Report: Optimising NYC Taxi Operations By Madhusudhan GB

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

Dataset was too large, so to reduce size, we take small sample from each file. From every hourly data of each day, 5% rows are randomly selected. We did this for all 12 monthly files. After sampling, all monthly samples are combined into one final DataFrame. This help in keeping trend pattern same while reducing file size.

2. Data Cleaning

2.1. Fixing Columns

2.1.1. Fix the index

Date and hour column we have removed as those were used in only sampling the dats

2.1.2. Combine the two airport_fee columns

2.1.3.

There were two columns with similar names -Airport_fee and airport_fee Both had the same meaning. So, we combined them into one new column by adding their values. After that, we removed the old two columns to avoid confusion.

2.2. Handling Missing Values

2.2.1. **Find the proportion of missing values in** The missing values are in passegner_count,ratecodeid,store_and_fwd_flag ,congestion_surcharge

2.2.2. **Handling missing values in passenger_count**

Missing values in passenger_count were filled using the median value to keep the data consistent

2.2.3. **Handle missing values in RatecodeID**

Null values in RateIDcode were replaced with the median to fix gaps and keep values real

2.2.4. **Impute NaN in congestion_surcharge**

Missing values in Congestion_surcharge were filled with the median value to complete the data.

2.3. Handling Outliers and Standardising Values

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

Outliers showed few rows had payment_type as 0, which is not valid. Need for standardization

3. Exploratory Data Analysis

3.1. General EDA: Finding Patterns and Trends

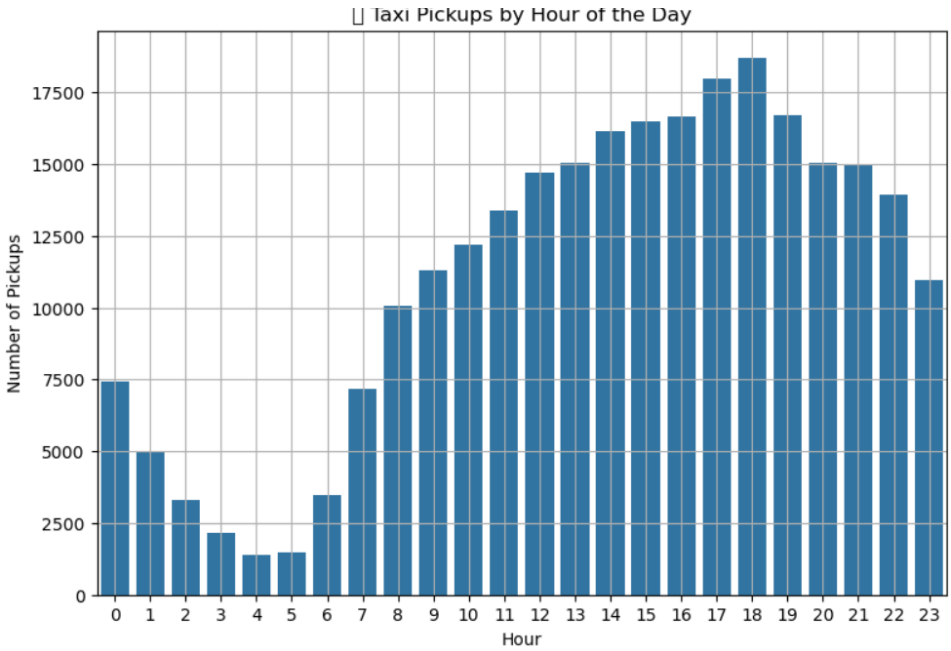
3.1.1. **Classify variables into categorical and numerical**

Numerical:trip_distance,faare_amount,trip_amount,total_amount,passenger_amunt

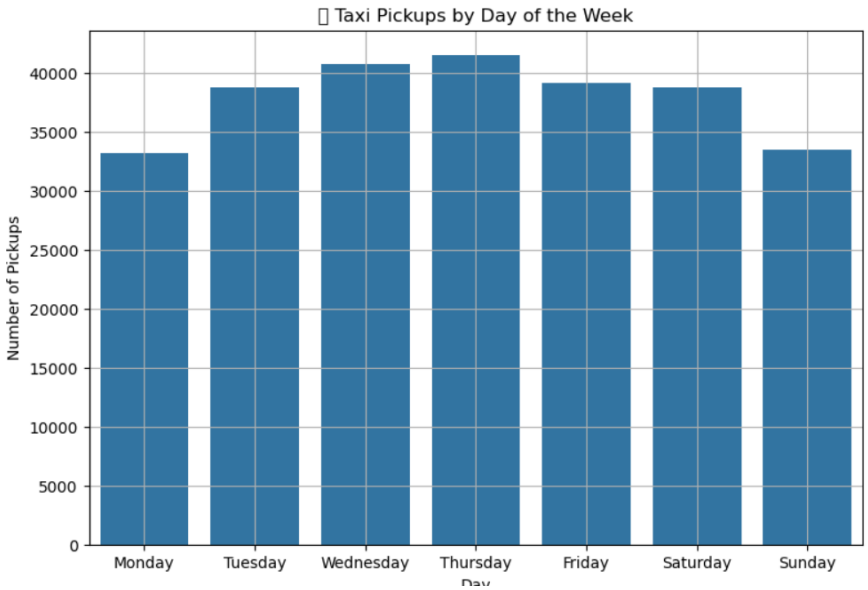
categorical: VendorID,RateCodeID,PULocationid,payment_type

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

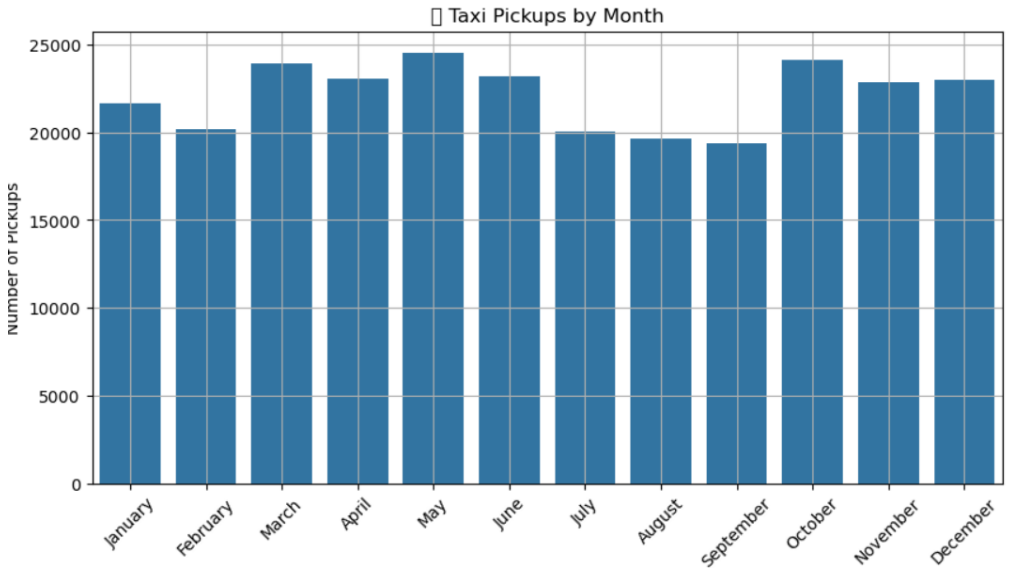
By hour :Number of pickups during 4pm to 7pm are high



By day: Thursday are high and Sunday are low



By month : In may pickup hours are high



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

3.1.4. Analyse the monthly revenue trends

Monthly revenue trends are high during months of may,september,october due to likely a holiday season

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

Quarter 1: 23.12

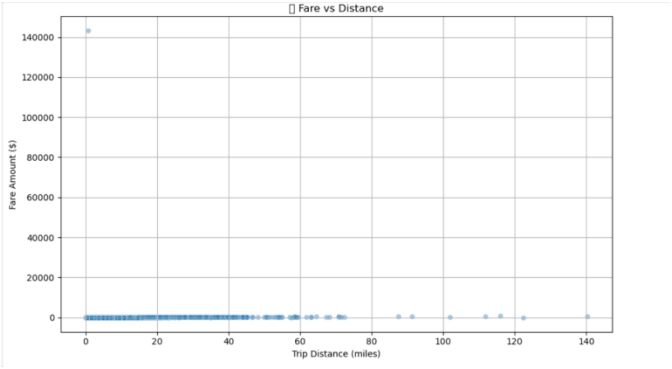
Quarter 2: 26.48

Quarter 3: 23.11

Quarter 4: 26.34

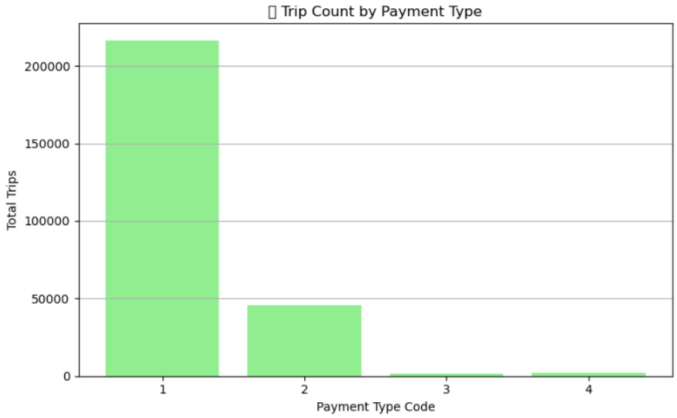
3.1.6. Analyse and visualise the relationship between distance and fare amount

Most trips showed a direct link between distance and fare — as thedistance increases, fare also increases. A scatter plot was used to show this relation clearly.



3.1.7. Analyse the distribution of different payment types

Most people paid using card, followed by cash payments. A very small number of trips used other payment types like No Charge or Dispute.



3.1.8. 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...

3.1.9.

3.1.10. Merge the zone data with trips data

The zone shapefile was merged with the trip records using PULocationID and LocationID This helped match each pickup trip with its corresponding zone name, which was later used for zone-based analysis and map visualizations.

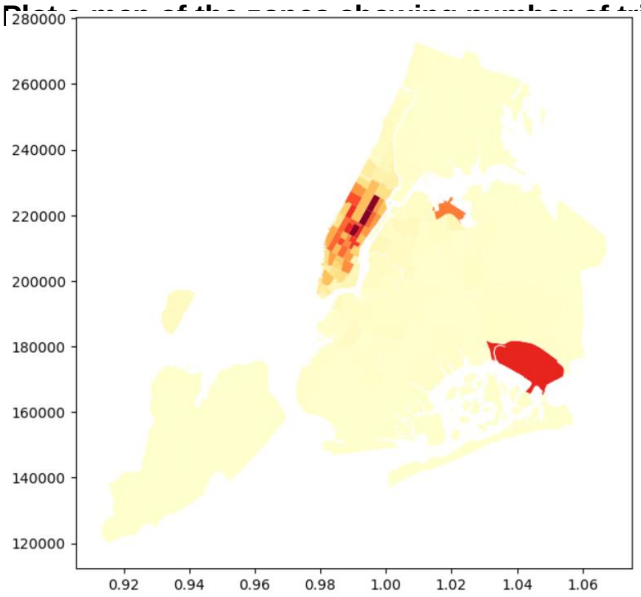
3.1.11. Find the number of trips for each zone/location ID

The number of trips was counted for each zone by combining both pickup and dropoff location IDs. This gave the total activity for each zone and helped identify which locations had the most taxi traffic.

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

The total trip counts for each zone were merged into the zones GeoDataFrame using LocationID. This allowed us to connect trip volume with geographic locations and prepare the data for visual mapping.

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



3.1.14. Conclude with results

From the zone-wise analysis, we found that a few zones had much higher trip counts than others. By merging trip counts with the zone data, we were able to see which areas had the most taxi activity. This information is useful for understanding demand across the city and can help with better planning and taxi distribution.

3.2. Detailed EDA: Insights and Strategies

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

For each pickup and dropoff zone pair, the average trip duration and distance were used to calculate speed. Routes with very low average speeds were marked as slow. These slow routes mostly happened during busy hours or in high-traffic zones, showing possible delays or traffic congestion.

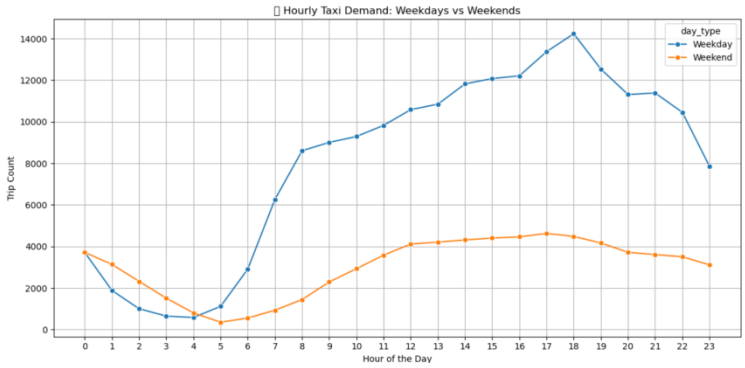
3.2.2. Calculate the hourly number of trips and identify the busy hours

Trips were grouped by pickup hour to find when taxis were used the most. The analysis showed that the evening hours, especially between 5 PM to 8 PM, had the highest number of trips. These hours were identified as the busiest for taxi demand.

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

Since the dataset was sampled, the number of trips was scaled up using the sampling ratio. This helped to estimate the real number of trips for each hour. The busiest hour after scaling still remained the same, confirming the pattern

3.2.4. Compare hourly traffic on weekdays and weekends



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

Top 10 pickup zones

	PULocationID	pickup_trip_count	LocationID	zone
0	132	15154	132	JFK Airport
1	237	13830	237	Upper East Side South
2	161	13553	161	Midtown Center
3	236	12296	236	Upper East Side North
4	162	10341	162	Midtown East
5	138	10054	138	LaGuardia Airport
6	186	9990	186	Penn Station/Madison Sq West
7	230	9701	230	Times Sq/Theatre District
8	142	9665	142	Lincoln Square East
9	170	8719	170	Murray Hill

Top 10 drop zones

	DOLocationID	drop_trip_count	LocationID	zone
0	236	12874	236	Upper East Side North
1	237	12379	237	Upper East Side South
2	161	11226	161	Midtown Center
3	230	8885	230	Times Sq/Theatre District
4	170	8518	170	Murray Hill
5	162	8303	162	Midtown East
6	142	8225	142	Lincoln Square East
7	239	8120	239	Upper West Side South
8	141	7682	141	Lenox Hill West
9	68	7325	68	East Chelsea

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

Top 10 Pickup/Dropoff Ratio Zones:

	zone	pickup_count	dropoff_count	ratio
0	JFK Airport	15154	15154	1.0
121	Soundview/Castle Hill	20	20	1.0
153	Cambria Heights	12	12	1.0
154	Inwood	12	12	1.0
155	Roosevelt Island	12	12	1.0
156	Brighton Beach	12	12	1.0
157	Hammels/Arverne	11	11	1.0
158	Eastchester	11	11	1.0
159	Claremont/Bathgate	11	11	1.0
160	Bellerose	11	11	1.0

Bottom 10 Pickup/Dropoff Ratio Zones:

	zone	pickup_count	dropoff_count	ratio
0	JFK Airport	15154	15154	1.0
152	Spuyten Duyvil/Kingsbridge	12	12	1.0
153	Cambria Heights	12	12	1.0
154	Inwood	12	12	1.0
155	Roosevelt Island	12	12	1.0
156	Brighton Beach	12	12	1.0
157	Hammels/Arverne	11	11	1.0
158	Eastchester	11	11	1.0
159	Claremont/Bathgate	11	11	1.0
160	Bellerose	11	11	1.0

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

Top 10 Pickup Zones during Night Hours:

	zone	pickup_count
0	East Village	2413
1	JFK Airport	2179
2	West Village	1991
3	Clinton East	1650
4	Lower East Side	1479
5	Greenwich Village South	1433
6	Times Sq/Theatre District	1282
7	Penn Station/Madison Sq West	1105
8	Midtown South	909
9	East Chelsea	908

Top 10 Dropoff Zones during Night Hours:

	zone	dropoff_count
0	East Village	1290
1	Clinton East	1113
2	Murray Hill	964
3	Gramercy	926
4	East Chelsea	891
5	Lenox Hill West	866
6	Yorkville West	818
7	West Village	766
8	Times Sq/Theatre District	723
9	Upper West Side South	699

3.2.8. Find the revenue share for nighttime and daytime hours

Night time revenue share: 11.88%

Day time revenue share: 88.12%

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

Average Fare per Mile per Passenger:

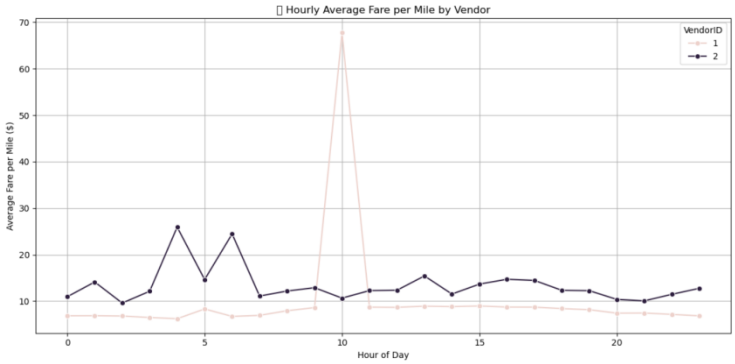
	passenger_count	fare_per_mile_per_passenger
0	1.0	11.546014
1	2.0	6.874245
2	3.0	4.531218
3	4.0	5.386175
4	5.0	1.837321
5	6.0	1.324068

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

Average Fare per Mile by Hour of the Day:

	pickup_hour	fare_per_mile
0	0	9.996458
1	1	12.449127
2	2	8.967867
3	3	10.827387
4	4	20.877157
5	5	12.816341
6	6	18.957024
7	7	9.846440
8	8	10.945063
9	9	11.669955
10	10	26.965271
11	11	11.255365
12	12	11.310190
13	13	13.610679
14	14	10.782792
15	15	12.348694
16	16	13.072404
17	17	12.904304
18	18	11.306220
19	19	11.209213
20	20	9.640957
21	21	9.458627
22	22	10.455081
23	23	11.361589

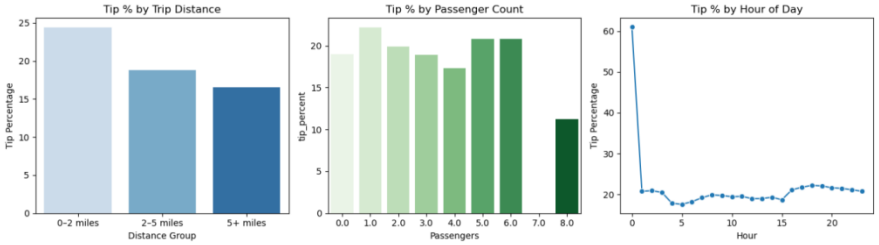
3.2.11. Analyse the average fare per mile for the different vendors



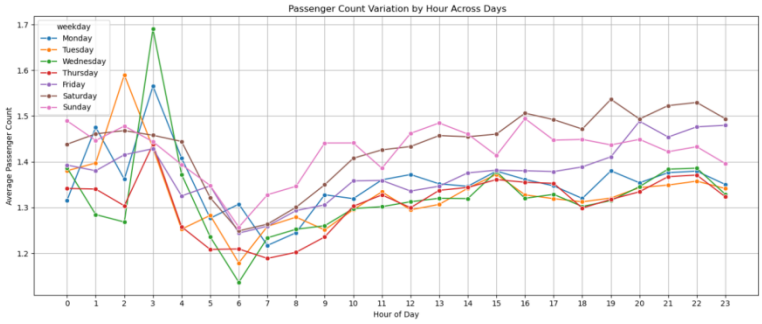
3.2.12. Compare the fare rates of different vendors in a distance-tiered fashion

	VendorID	distance_category	fare_per_mile
0	1	0-2 miles	15.084266
1	1	2-5 miles	6.393871
2	1	5+ miles	4.435301
3	2	0-2 miles	18.372741
4	2	2-5 miles	6.546712
5	2	5+ miles	4.501484

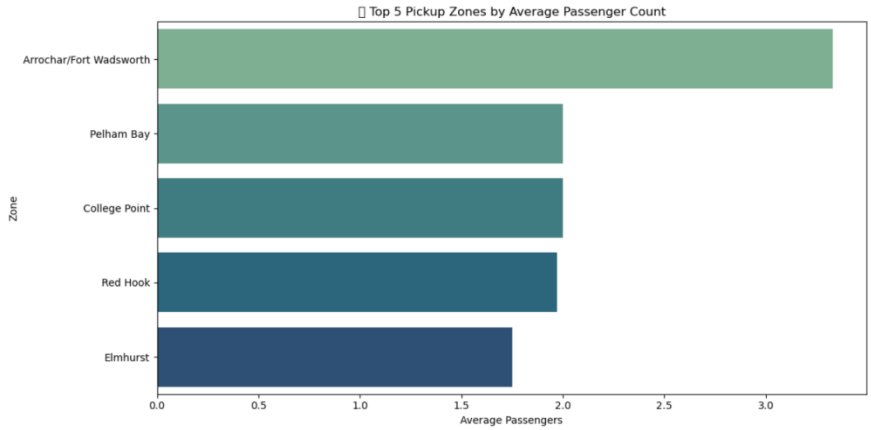
3.2.13. Analyse the tip percentages



3.2.14. Analyse the trends in passenger count



3.2.15. Analyse the variation of passenger counts across zones



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

Frequency of Each Surcharge Being Applied:

	Applied_Count
improvement_surcharge	265591
mta_tax	263287
congestion_surcharge	245322
extra	165252

4. Conclusions

4.1. Final Insights and Recommendations

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

need to increase in cab between 5–8 PM on weekdays by deploying more as peak demand occurs between 5pm and 8pm

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.

Demand is highly deployed around business hours. Avoid over-supplying cabs during mid-day in low-demand business zones.

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

Implement or dynamic surcharges during peak hours and weekends when demand is more than supply.