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

1.1.1. Sample the data and combine the files

In the section, I tried checking for different columns and their unique values like store_and_fwd_flag, airport_fee vs Airport_fee presence in different parquet files etc.

The parquet files provided are per month files which have data provided which has pickup time column called ‘tpep_pickup_datetime’. We are going to use take a file which is having data for whole month. Then we are going to create a DataFrame with that.

Now from the data frame we are picking the pick per hour data for whole month and take a sample percentage of 5% hourly data for all the dates in the month, and putting that in the sample. This single hour sample keep getting concatenated to the sampled_data, which is per file sample. This per file sampled_data is then appended to the overall dataframe which his keeping sampled_data for all the files.

Now there’s another requirement to keep only 250k to 300k entries in the final data frame. So to meet that I had to sample the sampled_data obtained in the last stage by taking only 15% of that, which provides the number of columns in the required range.

There are around **37.9m** rows accross all files for 2023.

Number of entries in the sampled data **284460 rows**

Sampling percentage is **0.75**

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

Used the function:

```
# Fix the index and drop any columns that are not needed
df = df.reset_index(drop=True)
df.head()
```

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationID	DOLocationID	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improve
0	1	2023-06-25 14:02:50	2023-06-25 14:54:33	1.0	12.60	1.0	N	132	181	1	60.4	0.0	0.5	12.35	0.0	
1	2	2023-09-09 11:47:33	2023-09-09 11:55:09	1.0	1.05	1.0	N	239	142	1	9.3	0.0	0.5	3.33	0.0	
2	2	2023-10-13 02:54:45	2023-10-13 03:04:02	1.0	1.85	1.0	N	114	164	1	11.4	1.0	0.5	3.28	0.0	
3	1	2023-07-10 12:08:14	2023-07-10 12:17:15	1.0	1.80	1.0	N	239	75	1	11.4	2.5	0.5	4.60	0.0	
4	2	2023-11-08 15:38:53	2023-11-08 15:56:51	1.0	2.61	1.0	N	186	113	1	17.7	0.0	0.5	4.34	0.0	

Also printed the columns in the sampled file which have negative values:

```
print_negatives(df)
```

```
mta_tax - Negative #: 13
improvement_surcharge - Negative #: 14
total_amount - Negative #: 14
congestion_surcharge - Negative #: 10
Airport_fee - Negative #: 1
```

```
get_null_data_cols(df)
```

```
Columns with null values: ['passenger_count', 'RatecodeID', 'congestion_surcharge', 'Airport_fee', 'airport_fee']
```

2.1.2. Combine the two airport_fee columns

We see there are two airport fee columns, one with name 'Airport_fee' in most of the data files and other with all small case column name 'airport_fee' in the file '2023-1.parquet'.

So, as the approach to combine them together, we look for each row and take the value for both the airport fee columns. Incase any of the column value is nan, we ignore that and check for the other one and take that instead. Created a new column called 'airport_fee_combo' and put values in that. After all the processing is done, dropped both the airport fee columns first. As a next step, renamed the 'airport_fee_combo' column as 'airport_fee'.

2.2. Handling Missing Values

2.2.1. Find the proportion of missing values in each column

Found the null values mean and multiplied it with 100 to get the proportion. Below the code and output :

See only 4 columns having missing values, store_and_fwd_flags, RatecodeID, passenger_count, congestion_surcharge and all of them has 3.47% of null values out of all of its column's data.

2.2.2. Handling missing values in passenger_count

Using mode here as passenger_count replacement for missing data as we are attempting fill it most frequent data. Also the fare calculation doesn't depend upon this passenger_count so this is okay to use this.

2.2.3. Handle missing values in RatecodeID

Using the mode for the RatecodeID column again to fill the null values in the column.

2.2.4. Impute NaN in congestion_surcharge

Taking the mode of congestion_surcharge to fill the null values for this column. This it to fill its value with the most common data in the column instead of mean or median.

2.2.5. Other missing columns

Noticed missing values for 'store_and_fwd_flag' column. Again used the mode() value for replacing the missing values for this column as well.

Also printed after these assignments, there are no nan values.

```
print(df.isna().sum())
```

```
VendorID                0
tpep_pickup_datetime    0
tpep_dropoff_datetime   0
passenger_count         0
trip_distance           0
RatecodeID              0
store_and_fwd_flag      0
PULocationID            0
DOLocationID            0
payment_type            0
fare_amount             0
extra                   0
mta_tax                 0
tip_amount              0
tolls_amount            0
improvement_surcharge   0
total_amount            0
congestion_surcharge    0
airport_fee             0
dtype: int64
```

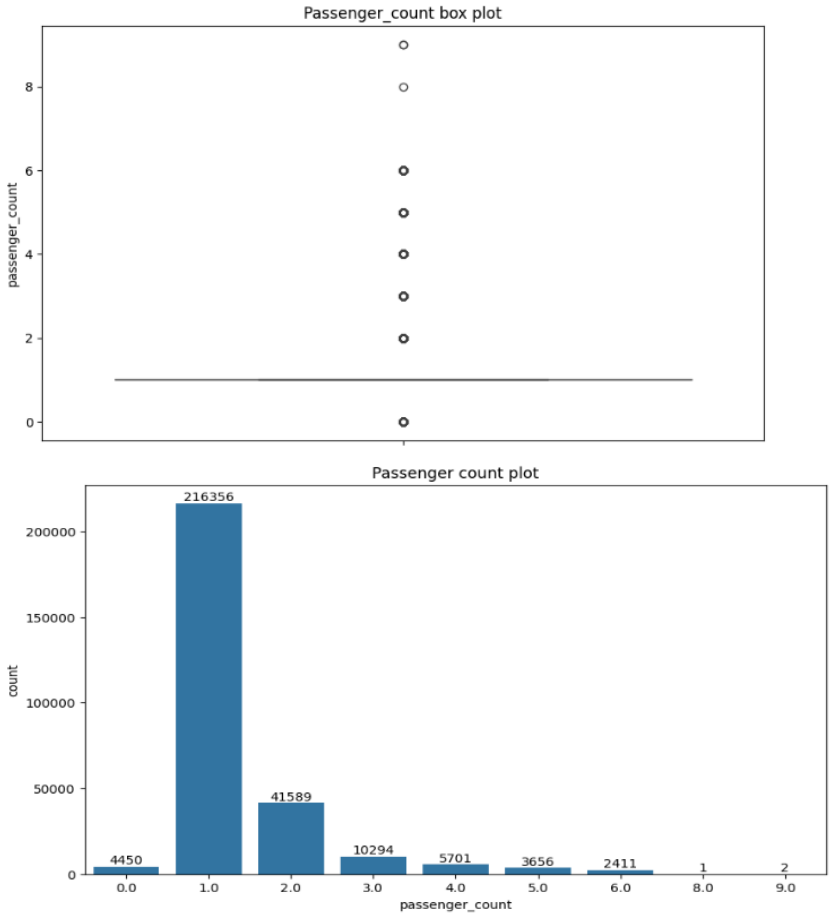
2.3. Handling Outliers and Standardising Values

Outliers noticed in the data

- Passenger count more than 6
- RatecodeID is 99
- Trip distance is more than 305 (sq miles radius of new york city)
- Trip distance is almost 0 and fare amount greater than 300
- Payment_type is 0
- Fare amount and trip distaces are 0 in 50 entries.
- Trip Duration value of 3000
- Tip amount around 90/110/115 for fare amount of 0
- Toll amount of 143 is an outlier
- VendorID with ID 6

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

2.3.1.1. The passenger_count is showing higher than 6 values which are outliers

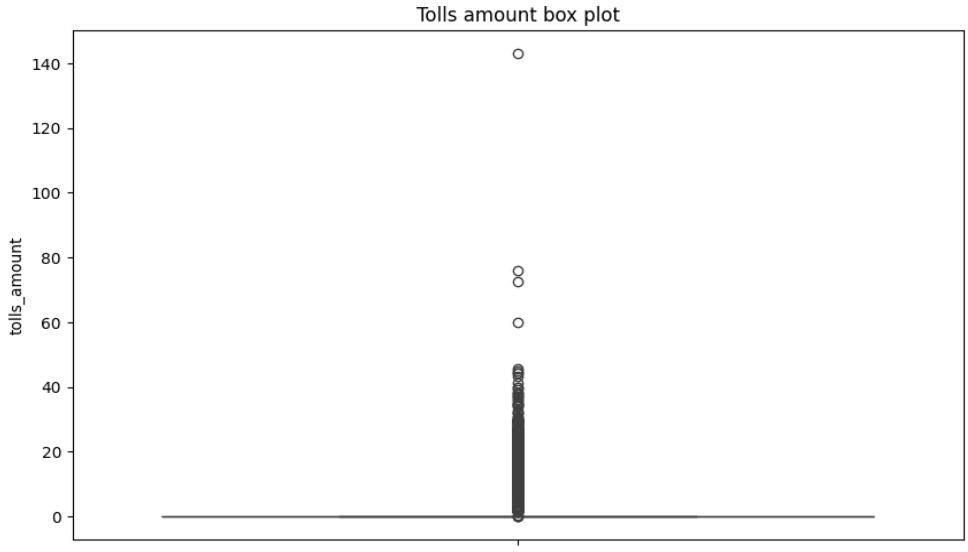


```
df['passenger_count'].value_counts()
```

passenger_count	count
1.0	216356
2.0	41589
3.0	10294
4.0	5701
0.0	4450
5.0	3656
6.0	2411
9.0	2
8.0	1

dtype: int64

2.3.1.2. There are outliers in tolls amount with the amount of 143, as the other toll amount are less than 60. The describe output on the toll amount confirms that too.



```
df['tolls_amount'].describe()
```

	tolls_amount
count	284460.000000
mean	0.593433
std	2.190028
min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
max	143.000000

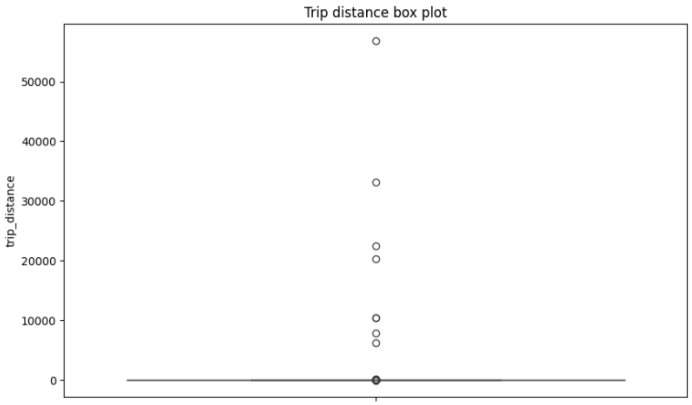
dtype: float64

2.3.1.3. Payment type of 0 is not defined in the data description.

```
df.groupby('payment_type')['fare_amount']
```

payment_type	fare_amount
0	**222656.36**
1	4430179.42
2	927309.32
3	21669.14
4	41229.44

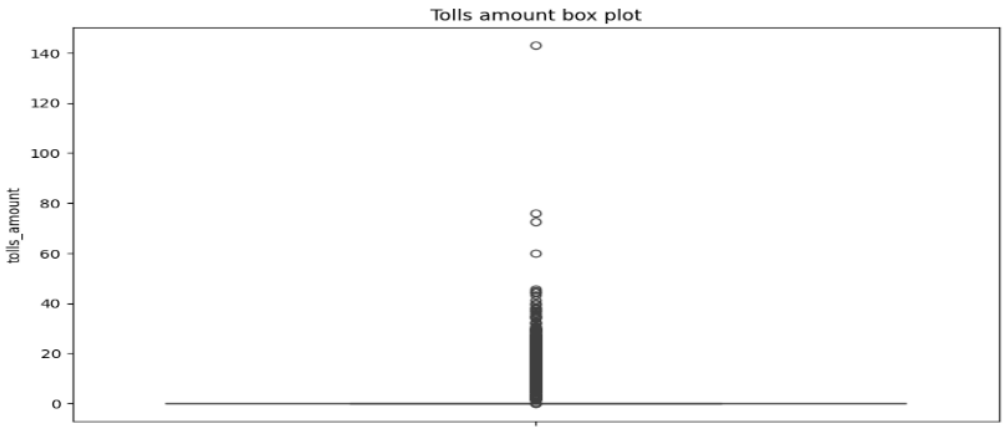
2.3.1.4. Trip distance has mean value of 4 with 75 percentile values are less than 3.4miles. After some more look up, there are trips which are more than 104.3m have their trip distance more than 6262 and more. So, they can be considered as outlier.




```
df['trip_distance'].describe()
```

	trip_distance
count	284460.000000
mean	4.035872
std	139.891866
min	0.000000
25%	1.040000
50%	1.790000
75%	3.400000
max	56823.800000

2.3.1.5. Tip amount analysis also shows presence of outliers. The describe() operation on the column shows that 75 percentiles of its values are less than $4.45. But the box plot shows that this has values which are more than 100 and max is around 140. There are few rows in that which are showing 0 fare amount but tip amount more than $110. These look suspicious.




```
df['tip_amount'].describe()
```

	tip_amount
count	284460.000000
mean	3.547149
std	4.021471
min	0.000000
25%	1.000000
50%	2.850000
75%	4.450000
max	150.000000

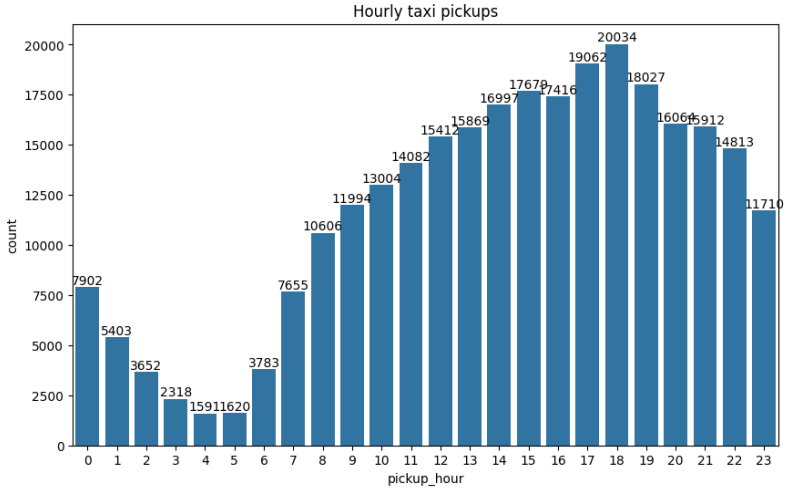
3. Exploratory Data Analysis

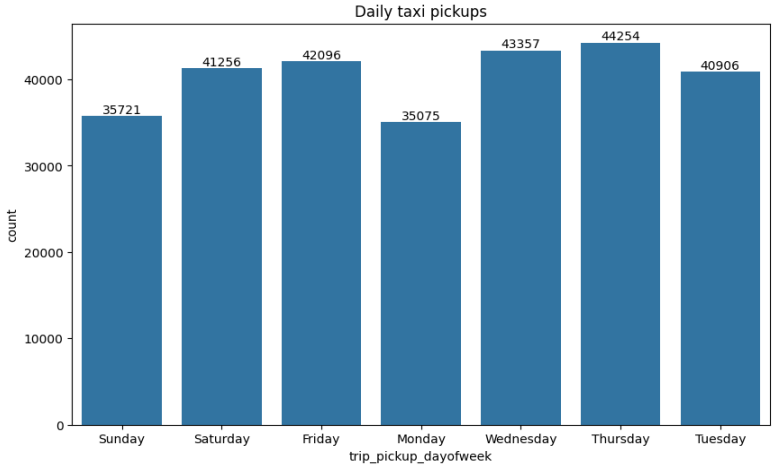
3.1. General EDA: Finding Patterns and Trends

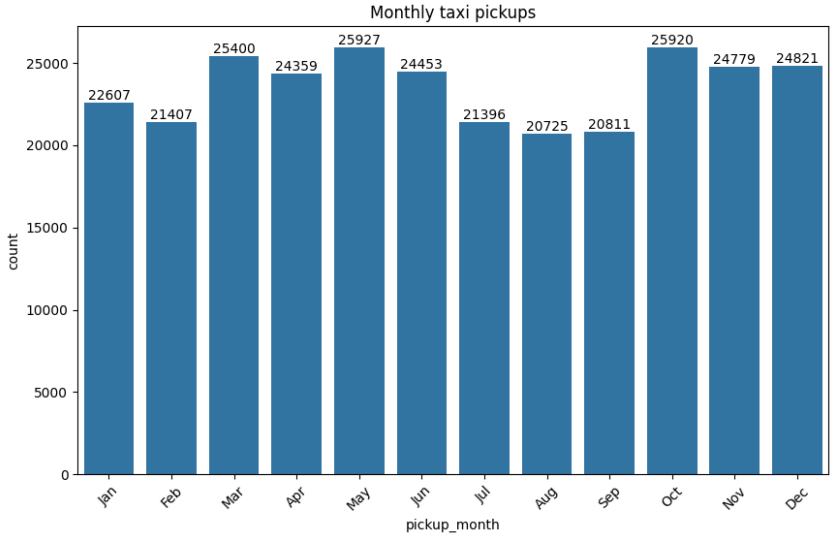
3.1.1. Classify variables into categorical and numerical

```
# List of continuous (numerical) columns in the dataset
category_cols = ["VendorID", "PULocationID", "DOLocationID", "payment_type", "store_and_fwd_flag", "RatecodeID" ]

# Fare related columns
numerical_cols = ["passenger_count", "trip_distance", "fare_amount", "RatecodeID", "extra", "mta_tax", "tip_amount", "toll

# Time related columns
time_cols = ["tpep_pickup_datetime", "tpep_dropoff_datetime"]
```

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







Analysis

- **Hourly Pickup trend**

  Hourly pickup shows evening hours are much more busy 5PM to 7PM. Early morning hours 2AM to 6AM are low traffic hours with lowest hour reaching around 4AM and 5AM.

- **Daily Pickup trend**

  Weekday traffic analysis shows that Wednesday and Thursday are higher traffic days. Then Sunday and Monday shows lowest traffic days during the week.

- **Monthly pickup trend**

  Monthly revenue analysis shows higher traffic with highest in May, followed by October and March. Lower traffic months are seen in August, September followed by Feb.

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

	fare_amount	tip_amount	total_amount	trip_distance
1399055	60.4	12.35	74.25	12.60
41775	9.3	3.33	16.63	1.05
1149477	11.4	3.28	19.68	1.85
805641	11.4	4.60	20.00	1.80
1774725	17.7	4.34	26.04	2.61

Answer:

Yes, it will be useful to seperate these financial parameter out to understand their zeros and negative entries.

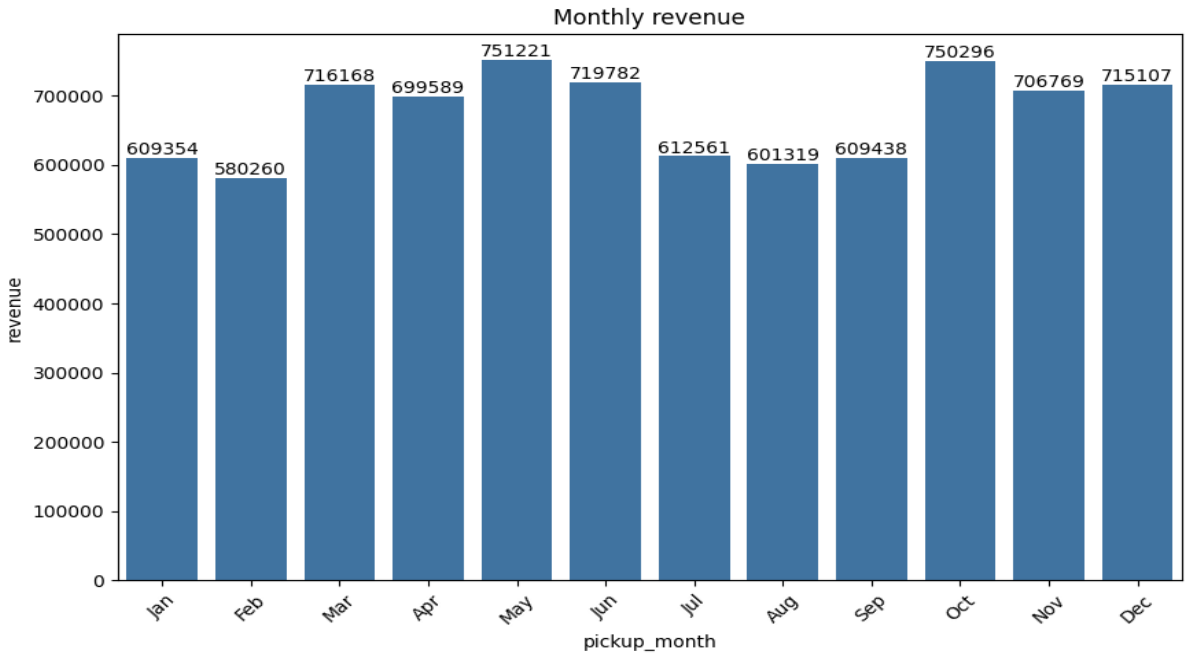
- The trip distance and tip amount of zero is normal for a same location id and trips which may not have tips offered by customer. So those are valid entries and should be kept for analysis.
- Also if there are lots of trips with fare_amount is zero, those should be kept to understand the reason for such trips, to avoid losses if trip distance is also high.

Explanation

After analysing the entries, found there are quite a few of entries which were having trip_distance of 0 and have PULocationID and DOLocationID as different. Removed those entries as those are incorrect entries.

There are quite a few entries which has trip_distance as 0 still and those are from the same location. As same zone entries may have zero distance covered, so leaving them as it is. There are around 2379 rows for such entries.

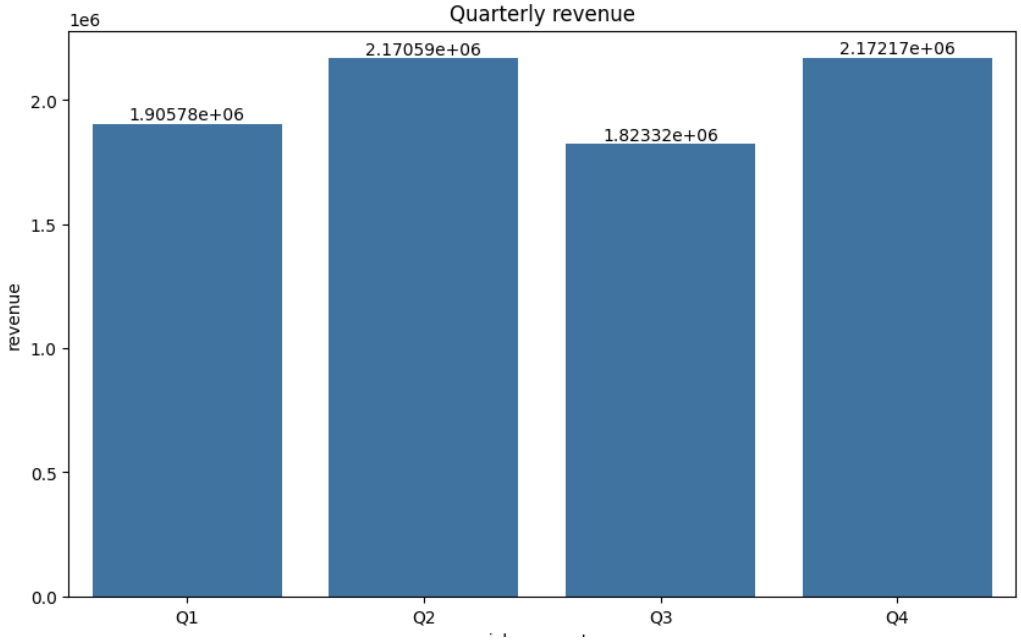
3.1.4. Analyse the monthly revenue trends



Explanation

Monthly revenue shows interesting data. The month of May has highest revenue, followed by Oct. This could be due to holiday seasons in October, with not much clear explanation for month of May. Also the month of Feb and Jan are quite low and that could be due to winter season.

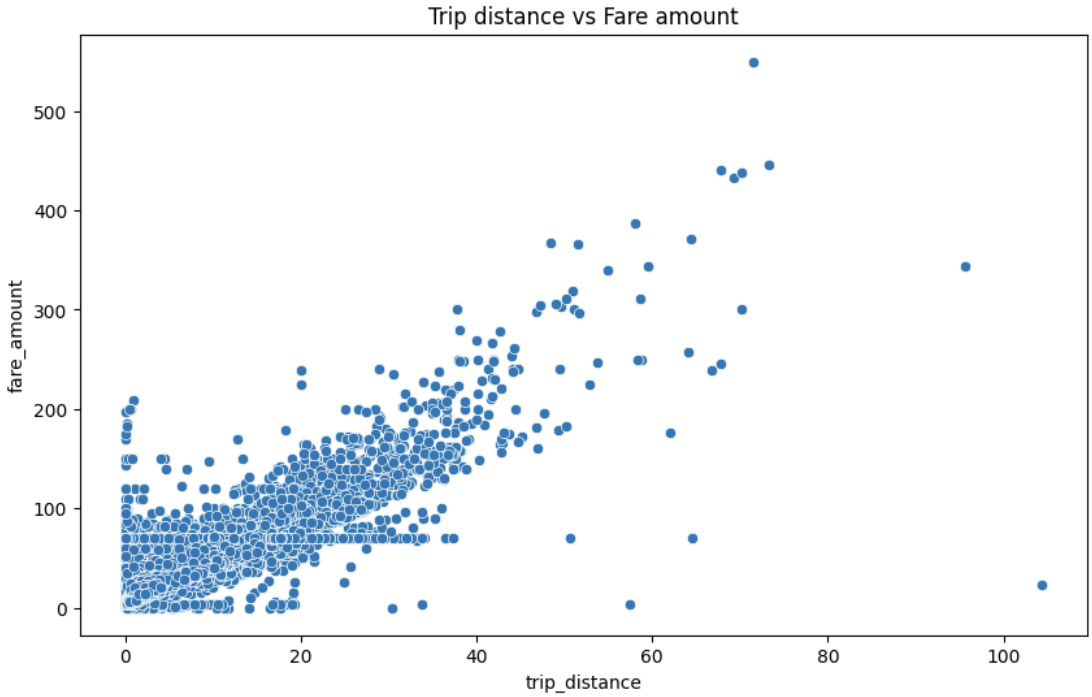
3.1.5. Find the proportion of each quarter's revenue in the yearly revenue



Explanation

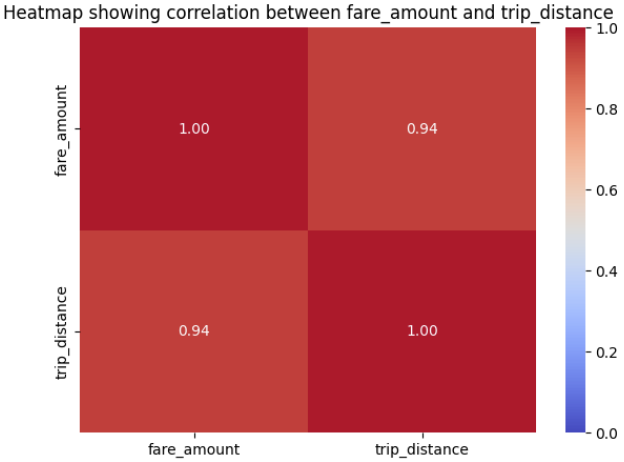
The quaterly revenue chart shows highest revenue earnings in the Q4 followed by Q2. There is a considerable dip in Q3 in comparison with other quarters.

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



Explanation

This scatter plot shows that fare_amount is increasing with trip distance. There are quite a few of fare_amount which are marked zeros, which are also showing up in the graph. Also there are quite a few of trips which are in the range of trip distance less than 40miles. Only a few of trips have higher trip distance of 60miles and only a couple of them around 100miles.



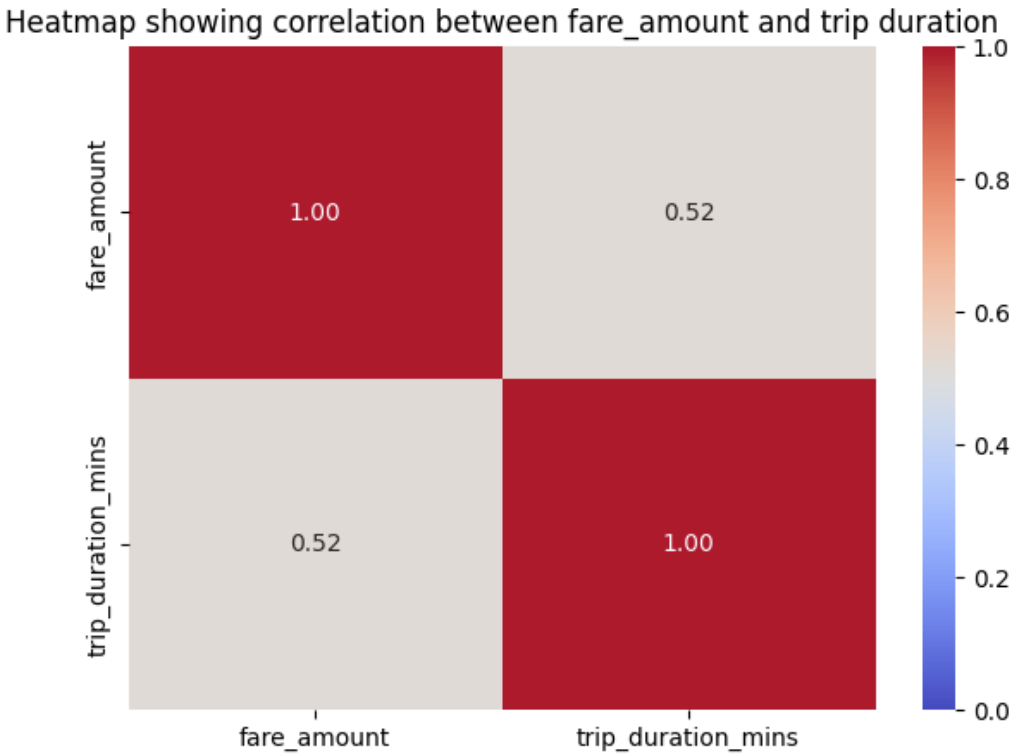
Explanation

The correlation heatmap between fare_amount and trip_distance is highly correlated. It shows that with increase of trip_distance the fare_amount also increases which is quite implicit too.

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

3.1.7.1. Heatmap showing correlation between fare_amount and trip_duration

3.1.7.2.

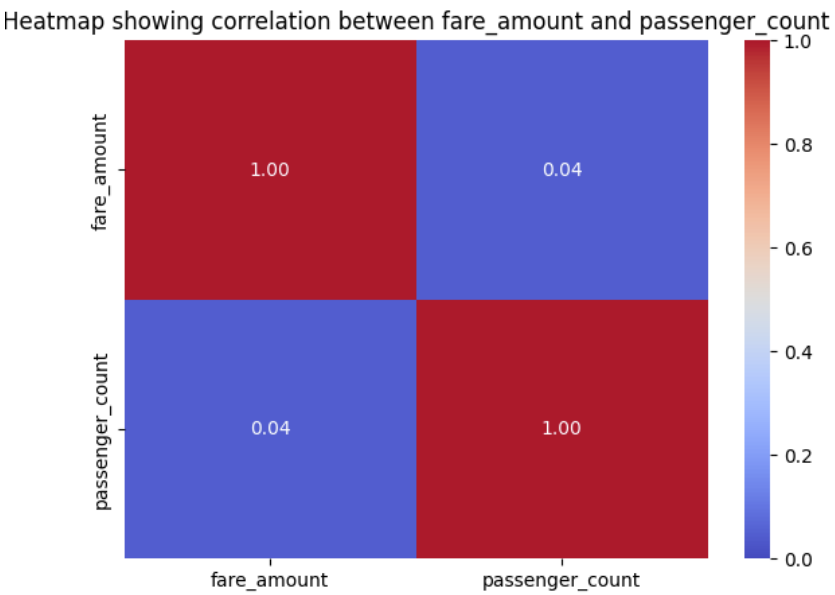



```
<Figure size 640x480 with 0 Axes>
```

Analysis

The above chart shows that the fare_amount and trip_duration are moderately correlated.

3.1.7.3. Heatmap showing correlation between fare_amount and passenger_count

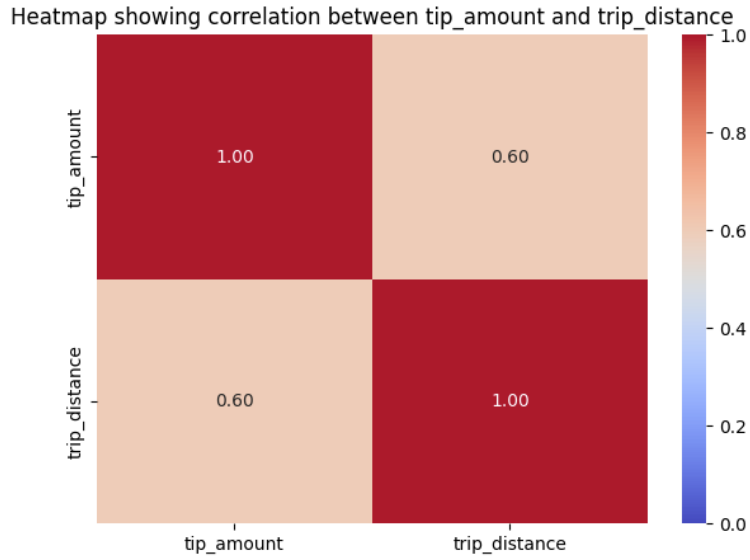


<Figure size 640x480 with 0 Axes>

Analysis

This figure shows that the fare_amount and passenger_count are having less correlation.

3.1.7.4. Heapmap map showing correlation between tip_amount and trip_distance

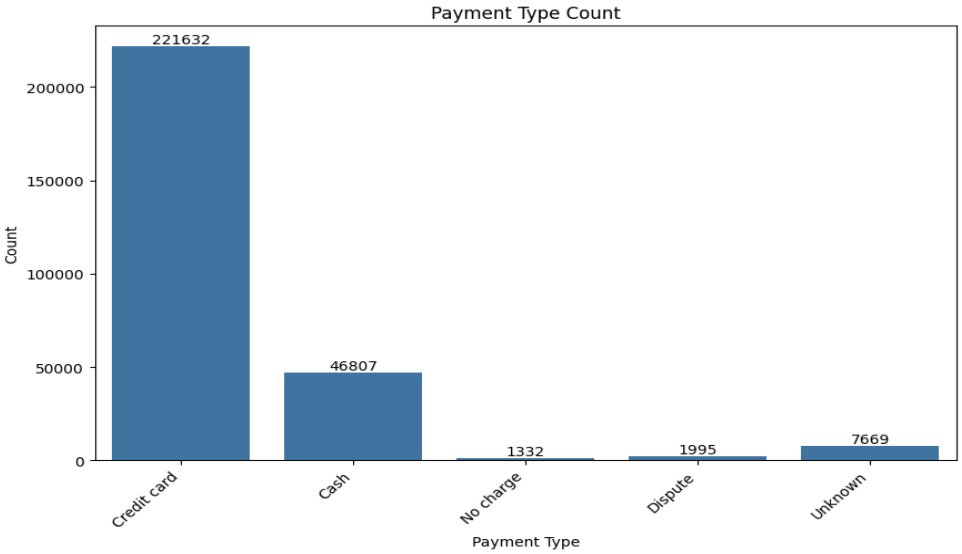


<Figure size 640x480 with 0 Axes>

Analysis

This figure shows that the tip_amount and trip_distance are having moderate correlation.

3.1.8. Analyse the distribution of different payment types



Explanation

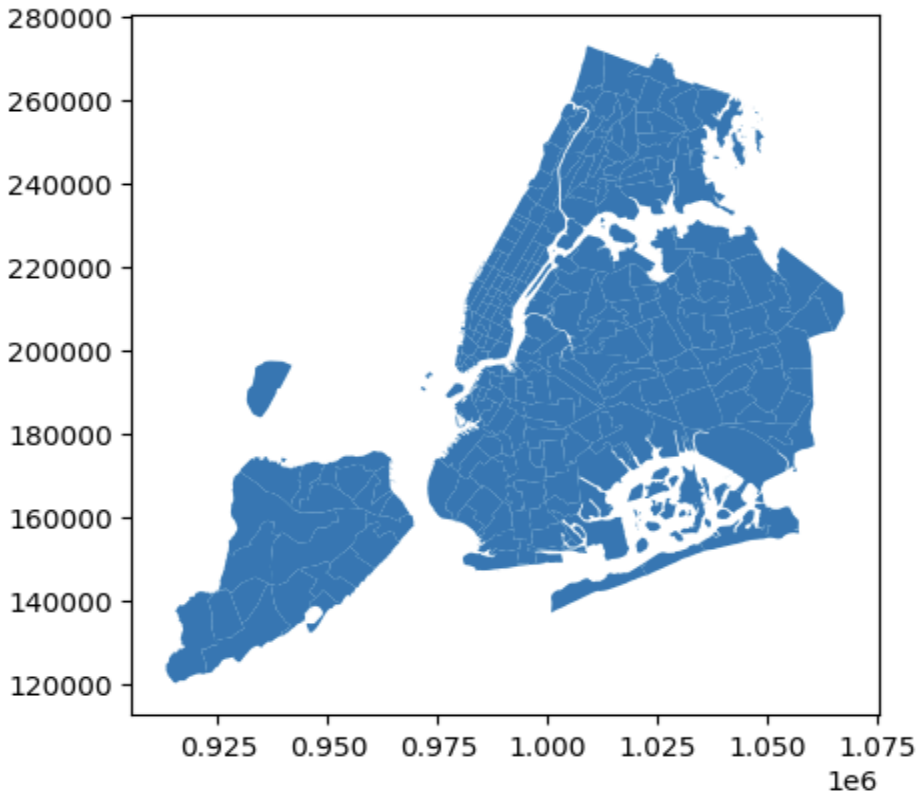
This shows that credit card payment type is used maximum number of times. There are a few cash payment type entries as well. Then the next highest is the unknown type which is also considerable count.

There are few dispute entries present which is for payments which were disputed later.

3.1.9. Load the taxi zones shapefile and display it

```
import geopandas as gpd


# Read the shapefile using geopandas
zones = gpd.read_file('/content/drive/MyDrive/Assignments/EDA/data_NYC_Taxi/taxi_zones/taxi_zones.shp') # read the .shp fil
zones.head()
```

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...



3.1.10. Merge the zone data with trips data

```
# Merge zones and trip records using locationID and PULocationID
df_merged = pd.merge(df_new, zones, left_on='PULocationID', right_on='LocationID', how='left')
df_merged.head()
```

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	PULocationI
0	1	2023-06-25 14:02:50	2023-06-25 14:54:33	1.0	12.60	1.0	0	13
1	2	2023-09-09 11:47:33	2023-09-09 11:55:09	1.0	1.05	1.0	0	23
2	2	2023-10-13 02:54:45	2023-10-13 03:04:02	1.0	1.85	1.0	0	11
3	1	2023-07-10 12:08:14	2023-07-10 12:17:15	1.0	1.80	1.0	0	23
4	2	2023-11-08 15:38:53	2023-11-08 15:56:51	1.0	2.61	1.0	0	18

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

```
df_location_based_trip_count = (
    df_merged.groupby('LocationID')['PULocationID']
             .size()
             .reset_index(name='trip_count')
             .sort_values(by='trip_count', ascending=False)
             .reset_index(drop=True)
)
df_location_based_trips = df_location_based_trip_count.merge(zones, on='LocationID', how='left')
df_location_based_trips.head()
```

	LocationID	trip_count	OBJECTID	Shape_Leng	Shape_Area	zone	borough	geometry
0	132.0	14491	132	0.245479	0.002038	JFK Airport	Queens	MULTIPOLYGON (((1032791.001 181085.006, 103283...
1	237.0	13094	237	0.042213	0.000096	Upper East Side South	Manhattan	POLYGON ((993633.442 216961.016, 993507.232 21...
2	161.0	13052	161	0.035804	0.000072	Midtown Center	Manhattan	POLYGON ((991081.026 214453.698, 990952.644 21...
3	236.0	11951	236	0.044252	0.000103	Upper East Side North	Manhattan	POLYGON ((995940.048 221122.92, 995812.322 220...
4	162.0	9920	162	0.035270	0.000048	Midtown East	Manhattan	POLYGON ((992224.354 214415.293, 992096.999 21...

Explanation

The locationId of 132 (JFK Airport, Queens) shows highest number of pickups, followed by 237 (Upper East Side South, Manhattan) and 161 (Midtown, Manhattan) location.

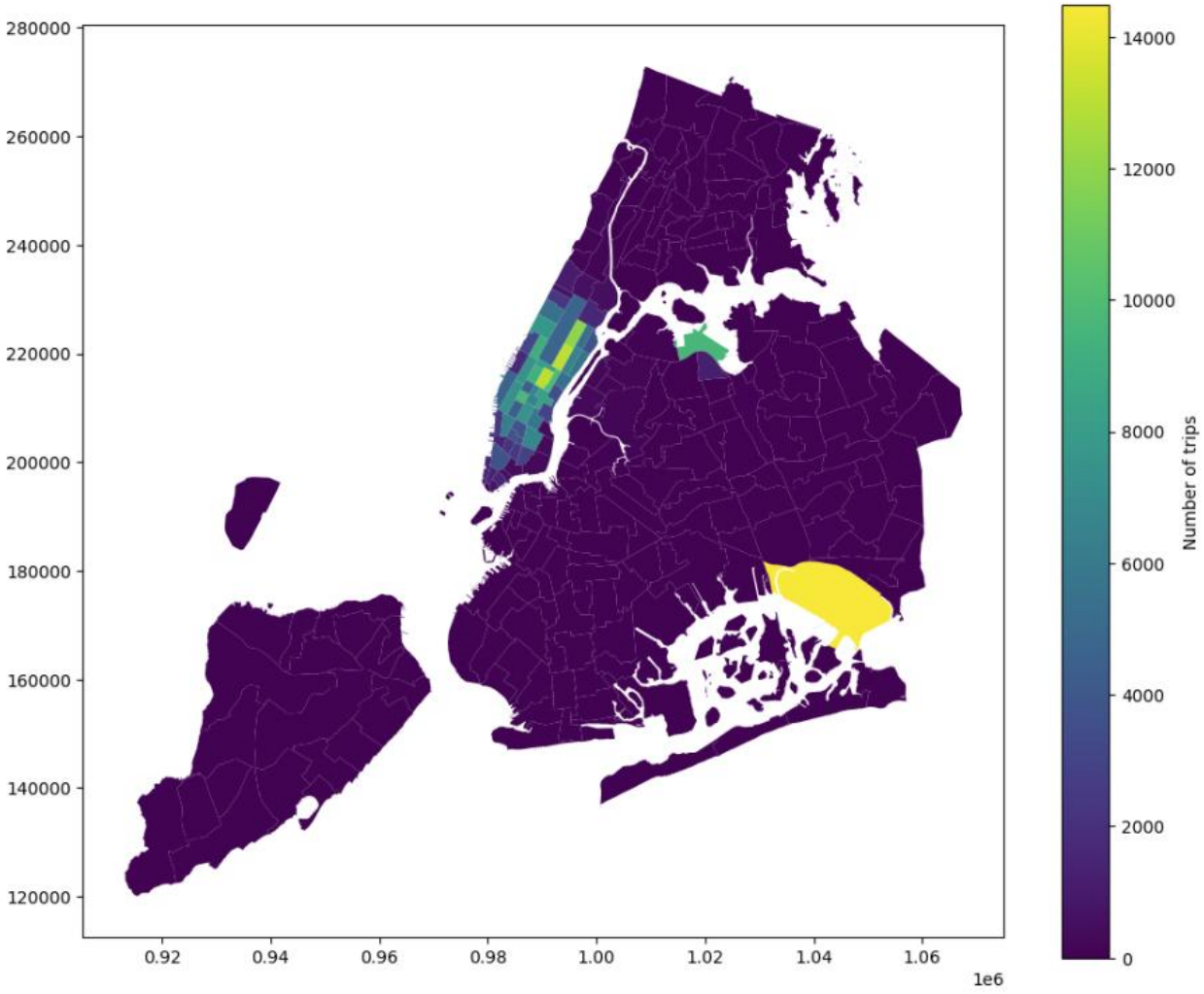
3.1.12. Add the number of trips for each zone to the zones dataframe

```
# Merge trip counts back to the zones GeoDataFrame

# Calculate trip counts per PULocationID from df_merged
trip_counts = df_merged.groupby('LocationID')['PULocationID'].count().rename('trip_count')

# Merge these trip counts back to the zones GeoDataFrame
zones = zones.merge(trip_counts, on='LocationID', how='left')

# Fill any NaN values in 'trip_count' with 0 (for zones with no recorded trips)
zones['trip_count'] = zones['trip_count'].fillna(0).astype(int)

zones.head()
```

	OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	trip_count
0	1	0.116357	0.000782	Newark Airport	1	EWR	POLYGON ((933100.918 192536.086, 933091.011 19...	26
1	2	0.433470	0.004866	Jamaica Bay	2	Queens	MULTIPOLYGON (((1033269.244 172126.008, 103343...	0
2	3	0.084341	0.000314	Allerton/Pelham Gardens	3	Bronx	POLYGON ((1026308.77 256767.698, 1026495.593 2...	1
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...	370
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...	0

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

```
# Define figure and axis
fig, ax = plt.subplots(1, 1, figsize = (12, 10))

# Plot the map and display it
zones.plot(column='trip_count', ax=ax, legend=True, legend_kwds={'label': "Number of trips", 'orientation': "vertical"})
plt.show()
plt.savefig('/content/drive/MyDrive/Assignments/EDA/data_NYC_Taxi/newcharts/zones_with_trips.png')
```



3.1.14. Conclude with results

- **Busiest hours, days and months**
  - Hourly: The busiest hours for taxi pickups are consistently between 4 PM and 7 PM (16:00-19:00) on weekdays, with the absolute peak at 6 PM (18:00). Early morning hours (3 AM - 5 AM) are the quietest.
  - Daily: Weekdays (especially Thursday and Wednesday) exhibit higher demand compared to weekends. Monday typically has the lowest demand.
  - Monthly/Quarterly: October, May, and March are the months with the highest pickup counts and corresponding highest revenues. Q2 (April-June) and Q4 (October-December) are the most lucrative quarters, with slight dips in summer months (July-August) and February.
- **Trends in revenue collected**

  Monthly revenue collection analysis shows that month of May is having highest revenue collected, followed by October. This could be due to holiday seasons. There are low revenue collection months in Feb and January. This could be due to winter season.
- **Trends in quarterly revenue** Quaterly revenue charts and analysis shows that Q4 has highest revenue collection done followed by Q2. The Q3 shows lowest collection in terms of revenue.

- **How fare depends on trip distance, trip duration and passenger counts**
  - Fare collection fare_amount is highly correlated with trip_distance (0.94) and moderately correlated with trip_duration_mins (0.53), as expected.
  - Fare collection shows very little correlation with passenger_count (0.04).
- **How tip amount depends on trip distance** Tipping amount shows higher collection with higher trip distance. There are few higher tip collection in moderate distance as well from the charts.
- **Busiest zones**
  - JFK Airport (LocationID 132) stands out as the zone with the highest number of trips, indicating its critical role as a major hub.
  - Midtown Center (161), Upper East Side South (237), Upper East Side North (236), and Midtown East (162) are consistently high-activity zones, particularly in Manhattan.
  - Pickup/Dropoff Ratios: Zones like East Elmhurst (70) and airports (JFK 132, LaGuardia 138) show high pickup-to-dropoff ratios, indicating they are primary origin points where taxis are picked up and then leave the area.

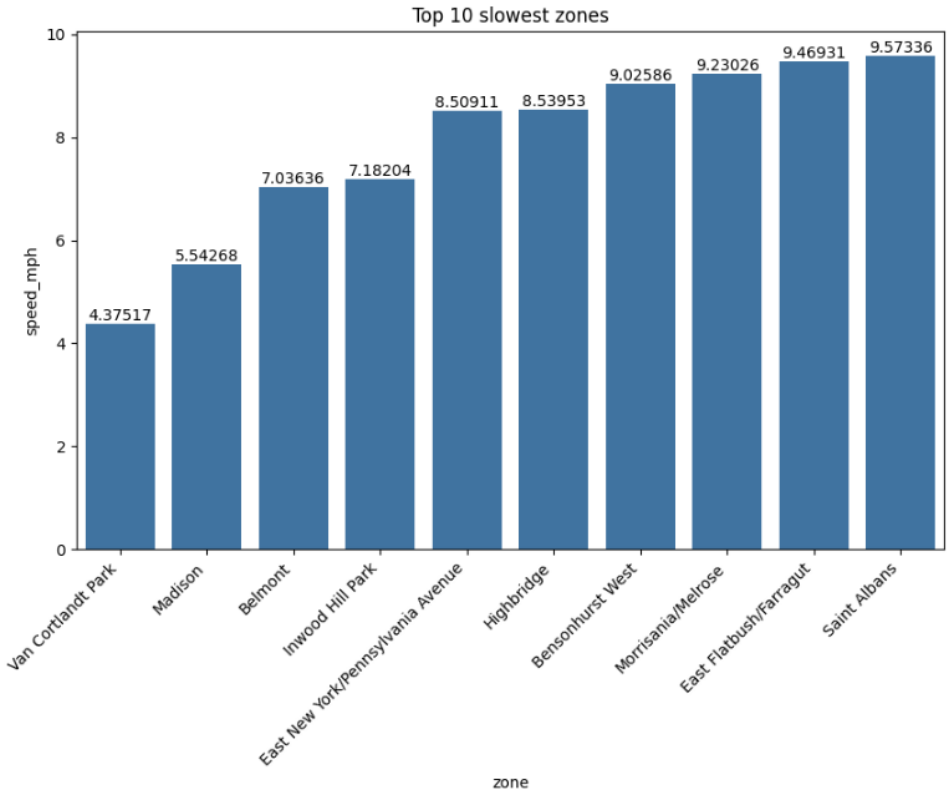
3.2. Detailed EDA: Insights and Strategies

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

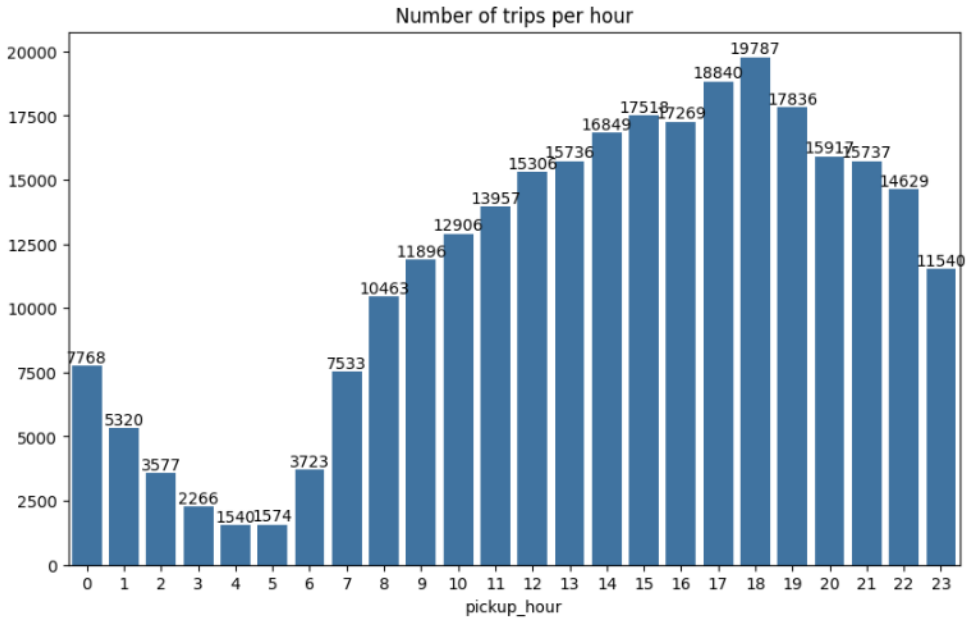
Analysis

By grouping in terms of pickup_hour, it will enable us compare the travel speed between two zones during during different hour of day. This would help us in understanding the congestion situation between the route and help in applying that during peak hours.

This also provide us insight on high traffic routes and suggesting in providing more cabs in that area. From the graph we can see that high demand timing of 1600hr to 1800hr shows relatively lower speed mph. So we need to put more cabs in the area to cover for higher demand.



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



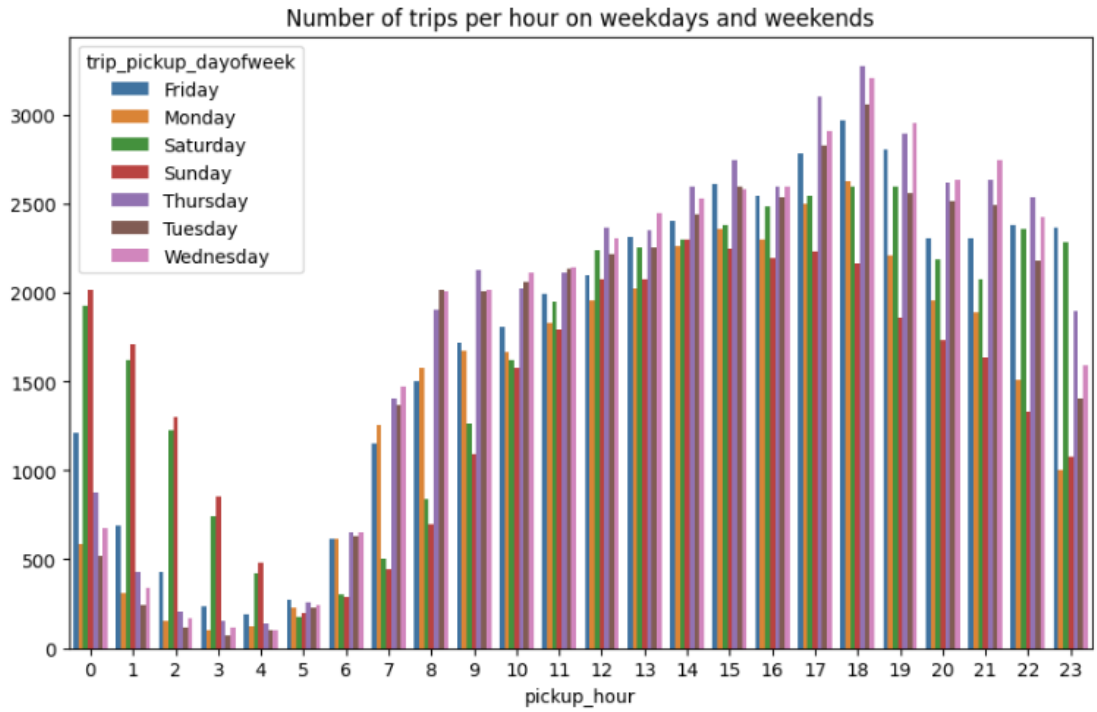
Analysis - The busiest hour of day is coming as 1800 hour which as 19899 trip. The trip counts starts increasing during 1700hr onwards itself and after hitting the peak, goes down from 1900hrs onwards.

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

```
# Scale up the number of trips

# Fill in the value of your sampling fraction and use that to scale up the numbers
# In the sampling done did sample of 5% of sample out of each parquet file. Then from the overall sample
# took 15% sample to get the data in the range of 250k to 300k rows. So overall sample fraction is 0.75%

df_scld_up_by_sampl_frac = df_merged.groupby('pickup_hour').size().reset_index(name='trip_count')
df_scld_up_by_sampl_frac['est_trip_count'] = (df_scld_up_by_sampl_frac['trip_count'] / 0.0075).round().astype(int)
df_scld_up_by_sampl_frac.sort_values(by='trip_count', ascending=False).head()
```

	pickup_hour	trip_count	est_trip_count
18	18	19785	2638000
17	17	18839	2511867
19	19	17835	2378000
15	15	17517	2335600
16	16	17266	2302133

3.2.4. Compare hourly traffic on weekdays and weekends



Explanation

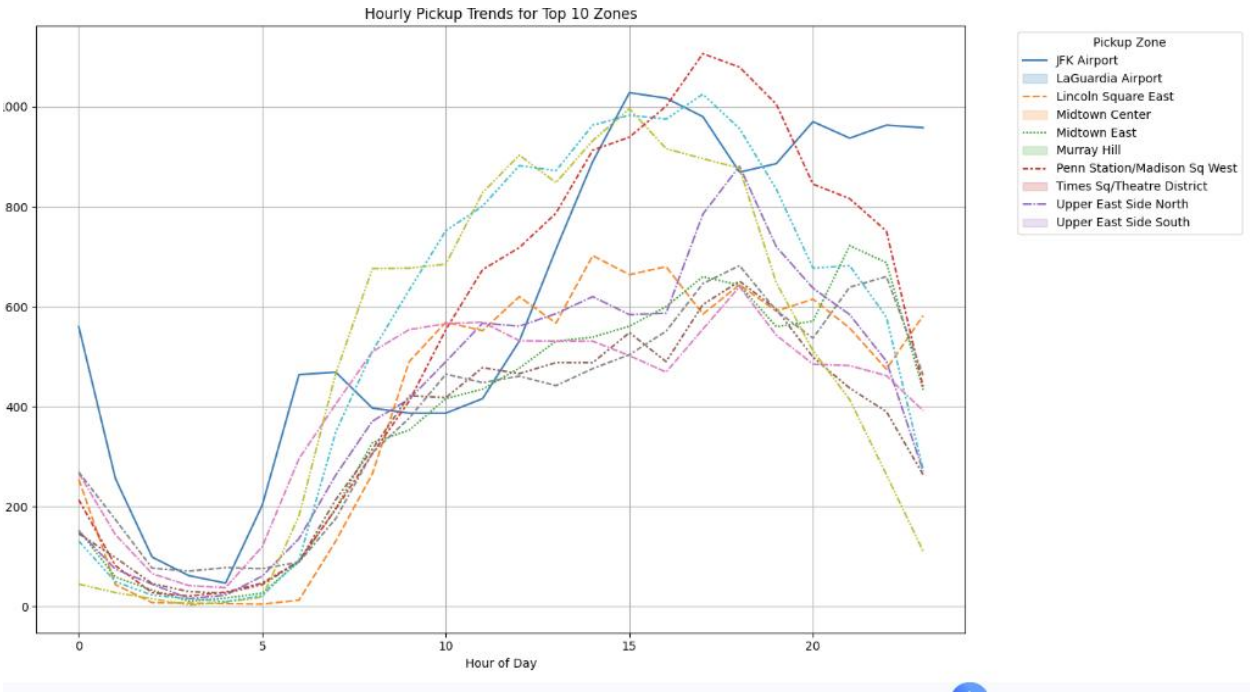
From the above graph, we can determine that:

- Weekdays are having higher traffic during morning hours of day increasing from 0700 (7AM) onwards and goes on till late evening till 2000hr
- Weekends (sat and sun) have higher traffic in early morning from 00hr to 0400hr.
- Also we see higher late night traffic 2200hr on saturdays.
- During weekdays thrusdays and weds have higher traffic than other weekdays.
- The peak traffic hours of 1700hr and 1800hr also shows that wed and thursdays have highest traffic

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

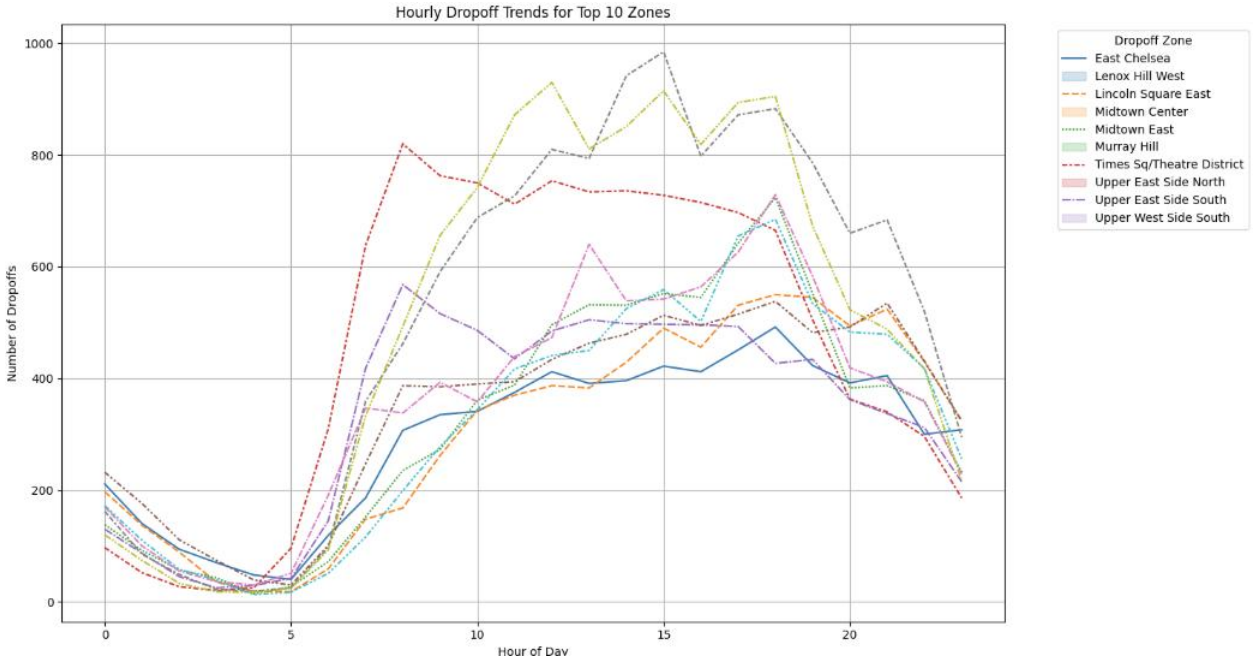
Hourly Pickup Trends :

There's a clear dip in pickups across all top zones during the early morning hours (around 3 AM to 5 AM). Pickups gradually increase from the morning, reaching a significant peak in the evening (typically between 4 PM and 7 PM) for most central Manhattan zones like Midtown Center (161), Upper East Side South (237), and Upper East Side North (236). JFK Airport (Zone 132) shows a consistently high number of pickups throughout the day, with a notable morning peak and then sustained high activity into the evening.



Hourly Dropoff Trends (Figure 2):

Similar to pickups, dropoffs are lowest in the very early morning. Dropoffs also show a build-up during the day, with many zones experiencing their highest dropoff volumes in the late afternoon and early evening, aligning with the end of the workday. Specific zones like Midtown Center (161), JFK Airport (132), and other Manhattan areas maintain high dropoff activity, indicating constant movement of passengers to and from these key locations.



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

This shows the zone 70 which is East Elmhurst, Queens has highest pickups to drop off ratio.

With this top ten location which has higher pickup, we should try to have high taxi availability in this areas.

For the location which has higher dropoff location, we should look for move them to the higher pickup ratios. This is to get quicker turn around of the trip assignments and not loose on trip assignments due to no available taxis in the areas with higher picksup. This would provide option to increase the revenue.

Top 10 zones pickup / dropoff ratio

	zone_id	pickup_count	dropoff_count	pickup_dropoff_ratio	LocationID	zone	borough
69	70.0	1259.0	177	7.112994	70	East Elmhurst	Queens
125	132.0	14491.0	3394	4.269593	132	JFK Airport	Queens
131	138.0	9628.0	3664	2.627729	138	LaGuardia Airport	Queens
179	186.0	9701.0	6094	1.591894	186	Penn Station/Madison Sq West	Manhattan
107	114.0	3781.0	2701	1.399852	114	Greenwich Village South	Manhattan
42	43.0	4794.0	3452	1.388760	43	Central Park	Manhattan
241	249.0	6258.0	4792	1.305927	249	West Village	Manhattan
155	162.0	9920.0	7987	1.242018	162	Midtown East	Manhattan
97	100.0	4560.0	3748	1.216649	100	Garment District	Manhattan
154	161.0	13052.0	11029	1.183426	161	Midtown Center	Manhattan

Bottom 10 zones pickup / dropoff ratio

	zone_id	pickup_count	dropoff_count	pickup_dropoff_ratio	LocationID	zone	borough
148	155.0	1.0	56	0.017857	155	Marine Park/Mill Basin	Brooklyn
249	257.0	2.0	106	0.018868	257	Windsor Terrace	Brooklyn
164	171.0	1.0	47	0.021277	171	Murray Hill-Queens	Queens
244	252.0	1.0	41	0.024390	252	Whitestone	Queens
227	235.0	1.0	37	0.027027	235	University Heights/Morris Heights	Bronx
90	92.0	5.0	175	0.028571	92	Flushing	Queens
112	119.0	1.0	35	0.028571	119	Highbridge	Bronx
200	208.0	1.0	34	0.029412	208	Schuylerville/Edgewater Park	Bronx
124	131.0	2.0	67	0.029851	131	Jamaica Estates	Queens
28	29.0	1.0	33	0.030303	29	Brighton Beach	Brooklyn

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

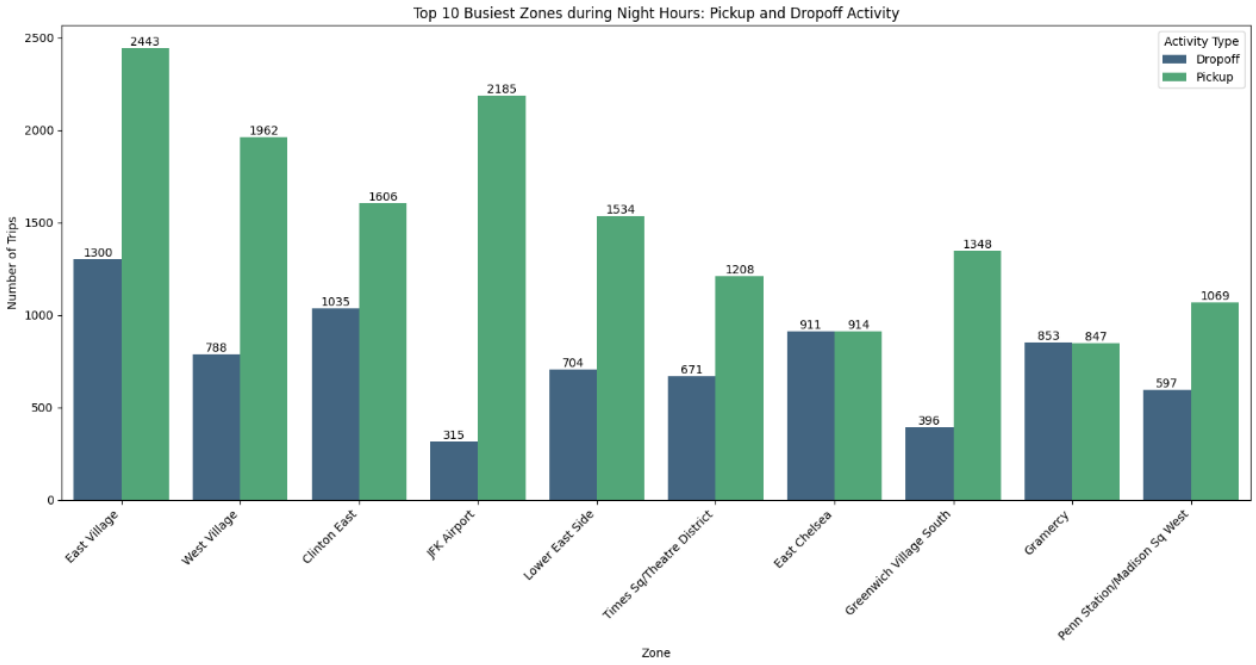
Explanation This chart visualizes the top 10 pickup and dropoff zones during night hours (11PM to 5AM)

Key observations

East Village (LocationID 79, Manhattan) clearly stands out as the most active zone during night hours, leading both pickup and dropoff counts. This highlights its significant role as a nightlife destination and a residential hub with active night-time taxi usage.

JFK Airport (LocationID 132, Queens) remains a crucial pickup point, securing the second spot, demonstrating consistent demand for airport departures even at night.

Other Manhattan neighborhoods like **West Village (LocationID 249)**, **Clinton East (LocationID 48)**, and** Lower East Side (LocationID 148)** show high activity for both pickups and dropoffs, reinforcing the strong night-time demand within central and downtown Manhattan.

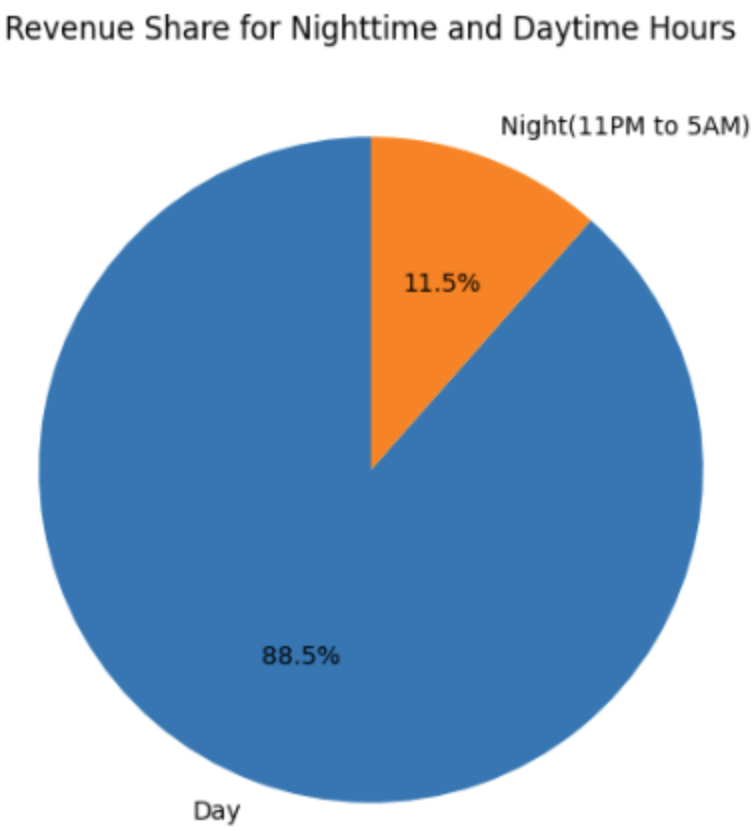


3.2.8. Find the revenue share for nighttime and daytime hours

Explanation:

Day time traffic revenue is 88.5% and night time traffic revenue is very low with 11.5%.

Using this data, we can reduce the taxi by giving instructions to reduce the number of taxis. This would help drivers in maximising their earning as well as customers would be less in the night hours. So they can focus on starting the run during early morning time after 5AM.



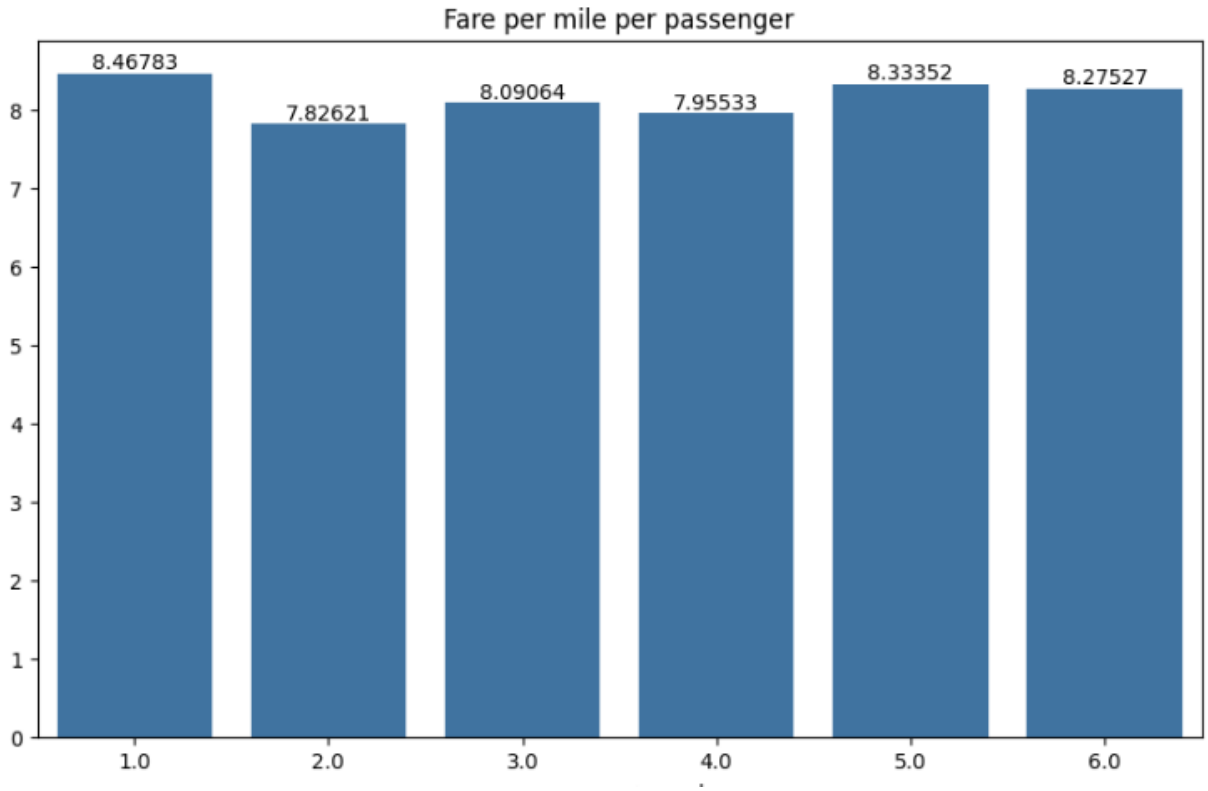
3.2.9. For the different passenger counts, find the average fare per mile per passenger

Explanation

The above chart shows the average fare per mile per passenger.

For passenger count 1, the chart shows that there's a hight fare per passenger. This could be due to too many short trips by single passenger. Other high fare / mile data is seen for passenger count of 5 and 6. These could be higher distance trips which from previous analysis has higher fare amount.

The fare / mile data is mile data is lowest for passenger count 2 followed by passenger count 4.



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

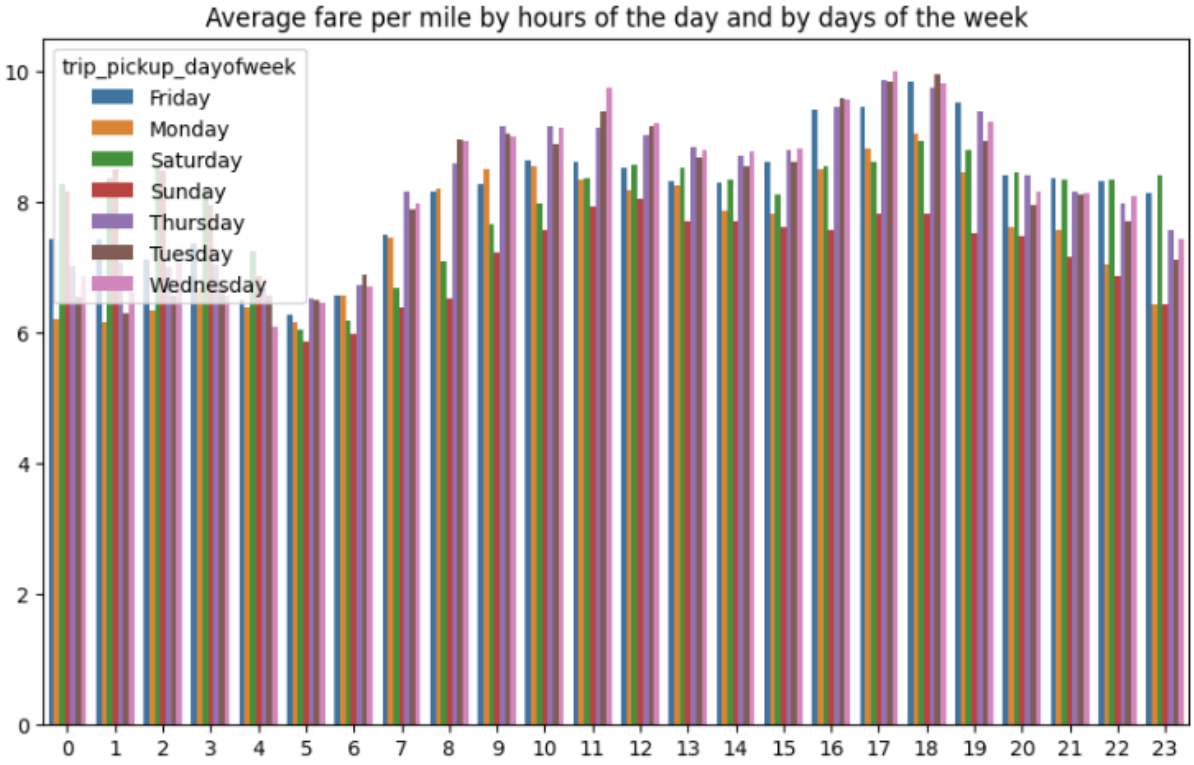
Explanation:

The above chart shows fare per mile by hour and days of week.

The chart relates to the other chart we saw for day of week with respect to taxi count. The week day Wednesdays, Thursdays, Tuesdays and Fridays have highest fare / mile data and that is highest around 1600hrs (4PM) to 1800hr (6PM) in the evenings with highest values hitting around $9/mile around 1700hrs (5PM).

Similar higher fare/mile data is seen in morning hours of weekdays around 0900hrs(9AM) to 1200hr (12PM) during weekdays.

For weekends, that is, Saturdays and Sundays, we see higher fare / mile data from 2100hr to early morning 3:00AM.



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

Explanation

Above plots show average fare per mile for different vendors (Hourly)

This bar plot compares the average fare per mile charged by two vendors, 'Creative Mobile Technologies, LLC' (Vendor 1) and 'VeriFone Inc.' (Vendor 2), for each hour of the day.

Key Observations:

Vendor 1 (Creative Mobile Technologies, LLC): Generally appears to have a slightly higher average fare per mile across most hours compared to Vendor 2. Their peak average fare per mile often reaches above $9.0/mile, particularly in the mid-morning and late afternoon/early evening.

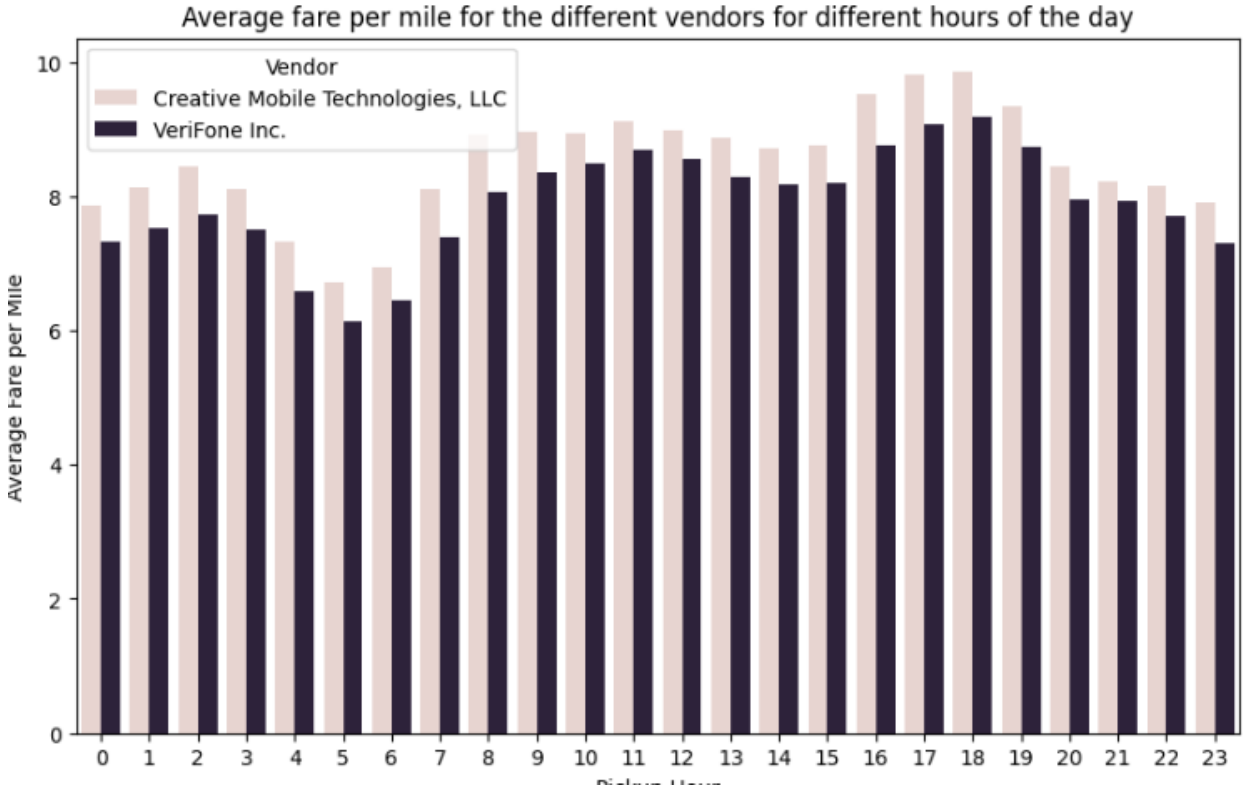
Vendor 2 (VeriFone Inc.): Consistently shows a slightly lower average fare per mile than Vendor 1. Their fares also follow a similar diurnal pattern but at a slightly reduced rate.

Hourly Trends: Both vendors exhibit similar trends throughout the day: Early Morning (0-5 AM): The average fare per mile is relatively moderate, but can fluctuate. There are times when it drops slightly before picking up.

Morning Rush (6-10 AM): Fares generally increase, likely due to higher demand during commuting hours. Midday/Afternoon (11 AM - 4 PM): Fares remain high, with some fluctuations.

Evening Peak (5 PM - 8 PM): Both vendors often show their highest average fares per mile during these hours, correlating with evening rush hour and social activities.

Late Night (9 PM - 11 PM): Fares tend to gradually decrease. Consistency: Despite the absolute difference in average fare per mile, the pattern of fare changes across the hours is quite consistent between the two vendors. This suggests that both vendors are reacting to similar demand and supply dynamics throughout the day, but with different base pricing strategies or operational costs reflected in their per-mile rates.



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

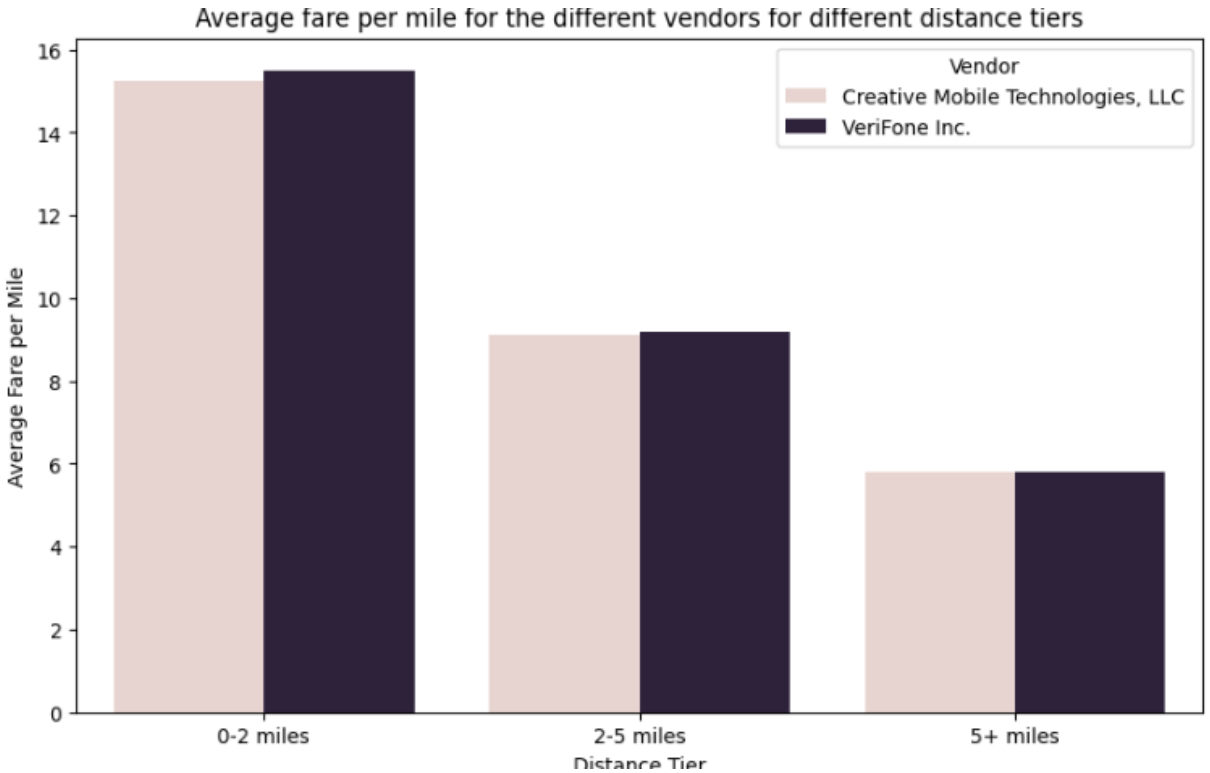
Explanation

The above chart shows average fare per mile for both the vendors for three different miles tiers of 0-2 miles, 2-5miles, and 5+ miles.

Key observations

- For 0-2 miles tier, the fare / mile is highests across vendors where both of them has around $15 / mile with Creative Mobile Technologies has slight higher value than Verifone.
- For 2-5 miles tier, the fare / mile data is almost same value of $9/mile for both the vendors.
- For 5+ miles tier, the fare / mile is lowest almost hitting around $6/mile.

This shows that 0-2 mile tier is much more profitable than other two tiers of 2-5miles and 5+miles. So both vendors should have taxis available for locations which has higher trips in the 0-2 miles travel.



3.2.13. Analyse the tip percentages

Explanation

Above charts are a subplot display of tip percentage for trip_distance, passenger_count and pickup_hours.

Key Observations

1. Tip Percentage by Trip Distance:

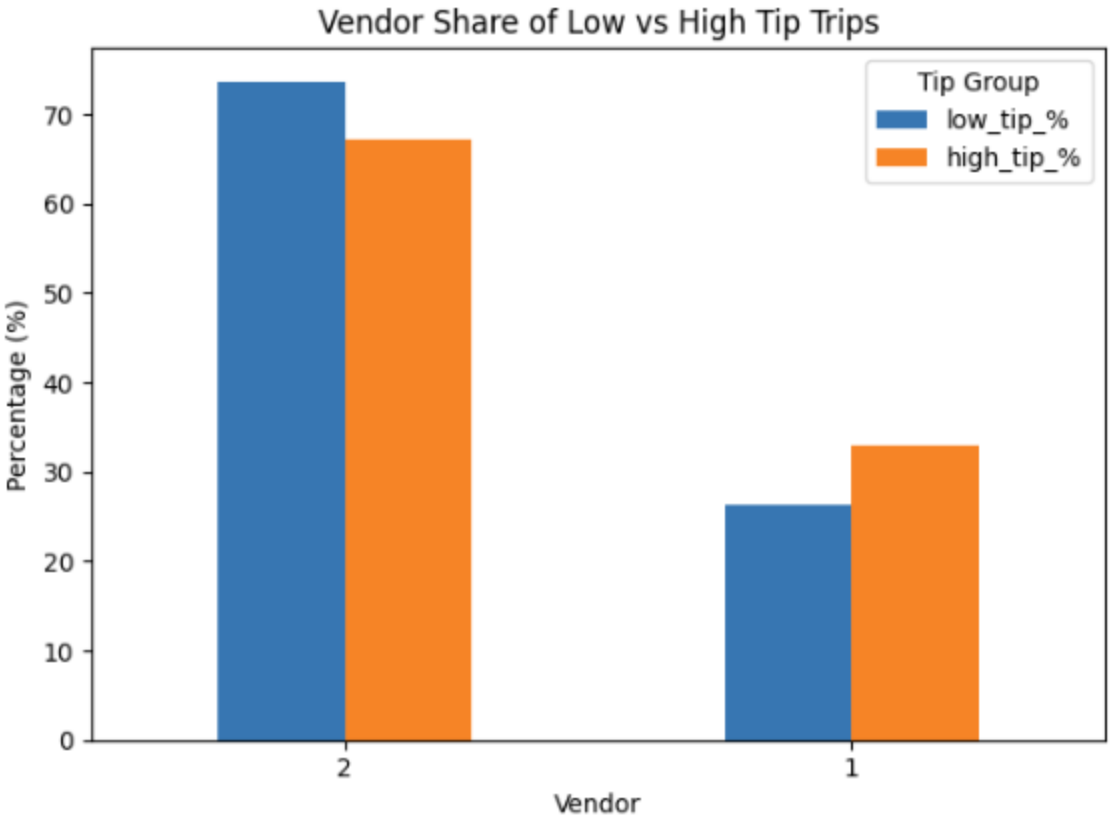
   The data shows some fluctuation in tip percentages for very short distances (e.g., 0.093 for 0.0 miles, 0.08 for 0.01 miles). It's hard to draw definitive conclusions from these initial few rows alone, as tiny distances might represent specific scenarios (e.g., waiting time, cancellations with minimal movement). However, generally, for longer trips, we would expect a relatively stable or slightly increasing tip percentage as the fare amount is higher. Further analysis across the full range of trip distances would be needed to see if there's a clear trend (e.g., if longer trips consistently yield higher tip percentages).

2. Tip Percentage by Passenger Count:

   Single passengers (1.0) tend to give the highest average tip percentage at 12.64%. As the passenger count increases, the tip percentage generally sees a slight decrease, with 2.0 passengers at 12.12%, 3.0 at 11.66%, and 4.0 at 10.68%. Interestingly, for 5.0 passengers, the tip percentage slightly rises again to 12.52%, indicating that larger groups might also be inclined to tip more generously or perhaps these trips involve specific scenarios.

3. Tip Percentage by Pickup Hour:

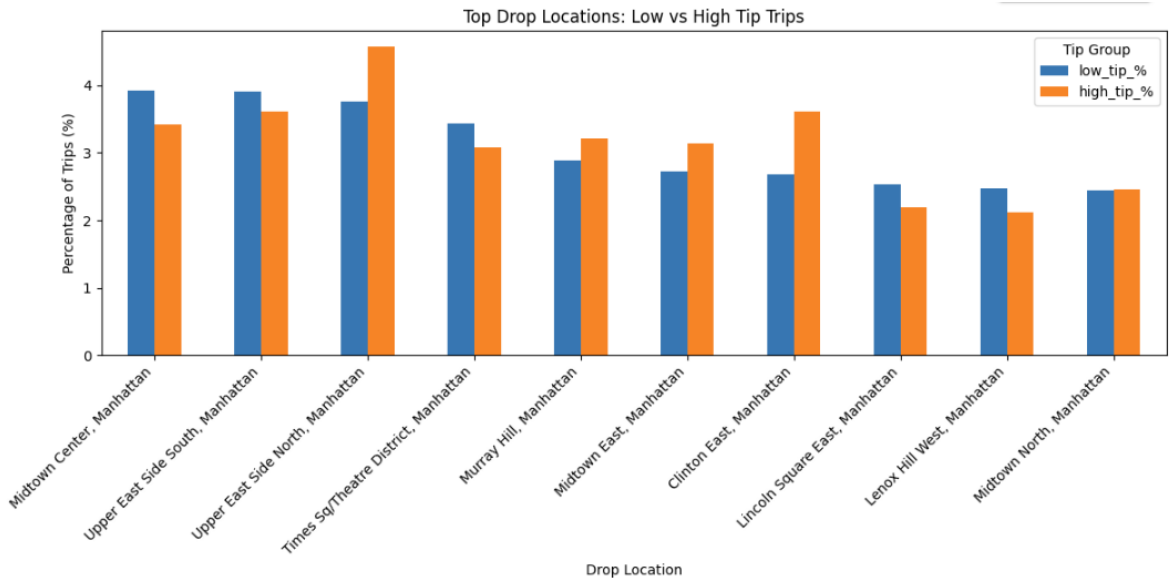
   The early morning hours (0, 1, 2 AM) show relatively high and consistent tip percentages (around 11.98% to 12.16%). This could be due to late-night trips often being longer, or passengers being more inclined to tip after social outings, or perhaps less traffic leading to better service perception. As the morning progresses, the tip percentage seems to slightly dip around 3 AM and 4 AM (11.29% and 11.36% respectively). A full hourly breakdown would be necessary to see the full diurnal pattern, but these initial values suggest that late-night/early-morning trips are quite good for tipping**.**



Vendor 2 has higher tip percentage in comparison to vendor 1 in both the categories. So clearly, vendor 2 is taking some good measures for customer satisfactions which can be taken up by Vendor 1 to improve driver earning in terms of tip amount.

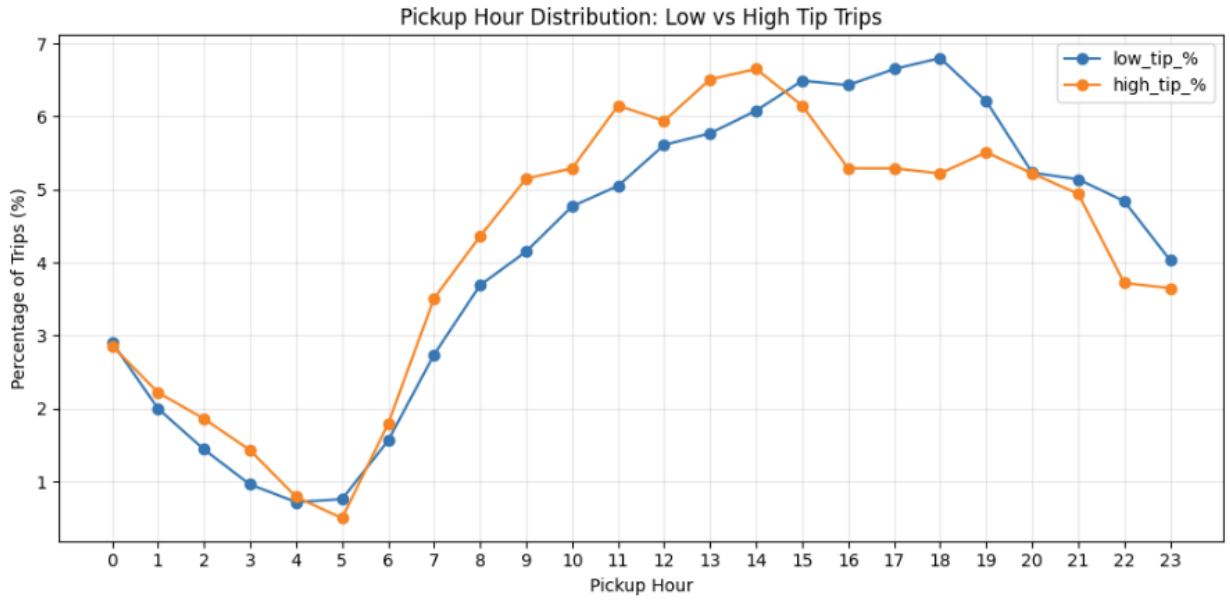
For vendor 2, tip percentage of 10% is slightly higher than 70% in comparison to the 25% tip amount which is falling aroung 65%.

For Vendor 1, the 25% tipping amount is higher than 10% tipping percentage



Above chart shows the tipping percentage variations for top 10 drop locations.

- Upper East Side North, Manhattan - shows higher percentage tipping behavior
- Midtown Center, Manhattan', 'Upper East Side South, Manhattan' and Upper East Side North, Manhattan location has hight value of lower tipping percentage (10%)
- 'Lincoln Square East, Manhattan', 'Lenox Hill West, Manhattan' shows lower tipping behavior in both higher and lower tipping percentage category.



Explanation

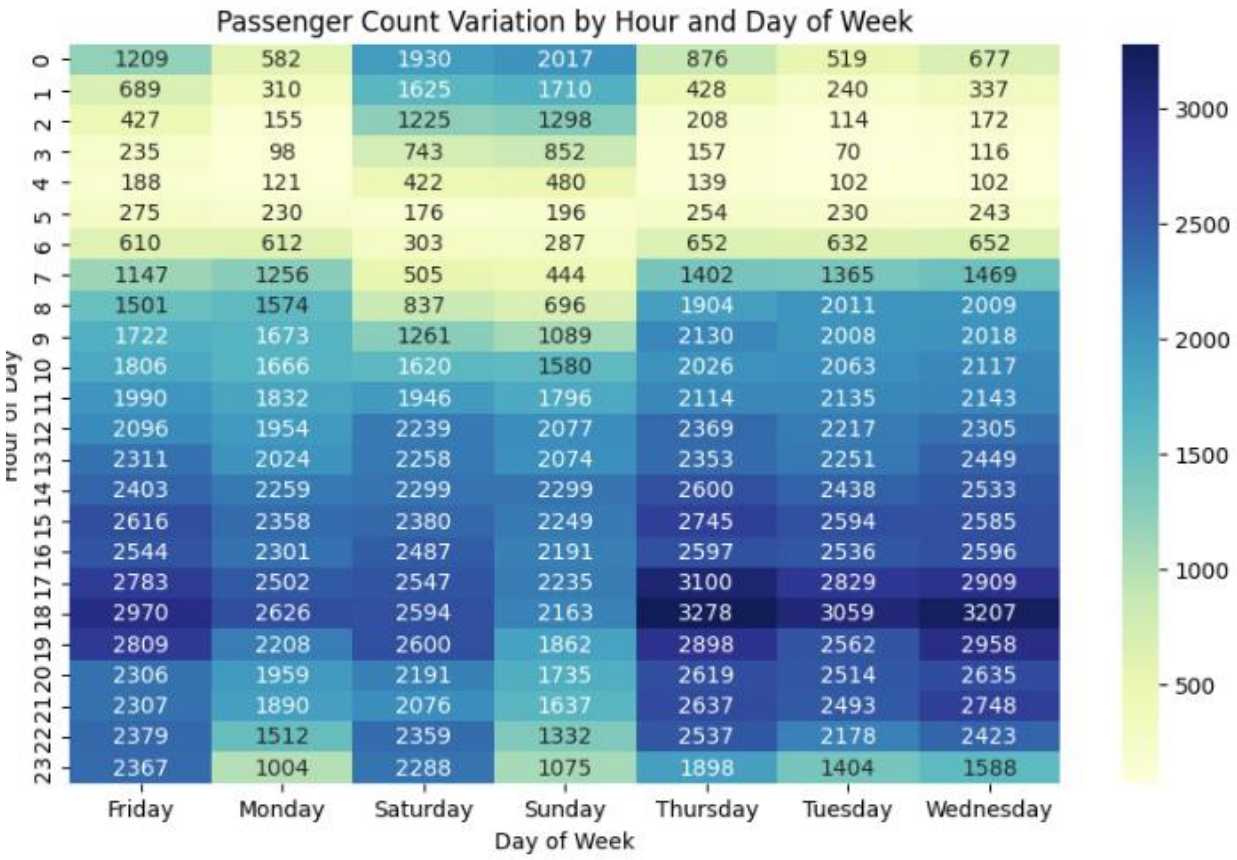
The above chart provides comparison of lower (10%) and higher(25%) tipping percentage in terms of pickup hours.

Key observations

Chart shows slight higher tipping behavior for early morning trips from 1am to 4am, dipping sharply for 5am trips. There's a sligh overlapping between the two tipping percentages for trip between 5am to 6am, but then again from 6am till 2pm shows higher tipping percentage of 25% behavior shown than lower tipping percentage.

For later part of day, from 3pm onwards, the tipping behavior shows lower tipping percentage more prevalent till late night of 2300hrs.

3.2.14. Analyse the trends in passenger count



Explanation

Above chart is a heatmap depicting passenger count for weekdays on hourly basis.

Dark blue color shows higher passenger count mentioning count of passenger overall and ligher color similarly shows lesser count of passernger during that time of day on that day of week.

Key observation

- This chart clearly shows highest passenger count 3278 on thursdays at 1800hrs, which corresponds to the pervious analysis as well. This follows by slightly lower number of passenger on Wednesdays and then tuesdays. Next in line in terms of passenger count is seen on Fridays as well.

- Thursdays have generally higher passenge count beginning from 7am going till late night. Similar slightly lower numbers seen on Wednesdays and Tuesdays well.
    - Weekends, that is Sat and Sunday shows higher early morning traffic 00hr (12AM) to 0200hr (2AM).

3.2.15. Analyse the variation of passenger counts across zones

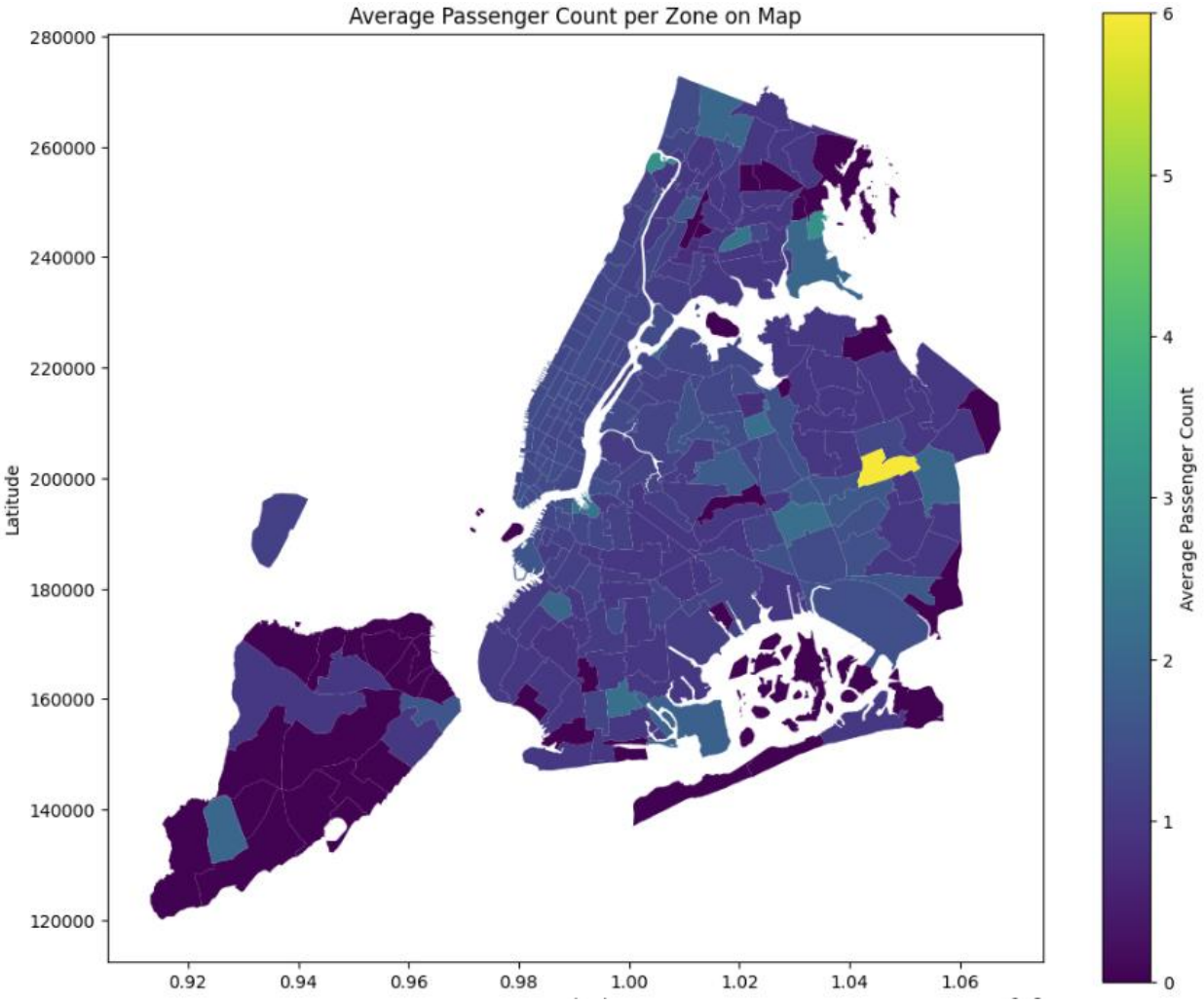
	PULocationID	avg_psngr_count
0	1	1.500000
1	3	2.000000
2	4	1.360656
3	6	1.600000
4	7	1.239726

Explanation

Key Observations:

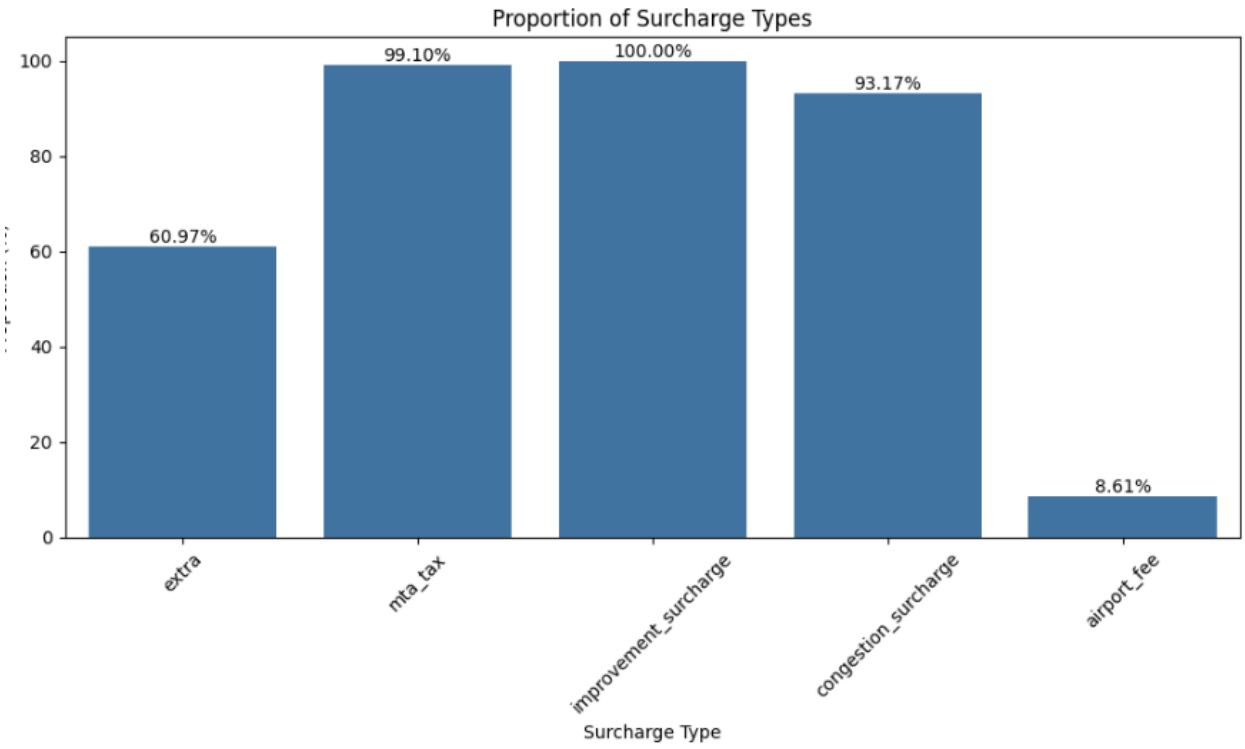
- **High Average Passenger Counts**: Some areas in Queens (visible as brighter green/yellow zones) appear to have relatively higher average passenger counts (approaching 3.0-3.5 passengers per trip). This could include areas around airports or major transportation hubs where groups might travel together.
- **Moderate Average Passenger Counts**: Many zones, particularly across Manhattan, show moderate average passenger counts (around 1.5 to 2.5 passengers per trip), represented by shades of blue and green. This is consistent with a mix of single travelers and small groups.
- **Low Average Passenger Counts**: Several zones, especially in Staten Island and some peripheral areas of other boroughs, display very low average passenger counts (below 1.0 or close to 0), indicated by the darker purple shades. These might be less populated areas or zones with fewer taxi pickups overall.
- **Overall Distribution**: While single-passenger trips are the most common overall, this map highlights the geographical variations where larger groups are more frequently picked up. These areas with higher average passenger counts might indicate locations like tourist

attractions, event venues, or specific residential areas where group travel is more common.



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

	Surcharge Type	Sum	Proportion
0	extra	170364	60.965782
1	mta_tax	276924	99.098919
2	improvement_surcharge	279431	99.996064
3	congestion_surcharge	260355	93.169602
4	airport_fee	24048	8.605721



Key Observations and Explanations:

- **Improvement Surcharge (99.996%):** This surcharge was applied to 99.996% of all trips. The fact that it's applied to all trips indicates it's a universal, mandatory fee for every ride.
- **MTA Tax (99.10%):** The MTA tax was applied to 99.10% of trips. Its near-universal application suggests it's a standard tax for almost all taxi rides, with a very small fraction of trips perhaps being exempt due to specific conditions or data anomalies.
- **Congestion Surcharge (93.17%):** This surcharge was applied in 93.17% of trips. The high proportion indicates that a vast majority of taxi trips in NYC occur within, or pass through, the designated congestion pricing zone, making this a very common additional cost for passengers.
- **Extra (60.97%):** The 'extra' charge was applied in 60.97% of trips. This category includes "Miscellaneous extras and surcharges. Currently, this only includes the 0.50 and 1 USD rush hour and overnight charges." The fact that it's applied to over 60% of trips suggests that rush hour and/or overnight travel are very common, significantly impacting the total fare for a large portion of passengers.
- **Airport Fee (8.61%):** The airport fee was applied in 8.61% of trips. The relatively lower percentage is expected, as not all taxi trips involve pickups from these specific airports.

4. Conclusions

4.1. Final Insights and Recommendations

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

- **Dynamic Fleet Positioning for Peak Demand:**
    - **Weekdays (4 PM - 7 PM):** Deploy a higher concentration of taxis to Manhattan's Midtown, Upper East Side, and Financial District areas. These hours and locations consistently show the highest pickup volumes.
    - **Weekends (Late Night/Early Morning, 11 PM - 2 AM):** Shift resources to entertainment districts and airport zones, where demand remains high even after traditional weekday peak hours.
    - **Airports (JFK, LaGuardia):** Maintain a steady and significant supply of taxis, especially during morning and evening rush hours, and throughout the night, given their consistently high pickup numbers and high pickup-to-dropoff ratios. Consider a dedicated airport fleet or incentivized airport runs.
- **Bottleneck Identification and Alternative Route Guidance:**
    - Utilize the "speed per minute" analysis for specific `PULocationID` to `DOLocationID` routes across different hours. Identify routes that are consistently slow during peak times.
    - Integrate this data into real-time dispatching systems to suggest alternative, faster routes to drivers, especially for trips originating or ending in congested areas. This improves efficiency and passenger satisfaction.
- **Optimize Off-Peak Operations:**
    - **Early Mornings (3 AM - 5 AM):** During these quietest hours, consider reduced operational staff, schedule vehicle maintenance, or incentivize drivers to be available in zones with slightly higher, albeit still low, demand (e.g., airports or 24/7 commercial areas).
    - **Load Balancing:** Implement dispatch algorithms that not only prioritize demand but also minimize deadheading (empty travel) by guiding drivers towards upcoming high-demand zones after dropping off passengers.
- **Day-of-Week Specific Strategies:**
    - **Thursday & Wednesday:** Recognize these as the busiest weekdays and ensure maximum driver availability.

- **Monday:** Implement targeted promotions or incentives to encourage more rides or driver presence, as it's the quietest weekday.
- **Weekend Shift:** Adapt staffing and positioning strategies to cater to leisure-focused travel patterns, which might include later peak hours and different popular zones compared to weekdays.

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.

1. **Hotspot Mapping & Predictive Positioning:**
   - Based on generated heat maps for pickup and drop off zones, we can predict higher traffic in zone. This map should highlight zones like JFK Airport (132), Midtown Center (161), Upper East Side South (237), Upper East Side North (236), and Midtown East (162).
   - From the montly and hourly graphs generated we can forecast demand surges in specific zones for the upcoming hour or several hours, especially factoring in month-specific trends (e.g., higher demand in October for certain areas).
   - Strategically position available taxis close to these predicted hotspots, even before a booking request comes in, to reduce pickup times and increase efficiency.
2. **Addressing Pickup/Dropoff Imbalances:**
   - For zones with very high pickup-to-dropoff ratios (e.g., East Elmhurst, JFK, East Village), ensure a continuous flow of taxis into these zones. After dropping passengers off in a high pickup-ratio zone, drivers should be immediately prompted to seek new passengers within that same zone or nearby high-demand zones.
   - For zones with low pickup-to-dropoff ratios (more dropoffs than pickups), incentivize drivers to quickly move out of these areas to more promising pickup locations to minimize idle time.
3. **Tiered Zone Prioritization:**
   - Categorize zones by their average revenue potential, demand consistency, and congestion levels.
   - Prioritize dispatching to high-value zones first, especially during peak hours.
   - Implement smart dispatching that considers the efficiency of the overall network rather than just individual trip profitability, optimizing for maximum total trips and revenue.
4. **Special Event and Seasonal Adjustments:**

- Integrate data on major events (concerts, sports, festivals) and seasonal tourist flows (e.g., more tourist traffic in spring/fall months like May and October) to anticipate localized demand spikes.
- Pre-position additional vehicles and drivers near event venues or tourist attractions accordingly.

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

1. **Dynamic Pricing based on Demand, Time, and Location:**
   - Peak Hour Surcharges: Automatically apply higher fare multipliers during peak hours (weekday 4 PM - 7 PM, weekend late nights) and in high-demand zones (Midtown, Upper East Side, Airports). This maximizes revenue when demand is inelastic.
   - Off-Peak Incentives: Experiment with minor discounts or promotions during off-peak hours (e.g., early mornings, Monday afternoons) to stimulate demand and utilize idle capacity, ensuring these don't cannibalize peak revenue.
   - Congestion-Based Pricing: Leverage insights from slow routes to apply targeted congestion surcharges for trips through known bottleneck areas during peak congestion times, reflecting the increased operational cost and time.
2. **Tiered Distance Pricing Refinement:**
   - The current structure already shows higher per-mile fares for shorter trips (0-2 miles). Maintain this strategy, as it captures value from quick, high-turnover rides.
   - Review the transition points for distance tiers (e.g., 2-5 miles vs. 5+ miles). Ensure the fare drop-off for longer distances is competitive and reflects operational costs.
   - Consider implementing a progressive fare per mile, where the rate slightly decreases beyond a certain distance to remain attractive for longer journeys.
3. **Competitive Vendor Analysis & Differentiation:**
   - Monitor Competitors: Continuously track Vendor 1 (Creative Mobile) and Vendor 2 (VeriFone) pricing, especially their average fare per mile. If Vendor 1 can charge slightly more and attract high-tip customers, it suggests opportunities for premium service offerings.
   - Value Proposition: If the company aims for market share, position itself with slightly lower, more competitive base rates (like Vendor 2). If targeting higher-value segments, focus on service quality (cleaner cars, better drivers) to justify a premium similar to Vendor 1.

- Transparent Surcharges: While congestion and improvement surcharges are common, ensure transparency in their application to maintain customer trust.

4. **Tipping Optimization Strategies:**
   - In-App Tipping Prompts: Implement smart in-app tipping suggestions based on trip distance, time, and service quality ratings to encourage higher tips, particularly during identified low-tipping periods (late morning/afternoon).
   - Driver Incentives: Offer bonus structures for drivers who maintain high customer satisfaction ratings, potentially leading to higher tips.
   - Targeted Service Improvements: Invest in driver training or vehicle upgrades for operations in zones or at times identified with lower tip percentages to improve the perceived value of the service.
5. **Passenger Count & Fare Structure:**
   - Given the low correlation between fare_amount and passenger_count, consider flat-rate charges for additional passengers beyond a certain number, or explore shared-ride options with dynamic pricing based on passenger count for efficiency.
   - The higher per-mile per-passenger fare for single passengers and groups of 5 suggests there's already some implicit optimization, but direct strategies could be more overt.