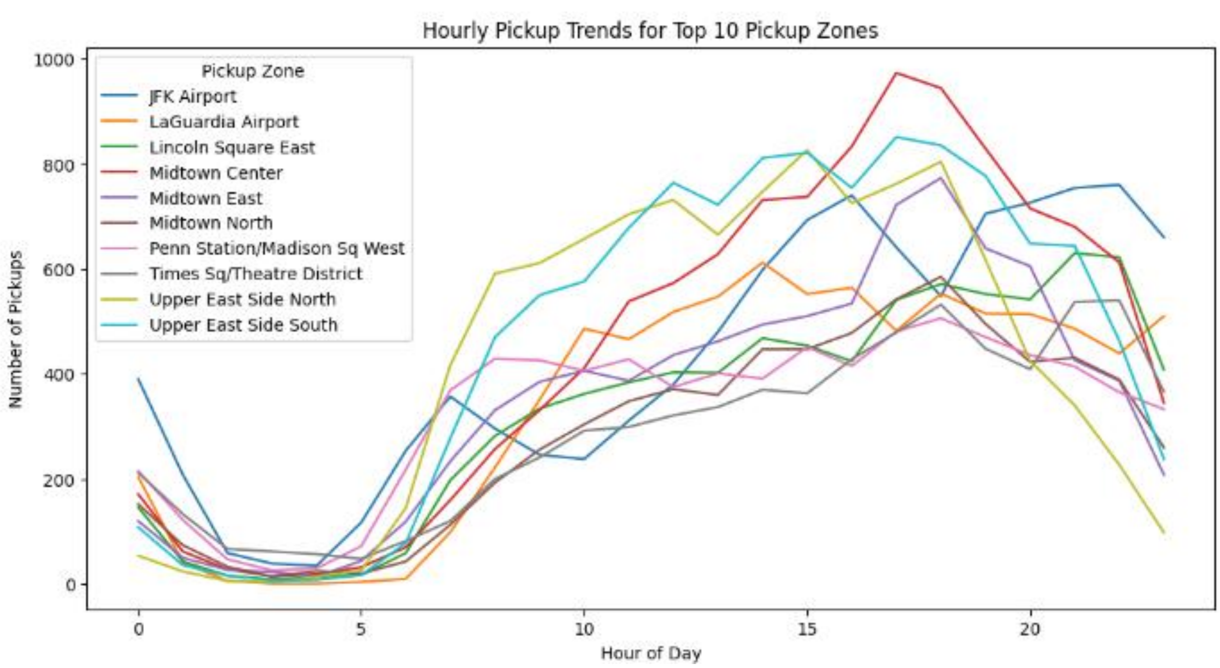
Done analysis on finding the number of trips in the 5 busiest hours:

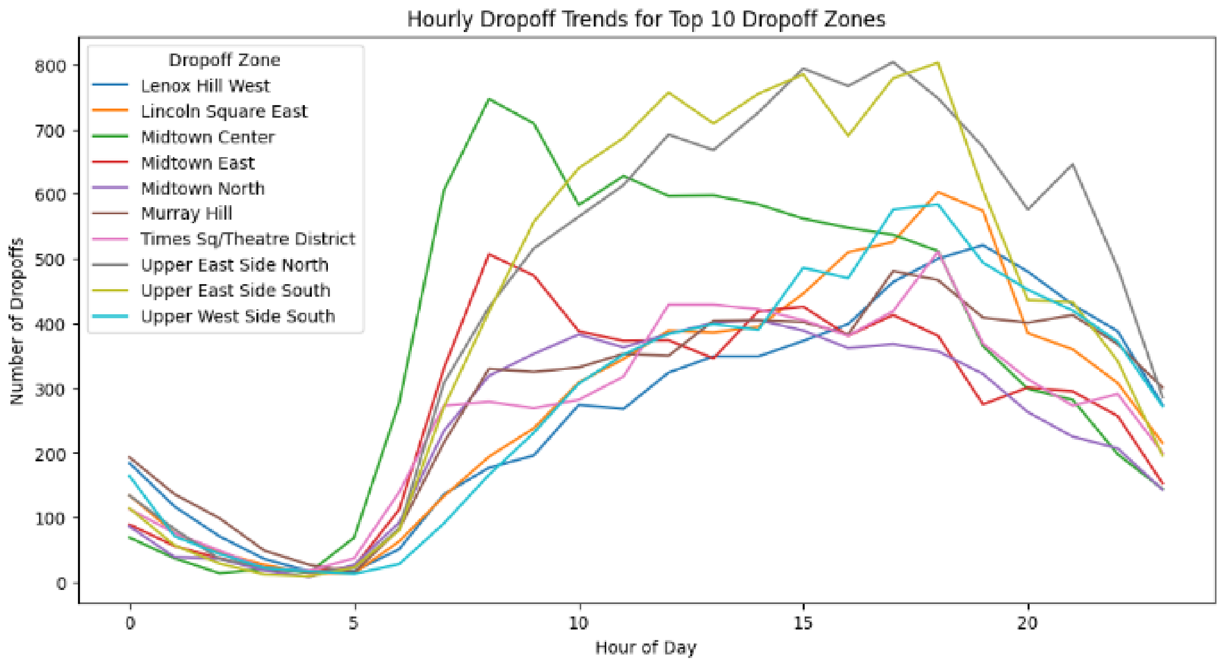
```
Estimated number of trips in the 5 busiest hours:
18    2056375
17    1945625
19    1853875
16    1737125
15    1728000
Name: pickup_hour, dtype: int32
```

Comparing the hourly traffic pattern on weekdays and weekends:



Done some analysi for finding the hourly pickup trends for top 10 pickup zones and also top 10 dropoff zones, below are the patterns or trends to showcase:





Enhanced Dispatch and Routing Optimization Using Demand Trends

Key Observations:

- **Daily Patterns:**
    - *Morning peak (7–9 AM)* and *evening peak (5–7 PM)* show high demand.
    - *Low activity* from 12–4 AM.
- **Weekly Trends:**
    - *Wednesday to Friday* have higher ride volumes.
    - *Weekend evenings/nights* spike due to social activities.
- **Seasonal Insights:**
    - *Sept–Nov* is peak season.
    - *Q4 (Oct–Dec)* drives 27% of annual revenue.
    - *July–Aug* sees a demand dip.

Strategic Recommendations:

- **Time-Based Dispatch:**
    - Boost driver availability during peak hours; reduce during low-demand windows.
- **Route Optimization:**

    - Use historical and real-time traffic data to avoid congestion.
- **Seasonal Fleet Scaling:**
    - Increase fleet and driver incentives in Q4; retain drivers with bonuses in Q3.
- **Predictive Modeling:**
    - Use demand trends in algorithms to pre-position vehicles effectively.

Optimization Approaches:

- **Dynamic Routing:**
    - Adapt to live traffic to reduce delays during peak times.
- **Dispatch Strategy:**
    - Focus on short trips during rush hours; prioritize longer trips in off-peak hours.
- **Proactive Forecasting:**
    - Pre-position vehicles in high-demand zones; leverage ML for accurate predictions.

Key Findings from Data Analysis

High-Demand Zones:

- **Top Pickup Areas:** LaGuardia Airport, Midtown Center, Upper East Side (North/South), Midtown East.
- **Top Drop-off Areas:** Upper East Side, Midtown Center, Upper West Side South, Murray Hill.

Late-Night Patterns:

- Active zones from **11 PM to 5 AM** include nightlife-heavy areas with bars and clubs.

Strategic Recommendations

1. Zone-Specific Dispatching:

- **Airport Strategy:** Increase driver presence near **LaGuardia and JFK** during peak arrival times (6–9 AM, 7–10 PM).

- **Midtown & UES:** Deploy more vehicles from **3–8 PM** to match commute and social demand.

2. Heatmap-Driven Positioning:

- Use real-time and historical data to generate **hourly demand heatmaps**.
- Guide drivers to **high-request areas** at optimal times.

3. Late-Night Strategy:

- Focus on **East Village, Midtown, Uptown** from **11 PM to 3 AM**.
- Adjust driver shifts based on historical pickup trends.

4. Pickup-Dropoff Balance:

- Identify zones with **pickup-dropoff imbalances**.
- Redirect idle vehicles to **high-pickup areas** to optimize coverage.

5. Implementation Tactics:

- **Tech Integration:** Embed heatmaps and live data into driver apps.
- **Driver Incentives:** Bonuses for covering hotspots or night shifts.
- **Continuous Updates:** Refresh demand maps to reflect events, holidays, or seasonal shifts.

Revenue and Correlation Analysis

Revenue Trends:

- **Peak Season:** Highest revenue from **Sept–Nov**; decline during **June–Aug**.
- **Late-Night Revenue:** 11 PM–5 AM shows **high earnings per trip** despite fewer rides.
- **Fare Variability:** Rates vary by **vendor** and **trip distance tiers**.

Correlation Insights:

- **Distance vs. Fare:** Strong correlation (~0.8); fare rises with distance.
- **Duration vs. Fare:** Moderate link (~0.6); less impact than distance.
- **Passenger Count:** Minimal effect (~0.1) on fare.

- **Tips vs. Distance:** Strong correlation; longer trips earn more tips.

Strategic Recommendations

1. Dynamic Pricing:

- **Late-Night Boost:** Slight fare hike from **11 PM–5 AM** to tap into high per-trip revenue.
- **Seasonal Surge:** Apply surge pricing in **Q4** to capitalize on peak demand.

2. Distance-Based Fare Strategy:

- **Short-Trip Premium:** Raise base fare for trips under 2 miles to cover operational costs.
- **Long-Trip Discounts:** Offer reduced per-mile rates for trips over 5 miles to drive volume.

3. Tip Optimization:

- **Encourage Tipping:** Use in-app prompts for long rides.
- **Driver Training:** Train drivers to enhance rider experience and boost tips.

4. Vendor Benchmarking:

- Track competitor fares to ensure **competitive pricing** without sacrificing margins.

Implementation Tips:

- **Smart Pricing Tools:** Automate fare changes based on time, season, and distance.
- **Driver Updates:** Keep drivers informed on fare changes and tip strategies.
- **Ongoing Monitoring:** Use live data to refine pricing and stay aligned with rider behavior.