



# NYC TAXI TRIPS



HIGH SEASON OF 2024 IS OCTOBER - DECEMBER

Presented by สปาเกตตี้พัสดุพริกแห้ง



# INTRODUCING TO NYC TAXI

ใช้ข้อมูล yellow taxi ปี 2024 ไตรมาสที่ 4 ได้แก่เดือน ตุลาคม พฤศจิกายน และ ธันวาคม เนื่องจากเป็นช่วง High season ของเมือง NYC และใช้ข้อมูล geo graphy เพื่อมาดู visualization

OCT 3,833,771 records, 19 columns

VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag
i32	datetime[ns]	datetime[ns]	i64	f64	i64	str
2	2024-10-01 00:30:44	2024-10-01 00:48:26	1	3.0	1	"N"
1	2024-10-01 00:12:20	2024-10-01 00:25:25	1	2.2	1	"N"
1	2024-10-01 00:04:46	2024-10-01 00:13:52	1	2.7	1	"N"
1	2024-10-01 00:12:10	2024-10-01 00:23:01	1	3.1		
1	2024-10-01 00:30:22	2024-10-01 00:30:39	1	0.0		
...	...	...	...	...		
2	2024-10-31 23:49:01	2024-11-01 00:04:31	null	3.49		
2	2024-10-31 23:35:15	2024-10-31 23:52:50	null	2.4		
2	2024-10-31 23:30:43	2024-11-01 00:08:12	null	12.28		
2	2024-10-31 23:00:00	2024-10-31 23:06:00	null	0.56		
2	2024-10-31 23:18:00	2024-10-31 23:51:00	null	6.25		

NOV 3,646,369 records, 19 columns

DEC 3,668,371 records, 19 columns

VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	store_and_fwd_flag	i32	datetime[μs]	datetime[μs]	i64	f64	i64	str
2	2024-12-01 00:12:27	2024-12-01 00:31:12	1	9.76	1	"N"	2	2024-12-01 00:12:27	2024-12-01 00:31:12	1	9.76	1	"N"
2	2024-11-30 23:56:04	2024-12-01 00:28:15	1	7.62	1	"N"	2	2024-12-01 00:50:35	2024-12-01 01:24:46	4	20.07	2	"N"
2	2024-12-01 00:18:16	2024-12-01 00:33:16	3	2.34	1	"N"	2	2024-12-01 00:18:16	2024-12-01 00:33:16	3	2.34	1	"N"
2024-12-01 01:18:25			1	5.05	1	"N"	2024-12-01 01:18:25			1	5.05	1	"N"
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2024-12-31 23:56:00			null	10.71	null	null	2024-12-31 23:56:00			null	10.71	null	null
2024-12-31 23:18:00			null	4.56	null	null	2024-12-31 23:18:00			null	4.56	null	null
2024-12-31 23:28:35			null	3.94	null	null	2024-12-31 23:28:35			null	3.94	null	null
2024-12-31 23:36:29			null	4.2	null	null	2024-12-31 23:36:29			null	4.2	null	null
2024-12-31 23:33:34			null	5.76	null	null	2024-12-31 23:33:34			null	5.76	null	null
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2024-11-30 23:11:15	2024-11-30 23:19:33	null	1.09	null	null	null	2024-11-30 23:11:15	2024-11-30 23:19:33	null	1.09	null	null	null
2024-11-30 23:49:30	2024-12-01 00:27:39	null	20.1	null	null	null	2024-11-30 23:49:30	2024-12-01 00:27:39	null	20.1	null	null	null
2024-11-30 23:31:46	2024-12-01 00:04:32	null	1.38	null	null	null	2024-11-30 23:31:46	2024-12-01 00:04:32	null	1.38	null	null	null
2024-11-30 23:41:21	2024-11-30 23:53:20	null	2.63	null	null	null	2024-11-30 23:41:21	2024-11-30 23:53:20	null	2.63	null	null	null
2024-11-30 23:21:52	2024-11-30 23:21:11	null	1.16	null	null	null	2024-11-30 23:21:52	2024-11-30 23:21:11	null	1.16	null	null	null

Total 11,148,511 records before clean

# DATA

# YELLOW TAXI 2024



	VENDORID	TPEP_PICKUP_DATETIME	TPEP_DROPOFF_DATETIME	PASSENGER_COUNT	TRIP_DISTANCE	RATECODEID	PULOCATIONID	DOLOCATIONID
1	1	2024-10-01 00:12:20	2024-10-01 00:25:25	10	2.20	1.0	48	236

FARE_AMOUNT	TIP_AMOUNT	TOLLS_AMOUNT	CONGESTION_SURCHARGE	AIRPORT_FEE
14.20	3.80	0.0	2.5	0.0

# MOTIVATION

- ปัญหา/คำถามที่สนใจคืออะไร  
บริษัทแท็กซี่ประสบปัญหาการคิดราคาค่าโดยสารทำให้เสียโอกาสและขาดทุน หากเราจึงต้องการพัฒนาโมเดลที่สามารถทำนายค่าโดยสารได้ และอีกเหตุผลคือบริษัทแท็กซี่ไม่ทราบว่าควรจะจัดแท็กซี่ในโซนไหน ช่วงวันและเวลาใดเพื่อเพียงพอต่อความต้องการของผู้บริโภค  
คำถาม: ราคาค่าโดยสารขึ้นอยู่กับปัจจัยใดบ้าง และควรเพิ่มจำนวนแท็กซี่ที่ zone ใด
  - ปัญหา/คำถามที่สนใจมีความสำคัญอย่างไร?ในส่วนการทำนายค่าโดยสารเราสามารถนำโมเดลนี้ไปต่อยอด พัฒนาเป็นแอปเรียก taxi ได้ในอนาคต นอกจากนี้บริษัทยังขาดข้อมูลเชิงลึกในการวางแผน เพื่อจัดสรรจำนวนแท็กซี่ให้เหมาะสมกับพื้นที่และช่วงเวลา กลุ่มเราระบุแนวคิดในการพัฒนาโมเดลทำนายค่าโดยสาร เพื่อช่วยเพิ่มประสิทธิภาพในการกำหนดราคาและการจัดสรรทรัพยากรให้สอดคล้องกัน

# MOTIVATION

- ข้อคาดการณ์ของท่านจะนำไปสู่การแก้ไขปัญหารือตอบคำถาม
- 1.วันใดในสัปดาห์ที่เป็นที่นิยมในการเรียกใช้บริการแท็กซี่
  - 2.ช่วงเวลา peak time ใน 24 hours
  - 3.เดือนที่มีการเดินทางมากที่สุด
  - 4.นับจำนวนการเดินทางต่อเขต เพื่อดูว่าเขตไหนมีการเดินทางมากที่สุด
  - 5.สถานที่ที่มีการปรับและไปส่งบ่อยๆ
  - 6.เดือนที่มีการเดินทางมากที่สุด



# DATA PREPARATION

จัดการค่าซ้ำ (Duplicates) 103 rows ✓

จัดการค่าว่างแต่ละคอลัมน์ ✓ #เติมmode

จัดการ outliers ✓

จัดการ noise ✓ #filtering

tpep_pickup_datetime	0
tpep_dropoff_datetime	0
passenger_count	1093860
trip_distance	0
RatecodeID	1093860
PULocationID	0
DOLocationID	0
fare_amount	0
tip_amount	0
P_place	29214
D_place	47161



tpep_pickup_datetime	0
tpep_dropoff_datetime	0
passenger_count	0
trip_distance	0
RatecodeID	0
PULocationID	0
DOLocationID	0
fare_amount	0
tip_amount	0
P_place	0
D_place	0

ก่อนกรอง: 11148408  
หลังกรอง: 10348975  
ตัดออก: 799433



# DATA

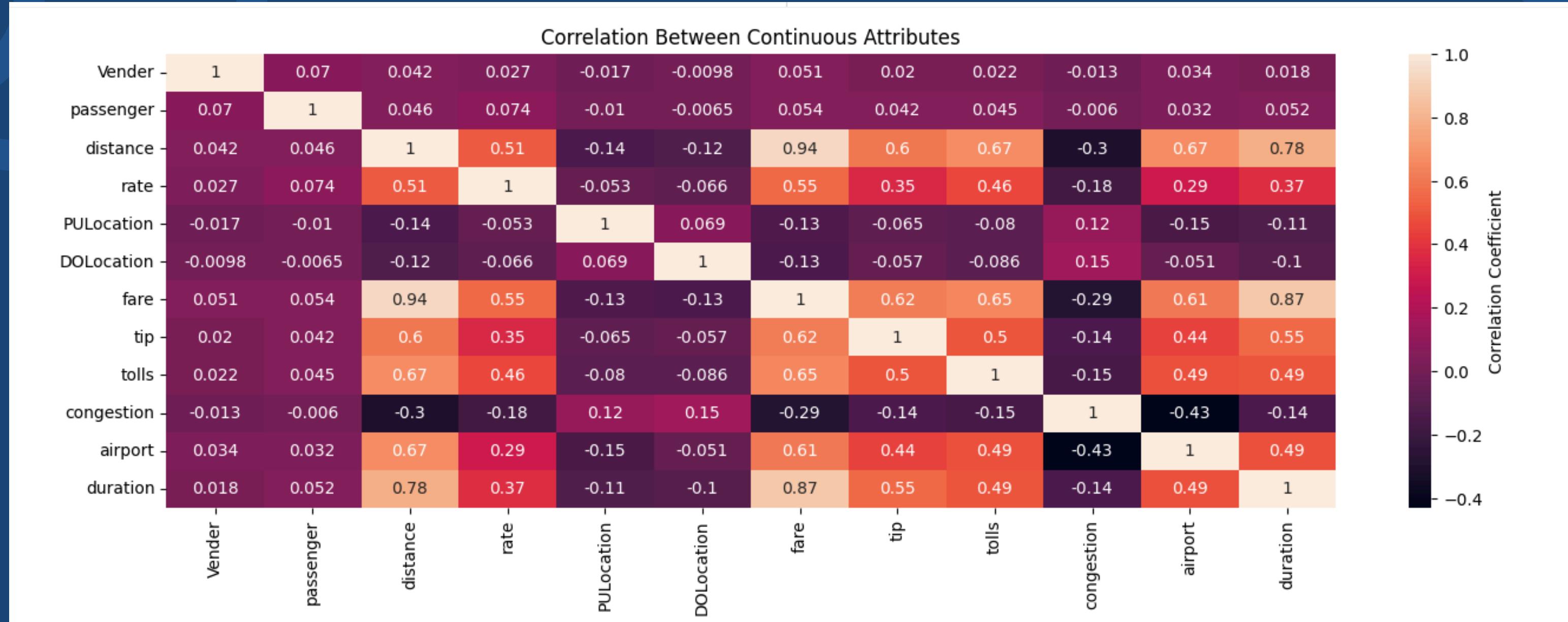
# VISUALIZATION ✨

## Descriptive statistics

	VendorID	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	trip_distance	RatecodeID	PULocationID	fare_amount	tip_amount	tolls_amount	congestion_surcharge	Airport_fee	date	time	duration
count	1.034898e+07	10348975	10348975	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	10348975	1.034898e+07	1.034898e+07
mean	1.781910e+00	2024-11-15 04:43:08.641376	2024-11-15 05:00:30.210602	1.314181e+00	3.283454e+00	1.053286e+00	1.652665e+02								
min	1.000000e+00	2002-12-31 22:17:43	2002-12-31 22:23:55	1.000000e+00	1.000000e-02	1.000000e+00	1.000000e+00								
25%	2.000000e+00	2024-10-23 23:33:38	2024-10-23 23:49:05.500000	1.000000e+00	1.030000e+00	1.000000e+00	1.320000e+02								
50%	2.000000e+00	2024-11-15 07:21:52	2024-11-15 07:36:57	1.000000e+00	1.740000e+00	1.000000e+00	1.620000e+02								
75%	2.000000e+00	2024-12-07 19:24:36	2024-12-07 19:44:35.500000	1.000000e+00	3.290000e+00	1.000000e+00	2.340000e+02								
max	6.000000e+00	2025-03-23 20:42:06	2025-03-23 22:52:56	6.000000e+00	9.690000e+01	6.000000e+00	2.630000e+02								
std	4.137415e-01	Nan	Nan	7.264837e-01	4.101254e+00	3.090792e-01	6.392425e+01	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	1.034898e+07	10348975	1.034898e+07	1.034898e+07
								1.969640e+01	3.502109e+00	5.502069e-01	2.379908e+00	1.309404e-01	2024-11-14 13:48:52.883496	1.440332e+01	1.735949e+01
								2.510000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2002-12-31 00:00:00	0.000000e+00	1.016667e+00
								9.300000e+00	7.200000e-01	0.000000e+00	2.500000e+00	0.000000e+00	2024-10-23 00:00:00	1.100000e+01	8.266666e+00
								1.420000e+01	2.800000e+00	0.000000e+00	2.500000e+00	0.000000e+00	2024-11-15 00:00:00	1.500000e+01	1.365000e+01
								2.260000e+01	4.480000e+00	0.000000e+00	2.500000e+00	0.000000e+00	2024-12-07 00:00:00	1.900000e+01	2.201667e+01
								2.500000e+02	5.725000e+02	1.500000e+02	2.500000e+00	1.750000e+00	2025-03-23 00:00:00	2.300000e+01	1.799000e+02
								1.614975e+01	3.991426e+00	2.023454e+00	6.440318e-01	4.478563e-01		Nan	5.838320e+00
															1.344583e+01



# TARGET VARIABLE ANALYSIS



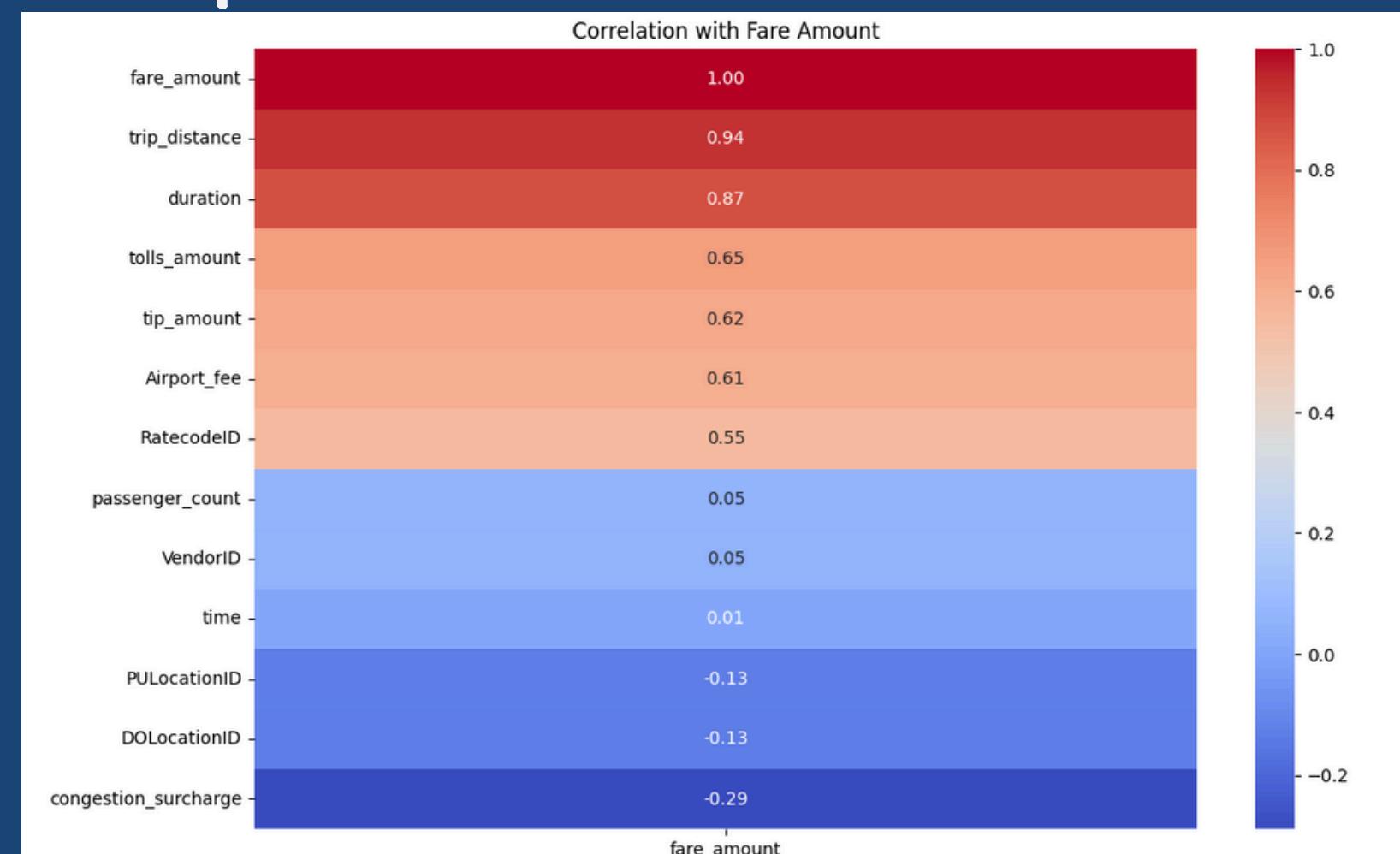
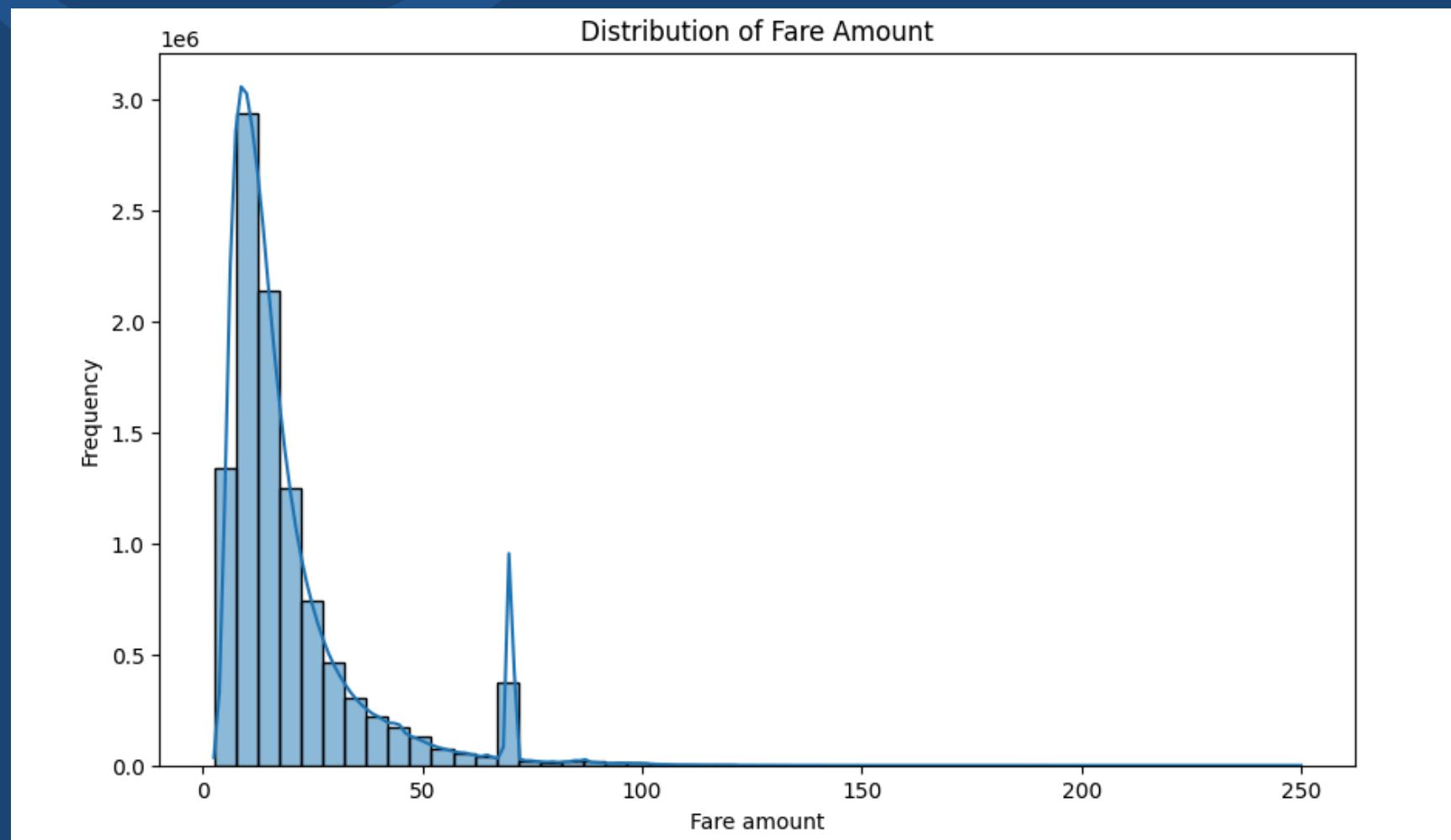
พบว่า fare\_amount (ค่าโดยสาร)

มีความสัมพันธ์เชิงบวกสูงมากกับ distance (0.94), duration (0.87), tolls\_amount (0.65)

และ tip\_amount (0.62) และแสดงว่า ระยะทางและระยะเวลาเดินทางเป็นปัจจัยหลักที่กำหนดค่าโดยสาร และการจ่ายทิปหรือต้องจ่ายค่าทางด่วนก็เพิ่มค่าโดยสารด้วย



# TARGET VARIABLE ANALYSIS

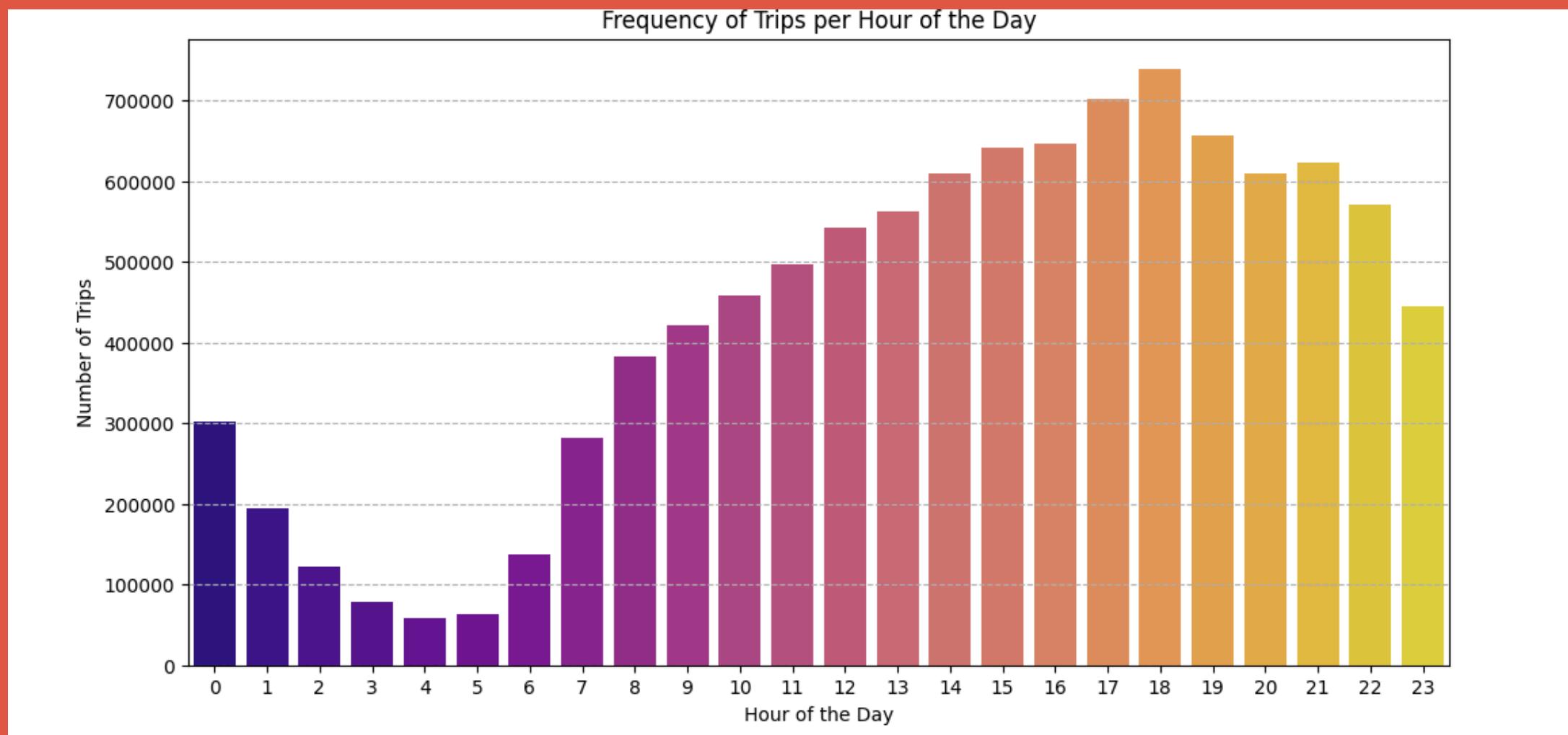


พบว่า กราฟ Fare\_amount (ค่าโดยสาร) เป็นขวัญชัดเจน และแสดงว่าส่วนใหญ่ของค่าโดยสารจะอยู่ในช่วงราคาต่ำ อยู่ระหว่าง 5 – 25 ดอลลาร์ และพบบ่อยที่สุดอยู่ที่ประมาณ 8-10 ดอลลาร์



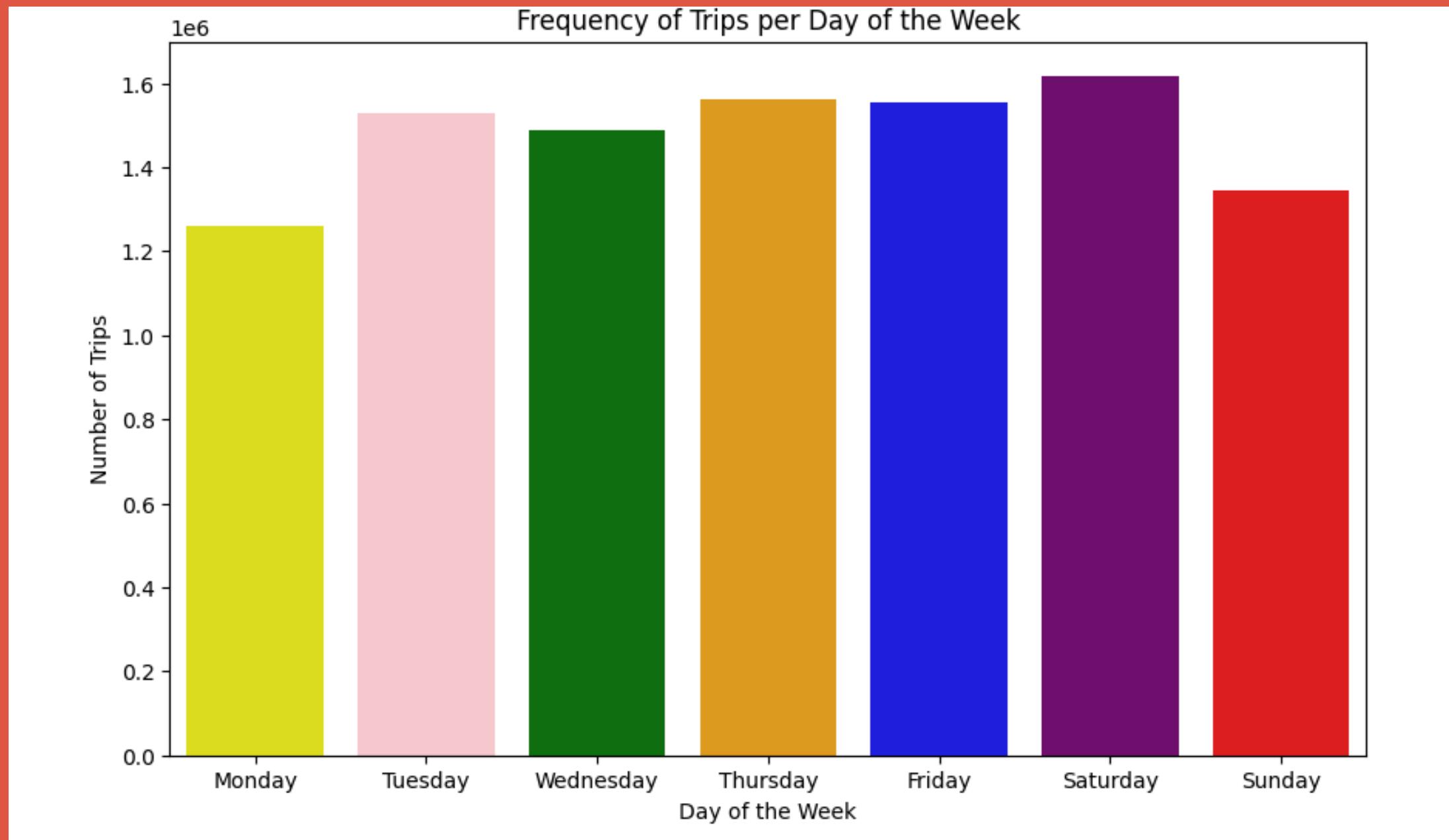
# DATA VISUALIZATION

## peak hour in 24 hours



พบว่า peak hour ช่วงที่มีการเดินทาง  
มากที่สุดคือช่วงเวลา 18.00 p.m.  
อีกทั้งช่วงเวลาตอนเย็น 16.00 - 20.00  
p.m. เป็นช่วงที่มีการเรียกใช้บริการ  
มากที่สุด

# FREQUENCY OF TRIPS PER WEEKDAY

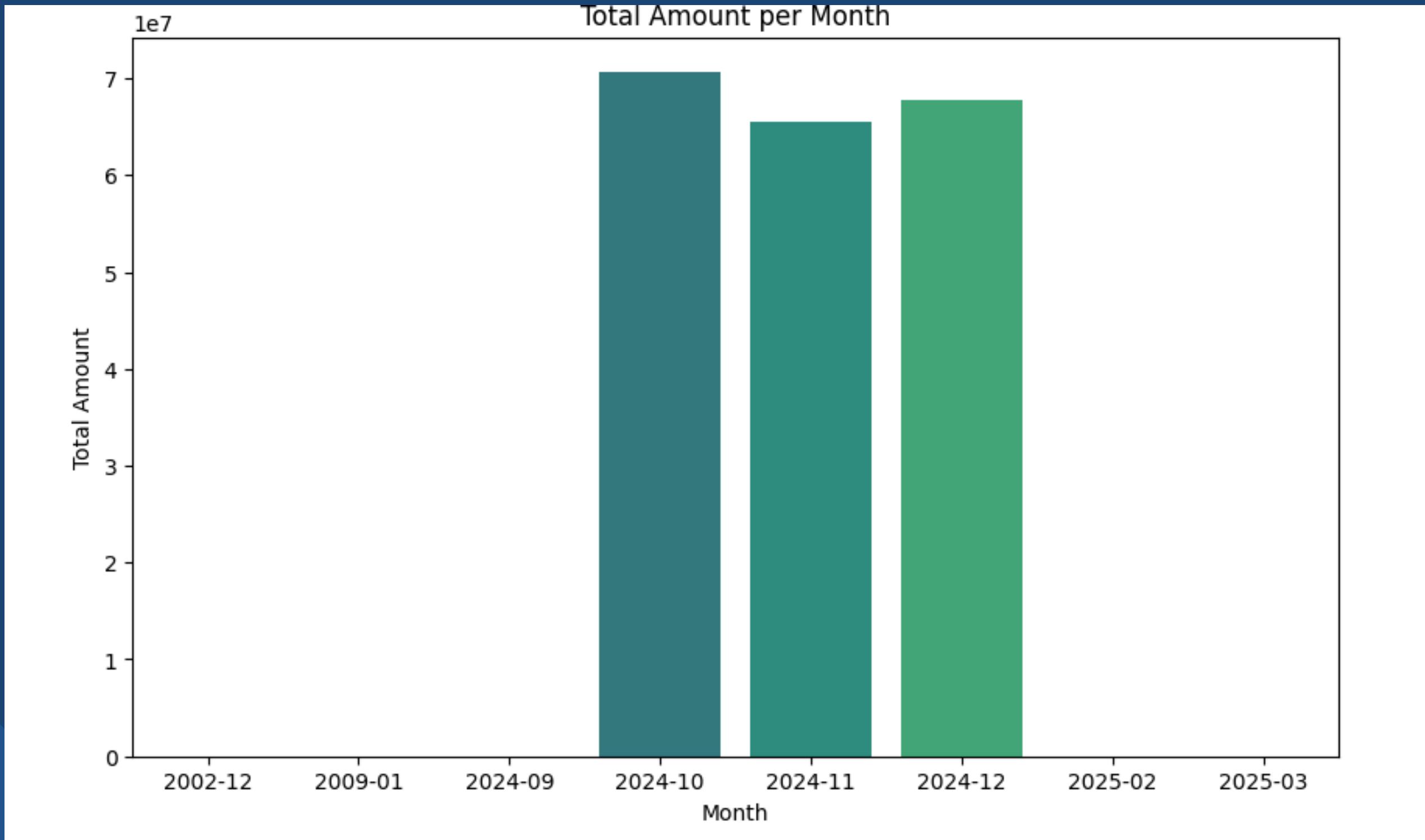


weekday	count
Friday	1553146
Monday	1259416
Saturday	1617888
Sunday	1344951
Thursday	1560045
Tuesday	1527842
Wednesday	1486501

พบร้า วันเสาร์เป็นวันที่มีการเดินทางและเรียกใช้บริการมากที่สุด



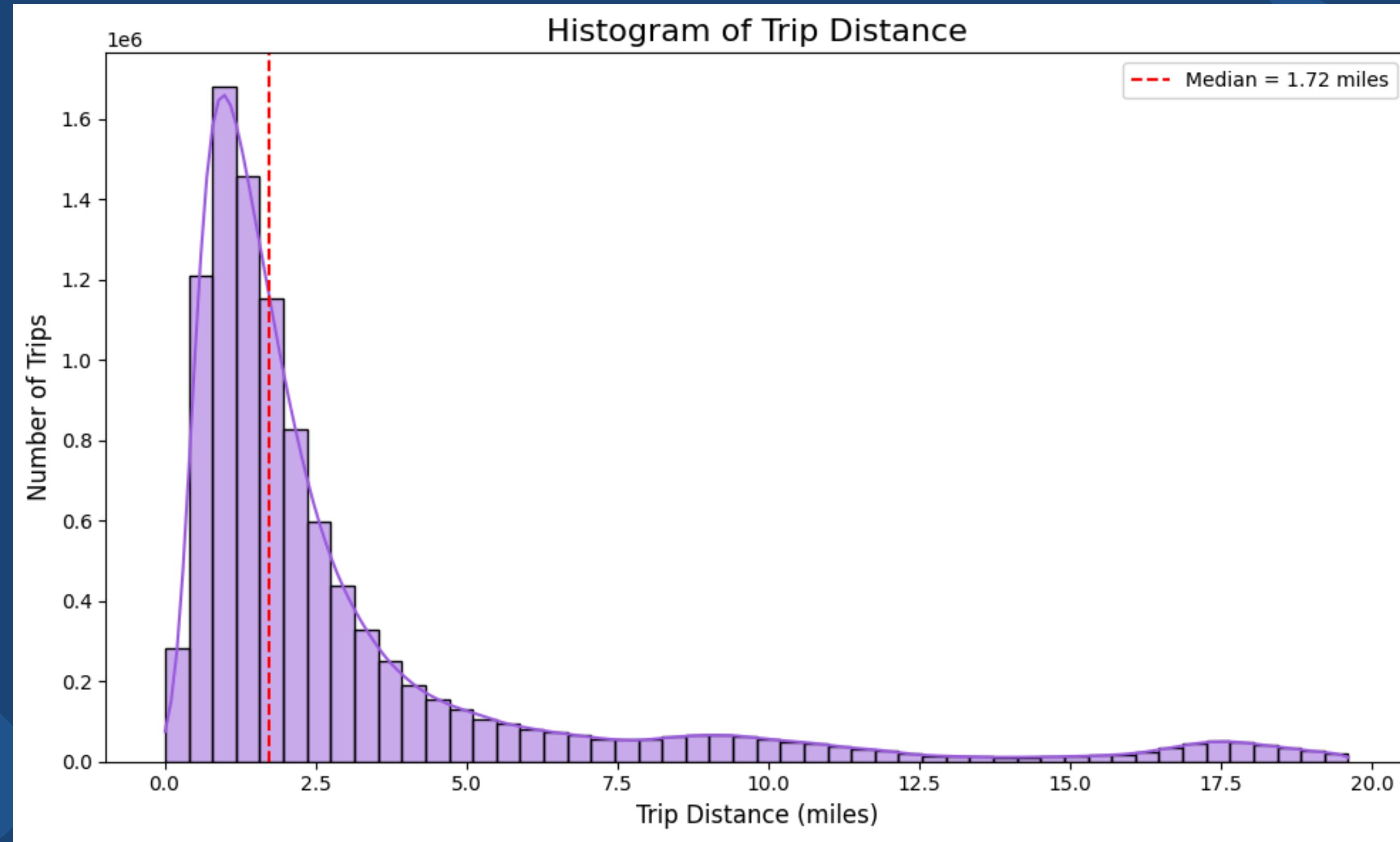
# เดือนที่มีรายได้รวมมากที่สุด



เดือนที่มีรายได้รวมของแท็กซี่มากที่สุด คือ เดือน ตุลาคม

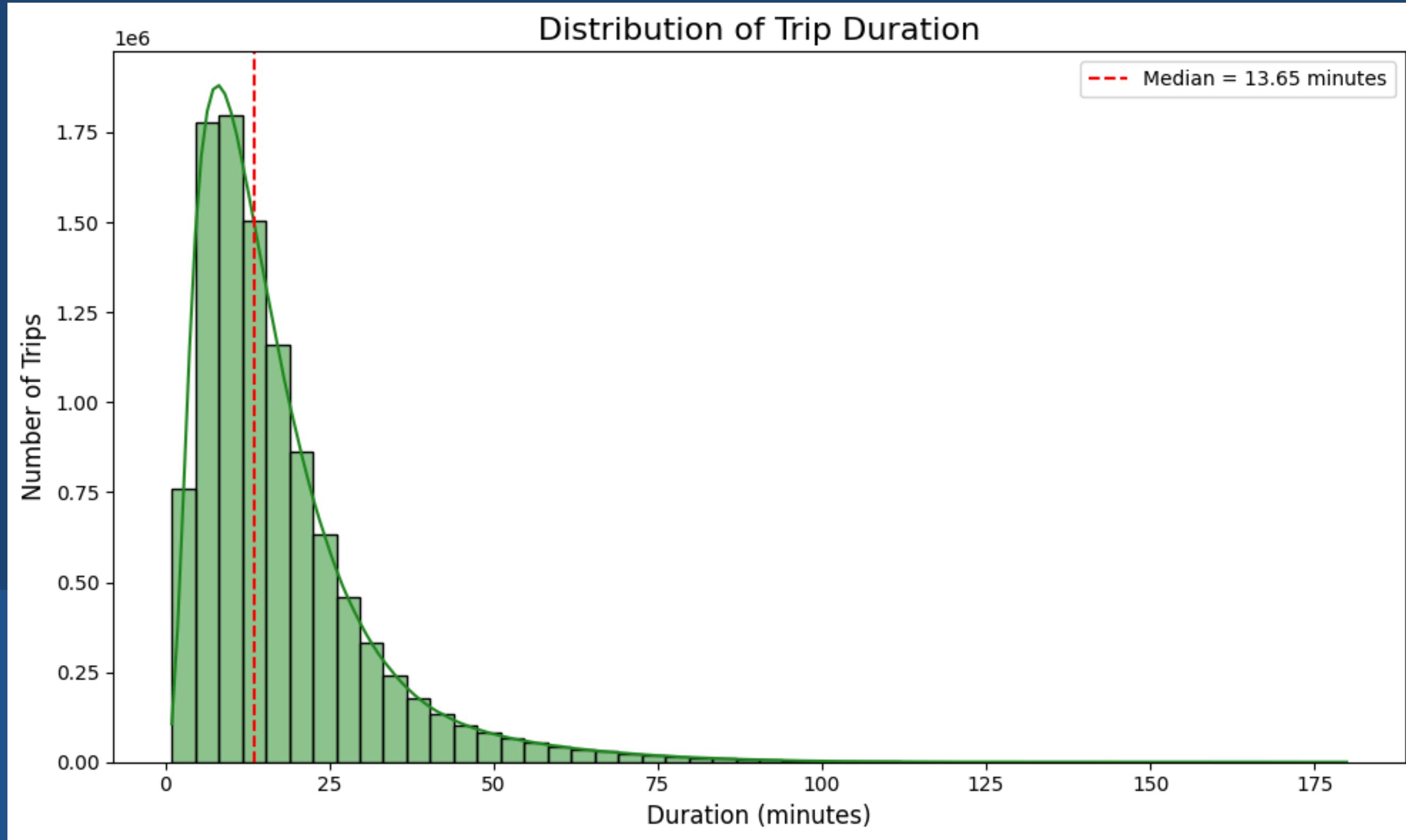


# กราฟแสดงการแจกแจงของระยะการเดินทาง



ช่วงของระยะทางที่ใช้บริการอยู่ที่ 0.01 - 5.0 miles ส่วนใหญ่อยู่ที่ 1.72 miles

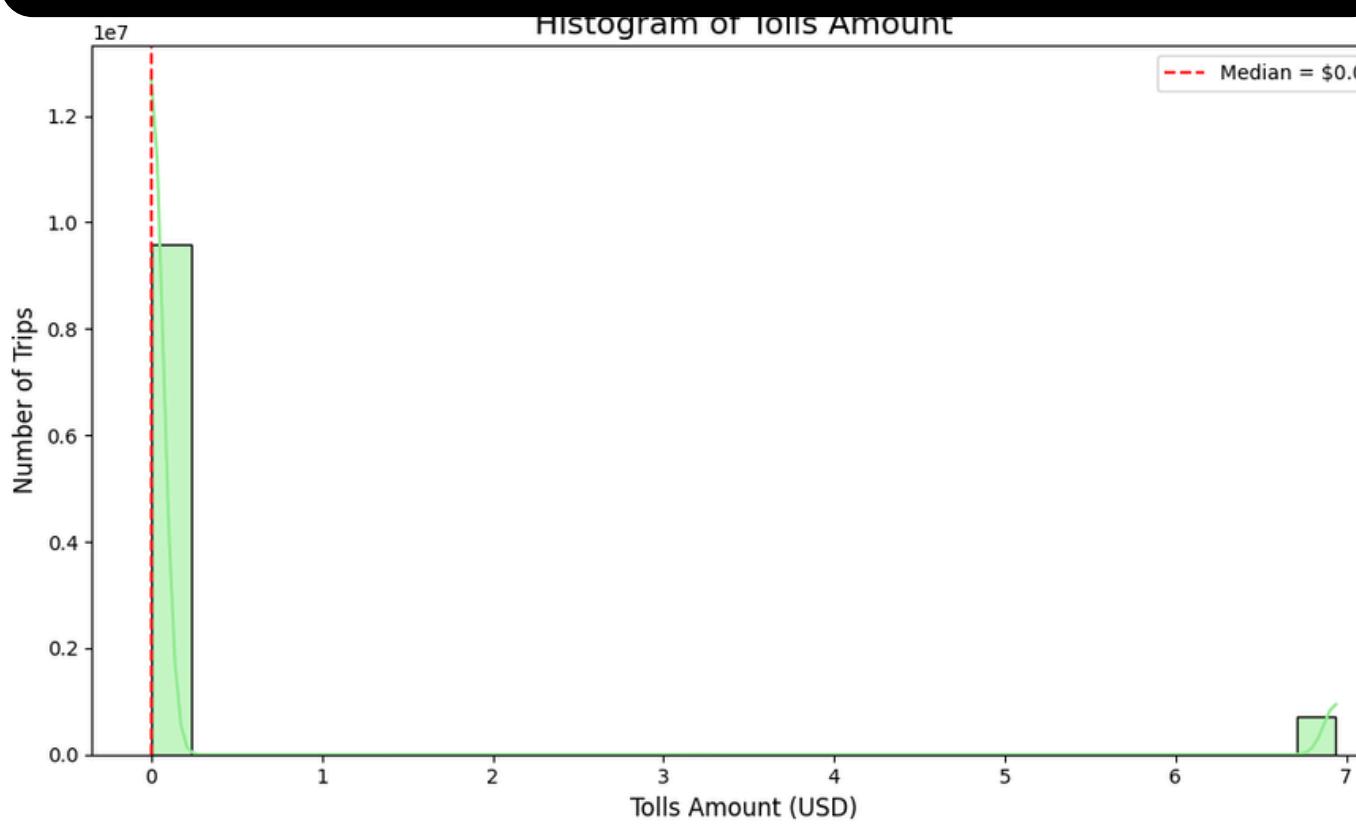
# กราฟแสดงการแจกแจงของระยะเวลาเดินทาง



พบว่า ระยะเวลาเดินทางอยู่ในช่วง 0 - 25 นาที และ ส่วนใหญ่ใช้เวลา 13.65 นาที



## กราฟแท่งแสดงการกระจายของค่าผ่านทาง



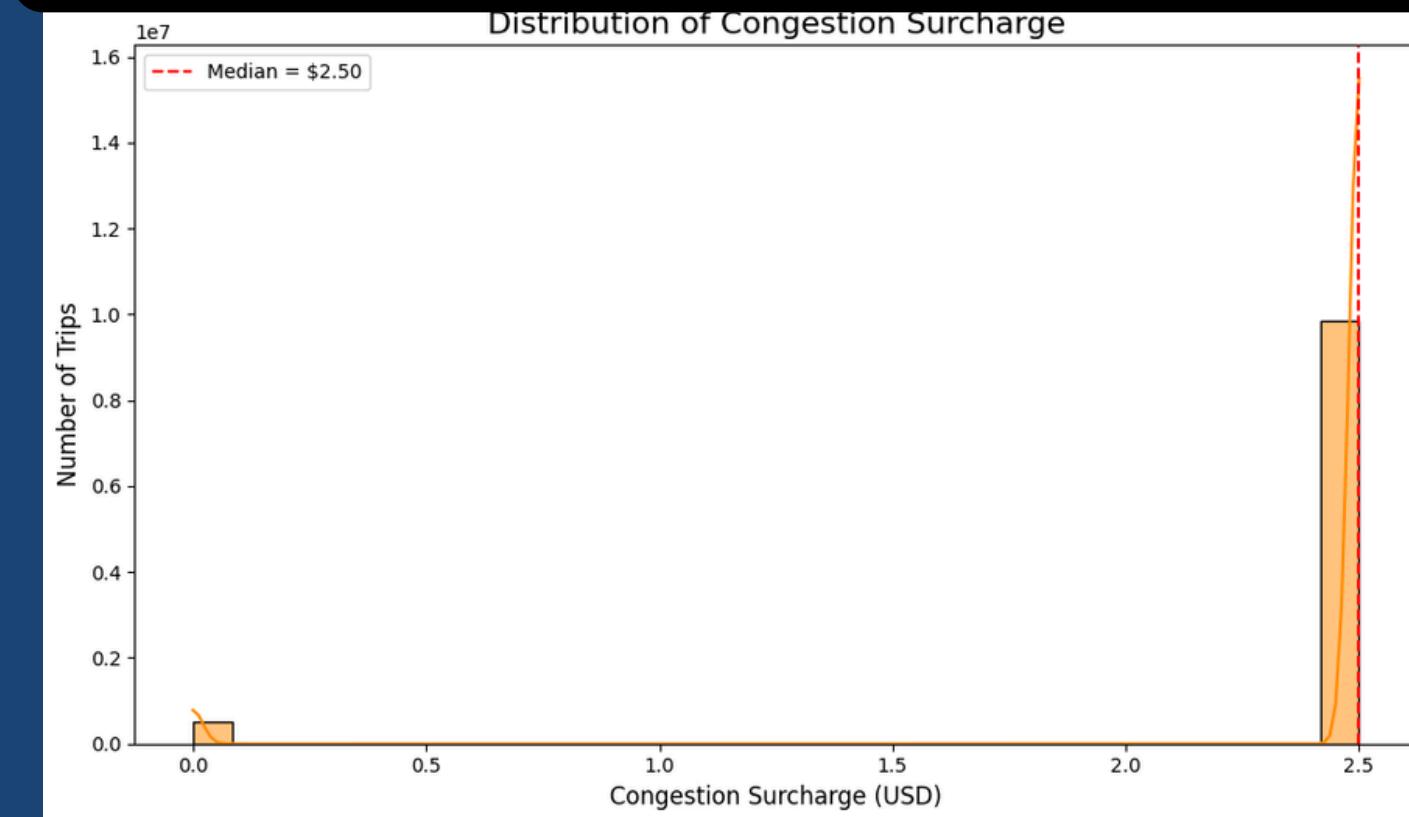
พบว่า ส่วนใหญ่ค่าผ่านทางเป็น 0 และรองลงมาคือ 6.94

## กราฟแสดงการแจกแจงของค่าธรรมเนียมสนามบิน



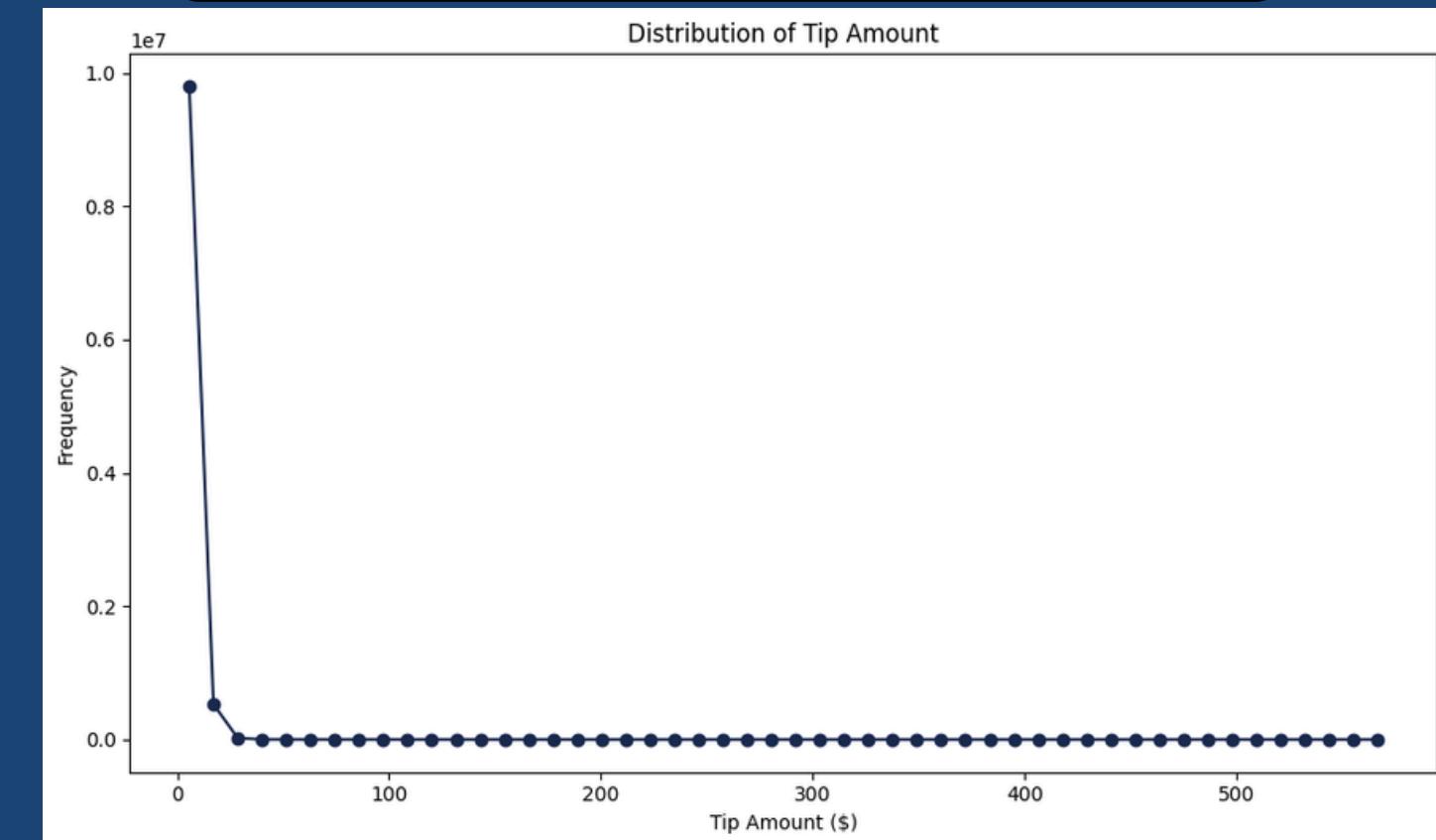
พบว่า ส่วนใหญ่ค่าธรรมเนียมสนามบิน เป็น 0 รองลงมาเป็น 1.75

## กราฟแสดงการแจกแจงของค่าธรรมเนียมเพิ่มเติม



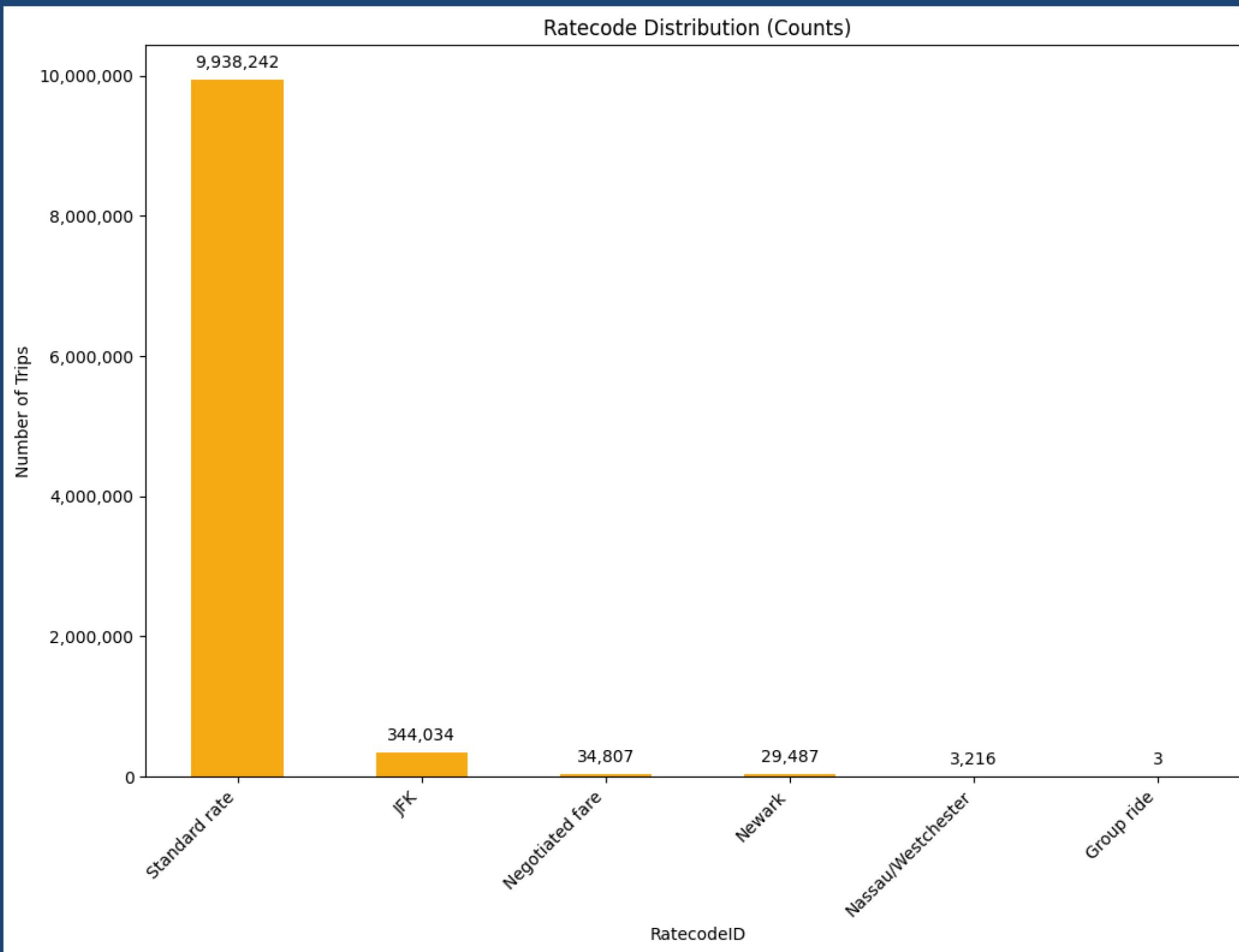
พบว่า ค่าธรรมเนียมเพิ่มเติม ส่วนใหญ่อยู่ที่ 2.50

## กราฟเส้นแสดงการกระจายของค่าทิป



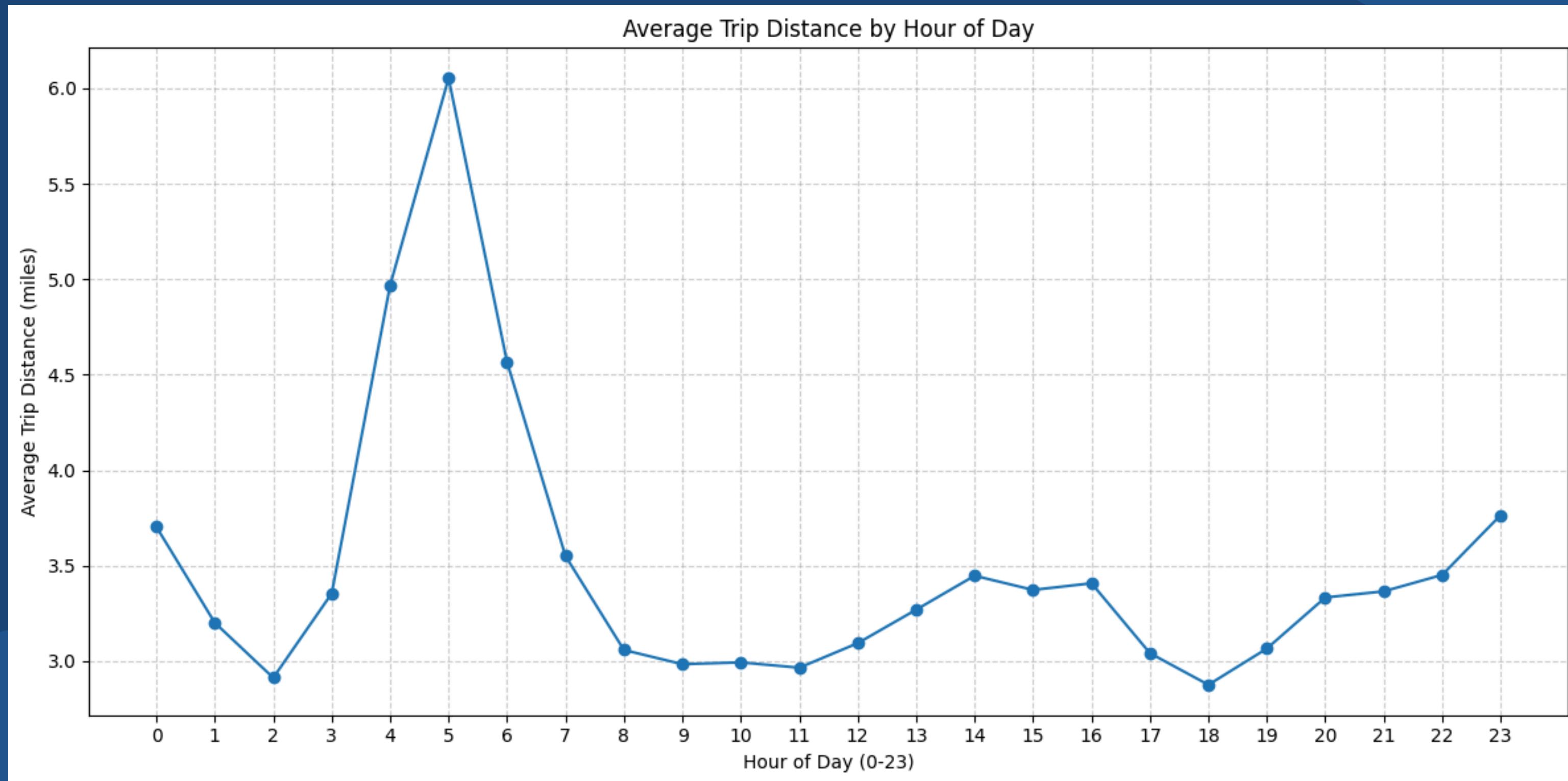
พบว่า ส่วนใหญ่ไม่มีการให้ทิป และทิปที่สูงสุดคือ 572 \$

# กราฟแสดงการใช้บริการในแต่ละประเภทการคำนวณค่าโดยสาร



**Standard rate(มิเตอร์ปกติ) มีจำนวนการใช้บริการมากที่สุด**

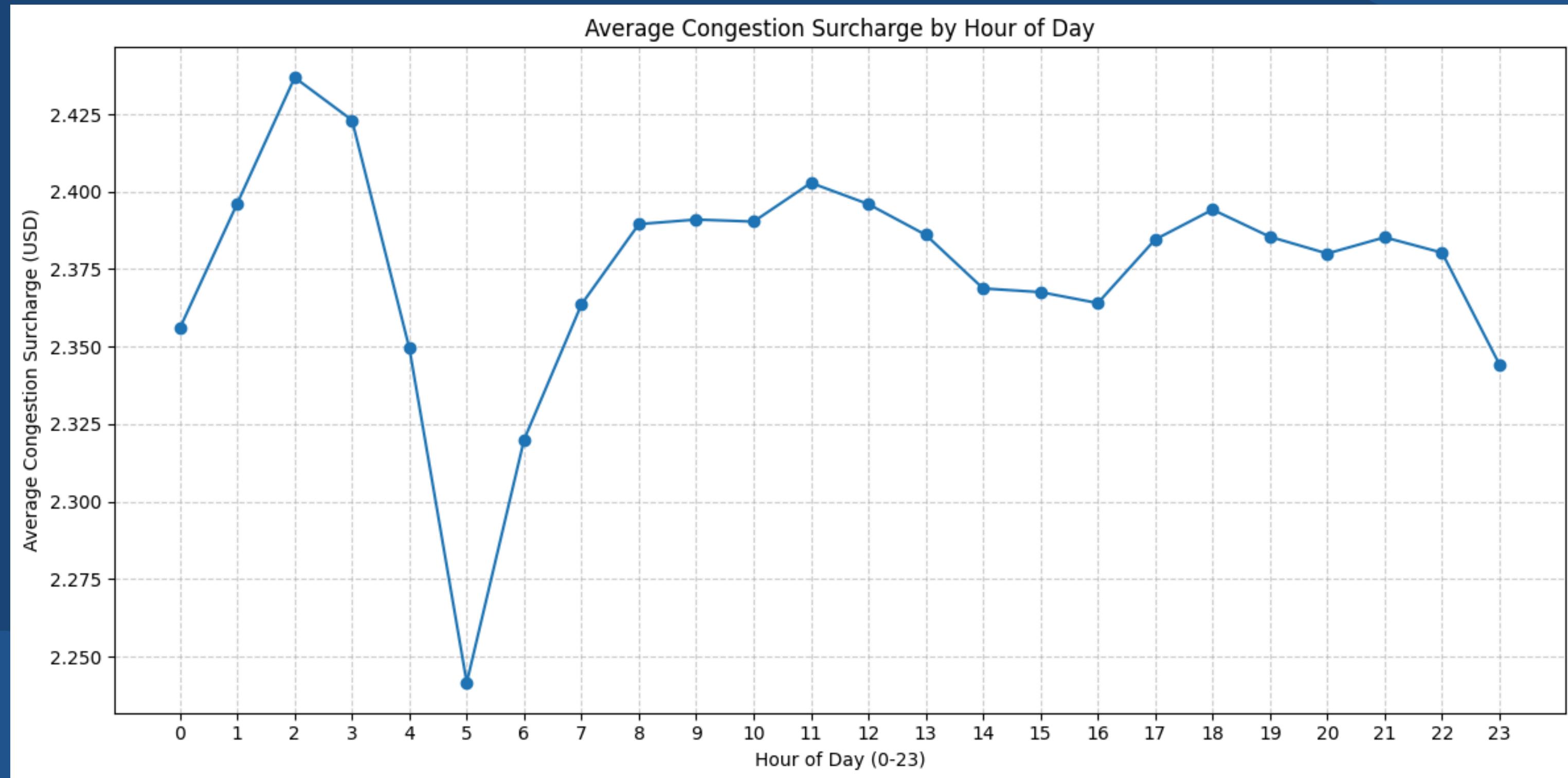
# กราฟแสดงค่าเฉลี่ยระยะทางในแต่ละช่วงเวลา



พบว่า ในช่วงเวลา 4 AM. - 6 AM. การเดินทางมีระยะทางไกล และช่วงเวลาที่การเดินทางมีระยะทางไกลที่สุดคือ 5 AM.



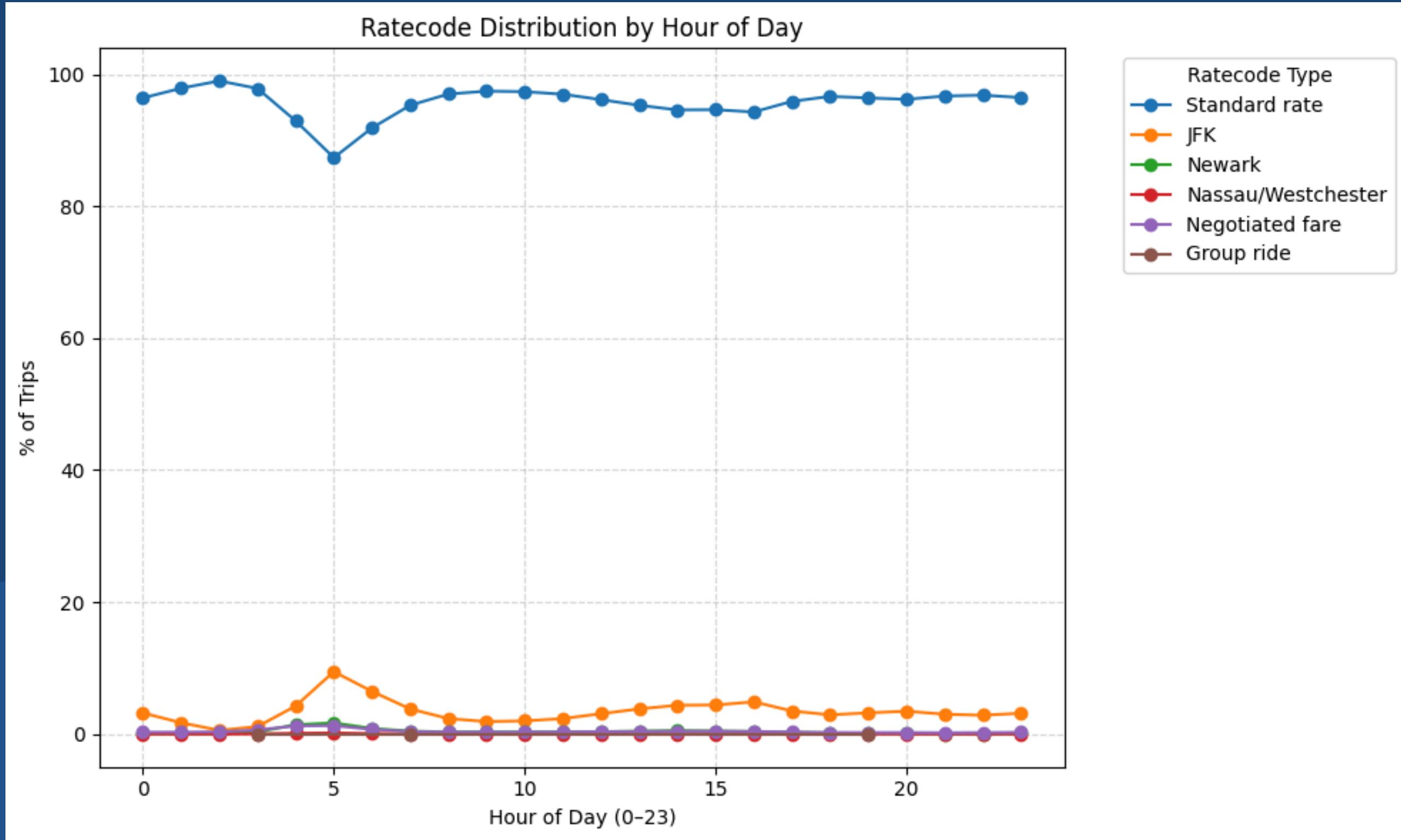
# กราฟแสดงค่าเฉลี่ยค่าธรรมเนียมเพิ่มเติมในแต่ละช่วงเวลา



พบว่า ในช่วงเวลา 4 AM. - 6 AM. ค่าเฉลี่ยค่าธรรมเนียมเพิ่มเติมต่ำ<sup>ที่สุด</sup>  
และช่วงเวลาที่ค่าเฉลี่ยค่าธรรมเนียมเพิ่มเติมต่ำสุดคือ 5 AM.



# กราฟแสดงจำนวนประเกตการคำนวณค่าโดยสารในช่วงเวลาการใช้บริการ



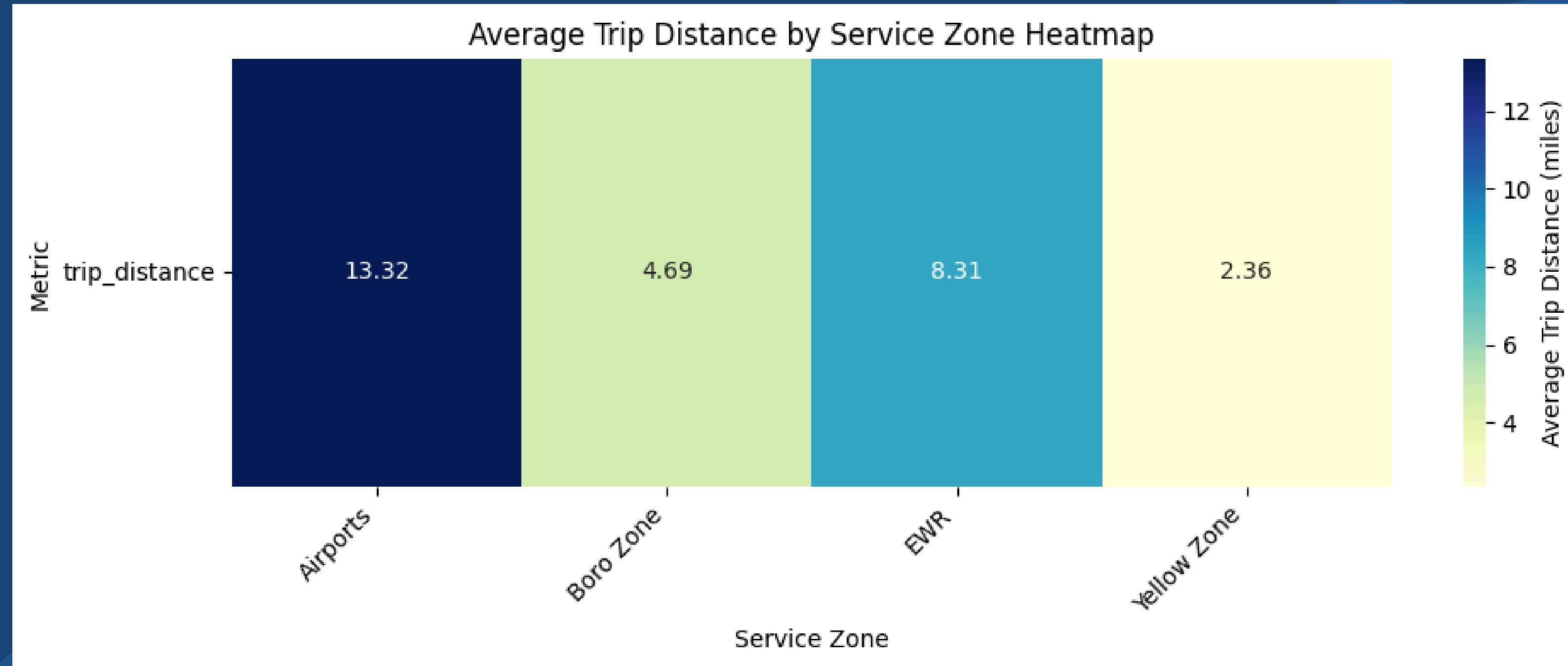
ในช่วงเวลา 5 AM. การคำนวณค่าโดยสารแบบ Standard rateลดลง แต่ JFK(แบบเหมาจ่าย) เพิ่มสูงขึ้น



# IMPORT GEO&ZONE LOCATION DATA TO VISUALIZE MAP

OBJECTID	Shape_Leng	Shape_Area	zone	LocationID	borough	geometry	Borough	Zone	service_zone
162.0	0.035270	0.000048	Midtown East	162.0	Manhattan	POLYGON ((992224.354 214415.293, 992096.999 21...)	Manhattan	Midtown East	Yellow Zone
48.0	0.043747	0.000094	Clinton East	48.0	Manhattan	POLYGON ((986694.313 214463.846, 986568.184 21...)	Manhattan	Clinton East	Yellow Zone
142.0	0.038176	0.000076	Lincoln Square East	142.0	Manhattan	POLYGON ((989380.305 218980.247, 989359.803 21...)	Manhattan	Lincoln Square East	Yellow Zone
233.0	0.048036	0.000116	UN/Turtle Bay South	233.0	Manhattan	MULTIPOLYGON (((993816.792 213230.43, 993857.4...))	Manhattan	UN/Turtle Bay South	Yellow Zone
137.0	0.046108	0.000116	Kips Bay	137.0	Manhattan	POLYGON ((991954.728 209026.462, 991949.076 20...))	Manhattan	Kips Bay	Yellow Zone

# กราฟแสดงค่าเฉลี่ยระยะทางในแต่ละโซนให้บริการ



พบว่า ค่าเฉลี่ยระยะทางในแต่ละโซนให้บริการ(miles) 1.Airports 13.32

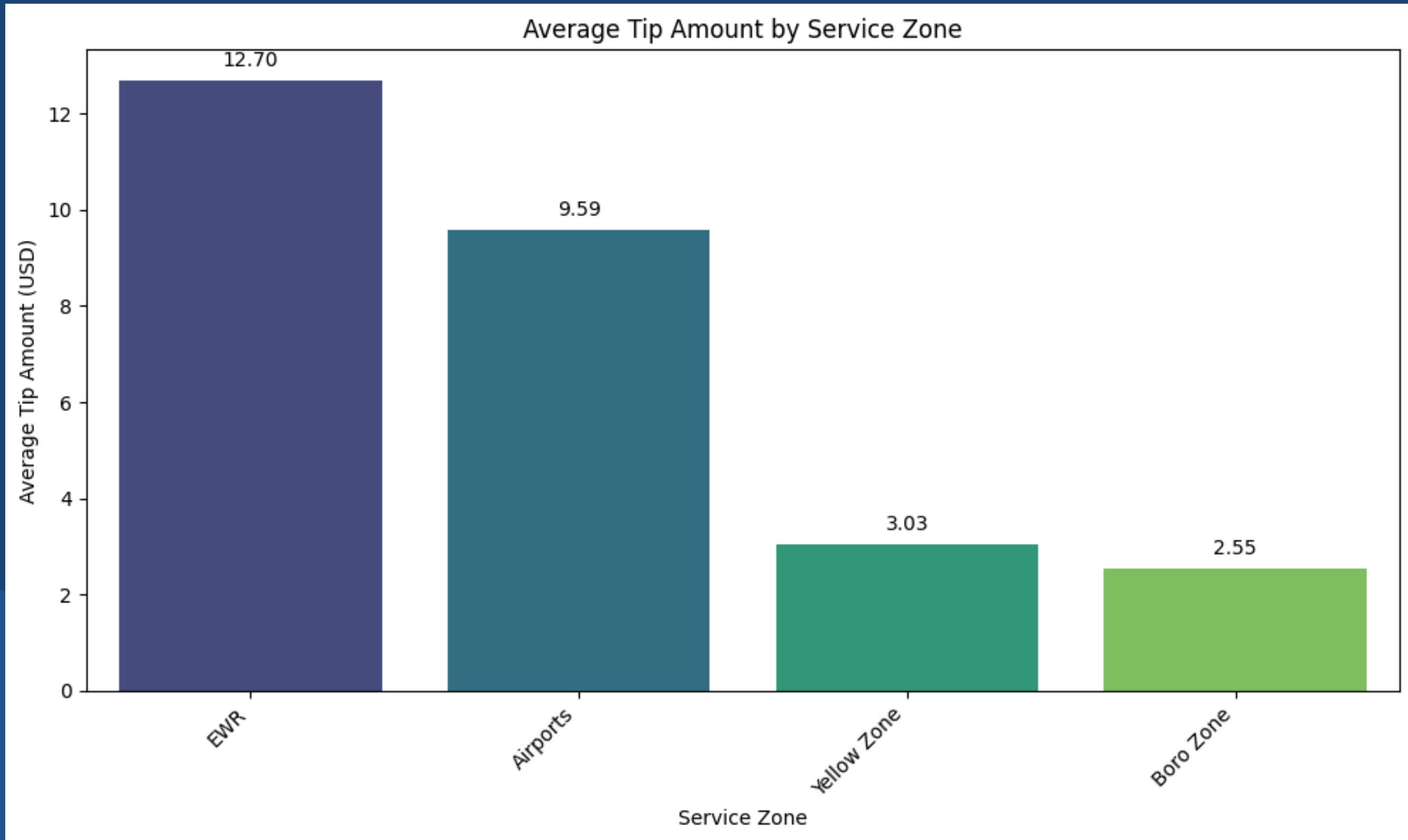
2.EWR 8.31

3.Boro Zone 4.69

และ 4.Yellow Zone 2.36



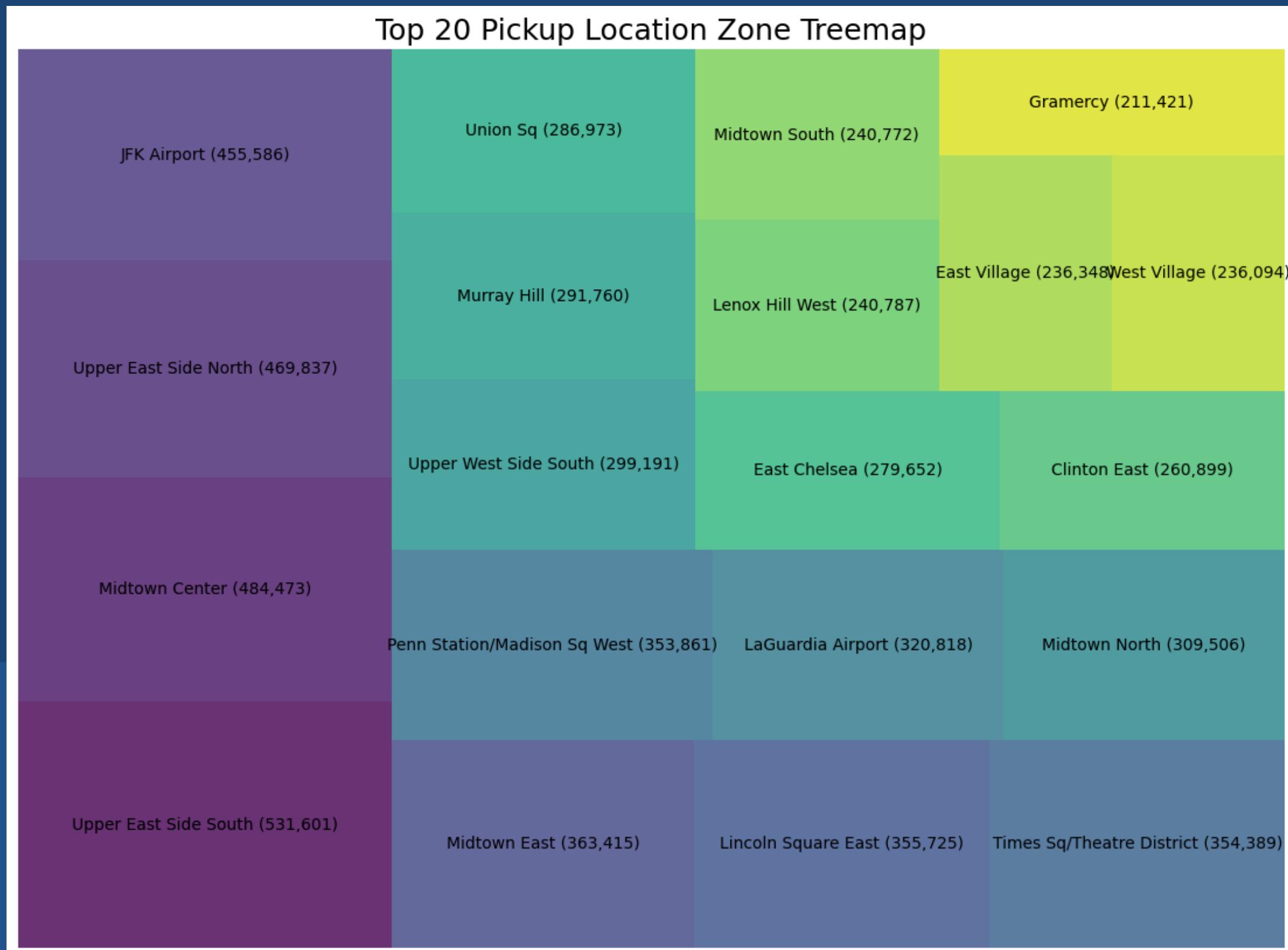
# กราฟแสดงค่าเฉลี่ยทิปในแต่ละโซนให้บริการ



พบว่า ค่าเฉลี่ยทิปในแต่ละโซน  
ให้บริการ(USD)  
1.EWR 12.70  
2.Airports 9.59  
3.Yellow Zone 3.03  
และ 4.Boro Zone 2.55



# TOP 20 ZONE ที่มีการรับผู้โดยสารสูงสุด

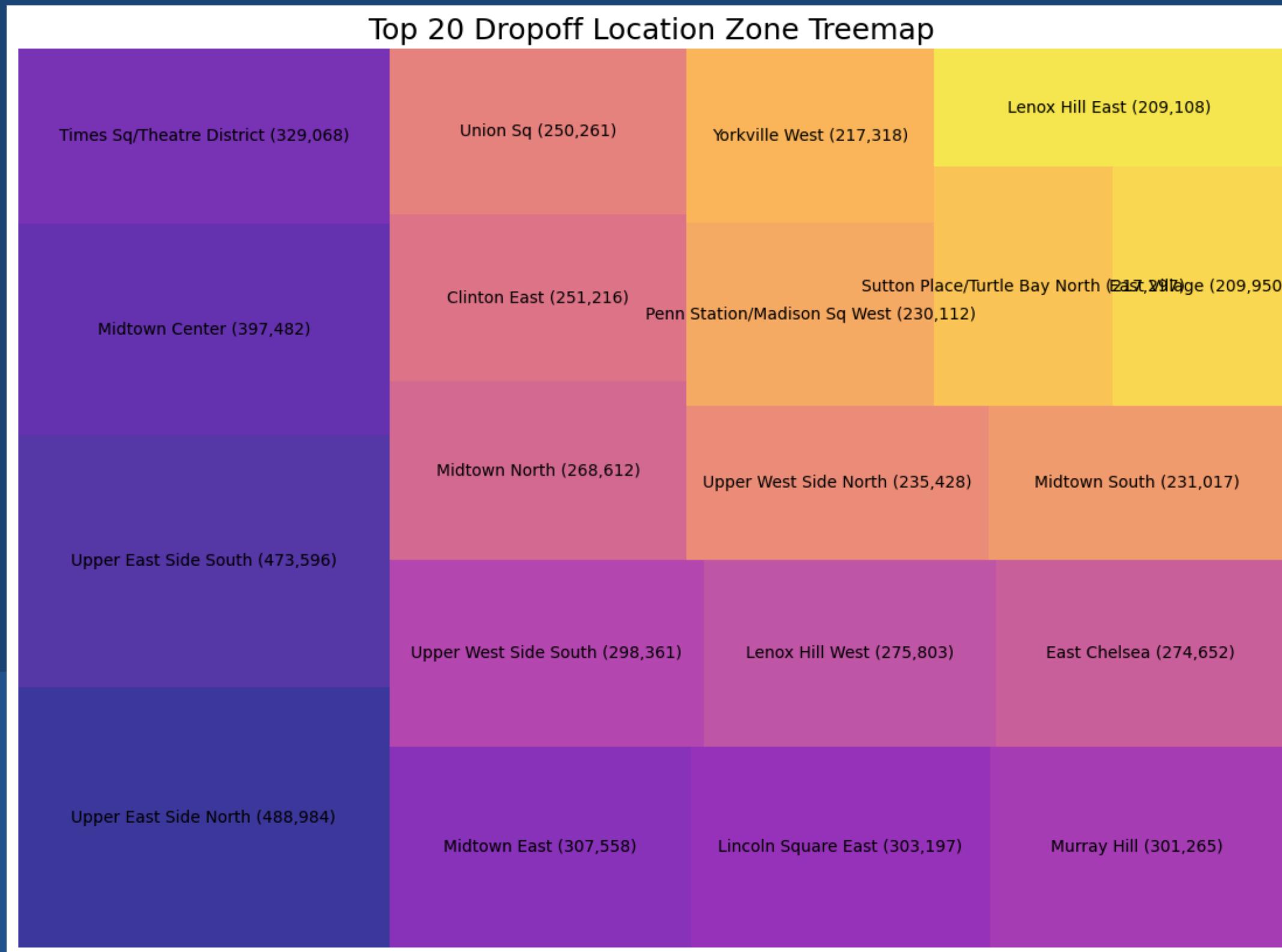


พบว่า Zone ที่มีการรับผู้โดยสารสูงสุด

1. **Upper East Side South 531,601 ครั้ง**
2. **Midtown Center 484,473 ครั้ง**
3. **Upper East Side North 469,837 ครั้ง**



# TOP 20 ZONE ที่มีการส่งผู้โดยสารสูงสุด

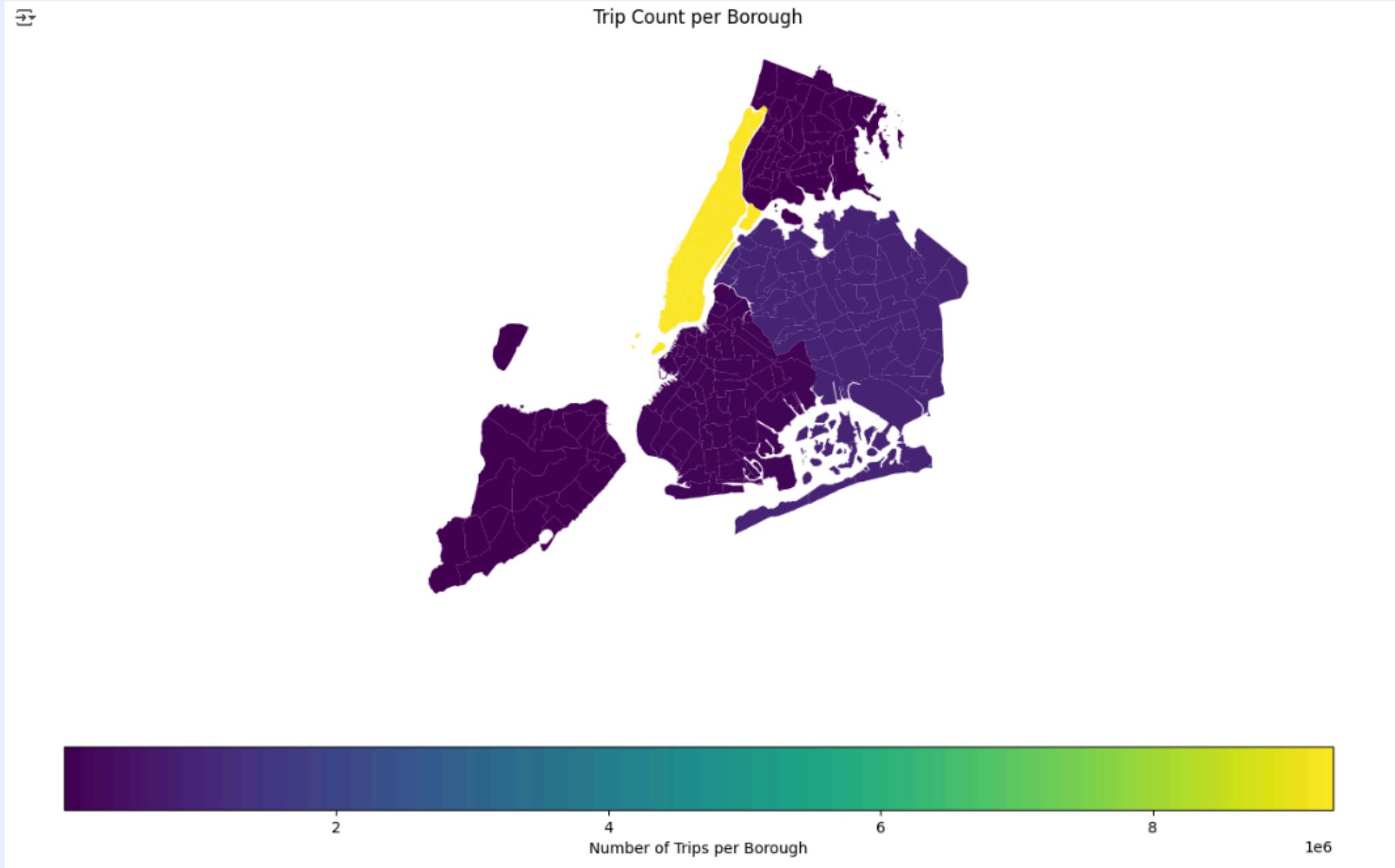


พบว่า Zone ที่มีการส่งผู้โดยสารสูงสุด

1. Upper East Side North 488,984 คน
2. Upper East Side South 473,596 คน
3. Midtown Center 397,482 คน



# เขตที่มีการเดินทางมากที่สุด



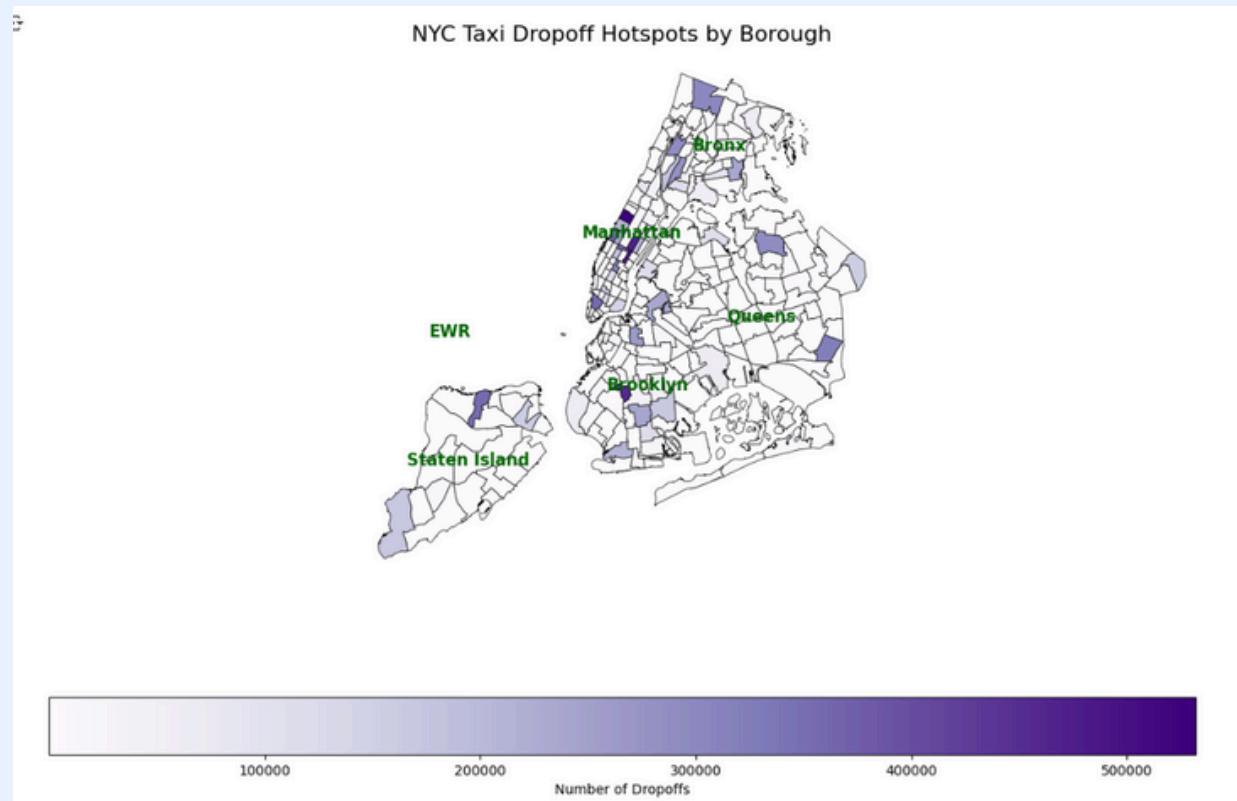
count

borough	count
Manhattan	9321118
Queens	892604
Brooklyn	118179
Bronx	17506
Staten Island	322
EWR	60

dtype: int64

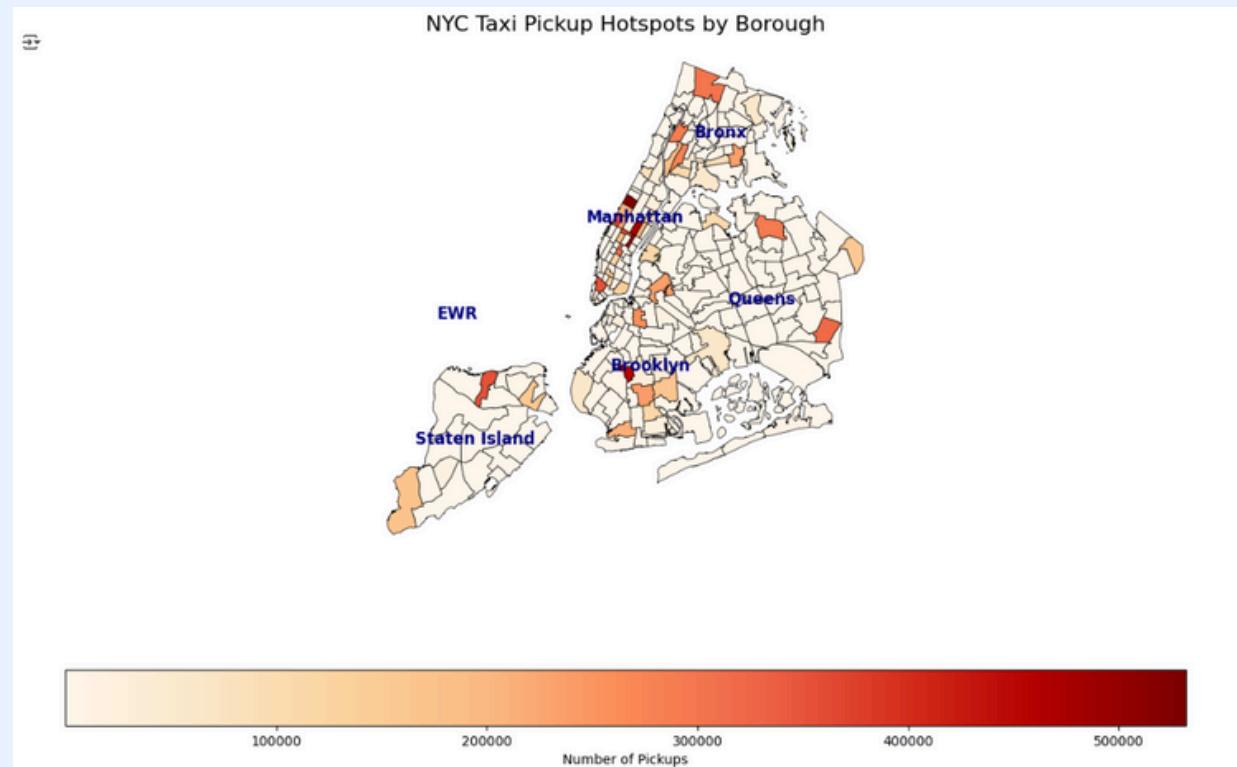
เขตที่มีการใช้บริการแท็กซี่มากที่สุด คือ Manhattan, Queens, Brooklyn, Bronx, Staten Island และ EWR ตามลำดับ

# เขตที่มีการปรับและเปลี่ยนแปลง



## Dropoff Hotspot

จุดหมายปลายทางยอดนิยมของแท็กซี่คือ Manhattan โดยเฉพาะ Brooklyn และ Bronx มากกว่าผู้โดยสารส่วนใหญ่จะการเดินทางในใจกลางเมืองธุรกิจ



## Pickup Hotspot

จุดรับผู้โดยสารหลักคือ Manhattan โดยเฉพาะเขต Brooklyn และ Bronx เช่นเดียวกับ Dropoff แต่ยังเห็นความเคลื่อนไหวบางส่วนที่สนามบินและเขตพักราชอาศัยรอบนอก

# MODEL JOURNEY

## STANDARDSCALER

```
[ [0.03085974 0.          0.06420462 ... 0.          0.          0.09326377]
[0.02260295 0.          0.04723423 ... 0.          0.          0.06745551]
[0.02776344 0.          0.04440583 ... 0.          0.          0.04518774]
...
[0.02982764 0.          0.07850823 ... 0.04626667 0.          0.10798472]
[0.04056146 0.          0.00787911 ... 0.          0.          0.13584274]
[0.04324492 0.          0.09923632 ... 0.          0.          0.11133885]]
```

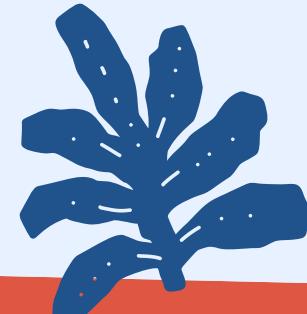
Y=FARE\_AMOUNT

X=TRIP\_DISTANCE,CONGESTION\_SURCHARGE,TIP\_AMOUNT  
,TOLLS\_AMOUNT,AIRPORT\_FEE,DURATION

## TRAIN-TEST SPLIT

```
7 # Select subsets for training and testing
8 X_train,X_test,y_train,y_test = model_selection.train_test_split(X,
9
10
11           y,
test_size=0.2,
random_state=123)
```

# MODEL



## MULTIPLE LINEAR REGRESSION

### EQUATION

$$\hat{y} = 0.12571864 + 0.01310097 \underline{X_1} - 0.07017756 \underline{X_2} \\ + 0.00130974 \underline{X_3} + 0.00291121 \underline{X_4} + 0.11478439 \underline{X_5} \\ + 0.01624025 \underline{X_6}$$

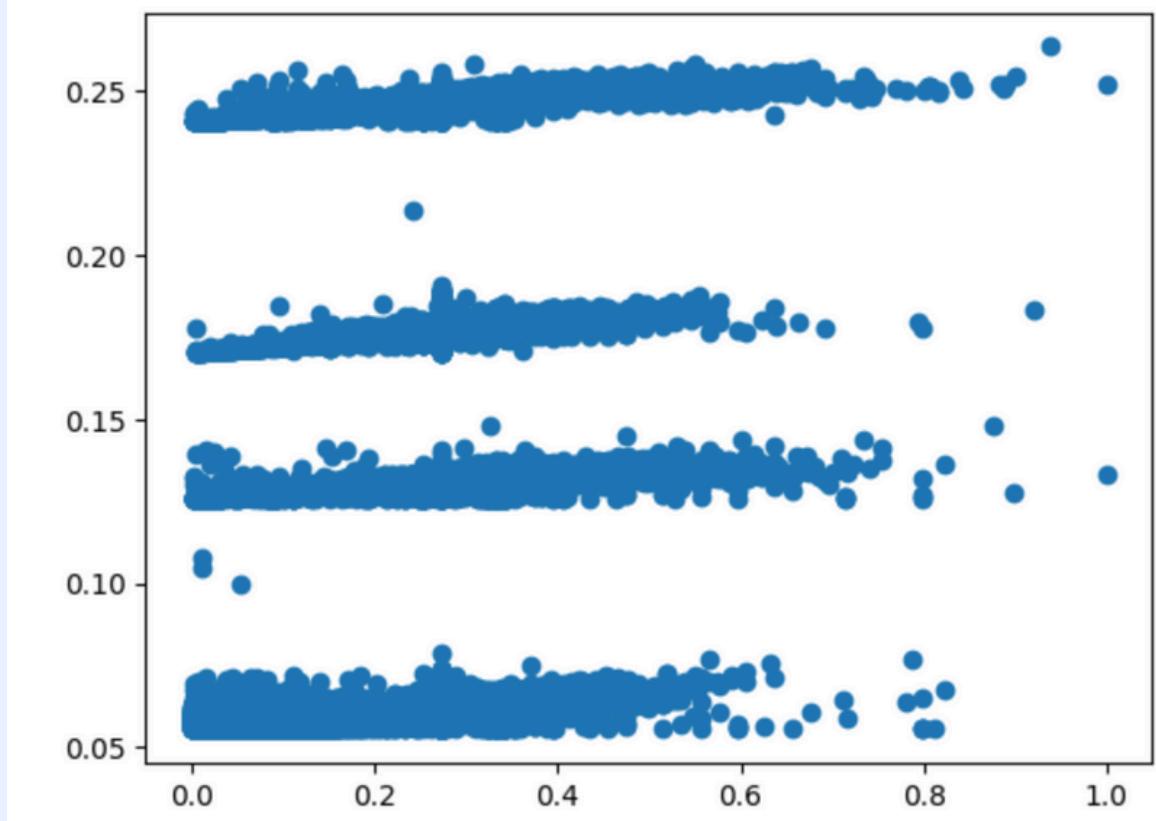
### SUMMARY

R2 score: 0.37

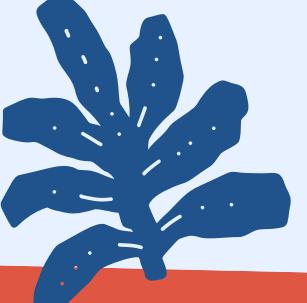
Mean Squared Error (MSE): 0.00

Root Mean Squared Error (RMSE): 0.05

### SCATTER PLOT



# MODEL



## LASSO

### EQUATION

$$\hat{y} = 0.02851359 -0.00712261\text{X}_2 +0.06980421\text{X}_5 \\ +0.46504983 \text{ X}_6$$

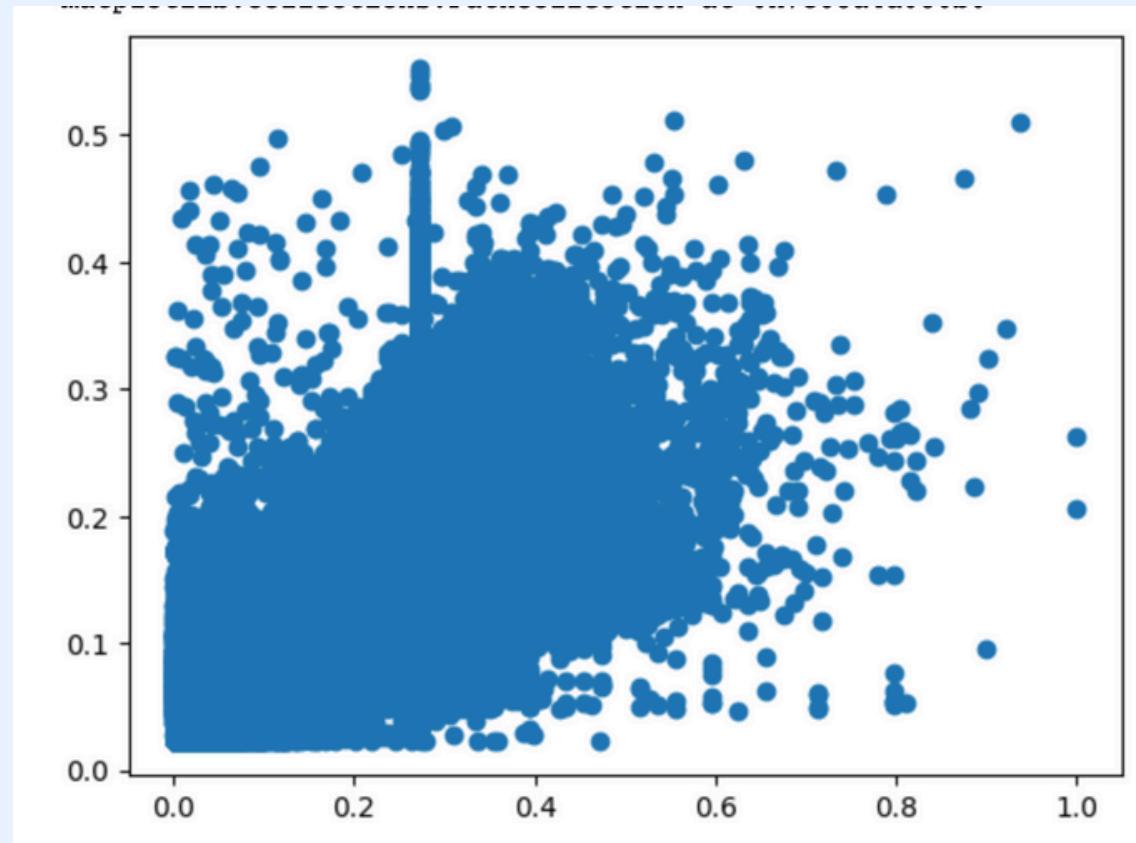
### SUMMARY

R2 score: 0.77

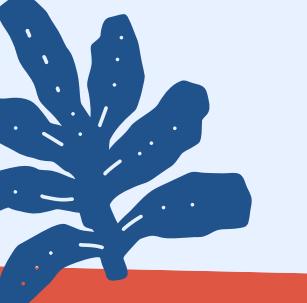
Mean Squared Error (MSE): 0.00

Root Mean Squared Error (RMSE): 0.03

### SCATTER PLOT



# MODEL



## RIDGE REGRESSION

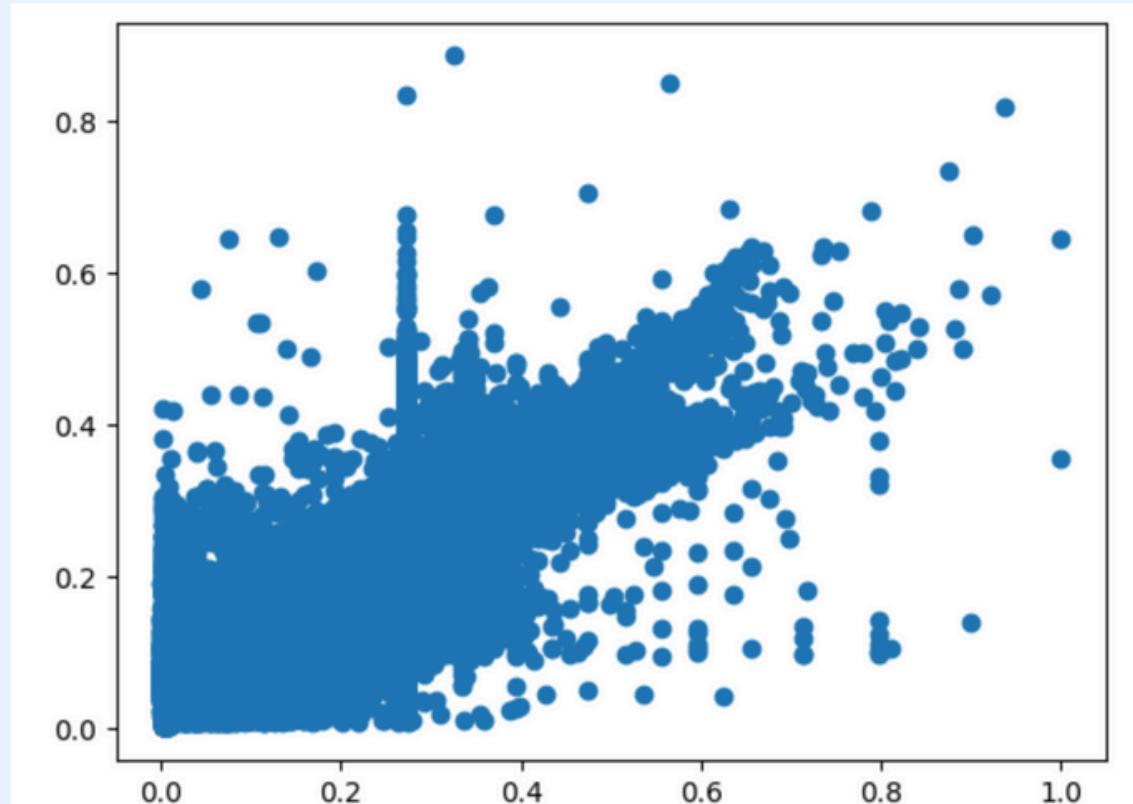
### EQUATION

$$\hat{y} = 0.0277616 + 0.9333397 \underline{X_1} - 0.02146341 \underline{X_2} + \\ 0.42634034 \underline{X_3} + 0.33360675 \underline{X_4} - 0.01699962 \underline{X_5} + \\ 0.3073652 \underline{X_6}$$

### SUMMARY

R2 score: 0.94  
Mean Squared Error (MSE): 0.00  
Root Mean Squared Error (RMSE): 0.02

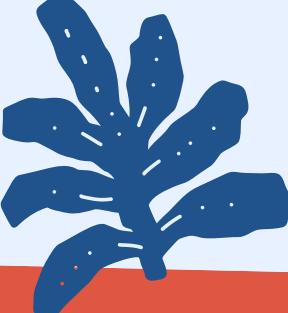
### SCATTER PLOT



### HYPERPARAMETER TUNING USE RANDOMIZEDSEARCHCV

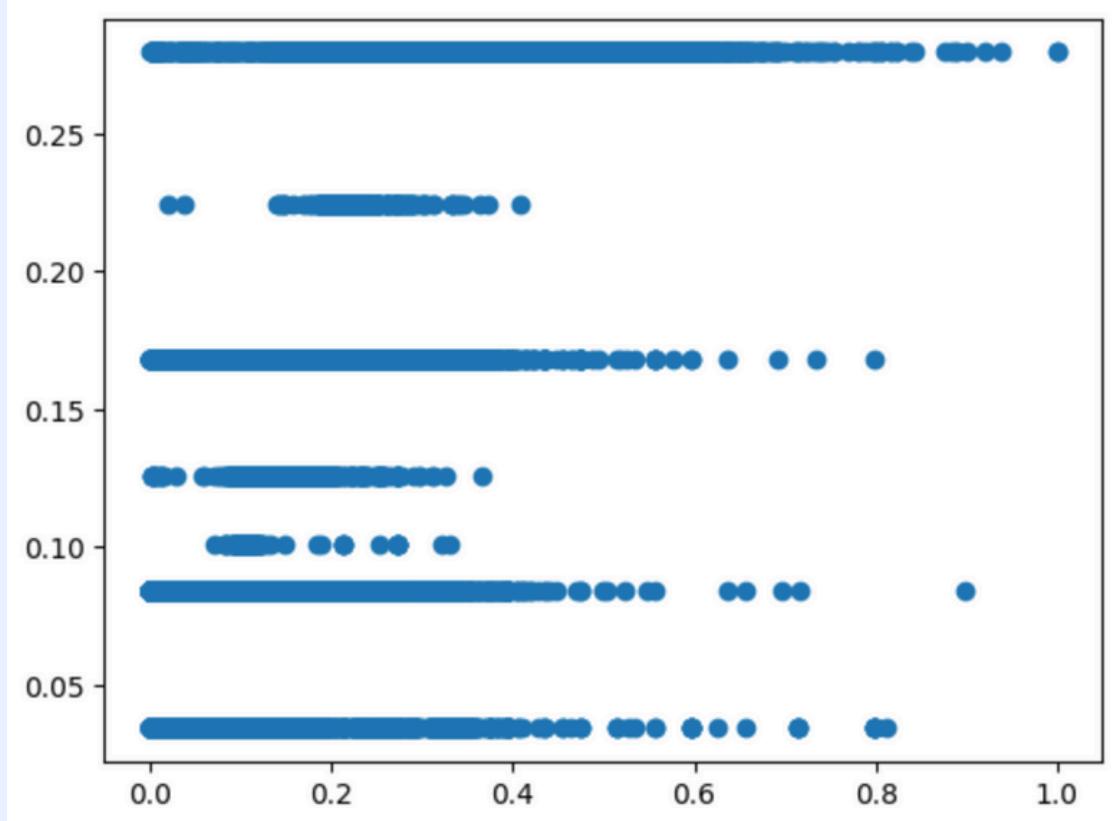
Best alpha: 9.808641983846154  
Best cross-validation R<sup>2</sup>: 0.9373316526412964

# MODEL



## RANDOM FOREST

### SCATTER PLOT



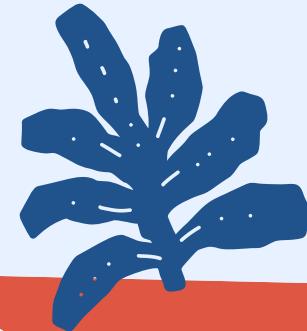
### SUMMARY

R2 score: 0.85

Mean Squared Error (MSE): 0.00

Root Mean Squared Error (RMSE): 0.03

# MODEL COMPARING



	Model	R2 Score	MSE	RMSE
0	Linear Regression	0.37	0.0027	0.0524
1	Lasso Regression	0.77	0.0010	0.0319
2	Ridge Regression	0.94	0.0003	0.0166
3	Random Forest Regression	0.85	0.0006	0.0253

# MODEL



## TESTING RIDGE REGRESSION

# ตัวอย่างข้อมูลใหม่ [trip\_distance, congestion\_fee, tip\_amount, tolls\_amount, Airport\_fee, duration]

X = 10.71, 2.5, 2.0, 0.0, 0.0, 24.00 (เอาข้อมูลไปสเกลต่อไป)

Predicted fare amount: \$0.03

Predicted fare amount (original scale): \$2.03

# THANK YOU

นางสาวศศิวิมล ภานุโชค 663020040-8  
นายวัฒนศักดิ์ คลังแสง 663020039-3  
นายกักรพล วรรณยศ 663020287-4  
นายวชากร สุขเก晦 663020293-9

