MTH209

Optimal Taxi Business Management Plan

Alpha Squad

Indraj Prajapat (C), MSc Statistics, (231080044)
Chandan Kumar Singh, MSc Statistics, (231080030)
Gaurav Tomar, MSc Statistics, (231080039)
Ketan Saini, BS SDS, (220523)
Vandan Neema, BS SDS, (221165)

Instructor :- **Subhajit Dutta**

ABOUT OUR PROJECT

Our project is committed to equipping taxi businesses with actionable insights for effective resource allocation, route optimization, and pricing strategies through an approach of data analysis. Our ultimate goal is to transform traditional taxi operations into data-driven, agile businesses that can adapt and grow in a constantly evolving market.

Our objective is to identify patterns and trends in the dataset using Descriptive Statistics, make inference and also do testing on it. This will enable a deeper understanding of the factors influencing trip dynamics and station load.

Our Dataset

We scrapped our dataset from : https://www.nyc.gov

The dataset has 19 Columns & 3066766 Rows.

It contains detailed trip-level data of taxi rides in New York City.

Key fields include **VendorID** (indicating the TPEP provider), **tpep_pickup_datetime** and **tpep_dropoff_datetime** (start and end times of the trip), **Passenger_count** (number of passengers), and **Trip_distance** (trip distance in miles).

It also includes location IDs for where the trip started and ended (PULocationID and DOLocationID), the final rate code (RateCodeID), and a flag indicating whether the trip record was held in vehicle memory before sending to the vendor (Store_and_forward_flag).

Payment details are also included, such as payment type, fare amount, extra charges, tip amount, tolls amount, total amount charged to passengers, and the congestion surcharge. This dataset provides a comprehensive view of taxi usage patterns in New York City.

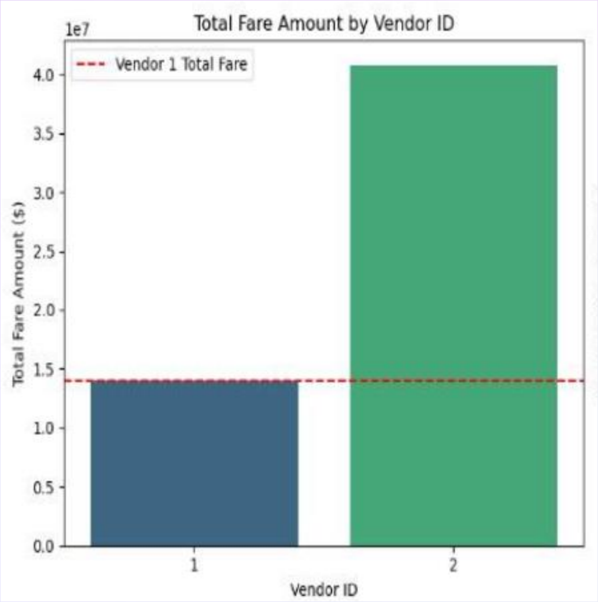
Data Cleaning

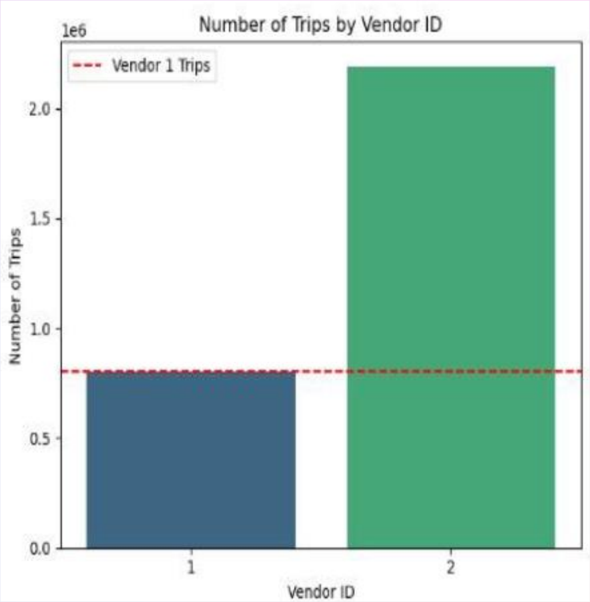
Total Fare Amount: Show that while VendorID 2 handles more trips, VendorID 1 generates a comparable total fare amount.

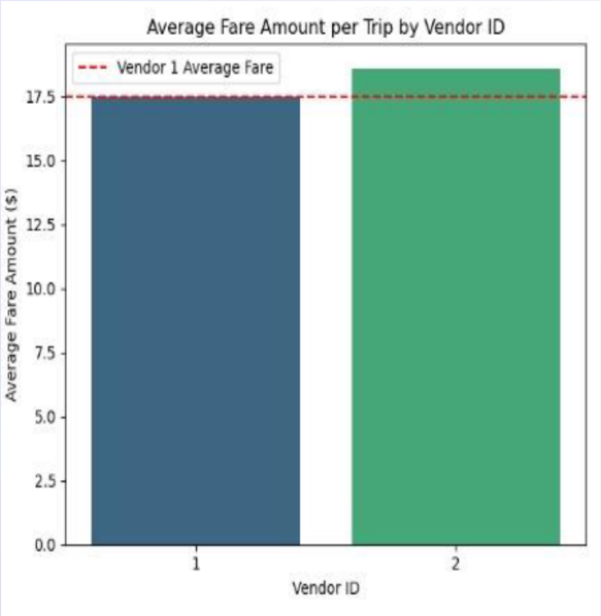
Average Fare Amount per Trip: Highlight that VendorID 1 has a higher average fare amount per trip compared to VendorID 2.

Number of Trips: Emphasize that although VendorID 1 serves fewer trips, it still generates a significant total fare amount due to the higher average fare per trip.

Finally, we chose VendorID 1 that contains 802710 rows.



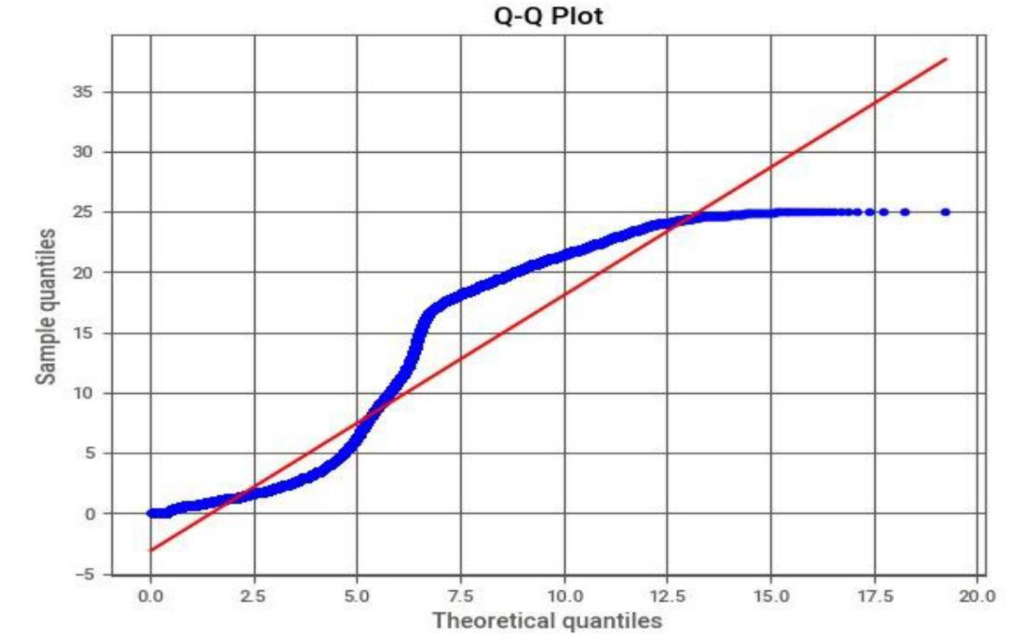


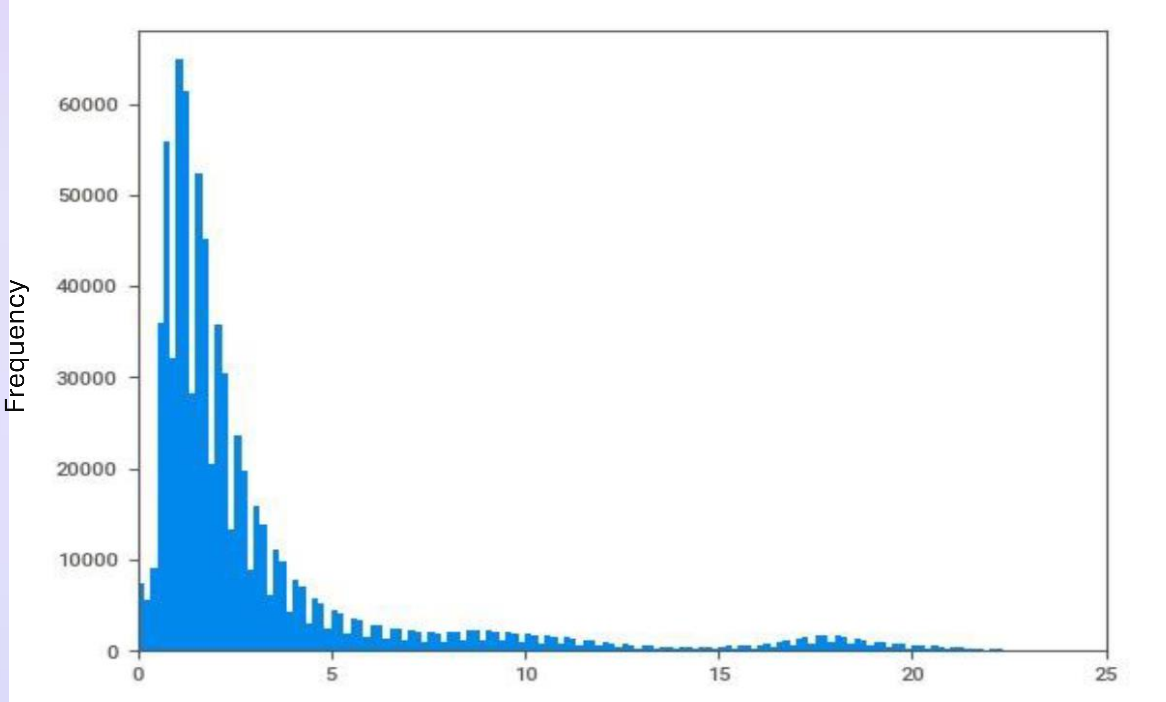


Data Visualisation

Trip Distance

Summary Statistics:
Mean: 3.226338205020213
Median: 1.8
Mode: 1
Minimum: 0.0
Maximum: 25.0
Range: 25.0
Variance: 16.271971044816883
Standard Deviation: 4.033853126331806
95% Confidence Interval for Trip Distance:
(3.2684261174955616, 3.2874082002951344)





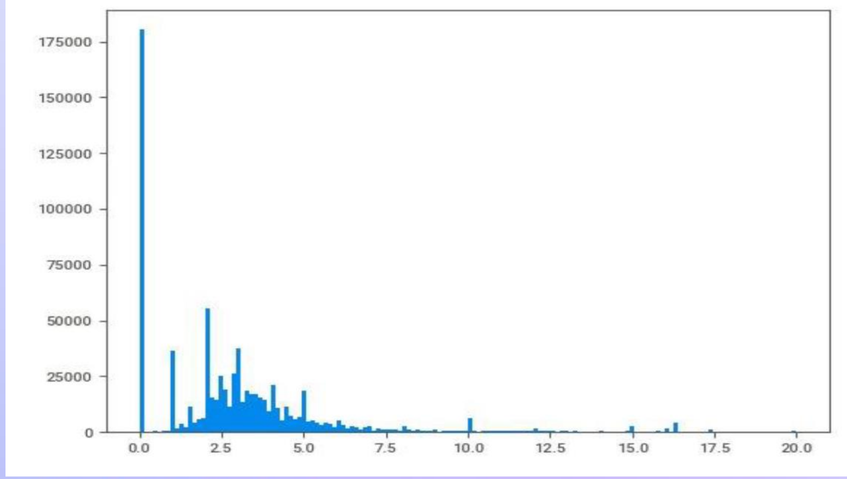
Frequency

Trip distance

Data Visualisation

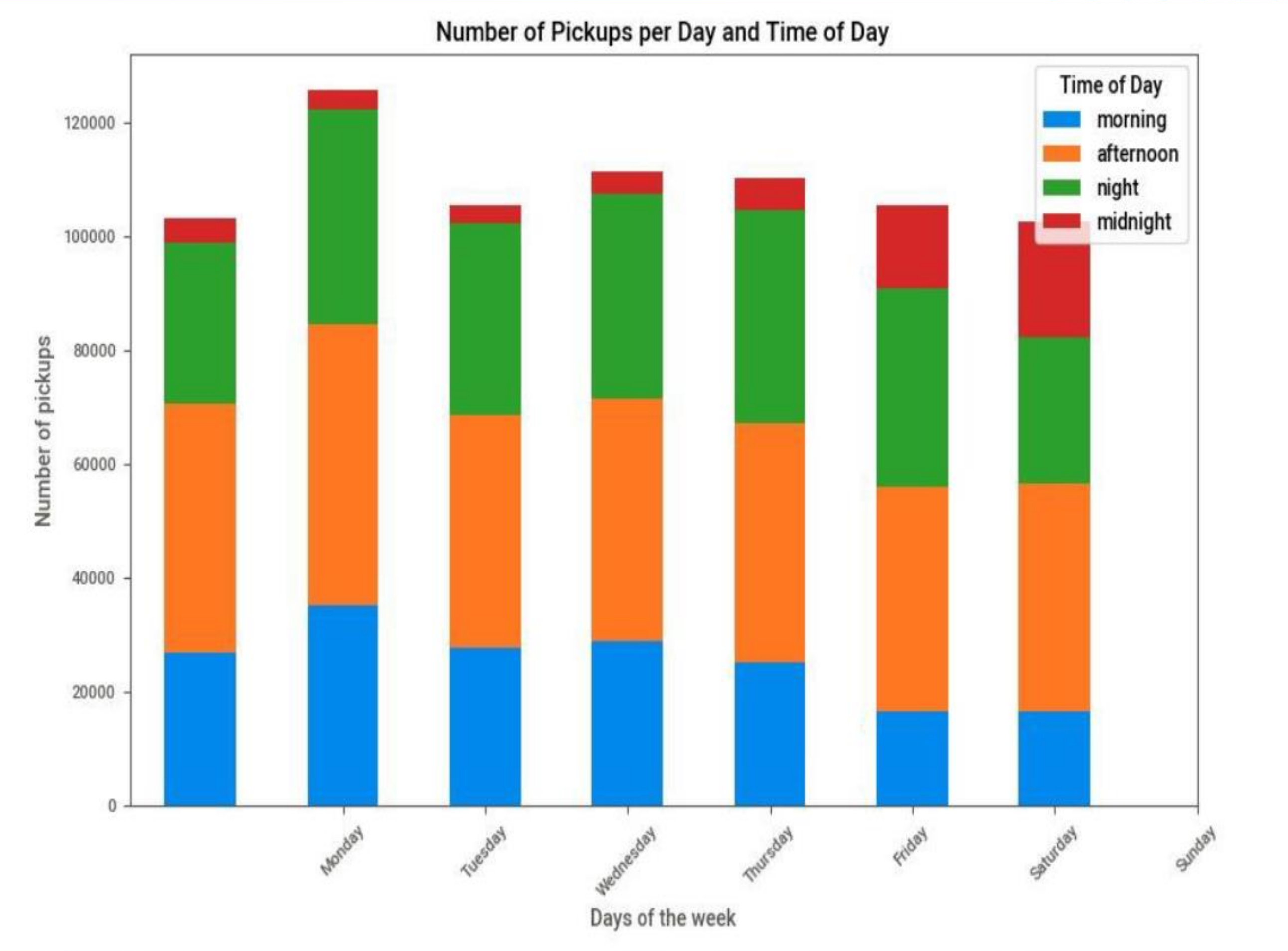
Summary Statistics (Tip Amount):
Mean: 3.079842405557762
Median: 2.65
Mode: 0
Minimum: 0.0
Maximum: 20.0
Range: 20.0
Variance: 10.094309884617488
Standard Deviation: 3.1771543690254473

Frequency



Tip Amount

From the graph we see that, We get most no. of pickups during daytime and lowest during midnight for all the days of the week.



Hypothesis Testing

Trip Distance

Null hypothesis : There is no significant difference in average trip distance between weekdays and weekends.

Alternative hypothesis : There is significant difference in average trip distance between weekdays and weekends.

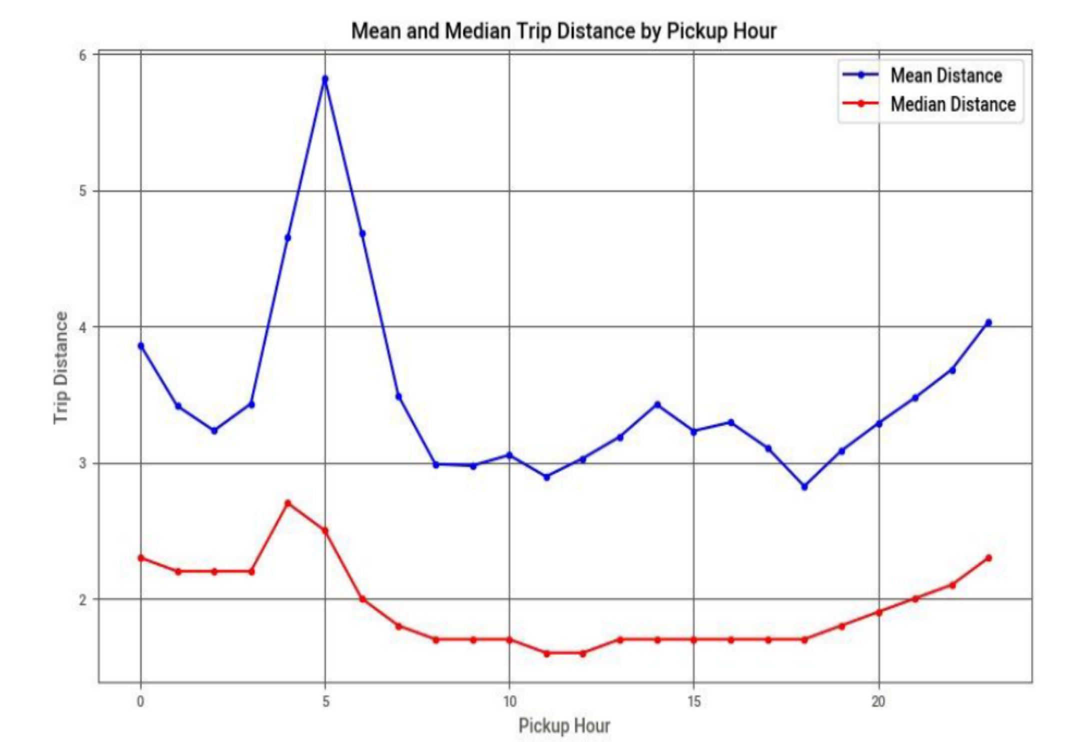
Here we are comparing the mean trip distance between weekday and weekends, and we do not know the variance of the population (Trip Distance) that why we use t-test here and also calculate the p-value.
We take alpha (Significance level) = 0.05.

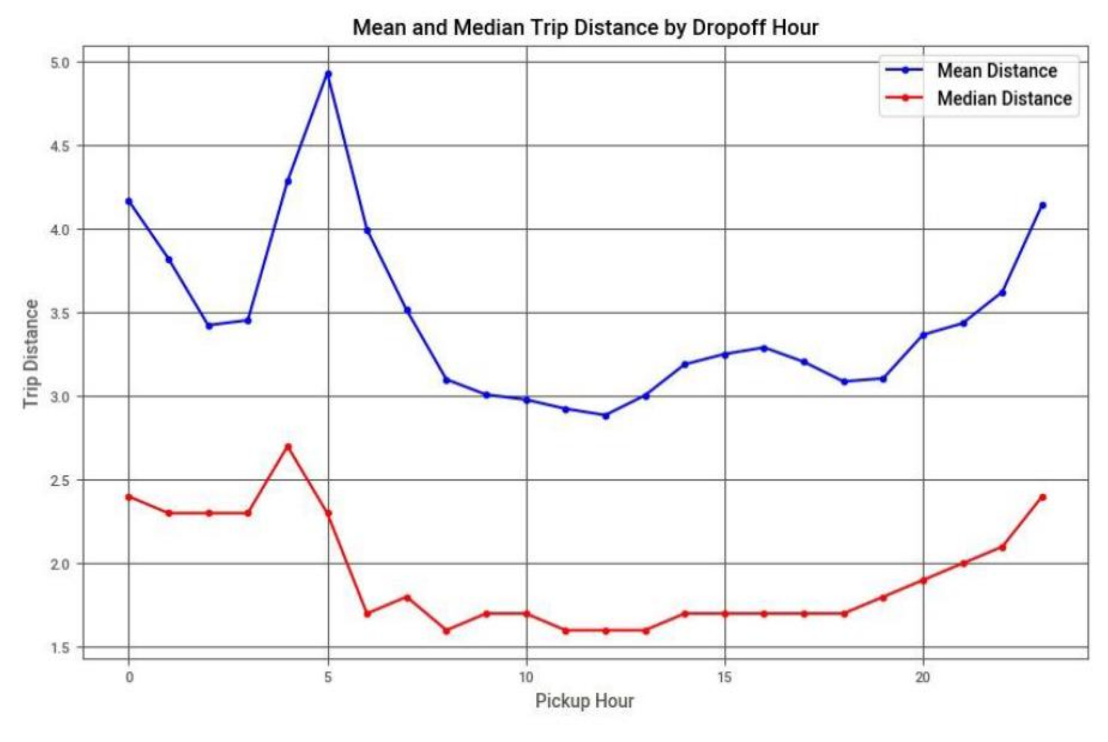
We find :

T-statistic: -8.736723447841165
P-value: 2.409194434789374e-18
Here we can see that p-value is nearly 0 < alpha = 0.05,
Critical value of t for one tailed test (alpha 0.05, df > 1000) is 1.645.

Thus, we conclude that there is a significant difference in trip distance between weekdays and weekends by rejecting the null hypothesis.

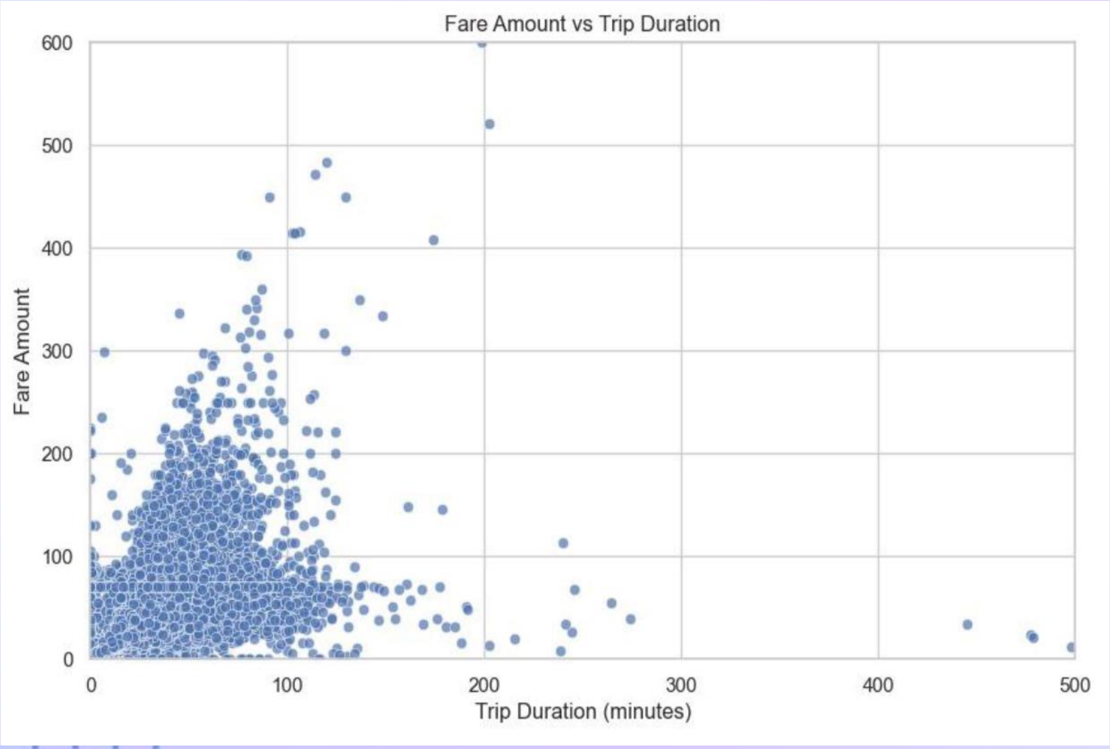
Mean and Median Trip Distance

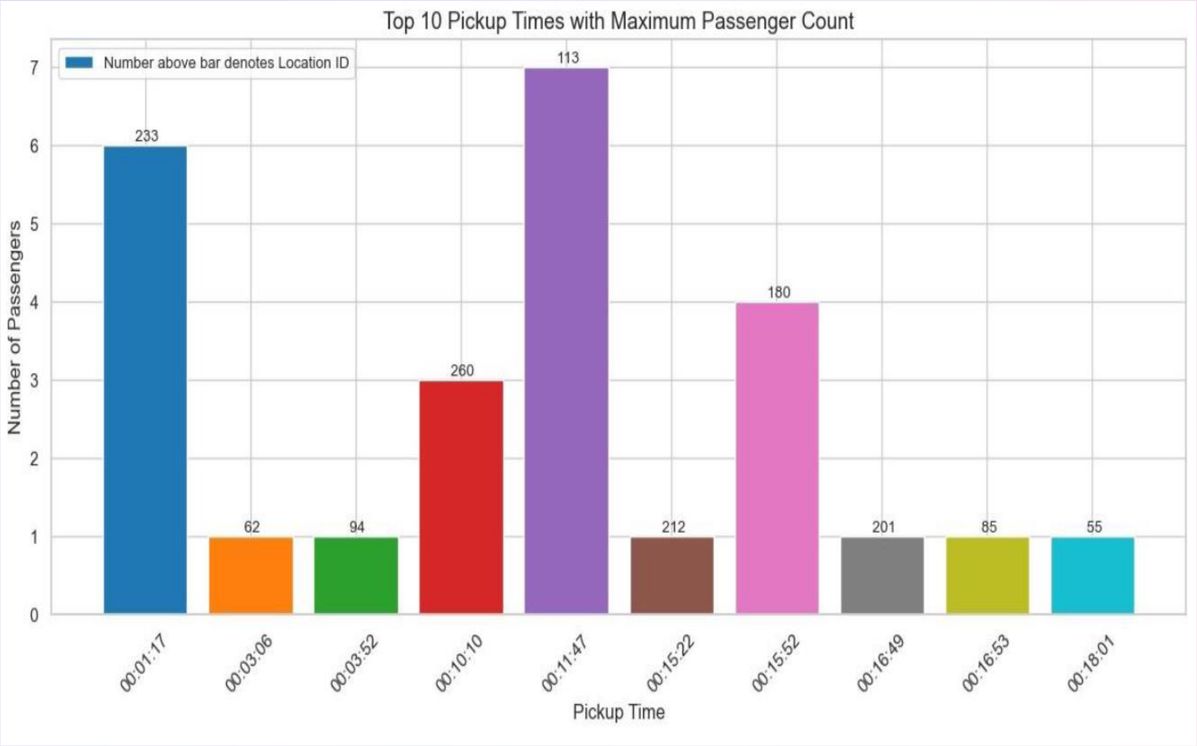




Trips started during 11:00 AM to 1:00 PM tended to be the shortest, and trips started between 4:00 AM and 6:00 AM were the longest. The surge in long-distance trips during the morning is likely driven by trips to the airports or other long-distance rides.

Pickup Time





Most of the trips are under 100 minute long

This plot show the station at which maximum passenger are recognized at specific time

Checking condition for Poisson Process

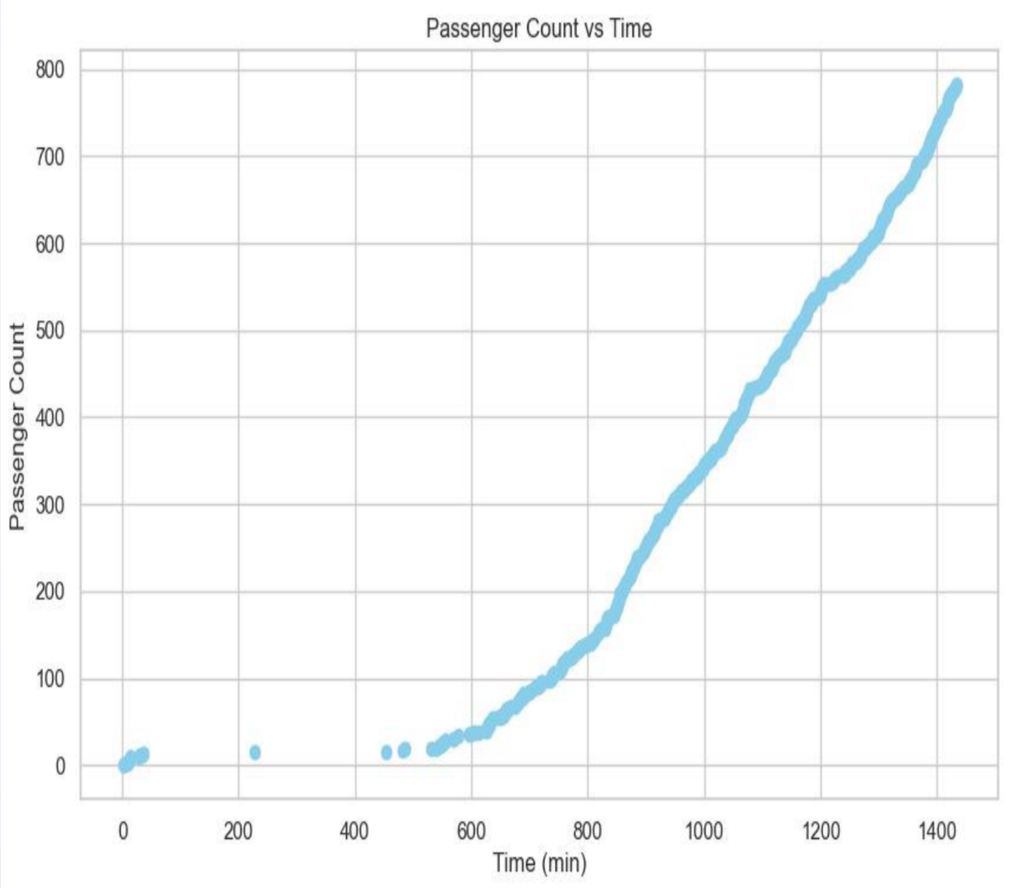
First, we assume that the passenger recognized when they got the taxi. We converted the column in such a way that we get counting of passenger.

1. $N(0) = 0$;
2. $N(t)$ has <u>independent</u> increments;
3. for any $t \in [0, \infty)$, we have

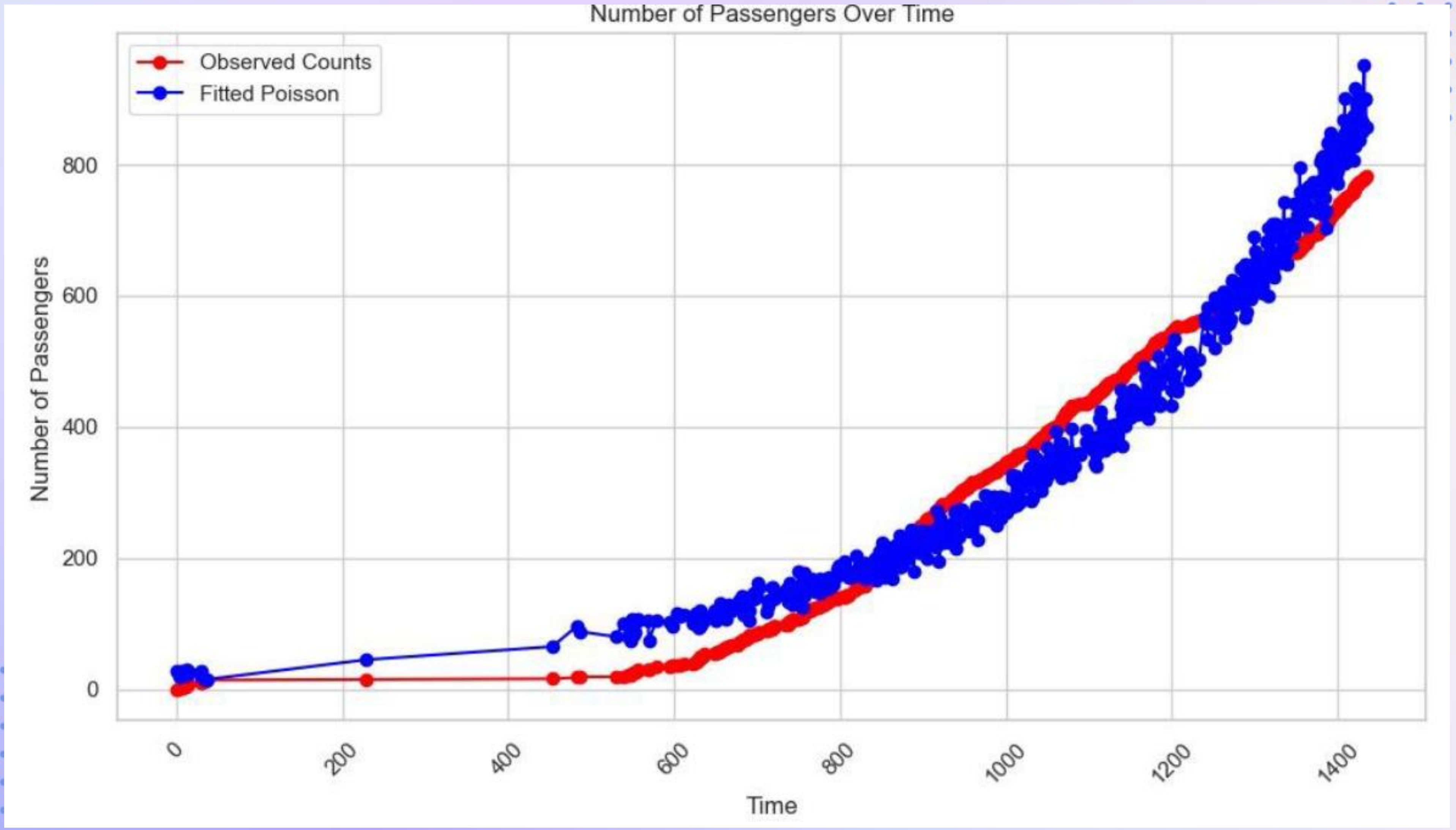
$$P(N(t+\Delta) - N(t) = 0) = 1 - \lambda(t)\Delta + o(\Delta),$$
$$P(N(t+\Delta) - N(t) = 1) = \lambda(t)\Delta + o(\Delta),$$
$$P(N(t+\Delta) - N(t) \geq 2) = o(\Delta).$$

$A \subseteq [0,\infty)$ $N(A)$ has the Poisson distribution with parameter $m(A)$

- First condition for poisson process is satisfied i.e. $N(0) = 0$

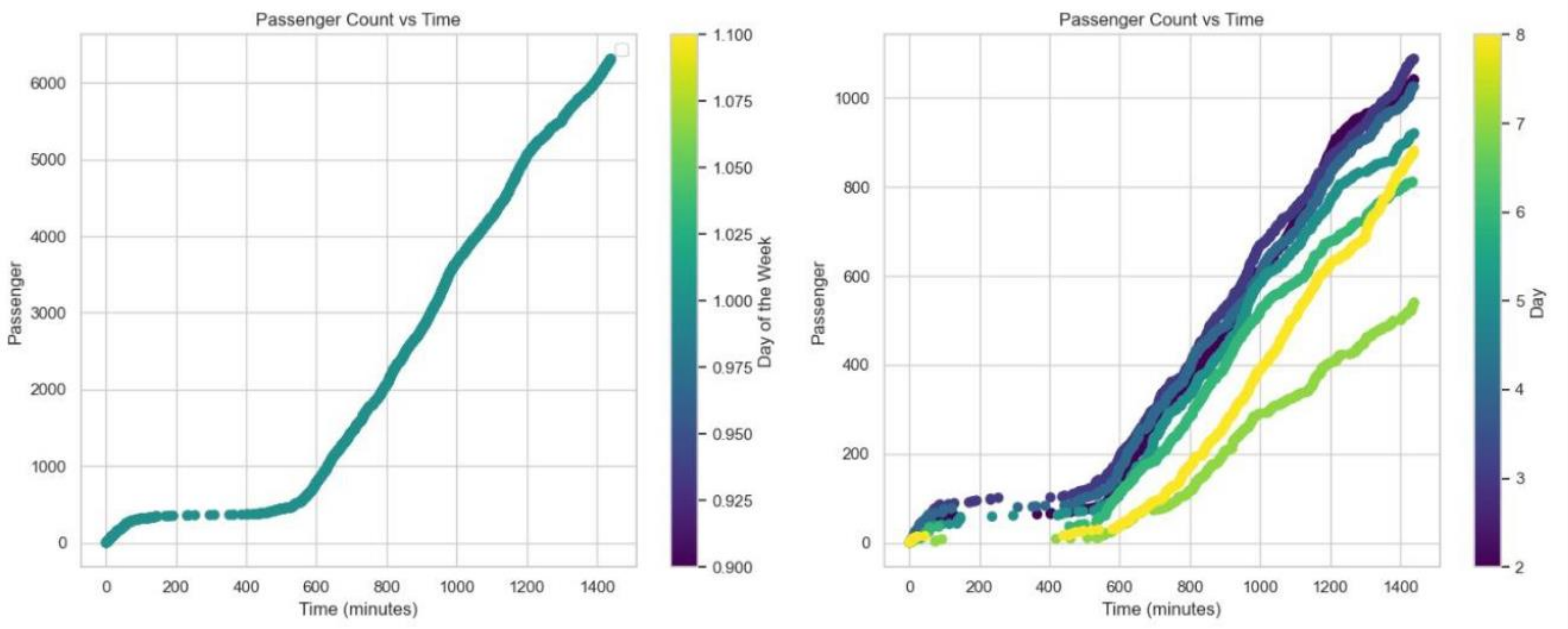


- If we fit poisson distribution then the observation will be as in graph



Estimating $\lambda(t)$ for non-homogenous poisson process

We estimated the rate using MLE and checked whether we estimate for a day or a week



If we estimate for a week, then by graph we can see that at time near 1400 (minute) we can see nearly 6000/7~ 900 counting is done but if we see for a day the difference is too much

Simulation

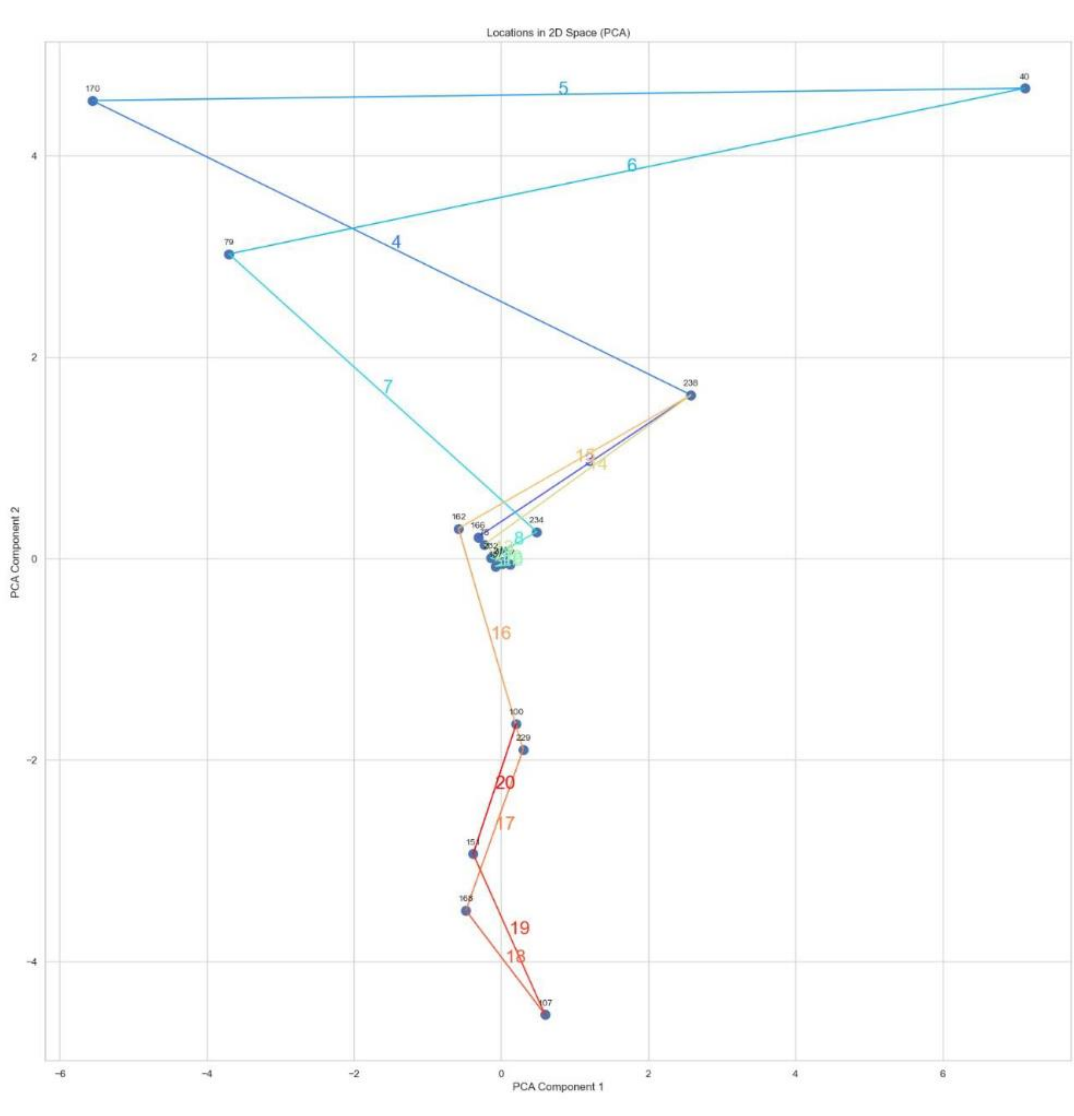
We make a function in which fare amount only depends on trip distance.
If we didn't get passenger we will wait

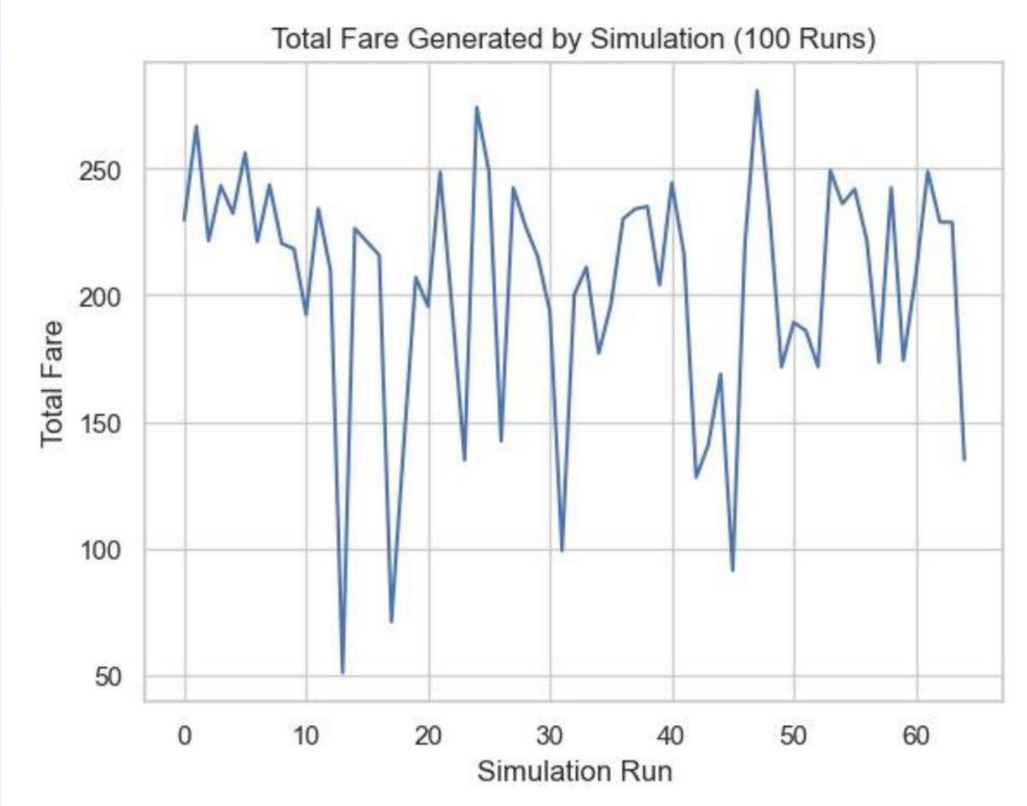
	PULocationID	DOLocationID	station_time	trip_distance	pass	fare
0	145	146	2.766667	0.5	1	4.090
1	146	161	10.500000	3.1	1	9.758
2	161	186	10.750000	1.5	1	6.270
3	186	107	8.983333	0.9	1	4.962
4	107	113	3.466667	0.5	1	4.090
...	...	...	...	...	...	...
186	53	53	356.700000	0.0	0	0.000
187	53	53	357.700000	0.0	0	0.000
188	53	53	358.700000	0.0	0	0.000
189	53	53	359.700000	0.0	0	0.000
190	53	53	360.700000	0.0	0	0.000

Now, optimizing for max profit with respect to Waiting or Moving to next station

	PULocationID	DOLocationID	station_time	trip_distance	pass
0	174	243	20.816667	0.0	1
1	243	244	10.916667	2.8	1
2	244	243	12.133333	1.7	0
3	243	166	14.516667	0.0	1
4	166	238	9.833333	1.4	1
5	238	170	27.700000	3.8	1
6	170	40	23.766667	7.6	1
7	40	79	12.200000	5.9	1
8	79	234	8.266667	1.1	1
9	234	232	15.016667	2.1	1
10	232	37	18.433333	4.5	1
11	37	137	45.866667	6.7	1
12	137	141	8.850000	2.1	1
13	141	74	15.716667	3.0	1
14	74	75	8.266667	0.9	1
15	75	238	9.933333	1.0	1

Visualizing Optimum Route





From the above graph,
We conclude that per day profit is around 193$

Conclusion

Basic Expenses and Rates (Current)

- According to the data acquired from nyc.gov, preferred car by taxi company cost around 29000$. Mileage of car is around 18$
- Average salary of a cab driver is around 130$/day in NYC
- Cost of Gas per Gallon averages around 4$

Our Prediction

- Mean Revenue is 193$/Day (Single Taxi).
- Average Distance travelled by a taxi is 70 miles.
- 70/18 (Fuel Consumption) x 4$ = 15.56$
- Daily Profit : 47.44$ (Removed Salary of Cab Driver + Fuel Charges)

Annual Profits

- 1423.2$ (Monthly Profit) x 12 = 17078.4$

Annual Maintenance Cost

- 2000$

THANK YOU