New York City Yellow Taxi EDA Report

NYC Yellow Taxi Trip Records - 2023

Assignment ID: EDA/02

Name: Shubham Modi

📊 Objective

To uncover insights from 2023 NYC Yellow Taxi trip data that can:

- Optimize taxi operations
- Improve service efficiency
- Maximize revenue
- Enhance passenger experience

This report follows the structure and expectations outlined in the assignment brief (*Assignment_1.pdf*).

It includes assumptions, methodology, visual insights, and business-focused recommendations.

🔧 1. Data Preparation

Data Sampling: Sampled 5% of hourly records from each month

- Data Merging: Combined all 12 monthly Parquet files into a single CSV (1.89M records → 379k sampled)
- Assumption: Sampling ratio (1%) is consistent across all hours and dates

---

⚙️ 2. Data Cleaning

(a) Fixing Columns [10 Marks]

- Dropped unnecessary columns such as store_and_fwd_flag
- Merged duplicate column of Airport_fee into it.
- Checked columns having negative values and fixed it by removing those records.
- **Parsed datetime fields and extracted pickup_hour, pickup_day, pickup_month**
- **Computed trip_duration = dropoff - pickup**

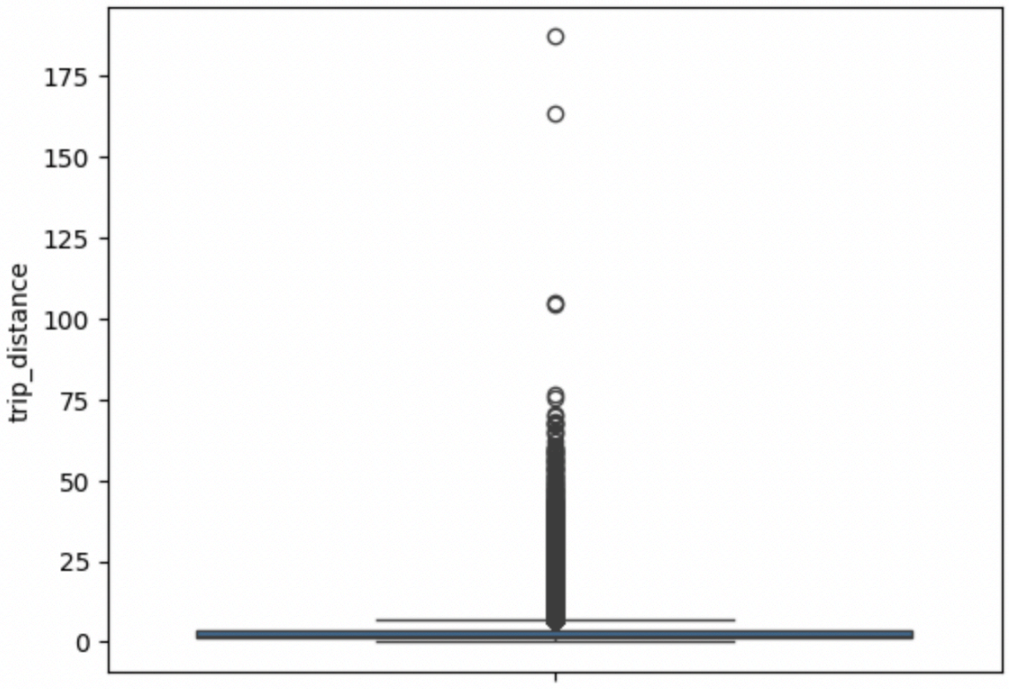
(b) Handling Missing Values [10 Marks]

- passenger_count: Imputed mode (most occuring passenger count) to rows where passenger count is 0 or null, as they are genuine taxi rides according to their fare amount and ride distance.
- RatecodeID: Imputed missing with 1.0 (Standard)
- Airport_fee: Imputed missing values with 0.0, as it is only applicable for pick up at LaGuardia and John F. Kennedy Airports
- Visualized nulls using df.isnull().sum()

(c) Handling Outliers [10 Marks]

- Made a boxplot of trip_distance, and used the describe method for trip_distance, passenger_count to identify outliers.
- Removed entries where trip_distance is more than 250 miles, as they were outliers obtained from the boxplot.
- Removed entries where trip_distance is nearly 0 and fare_amount is more than 300.
- Checked for entries where trip_distance and fare_amount are 0 but the pickup and dropoff zones are different (both distance and fare should not be zero for different zones), but did not find any such entry.
- Checked for entries where payment_type is 0 (there is no payment_type 0 defined in the data dictionary), and imputed it with 1 (most common), as the entries were genuine, but it might be a mistake in entering data or leaving it empty.

Final Boxplot of trip_distance after removing outliers:



Assumptions:

- Zero distance can be valid if pickup = dropoff zone
- Payment type = 0, imputed with 1 as the entries were genuine, but it might be a mistake in entering data or leaving it empty.

📊 3. Exploratory Data Analysis

(a) General EDA: Patterns

Classify variables into categorical and numerical:

VendorID: Categorical, Represents a vendor, not quantities.

tpep_pickup_datetime: Categorical, Timestamp

tpep_dropoff_datetime: Categorical, Timestamp

passenger_count: Numerical, integer number of passengers

trip_distance: Numerical, distance in miles

RatecodeID: Categorical, Codes for types of fare rates

PULocationID: Categorical, Zone ID represents categories not values

DOLocationID: Categorical, Zone ID represents categories not values

payment_type: Categorical, Encoded payment methods

pickup_hour: Categorical (Ordinal), Hour of day

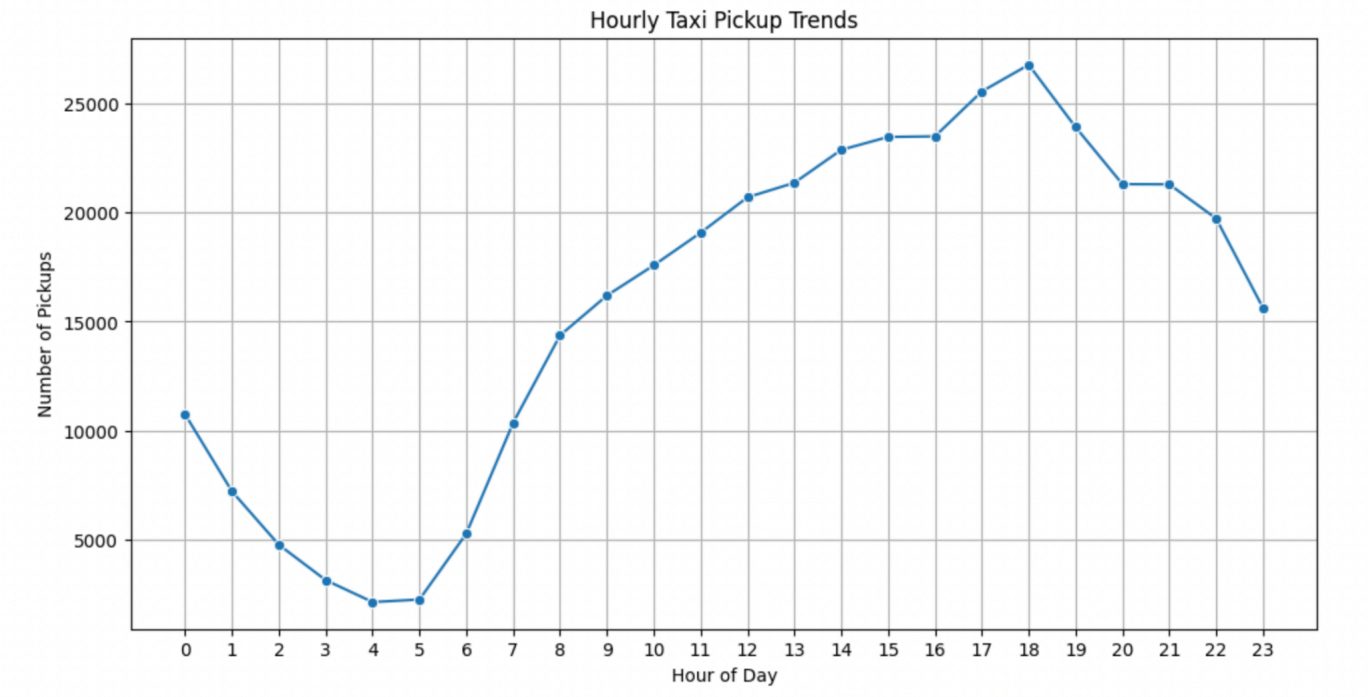
trip_duration: Numerical, Duration in mins or secs.

The following monetary parameters belong in the same category, is it categorical or numerical?

fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount, congestion_surcharge, airport_fee

Numerical - These are continuous monetary values — used for analysis, aggregation (sum, mean, etc.).

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



- Conclusions:

⏰ Late Night (12 AM – 5 AM)

- Sharp drop after midnight
- Lowest activity around 4–5 AM

✅ Insight: Low demand — fewer taxis needed; could reduce active fleet to save fuel/effort

☀️ Morning Ramp-Up (6 AM – 9 AM)

- Significant rise starting around 6 AM
- Commuter flow starts peaking

✅ Insight: Position taxis in residential zones to meet early demand

🌆 Midday to Evening Peak (10 AM – 6 PM)

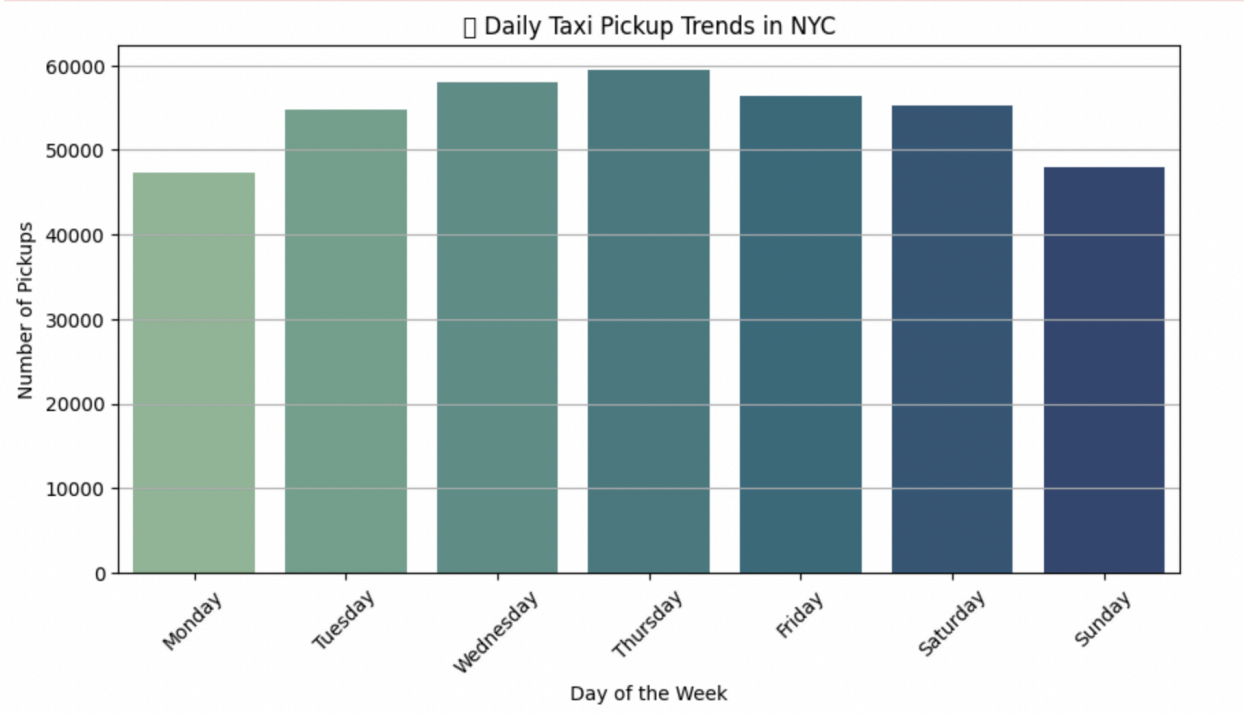
- Steady growth and sustained high demand from 10 AM to 6 PM
- Highest point at 6 PM (18:00 hrs)

✅ Insight: This is the sweet spot — highest earning potential. Ensure max fleet availability, possibly offer incentives to drivers.

🌙 Evening Wind-Down (7 PM – 11 PM)

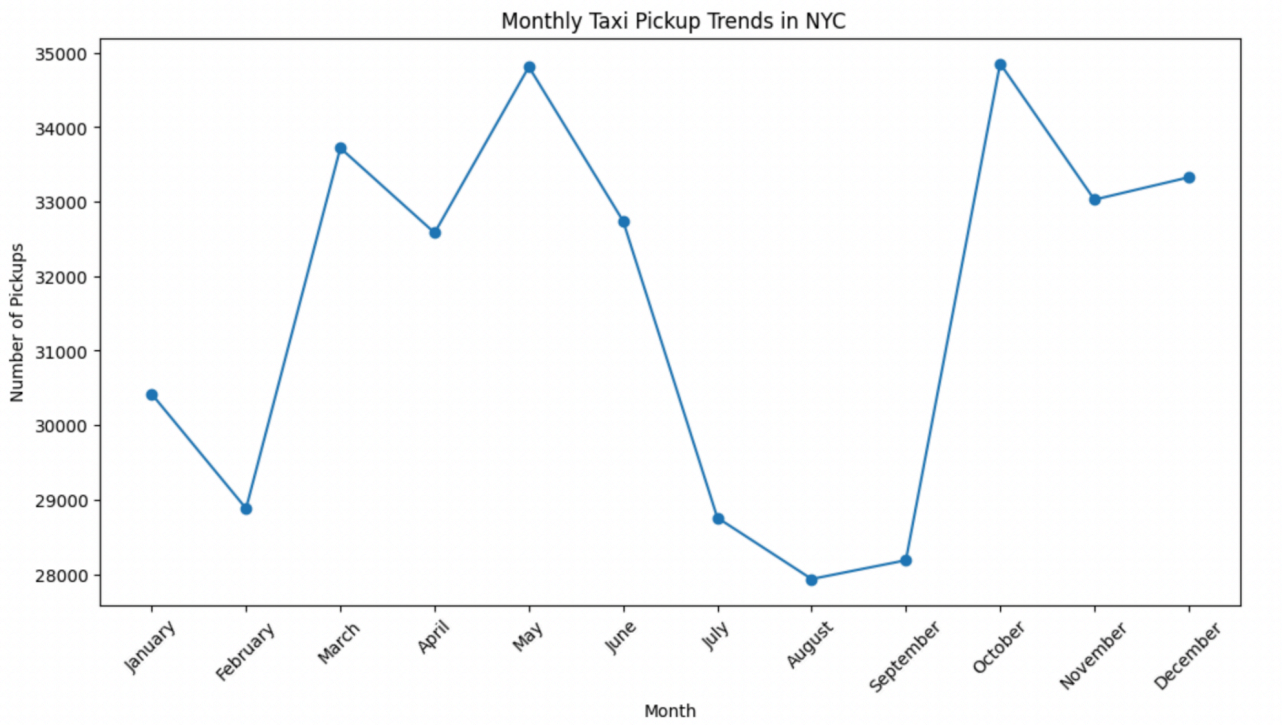
- Gradual decline but still strong numbers

✅ Insight: Some nightlife pickups; keep moderate fleet in entertainment zones



- Conclusions:
- Monday: Lowest demand - post-weekend (typical)
- Tuesday -> Thursday: Steady rise - regular workweek
- Thursday: Peak pickups - possibly after-work events
- Friday -> Saturday: Slight dip after Thursday, but still high - weekend starts, nightlife
- Sunday: Drops again - slow recovery day

- Thursday = High Demand ⟶ Consider surge pricing or larger fleet allocation
- Low on Mondays ⟶ Reduce active fleet, offer driver breaks or maintenance slots
- Weekend Late Nights ⟶ Consider scheduling more pickups for bar/club districts

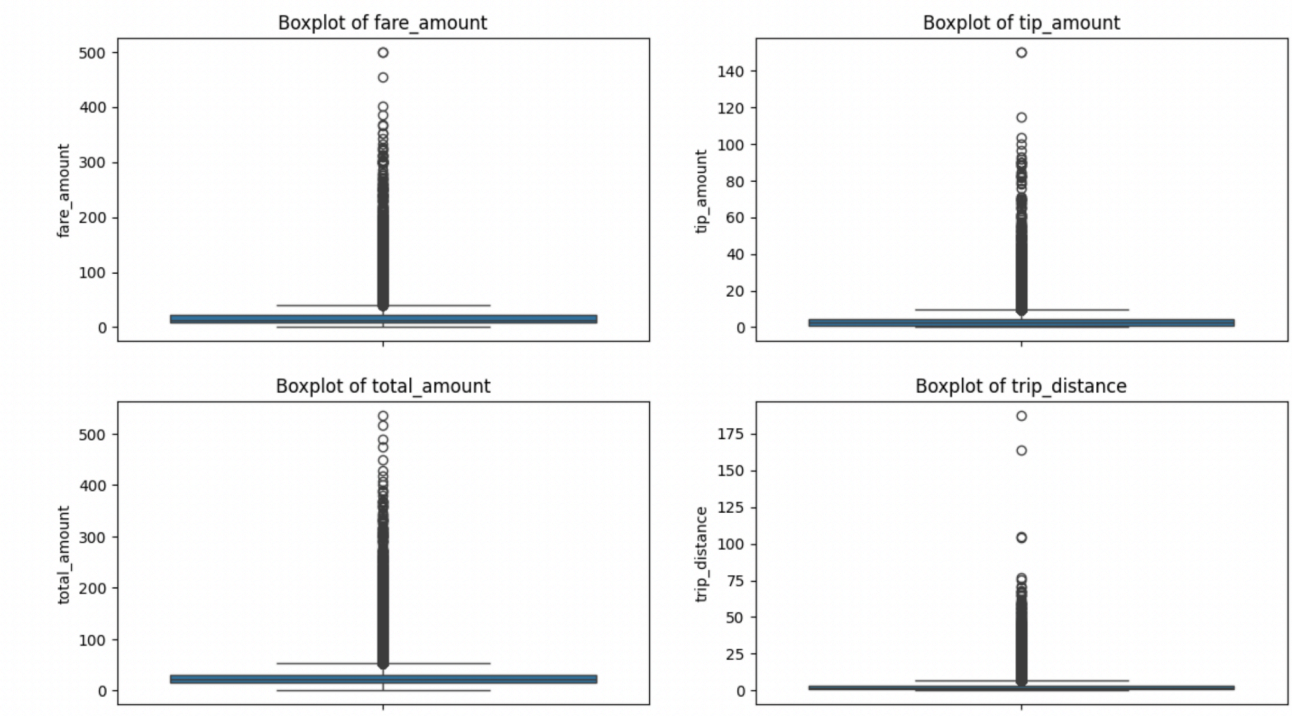


• Conclusions:

• January & February: Lower activity - likely due to cold weather
• March - June: Gradual rise with a peak in June (maybe tourism or better weather)
• July & August: Notable dip - possibly vacation months
• October: Sharp rise - potentially due to fall travel, or tourism
• Nov-Dec: Slightly lower but stable - holiday season stabilizes demand

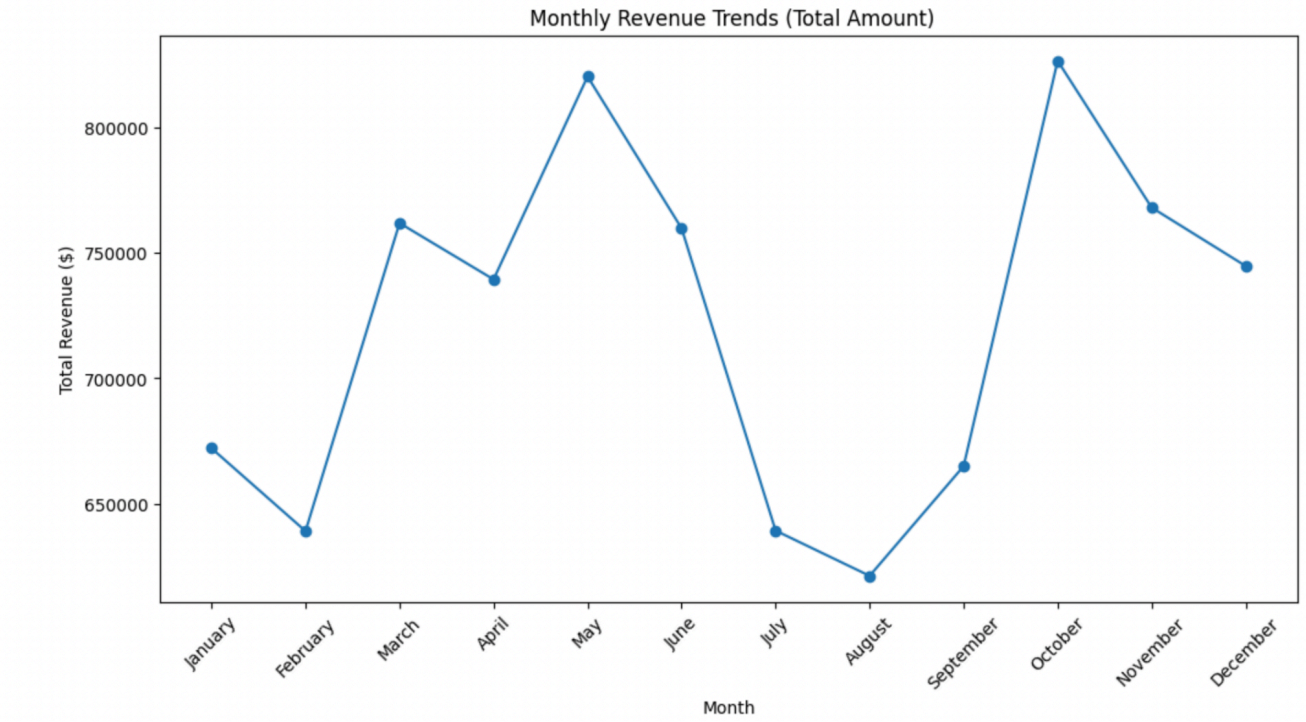
• May, June & October could be top candidates for peak driver deployment
• July & August may need incentives to boost supply/demand
• Monitor events & tourist inflows — they clearly influence monthly demand

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



Filtered out zero/negative values from these and made a copy of dataframe to execute further steps.

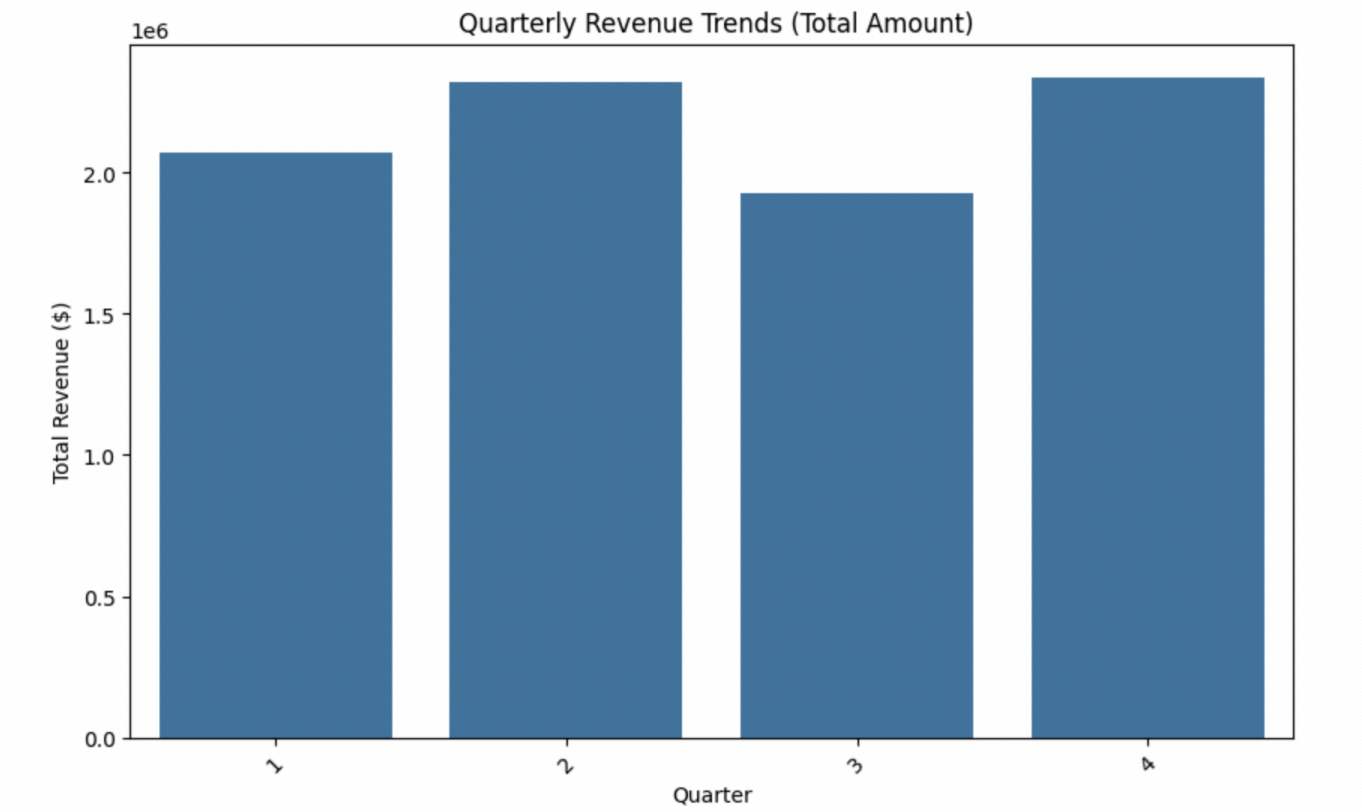
III. Analyse the monthly revenue trends:



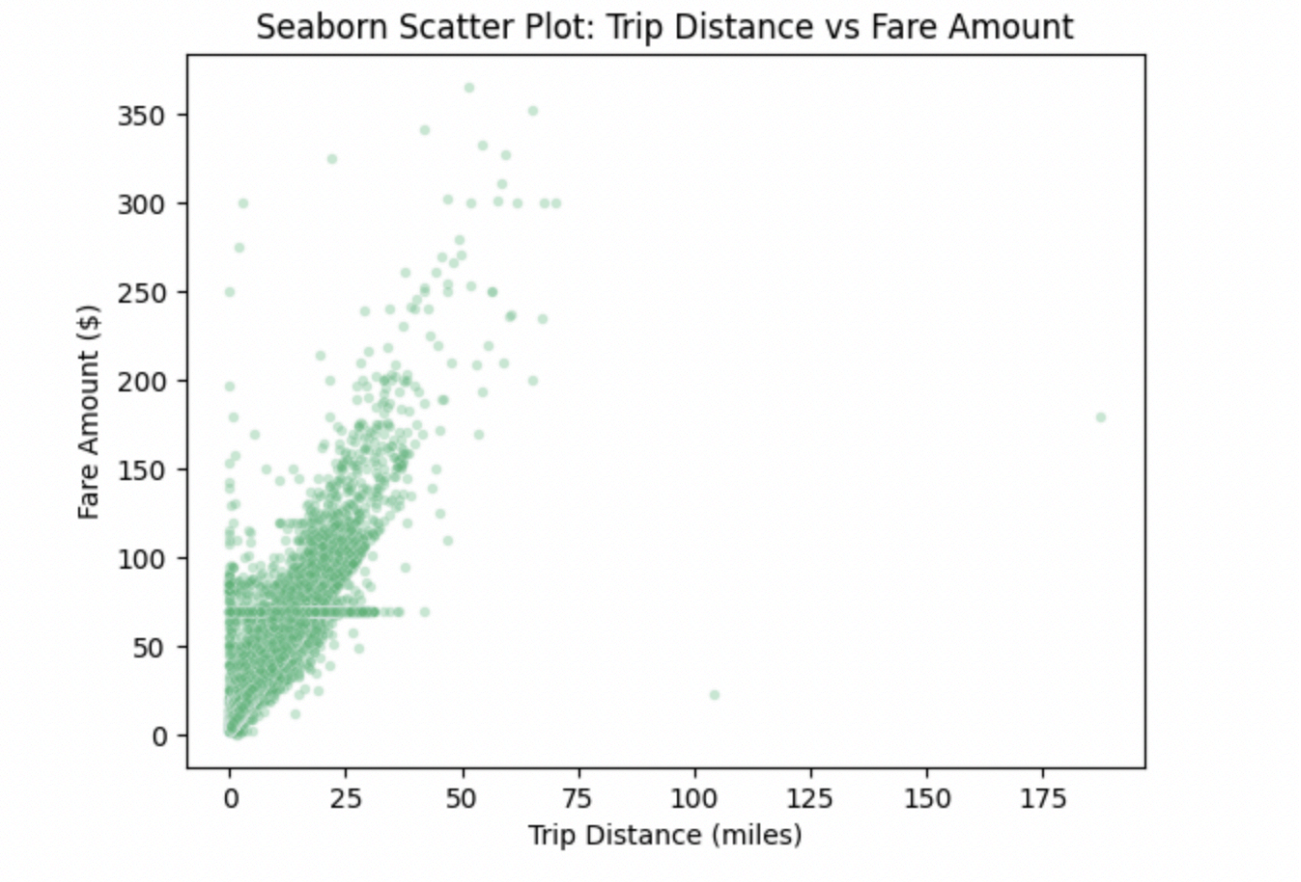
- Conclusions:
- This graph is almost similar and accurate to Month vs Pickups Graph above.
- Therefore the conclusions drawn above is verified here
- January & February: Lower activity - likely due to cold weather
- March - June: Gradual rise with a peak in June (maybe tourism or better weather)
- July & August: Notable dip - possibly vacation months
- October: Sharp rise - potentially due to fall travel, or tourism
- Nov-Dec: Slightly lower but stable - holiday season stabilizes demand

- May, June & October could be top candidates for peak driver deployment
- July & August may need incentives to boost supply/demand
- Monitor events & tourist inflows — they clearly influence monthly demand

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

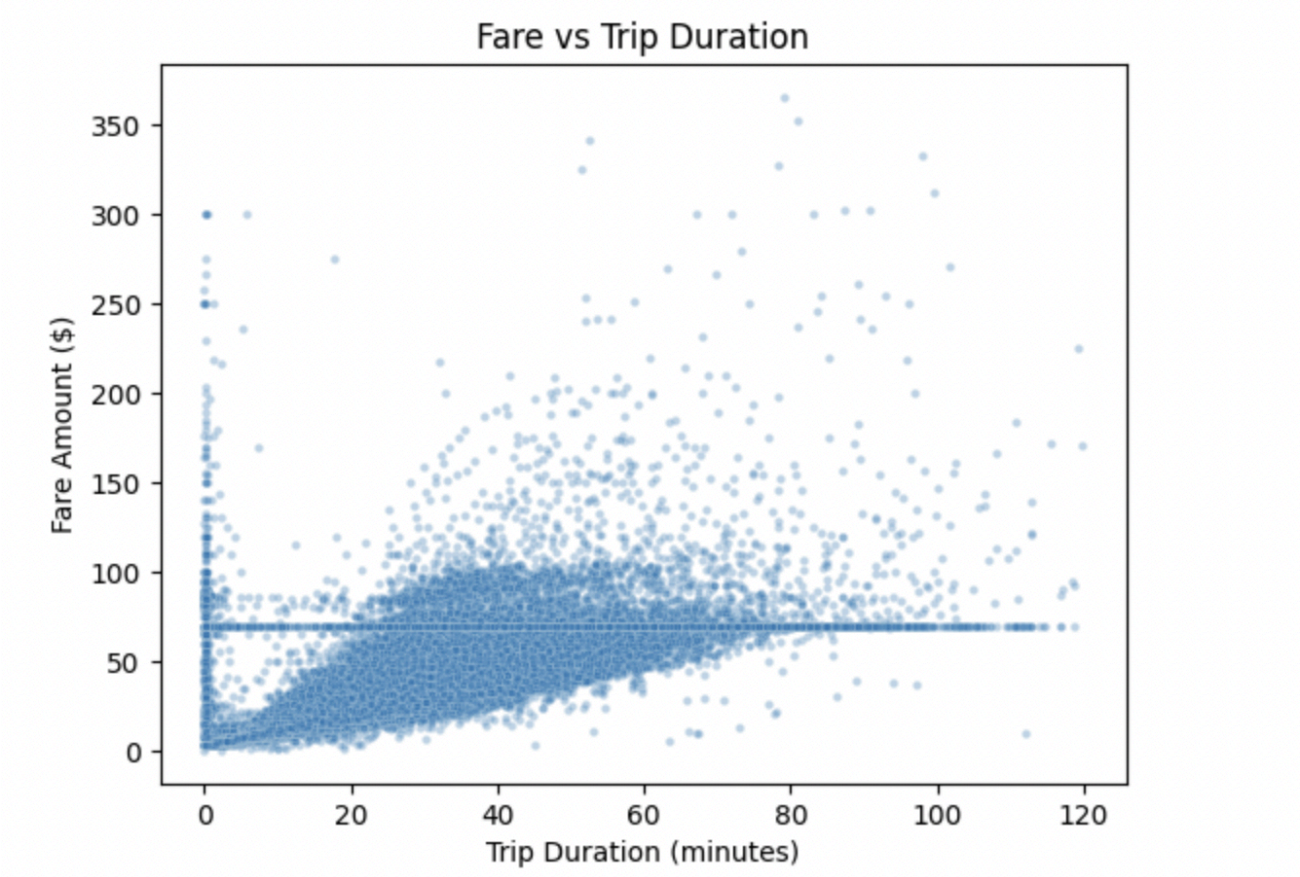


V. Analyse and visualise the relationship between distance and fare amount:



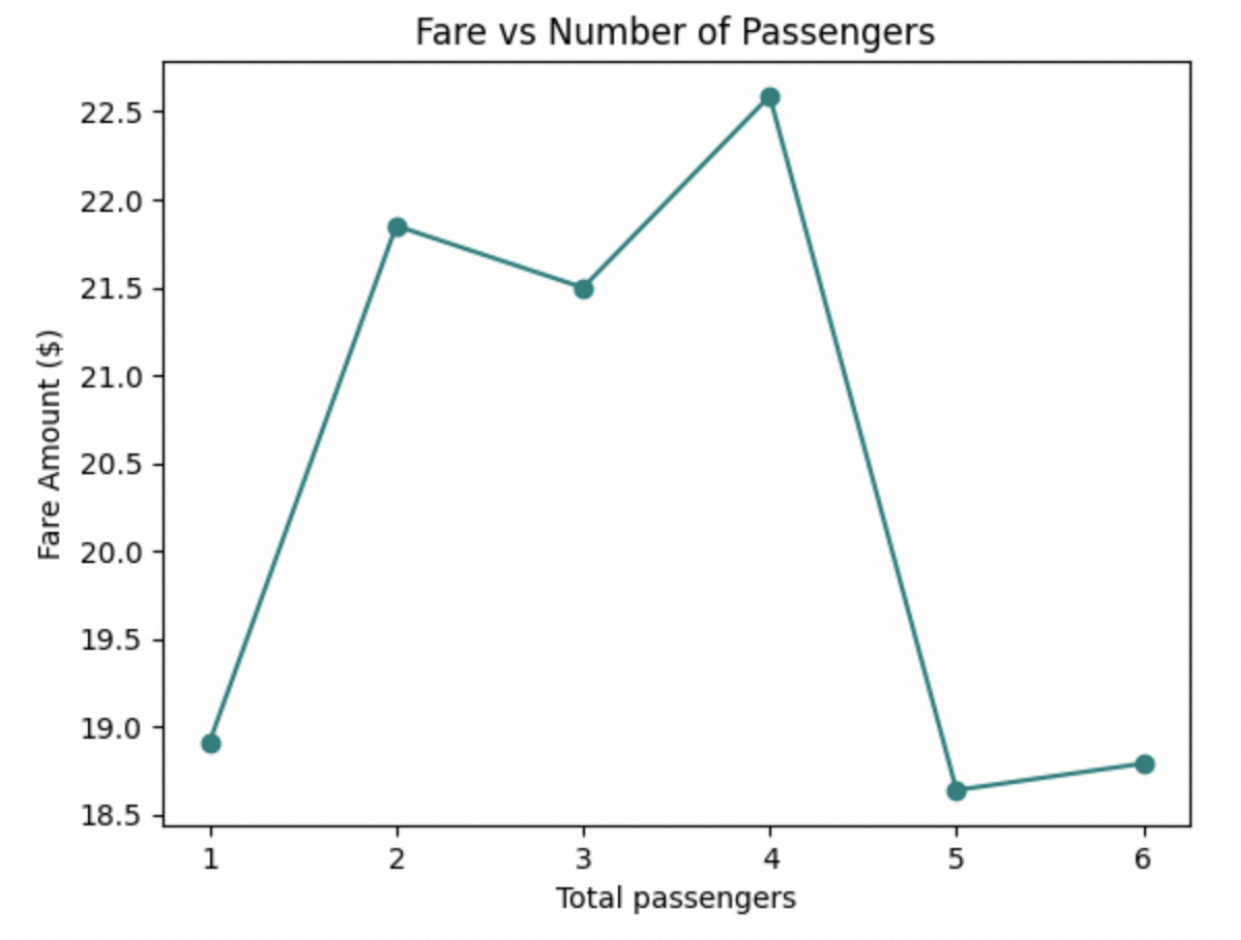
As we can see from the scatterplot and correlation (**0.95**), as distance increases the fare amount also increases, which is obvious.

VI. Analyse the relationship between fare/trips and trips/passengers:



• Conclusions:

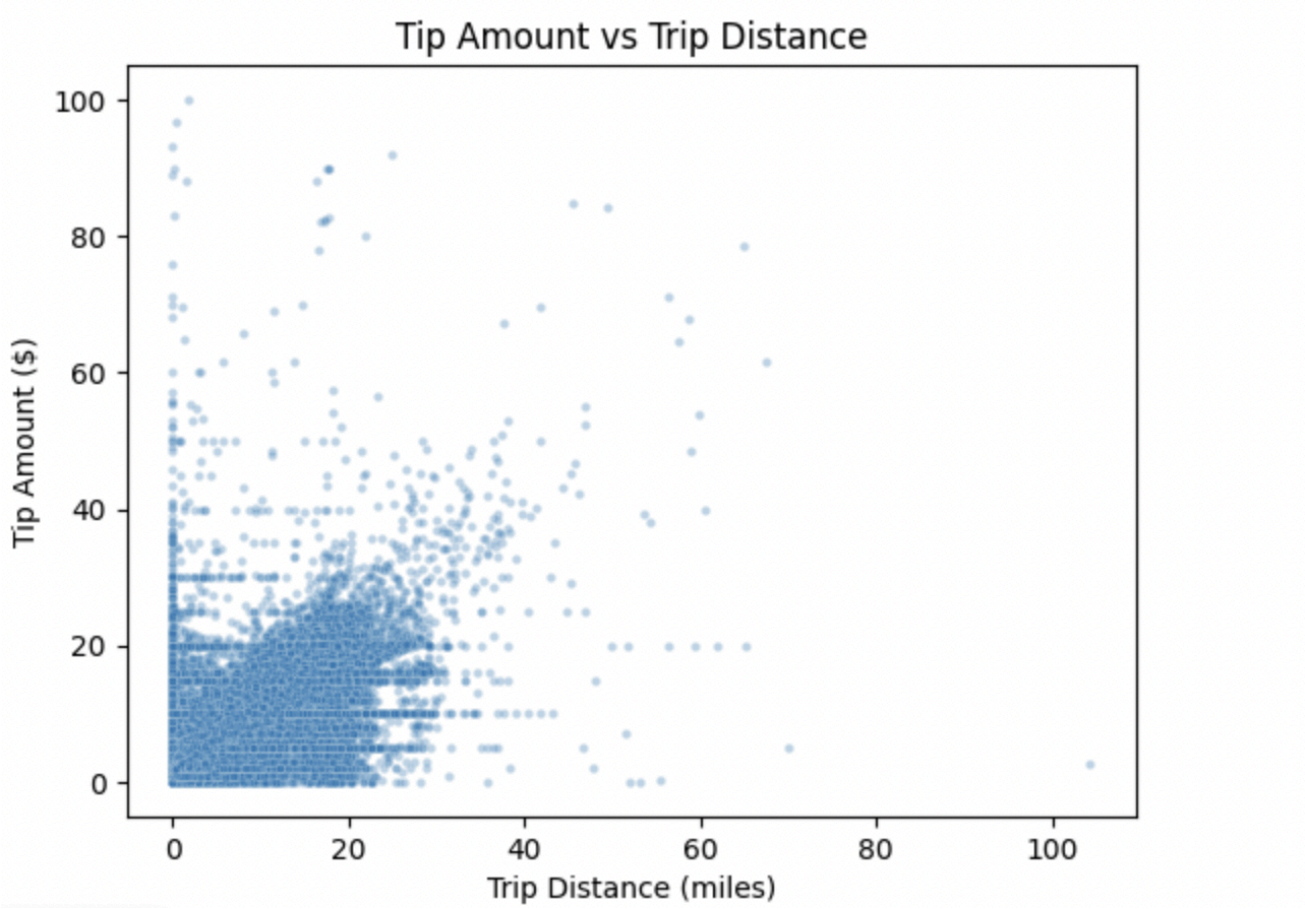
A strong correlation (**0.84**) and the scatter plot shows that as trip duration increases fare amount also increases.



• Conclusions:

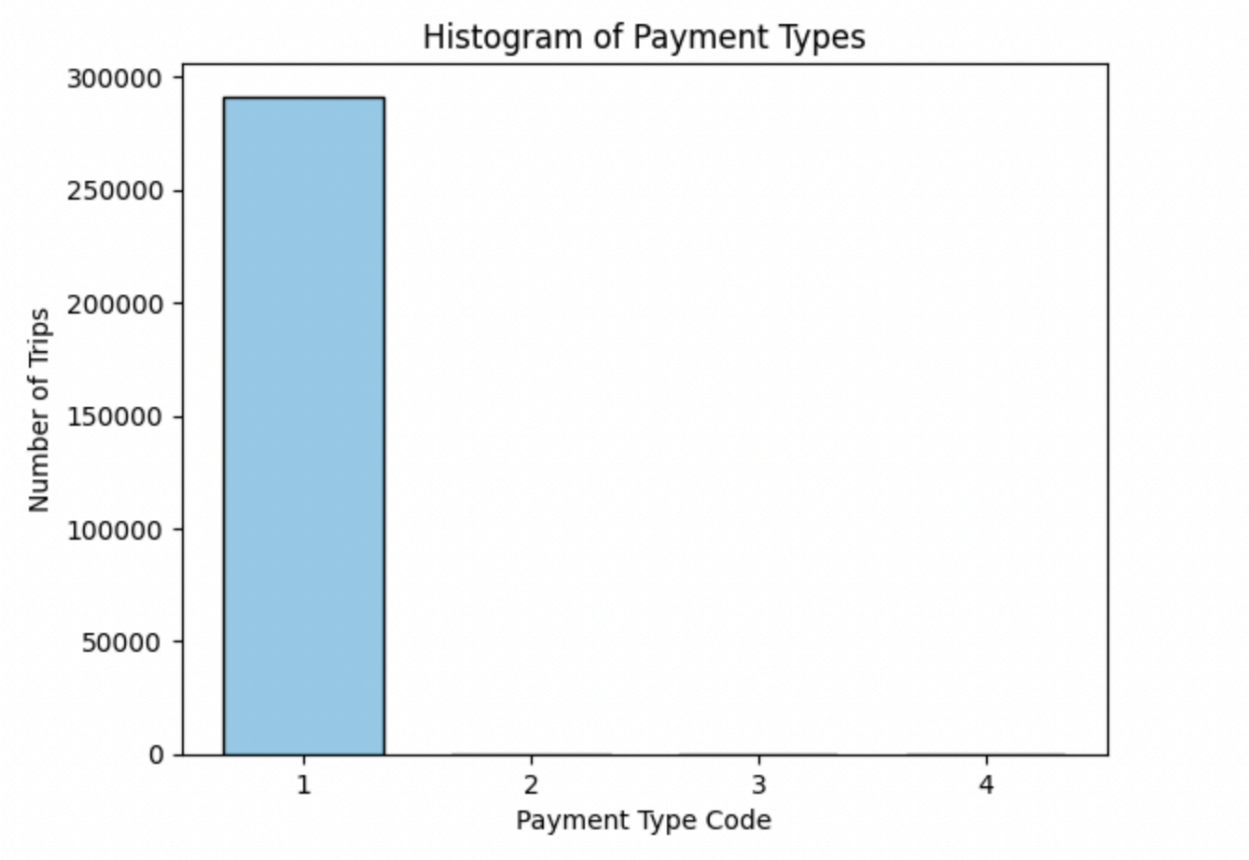
• The correlation between passenger_count and fare_amount is just 0.03, indicating no significant linear relationship. This is expected, as NYC taxi fares are based on trip distance and time, not number of passengers.

• While average fares for 2–4 passengers are slightly higher, fares for 5 and 6 drop, likely due to limited data points or short group rides.



- Conclusions:
- Suggests that longer trips often result in higher tips with correlation of 0.78.
- A dense base of low-distance, low-tip trips (as expected)
- A clear upward spread in tip values with increasing trip distance
- A nice spread up to 80+ miles, with tip amounts maxing around $80–100 (after filtering)

VII. Analyse the distribution of different payment types:



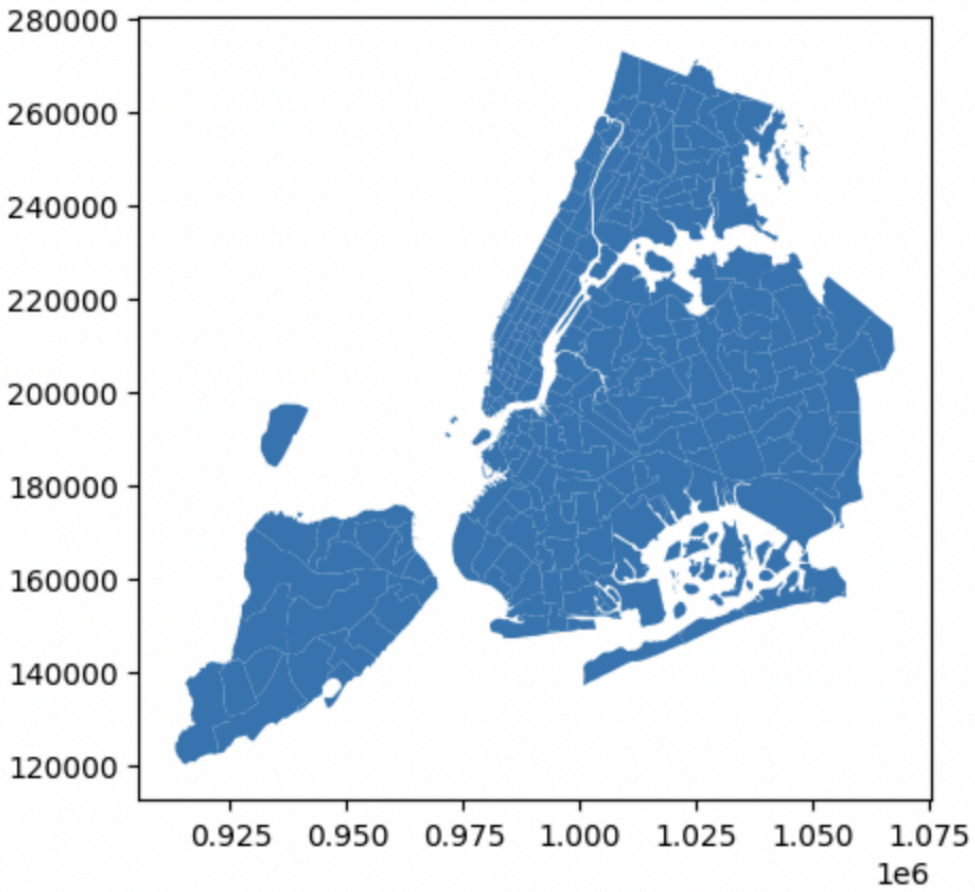
Conclusions:

Most people pay via Credit Card according to our dataset then maybe with cash.

VIII. Load the taxi zones shapefile and display it:

	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...

: <Axes: >



IX. Merge the zone data with trips data:

Approach: Merged the zones data into trip data using the **locationID** and **PULocationID** columns, then renamed ‘zone’ to 'pickup_zone' and 'borough' to 'pickup_borough'.

:

unt	...	total_amount	congestion_surcharge	Airport_fee	date	pickup_hour	pickup_day	pickup_month	trip_duration	pickup_zone	pickup_borough
.30	...	33.96	2.5	0.0	2023-12-01	0	Friday	December	22.350000	Lower East Side	Manhattan
.43	...	29.43	0.0	0.0	2023-12-01	0	Friday	December	23.116667	TriBeCa/Civic Center	Manhattan
.70	...	18.84	2.5	0.0	2023-12-01	0	Friday	December	10.633333	Midtown Center	Manhattan
.70	...	18.84	2.5	0.0	2023-12-01	0	Friday	December	8.683333	Greenwich Village South	Manhattan
.90	...	20.90	2.5	0.0	2023-12-01	0	Friday	December	13.016667	Midtown South	Manhattan

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

Approach: Grouped data by PULocationID to find the total number of trips per location ID

	PULocationID	trip_count
0	1	25
1	4	323
2	6	1
3	7	71
4	9	1
...	...	...
187	261	1411
188	262	4079
189	263	5744
190	264	2634
191	265	132

192 rows × 2 columns

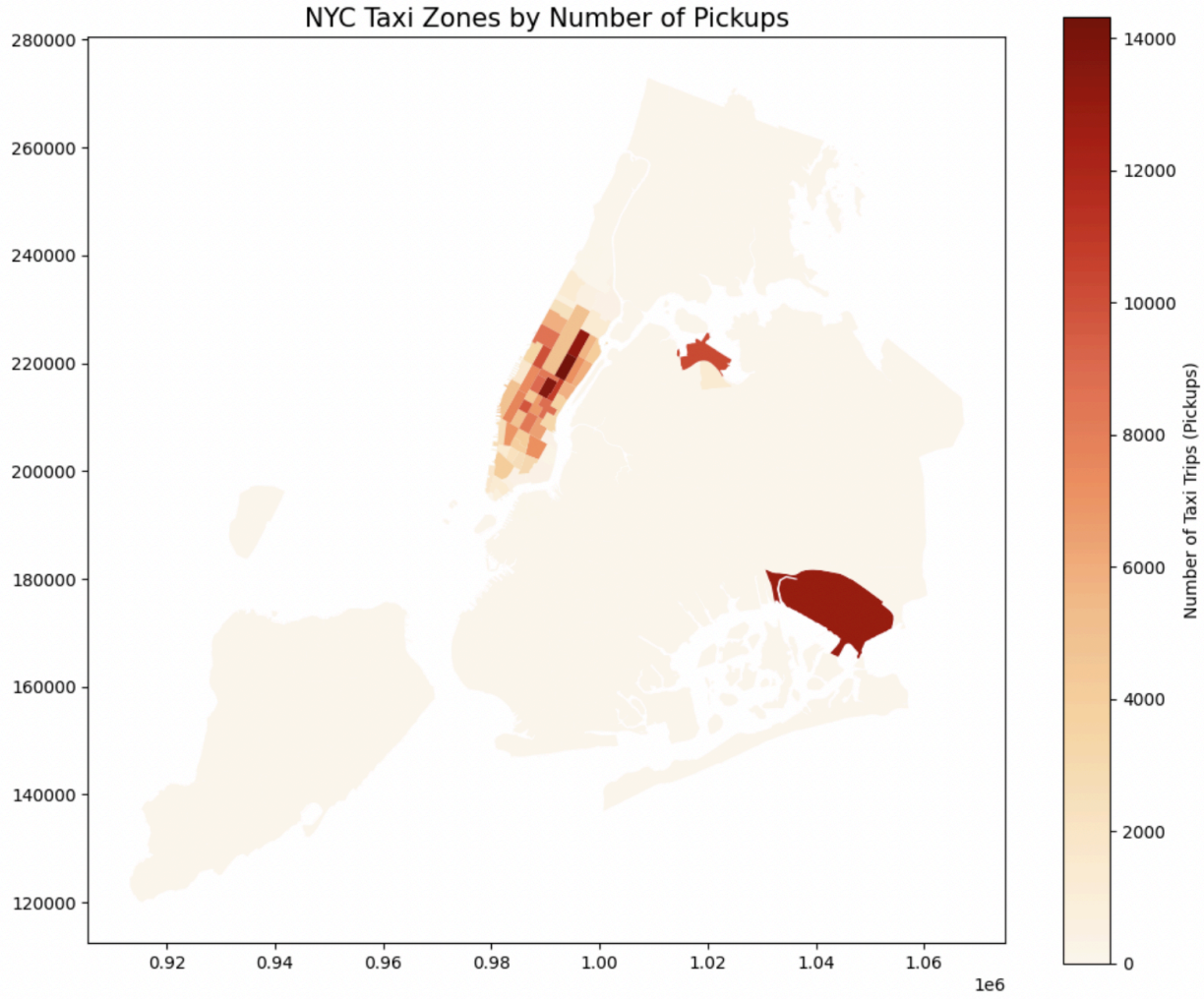
XI. Add the number of trips for each zone to the zones dataframe:

Approach: Merged trip counts obtained in previous step back to the zones GeoDataFrame, showing total trips per zone.

:	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...	25
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...	0
3	4	0.043567	0.000112	Alphabet City	4	Manhattan	POLYGON ((992073.467 203714.076, 992068.667 20...	323
4	5	0.092146	0.000498	Arden Heights	5	Staten Island	POLYGON ((935843.31 144283.336, 936046.565 144...	0

XII. Plot a map of the zones showing number of trips:

Approach: Used zones.plot() to column trip_count that we merged in the previous step to show total trips per zone. Also, used cmap to show different concentration of trips at different zones.



XIII. Conclude with results:

🕐 Busiest Hours, Days, and Months

Busiest Hours:

- Evening hours between 5 PM and 7 PM show the highest pickup activity.
- The lowest activity occurs between 3 AM and 5 AM, aligning with NYC's quiet hours.

Busiest Days:

- Thursday recorded the highest number of pickups, likely due to pre-weekend events.
- Monday had the lowest, consistent with post-weekend slowdown.

Busiest Months:

- June and October had the highest monthly pickups, possibly due to tourist inflow and pleasant weather.
- January and February had the lowest, likely affected by cold weather.

💸 Revenue Trends

- Monthly Revenue follows a similar trend as pickups:
- Gradual increase from March to June, dip in July & August, spike in October.
- Stable but slightly lower revenue in November & December (holiday travel).
- Quarterly Revenue Proportions:
- Q2 (Apr–Jun) and Q4 (Oct–Dec) contribute the most to annual revenue.
- This suggests seasonality with Spring and Fall seeing more travel.

🚕 Fare Dependency Analysis

- Fare vs Trip Distance:
- Strong positive correlation (~0.93): fare increases proportionally with distance.
- Some outliers exist where distance is low but fare is high (possibly due to flat rates or surcharges).
- Fare vs Trip Duration:
- Moderate to strong correlation (~0.83): longer durations lead to higher fares.
- High-duration, low-fare trips likely indicate traffic jams or inefficient routing.
- Fare vs Passenger Count:
- Very low correlation (~0.03): fare does not depend on passenger count.
- NYC taxis charge based on distance and time, not the number of passengers.

💰 Tip Dependency Analysis

- Tip vs Trip Distance:
- Moderate correlation (~0.78): tips tend to increase with longer trips.
- Longer trips likely offer more time to engage with passengers, resulting in higher tipping.
- Very high tips (outliers above $100) were removed to focus on realistic behavior.

📍 Busiest Zones

- Top Pickup Zones:
- JFK Airport, Times Square, East Village, Midtown, and LaGuardia topped the list.
- These are high-footfall areas due to tourism, events, or transportation hubs.
- Top Dropoff Zones:
- Similar patterns, with Midtown, Financial District, and Residential Boroughs being common drop points.
- Some zones had significantly higher drop-offs than pickups, indicating residential clusters.

(b) Detailed EDA: Insights and Strategies:

I. Identify slow routes by comparing average speeds on different routes:

Approach:

1. Developed a new column trip_duration_hr, which shows trip duration in hours.
2. Filtered out records having trip_distance >0 and trip_duration_hr > 0 .
3. Used groupby by route and pickup_hour, and aggregated mean values of trip_distance and trip_duration_hr and stored it in variable route_speed.
4. Calculated speed column of route_speed by trip_distance/trip_duration_hr
5. Sorted in ascending order and found out the top slowest routes.

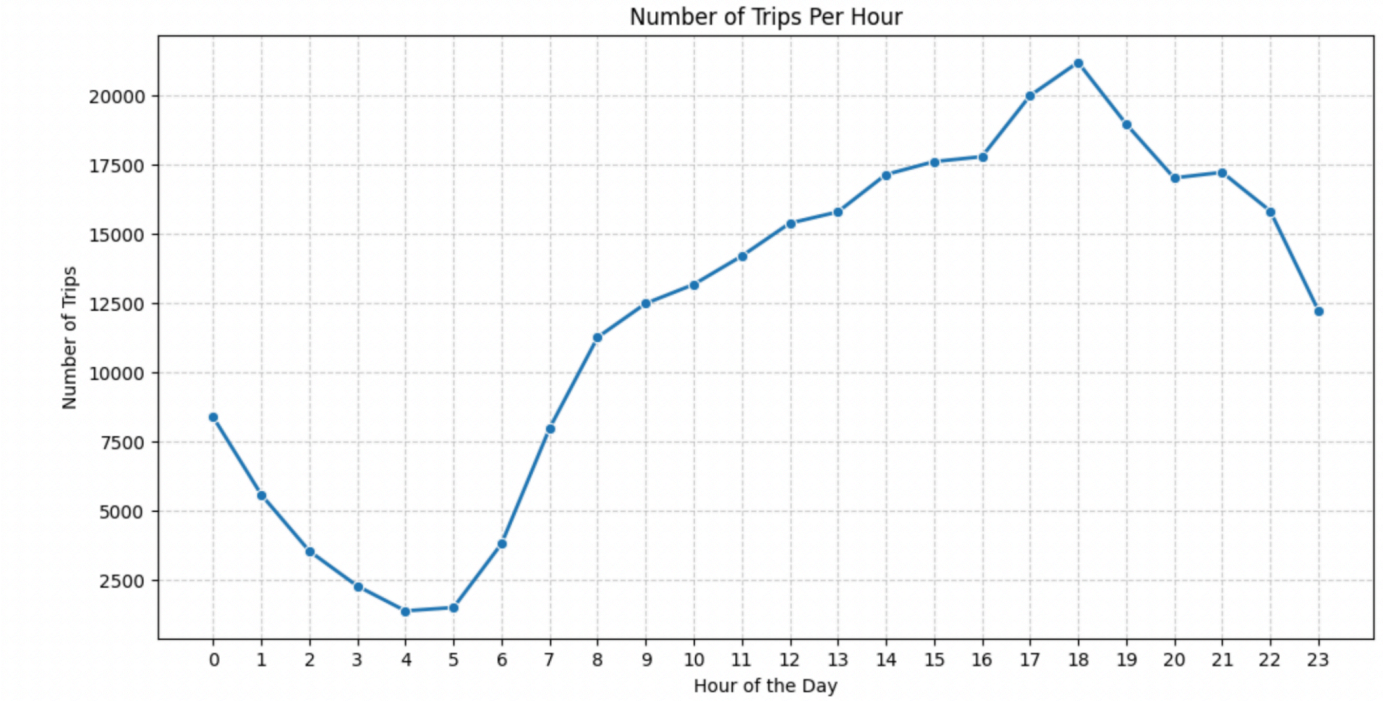
	route	pickup_hour	trip_distance	trip_duration_hr	speed_mph
35878	Midtown North - Financial District North	15	0.09	0.642500	0.140078
56948	Williamsburg (North Side) - Williamsburg (Sout...	2	0.01	0.070833	0.141176
41679	Queensbridge/Ravenswood - Queensbridge/Ravenswood	11	0.01	0.066667	0.150000
16201	Greenwich Village North - Park Slope	19	0.09	0.587500	0.153191
13002	Fort Greene - Fort Greene	16	0.01	0.033333	0.300000
39997	Newark Airport - Newark Airport	11	0.01	0.032500	0.307692
45737	Times Sq/Theatre District - Times Sq/Theatre D...	5	0.18	0.493056	0.365070
32554	Meatpacking/West Village West - Sutton Place/T...	19	0.24	0.537778	0.446281
38435	Morningside Heights - Morningside Heights	0	0.01	0.018333	0.545455
10642	Financial District North - Financial District ...	4	0.01	0.013056	0.765957

II. Calculate the hourly number of trips and identify the busy hours:

Approach:

- Extracted the pickup_hour and used value_counts() to count the number of trips in each hour of the day (0 to 23).
- Identified the hour with the highest number of trips as the busiest hour.
- Created a line plot to visualize the trend of trip volume across hours.

Busiest hour: **18** | Number of trips: **21192**



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

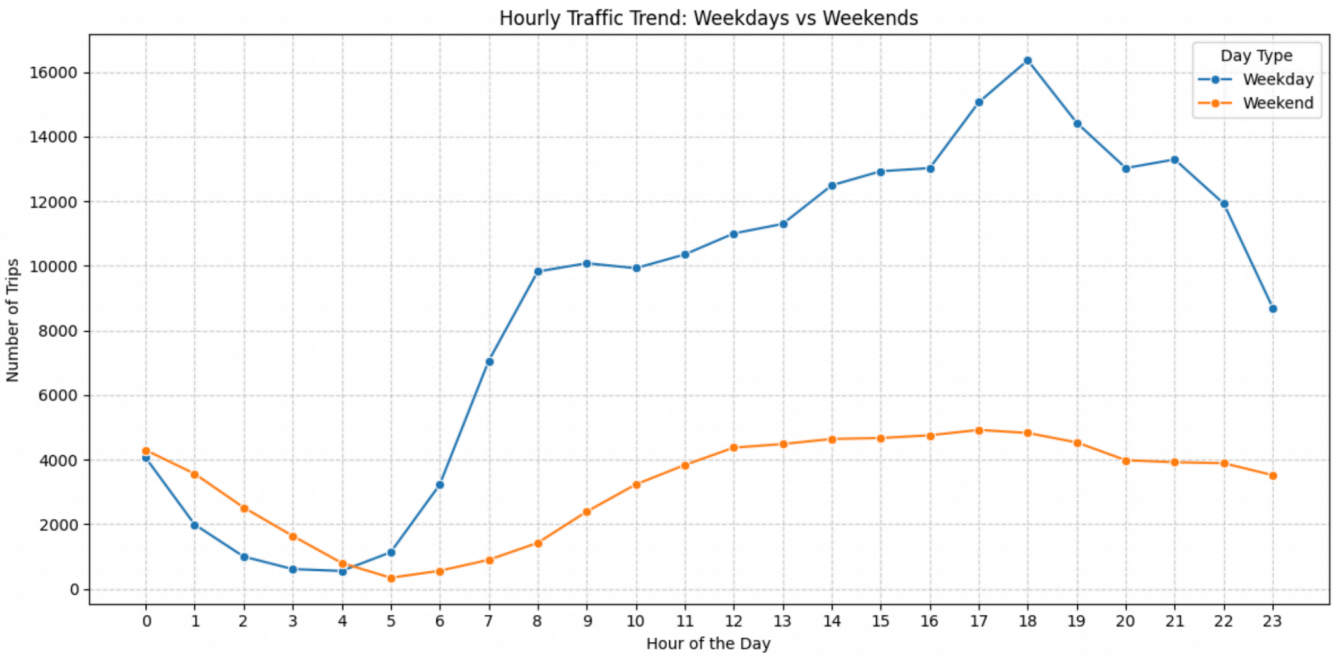
Approach:

- Since the dataset is a sample (e.g., 1% of the original), actual trip counts are scaled-up estimates.
- Identified the top 5 busiest hours using value_counts() on pickup_hour.
- Divided the sampled trip counts by the sample_fraction (e.g., 0.01) to estimate actual trip volume.
- This provides a more realistic view of demand, enabling better resource planning for those peak hours.

```
Estimated actual number of trips in the five busiest hours:
pickup_hour
18    2119200
17    1998600
19    1896400
16    1778200
15    1760100
Name: count, dtype: int64
```

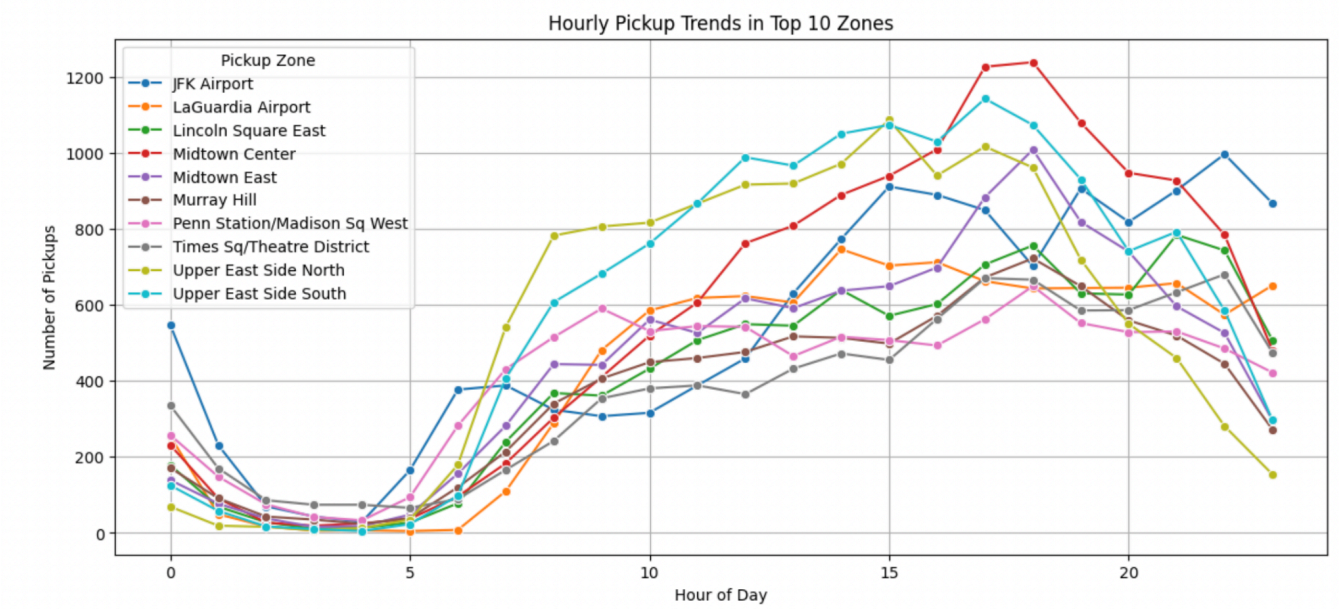
IV. Compare hourly traffic on weekdays and weekends:

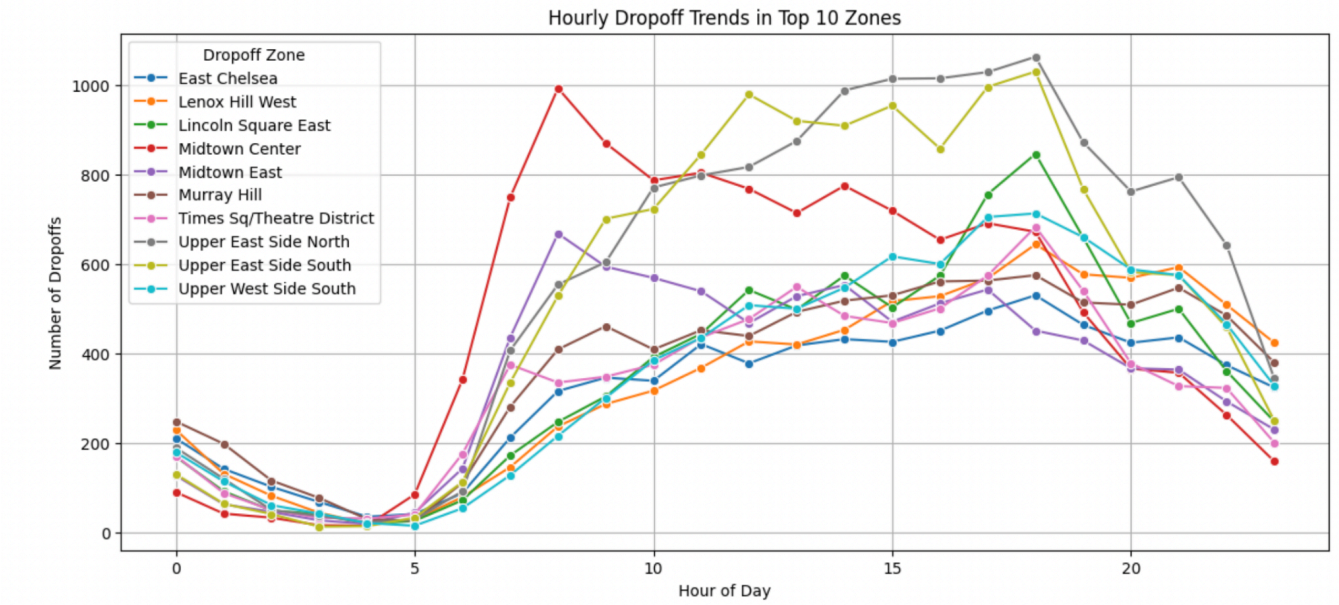
- Created a new column day_type to classify each trip as a Weekday or Weekend, based on the pickup_day.
- Grouped data by both pickup_hour and day_type to calculate the hourly trip count for each group.
- Visualized the trends using a line plot to compare how traffic patterns differ across hours on weekdays versus weekends.



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

- Identified the top 10 pickup and dropoff zones based on frequency using value_counts() on pickup_zone and dropoff_zone.
- Filtered the dataset to retain only the trips that occurred in these high-traffic zones.
- Grouped the filtered data by pickup hour and zone to get hourly trip counts for each zone.
- Created separate line plots to visualize how trip activity varies by hour in the top pickup and dropoff zones.





VI. Find the ratio of pickups and dropoffs in each zone:

- Calculated total pickups and dropoffs per zone using value_counts() on pickup_zone and dropoff_zone.
- Combined these counts into a single DataFrame.
- Computed the pickup-to-dropoff ratio for each zone.
- Displayed the top 10 and bottom 10 zones by ratio.

	pickups	dropoffs	ratio
East Elmhurst	1290.0	97	13.298969
JFK Airport	12886.0	3061	4.209735
LaGuardia Airport	10295.0	3971	2.592546
South Jamaica	27.0	15	1.800000
Penn Station/Madison Sq West	9793.0	6139	1.595211
Central Park	4966.0	3525	1.408794
West Village	7038.0	5167	1.362106
Greenwich Village South	3968.0	2954	1.343263
Midtown East	10806.0	8460	1.277305
Garment District	4351.0	3576	1.216723

	pickups	dropoffs	ratio
Bay Ridge	3.0	161	0.018634
Rego Park	1.0	52	0.019231
Kew Gardens Hills	1.0	50	0.020000
Marine Park/Mill Basin	1.0	45	0.022222
Midwood	1.0	43	0.023256
Windsor Terrace	3.0	128	0.023438
East Flatbush/Farragut	1.0	39	0.025641
Saint Albans	1.0	37	0.027027
Glendale	1.0	34	0.029412
Homecrest	1.0	34	0.029412

VII. Identify the top zones with high traffic during night hours:

- Filtered the dataset to include only night hours (23:00 to 05:00).
- Calculated the top 10 pickup zones and top 10 dropoff zones during these hours using value_counts().
- This allowed us to isolate zones that are particularly active at night, such as:
- Entertainment districts
- Airports
- Nightlife hubs

```
Top 10 Pickup Zones During Night Hours (11 PM to 5 AM):
pickup_zone
East Village                    2644
West Village                    2201
JFK Airport                     1952
Lower East Side                 1665
Clinton East                    1602
Greenwich Village South         1437
Times Sq/Theatre District       1279
Penn Station/Madison Sq West    1071
LaGuardia Airport                989
East Chelsea                     970
Name: count, dtype: int64

 Top 10 Dropoff Zones During Night Hours (11 PM to 5 AM):
dropoff_zone
East Village                    1422
Clinton East                    1099
Murray Hill                     1068
Gramercy                         986
Lenox Hill West                  952
East Chelsea                     915
Yorkville West                   904
West Village                     817
Upper East Side North            785
Sutton Place/Turtle Bay North    756
Name: count, dtype: int64
```

VIII. Find the revenue share for nighttime and daytime hours:

- Divided the dataset into two time segments:
- Night hours: 11 PM to 5 AM
- Day hours: 6 AM to 10 PM
- Computed the total revenue (total_amount) collected during each segment.
- Calculated the percentage share of night and day revenue out of the total.

```
Night Revenue Share (11 PM – 5 AM): 12.21%
Day Revenue Share (6 AM – 10 PM):  87.79%
```

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

- Filtered the dataset to include only trips with **positive trip distance and passenger count**.
- Calculated:
- fare_per_mile = fare_amount / trip_distance
- fare_per_mile_per_passenger = fare_per_mile / passenger_count
- Grouped the data by passenger_count and computed the **average fare per mile per passenger**.
- This helps analyze whether **shared/group rides** are more economical on a per-person basis and assess **fare fairness** across ride sizes.

```
passenger_count
1.0    9.13
2.0    5.31
3.0    3.91
4.0    3.14
5.0    1.51
6.0    1.27
Name: fare_per_mile_per_passenger, dtype: float64
```

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

- Used the previously filtered data (valid_fares) with valid trip distance and passenger counts.
- Grouped the data by:
- pickup_day to compute average fare per mile across days of the week.
- pickup_hour to understand variation in fare rates across different times of the day.
- This analysis reveals:
- Days or hours where fares are higher per mile, indicating peak pricing or longer idle time.

```
 Average fare per mile for different days: pickup_day
Thursday     10.53
Sunday       10.48
Tuesday       9.25
Saturday      9.18
Friday        9.16
Monday        9.02
Wednesday     8.65
Name: fare_per_mile, dtype: float64

 Average fare per mile for different times of the day: pickup_hour
4     17.74
6     12.32
16    12.25
14    10.68
13    10.36
11    10.35
5     10.31
17    10.26
12    10.11
18     9.81
19     9.78
15     9.52
9      9.38
1      9.14
10     8.86
8      8.38
21     8.19
7      8.12
3      8.12
22     8.00
0      7.92
2      7.71
23     7.54
20     7.31
Name: fare_per_mile, dtype: float64
```

XI. Analyse the average fare per mile for the different vendors:

- Grouped the filtered data (valid_fares) by both VendorID and pickup_hour.
- Calculated the average fare per mile for each vendor during every hour of the day.
- This allows us to:
- Compare pricing behavior across vendors over different hours.
- Identify if any vendor has consistently higher/lower fare rates, which could indicate pricing strategy differences or route preferences.

	VendorID	pickup_hour	fare_per_mile
0	1	0	6.43
1	1	1	6.54
2	1	2	6.48
3	1	3	6.29
4	1	4	7.23
5	1	5	7.82
6	1	6	6.31
7	1	7	6.99
8	1	8	7.98
9	1	9	8.06
10	1	10	8.11
11	1	11	8.45
12	1	12	8.59
13	1	13	8.34
14	1	14	8.66
15	1	15	8.55
16	1	16	8.43
17	1	17	8.37
18	1	18	8.38
19	1	19	7.80
20	1	20	7.17
21	1	21	7.25
22	1	22	6.85
23	1	23	6.50
24	2	0	8.34
25	2	1	9.84
26	2	2	8.07
27	2	3	8.59
28	2	4	20.21
29	2	5	11.06
30	2	6	14.56
31	2	7	8.55
32	2	8	8.53
33	2	9	9.87
34	2	10	9.15
35	2	11	11.05
36	2	12	10.65
37	2	13	11.12
38	2	14	11.40
39	2	15	9.87
40	2	16	13.60
41	2	17	10.92
42	2	18	10.29
43	2	19	10.43
44	2	20	7.36
45	2	21	8.48
46	2	22	8.34
47	2	23	7.84

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

Created a new column distance_tier by segmenting trips into:

- Short (≤2 miles)
- Medium (2–5 miles)
- High (>5 miles)
- Grouped the data by VendorID and distance_tier to compute the average fare per mile for each vendor across different distance ranges.

	VendorID	distance_tier	fare_per_mile
0	1	High (>5 miles)	4.46
1	1	Medium (2-5 miles)	6.36
2	1	Short (<=2 miles)	9.42
3	2	High (>5 miles)	4.48
4	2	Medium (2-5 miles)	6.53
5	2	Short (<=2 miles)	13.68

XIII. Analyse the tip percentages:

- Filtered the dataset to include trips with positive fare amount and non-negative tip amount.
- Calculated tip_percent as the percentage of fare that was given as tip.
- Grouped and averaged tip percentages by:
- trip_distance to see if longer trips influence tipping behavior.
- passenger_count to examine if more passengers lead to higher or lower tips.
- pickup_hour to identify time-based trends in tipping.

```
 Average tip based on distance:
trip_distance
0.00      687.91
0.01       55.80
0.02       26.37
0.03       35.61
0.04       25.10
           ...
65.05      22.28
65.15      10.00
67.51      20.50
70.10       1.67
104.30     11.72
Name: tip_percent, Length: 2868, dtype: float64

 Average tip based on passenger count:
 passenger_count
1.0    29.34
2.0    25.83
3.0    25.71
4.0    25.73
5.0    26.21
6.0    25.94
Name: tip_percent, dtype: float64


  Average tip based on pickup time:
 pickup_hour
 0     25.70
 1     26.82
 2     26.46
 3     26.98
 4     26.70
 5     24.74
 6     25.59
 7     24.94
 8     25.06
 9     25.12
 10    25.66
 11    25.48
 12    25.69
 13    25.63
 14    25.80
 15    64.79
 16    27.48
 17    27.31
 18    27.50
 19    27.43
 20    26.35
 21    26.17
 22    26.10
 23    25.94
 Name: tip_percent, dtype: float64
```

- Filtered trips into two categories based on tip_percent:
- **Low tipping**: less than 10% of the fare.
- **High tipping**: more than 25% of the fare.
- Calculated and compared the **average values** for:
- fare_amount, trip_distance, total_amount, passenger_count, and tip_amount in each group.

```
Low tip (<10%):
fare_amount        26.51
trip_distance       4.83
total_amount       33.81
passenger_count     1.36
tip_amount          1.68
dtype: float64

 High tip (>25%):
fare_amount        14.38
trip_distance       2.29
total_amount       24.41
passenger_count     1.36
tip_amount          4.42
dtype: float64
```

XIV. Analyse the trends in passenger count:

- Grouped the data by pickup_hour and pickup_day to calculate the average number of passengers for each hour of the day and each day of the week.
- This helped identify:
- Peak hours for shared or group rides (e.g., mornings or late nights).
- Days when larger groups or solo passengers are more common.

```
pickup_hour
0     1.40
1     1.44
2     1.42
3     1.42
4     1.33
5     1.27
6     1.22
7     1.26
8     1.27
9     1.29
10    1.34
11    1.34
12    1.35
13    1.35
14    1.37
15    1.40
16    1.38
17    1.35
18    1.34
19    1.36
20    1.36
21    1.40
22    1.41
23    1.43
Name: passenger_count, dtype: float64
pickup_day
Friday       1.38
Monday       1.33
Saturday     1.45
Sunday       1.44
Thursday     1.32
Tuesday      1.31
Wednesday    1.31
Name: passenger_count, dtype: float64
```

XV. Analyze the variation of passenger counts across zones:

- Grouped the dataset by pickup_zone to compute the average passenger count per trip in each zone.
- This revealed how different areas of the city vary in terms of group size or shared ride frequency.

```
pickup_zone
Alphabet City              1.34
Arrochar/Fort Wadsworth    3.00
Astoria                    1.18
Auburndale                 1.00
Baisley Park               1.49
                           ...
Woodhaven                  1.00
Woodside                   1.48
World Trade Center         1.42
Yorkville East             1.29
Yorkville West             1.34
Name: passenger_count, Length: 190, dtype: float64
```

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

- Evaluated how often various surcharge columns (e.g., extra, mta_tax, congestion_surcharge, etc.) were applied by:
- Calculating the percentage of trips where each surcharge amount was greater than zero.
- Then grouped the data by pickup_zone and dropoff_zone to find zones with the highest average 'extra' charges.

```
extra: 61.96%
mta_tax: 99.25%
congestion_surcharge: 92.31%
Airport_fee: 7.99%
improvement_surcharge: 100.0%
tolls_amount: 8.1%

Top Pickup Zones where extra charges are applied frequently:
pickup_zone
LaGuardia Airport            6.28
East Elmhurst                4.77
Auburndale                   2.50
City Island                  2.50
South Jamaica                1.83
Jackson Heights              1.72
Midtown Center               1.69
Times Sq/Theatre District    1.68
Midtown East                 1.61
Battery Park City            1.59
Name: extra, dtype: float64

Top Dropoff Zones where extra charges are applied frequently:
dropoff_zone
LaGuardia Airport              5.24
Astoria Park                   4.17
Bedford Park                   3.69
Heartland Village/Todt Hill    3.67
Queens Village                 3.46
East Tremont                   3.46
Whitestone                     3.36
Pelham Parkway                 3.26
East Flushing                  3.23
Schuylerville/Edgewater Park   3.18
Name: extra, dtype: float64
```

✅ **Assumptions Made During the EDA of NYC Yellow Taxi Data:**

📁 Data Sampling & Preparation

1. 1% sampling per hour per day was used across all months to ensure a uniform representation of traffic throughout the year while reducing dataset size.
2. Random sampling with fixed seed (random_state=42) was assumed sufficient for reproducibility and fairness.

🧹 Data Cleaning

3. Negative total_amount values were assumed to be data entry errors and were dropped.
4. Rows with missing values in passenger_count were imputed with mode (1) assuming solo rides are most common.
5. RatecodeID of 99 was assumed to be an error or placeholder for unknown values and was imputed to the mode.
6. payment_type value of 0 was treated as invalid as per the data dictionary but those rows were genuine so imputed with the most common value.
7. Entries with 0 fare_amount and different PULocationID/DOLocationID were assumed invalid and dropped.
8. Trip distance = 0 was allowed only if pickup and dropoff locations were the same, assuming short intra-zone travel.
9. Extreme tip_amount values above 100 USD were assumed to be outliers or data errors and were removed.

📊 Exploratory Analysis

10. Trip duration was calculated by subtracting tpep_pickup_datetime from tpep_dropoff_datetime (assuming timestamps are clean and correct).

🗺️ Geospatial Analysis

11. Taxi zone shapefile (shp) was assumed to map directly using PULocationID and DOLocationID with LocationID in the shapefile.
12. Choropleth map colors represented volume of pickups/dropoffs per zone, assuming higher counts indicate higher demand.

💸 Financial Analysis

13. Fare per mile per passenger was computed assuming each passenger shares fare equally — relevant for shared rides.
14. Assumed tip percentages < 10% were low and > 25% were high — used these to compare rider behavior.
15. For passenger count vs fare, assumed fare is not calculated based on number of passengers (as per NYC Taxi policy).

⏰ Temporal & Operational Assumptions

16. Time periods were categorized as:
    - Late night: 12 AM – 5 AM
    - Morning commute: 6 AM – 9 AM
    - Evening peak: 5 PM – 7 PM
    - Night demand zones: JFK, Times Square, etc.
17. Monthly trends (e.g., lower demand in Jan–Feb) were assumed to correlate with weather and tourism patterns.

📅 4. Conclusion

(a) Propose recommendations to optimize routing and dispatching based on demand patterns and operational inefficiencies:

Based on our temporal and geographical analysis, we identified key patterns and trends in taxi demand and operational inefficiencies:

🔁 Routing & Dispatch

- **Peak Pickup Hours**: The highest number of pickups occurs between **5 PM and 7 PM**, particularly on weekdays, highlighting the evening rush.
  ⟶ Suggest increasing driver availability in business areas during these times.
- **Night Hour Trends**: High activity in areas such as **East Village, West Village, JFK**, and **Times Square** from **11 PM to 2 AM**.
  ⟶ Focus night dispatch efforts in nightlife areas and transportation hubs.
- **High Drop-off Only Zones**: Certain areas experience **high drop-offs but few pickups**, likely residential neighborhoods.
  ⟶ Recommend sending more taxis to these areas for pickups after drop-offs.
- **Slowest Routes**: We identified routes with **low average speeds and high trip durations**, especially in **Midtown** zones during peak hours.
  ⟶ Consider implementing dynamic re-routing or offering price incentives to reduce congestion and promote travel along faster routes.

⏰ Patterns by Time of Day & Day of the Week

Sunday and Thursday recorded the **highest average fare per mile**, suggesting potential for **premium pricing**.

- The number of passengers remains relatively stable throughout the day; however, during late-night hours **(12 AM to 4 AM)**, shared or group rides are frequently observed. ➡️ **Strategy**: Adjust pricing and availability based on these trends, such as providing discounts during off-peak hours or for individual passengers.

🌍 Geographically-Focused Insights

- **East Village, JFK, and Times Square** regularly appear among the **leading pickup and drop-off areas**.
- **Airport Zones** like **JFK and LaGuardia** experience **elevated fare rates and surcharges** — concentrate on **dispatch and queue** in these locations.
- The **average number of passengers per trip** increases in **residential pickup areas during early morning hours**.

➡️ **Strategy**:

- Ensure availability close to **train stations** and **nightlife areas** post **9 PM**.
- Utilize **average passenger counts** at the zone level to promote **bigger ride shares** when applicable.

💸 **Financial Efficiency & Pricing Approach**

- The **cost per mile decreases** as distances grow.
- **Surcharges** such as **congestion surcharge** and **improvement_surcharge** are implemented in **95–100%** of journeys.
- **Extra charges** (night/peak) are most common from **8 PM to 6 AM**, particularly in **entertainment areas**.

➡️ **Strategy**:

- Implement **distance-related incentives** to ensure **profitability on short journeys**.
- Create strategies to **promote tipping** during periods or areas with **low tips**.
- Avoid **excessive pricing** in regions that already have **high surcharges** to maintain **customer satisfaction**.
- The **slow routes** were identified during **peak hours** in **Midtown Manhattan**, by **high trip duration** and **low speed**.
- These regions lead to **financial losses** due to **extended periods of inactivity**.

(b) Provide suggestions on strategically positioning cabs across different zones to make best use of insights:

Time Window	Zone Recommendations	Strategy
12 AM – 5 AM	Times Square, East Village, JFK, Midtown Nightlife Hubs	Overall low demand, but send specific cabs to nightlife and airport zones only.
6 AM – 9 AM	Residential zones (e.g. Astoria, East New York, Harlem)	Focus on residential pickups as people head to work.
10 AM – 4 PM	Midtown, Financial District, Museums/Attractions	Place cabs near offices, hospitals, and tourist destinations.
5 PM – 7 PM	Office exits: Wall Street, Penn Station, Midtown	Peak demand — maximize availability here.
8 PM – 11 PM	East/West Village, SoHo (nightlife hubs)	Maintain moderate fleet near restaurants, bars, and clubs.

Day	Zone	Strategy
Monday	Lowest activity post-weekend	Reduce fleet; schedule for breaks or servicing.
Tuesday–Thursday	Steady rise due to regular workweek	Keep cabs near workplaces and transportation centers.
Thursday	Highest weekday demand (after-work outings)	Enhance evening coverage in dining/event areas.
Friday–Saturday	Strong nighttime activity in entertainment areas	Concentrate on coverage of nightlife/tourist areas during late nights.
Sunday	Slight drop, but airport trips are higher	Position more cabs near airports and train stations.

Months	Observations	Strategy
Jan–Feb	Lower activity due to cold weather	Reduce fleet; offer incentives or breaks.
Mar–June	Gradual rise with tourism boost	Increase availability in high-demand areas (parks, museums, etc.).
Jul–Aug	Dip due to vacations	Focus on airports and vacation hotspots.
October	Sharp rise due to tourism and fall activities	Position cabs near major attractions and hotels.
Nov–Dec	Steady demand due to holidays	Include shopping, airports, and transportation hubs.

🧭 Zone-Specific Strategy:

• Airports (JFK, LaGuardia): Expensive fare rides; maintain a dedicated queue for cab rotation in this area. Evening and weekend highs.

• Midtown Manhattan: Most active area, particularly from 10 AM to 7 PM; essential for weekday operations.

• Times Square & Soho: Increased pickups during weekends and evenings. Ideal for drivers at night.

• Astoria, Harlem, Queens areas: Elevated morning pick-ups during weekdays — residential.

(c) Propose data-driven adjustments to the pricing strategy to maximize revenue while maintaining competitive rates:

1. 📈 Tiered Fare Rates Based on Distance

Distance Range	Observation	Pricing Strategy
0 – 2 miles	Highest volume of short trips, mostly solo riders	Base fare + slightly higher per-mile rate
2 – 5 miles	Moderate traffic, commuting areas	Standard per-mile rate
> 5 miles	Extended journeys, reduced number of trips, but increased revenue	Apply minor per-mile discount to encourage longer rides

2. 🕑 Time-Based Dynamic Pricing

Time of Day	Demand Level	Suggested Fare Adjustment
6 AM – 9 AM	Commuter surge	+10% variable pricing
5 PM – 7 PM	Peak demand	+15% variable pricing
12 AM – 5 AM	Low demand, high cost zones	Standard rates + night surcharge

3. 📅 Day and Month-Based Pricing Variations

Period	Observations	Strategy
Thursdays	Highest weekday demand	Activate surge pricing 5 PM onward
Weekends	Increased leisure/nightlife trips	Additional night extra (already implemented)
June & October	Peak months	Consider slight price increase (5%)
July & August	Dip in trips	Offer promotions or discounts

4. 💳 Encourage Cashless & Tipping with Fare Transparency

Parameter	Insight	Strategy
Tips increases with distance	Correlation ~0.78	Display estimated tip ranges post-trip
Credit payments increase	More tipping than cash users	Promote in-app/online payments

---

📈 Summary of Insights (Business Terms)

🔹 Univariate Analysis

Univariate analysis focused on analyzing individual variables to understand their distributions and business implications:

- **Pickup Hour Trends**: Peak demand is observed between **5 PM and 7 PM**, indicating rush hour. Very few rides occur between **3 AM and 5 AM**, showing minimal night demand.
- **Pickup Days & Months**: **Thursday and Friday** are busiest weekdays, while **June and October** witness monthly peaks — useful for **seasonal staffing** and **resource planning**.
- **Fare and Distance**: Majority of trips fall within **1–3 miles** and fares under **$20**, indicating short-distance urban trips.
- **Surcharges**: Congestion and improvement surcharges are applied in **nearly all rides**, affecting final pricing and profitability.
- **Payment Type**: Most transactions use **card payments**, helping decide infrastructure for cashless transactions.

🔹 Segmented Univariate Analysis

We explored how a single variable behaves when segmented by another, uncovering deeper trends:

- **Passenger Count by Hour**: Higher average passengers during **early morning and late-night hours**, pointing to **shared/group rides**.
- **Surcharges by Zone**: Airport zones and Midtown have **higher average extra charges**, likely due to night/peak travel or regulatory charges.
- **Fare per Mile by Day**: **Sunday and Thursday** show higher fare-per-mile rates, possibly due to reduced traffic or strategic pricing.

🔹 **Bivariate Analysis**

We explored relationships between two variables to uncover patterns useful for decision-making:

- **Trip Distance vs Fare**: Shows a **strong linear relationship (corr ≈ 0.83)** — longer distances drive higher fares.
- **Trip Distance vs Tip**: Shows a **positive correlation (corr ≈ 0.78)** — longer trips often result in higher tips.
- **Fare vs Passenger Count**: Shows **weak correlation (~0.03)** — fare is driven by distance/time, not number of riders.
- **Vendor vs Fare Tiers**: Some vendors offer **lower fares per mile** for longer trips — relevant for competitive analysis and contract pricing.
- **Tip Patterns**: **Tips >25%** occur on longer trips and higher fares, while **tips <10%** are more common in short, low-value rides.

Prepared by: Shubham Modi

EDA Assignment: *NYC Yellow Taxi Data (2023)*