NEW YORK TLC TRIP RECORD ANALYSIS

CAPSTONE PROJECT MODULE 2

DTI DS

Huwaida Nur Asysyifa Mufarrida

November 2024

TABLE OF CONTENT

01 Project Overview

02 Data Understanding

03 Exploratory Data Analysis (EDA)

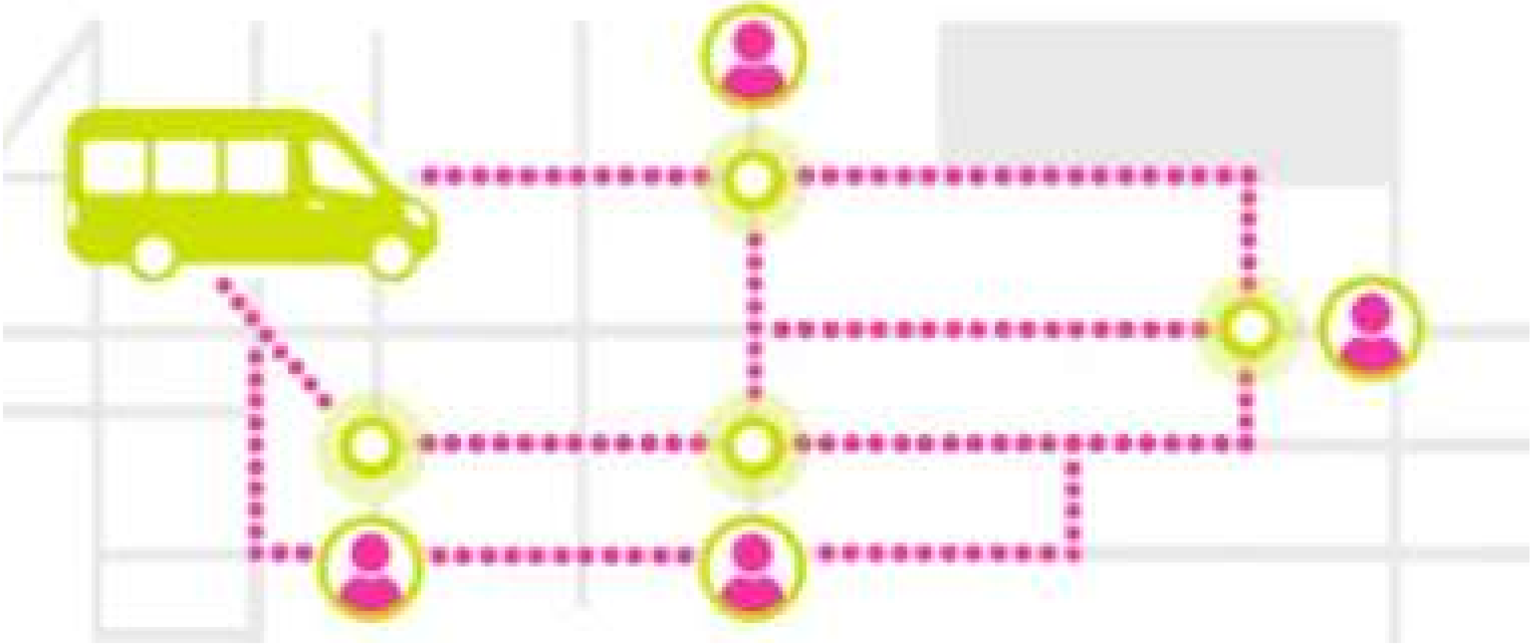
04 Project Objectives

05 Tableau Visualization and Statistic

06 Conclusion and Recommendation

1. PROJECT OVERVIEW

As demand for digital transportation grows, trip data helps improve service and efficiency.



As the demand for digital-based transportation services increases, data related to trips such as pickup and drop-off locations, duration, distance, time, and payment patterns become valuable assets for understanding customer behavior, route efficiency, and revenue potential.

This dataset of taxi trips in New York City contains detailed information about the time, location, distance, cost, and other details of each trip. It can help identify trends and patterns that may not be visible at first glance.
By using interactive data visualizations in Tableau, we can dive deeper into the data and produce actionable insights that service providers can use to improve their service quality and efficiency.

Without proper analysis, identifying patterns, demand areas, and revenue factors becomes challenging.

- Without proper data analysis and visualization, it is difficult to identify usage patterns, areas with the highest demand, the longest trip durations, and factors affecting revenue, such as payment type, fares, or tips.

- The large raw data and numerous variables can be a challenge in identifying key information that supports decision-making.

- Lack of insight into time patterns (such as rush hours and high-demand days) can result in suboptimal resource management, like fleet distribution.

Background

The New York Taxi company operates various trips in New York City using fleets like yellow taxis, green taxis, and other for-hire vehicles such as Uber. The TLC (Taxi and Limousine Commission) collects and publishes data to ensure transparency and provide insights into the city's transportation network. This company aims to increase revenue and the number of trips at specific strategic times and locations.

Problem Statement

The company wants to identify strategic times and locations that can boost its revenue.

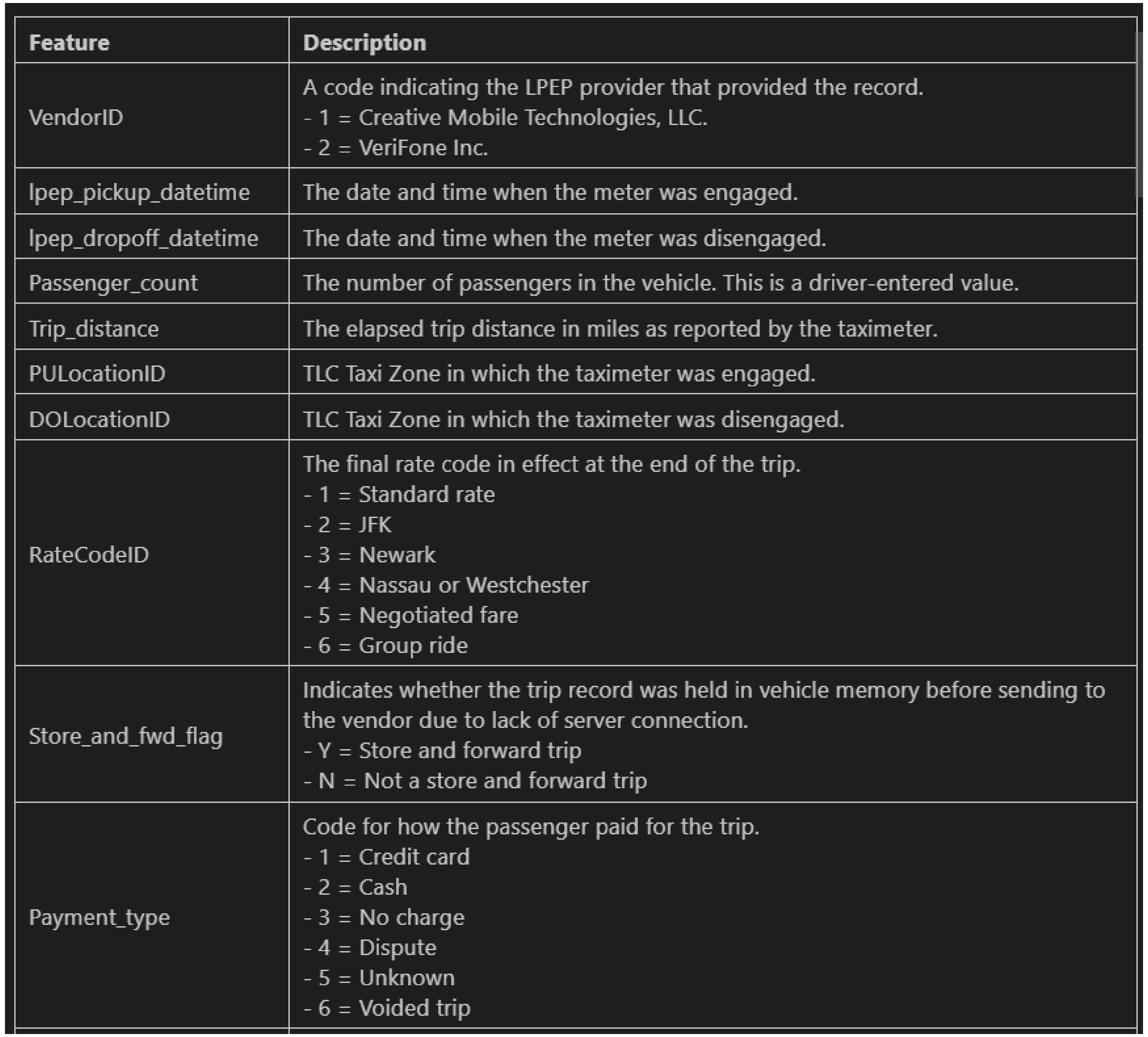
Goals

Present detailed information about the number of trips, locations, times, distances, and trip costs to identify:

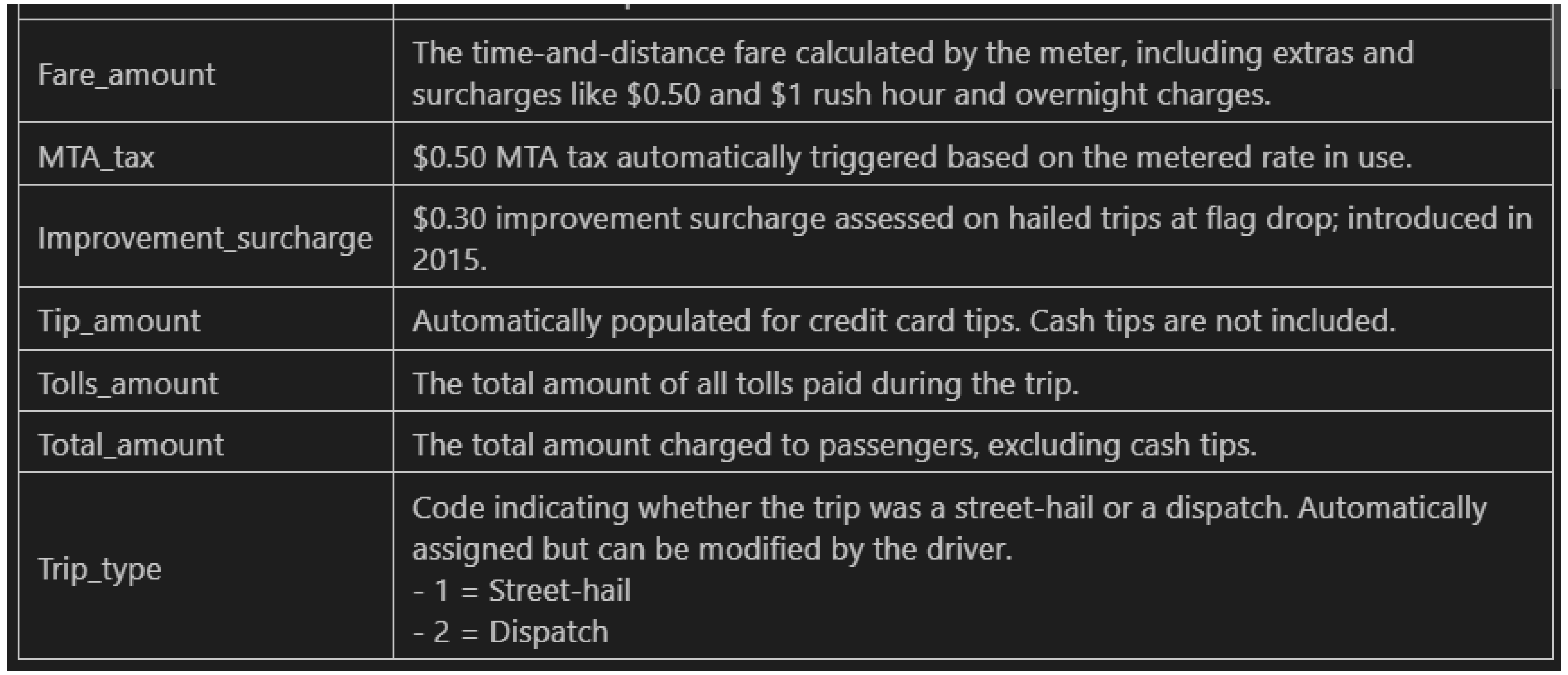
1. Generate operational insights that allow taxi companies to maximize their fleet in areas and times that require the most service.
2. Improve service efficiency by providing recommendations based on data on demand patterns, such as the most commonly traveled routes and busiest travel times.
3. Provide strategic guidance for business decisions, such as determining pricing plans based on time or location, introducing marketing strategies to increase service use during off-peak hours or days.
4. Offer a comprehensive Tableau dashboard that is user-friendly and can serve as a basis for data-driven decisions across various areas such as operations, marketing, and finance.

2. DATA UNDERSTANDING

Features in Raw Data

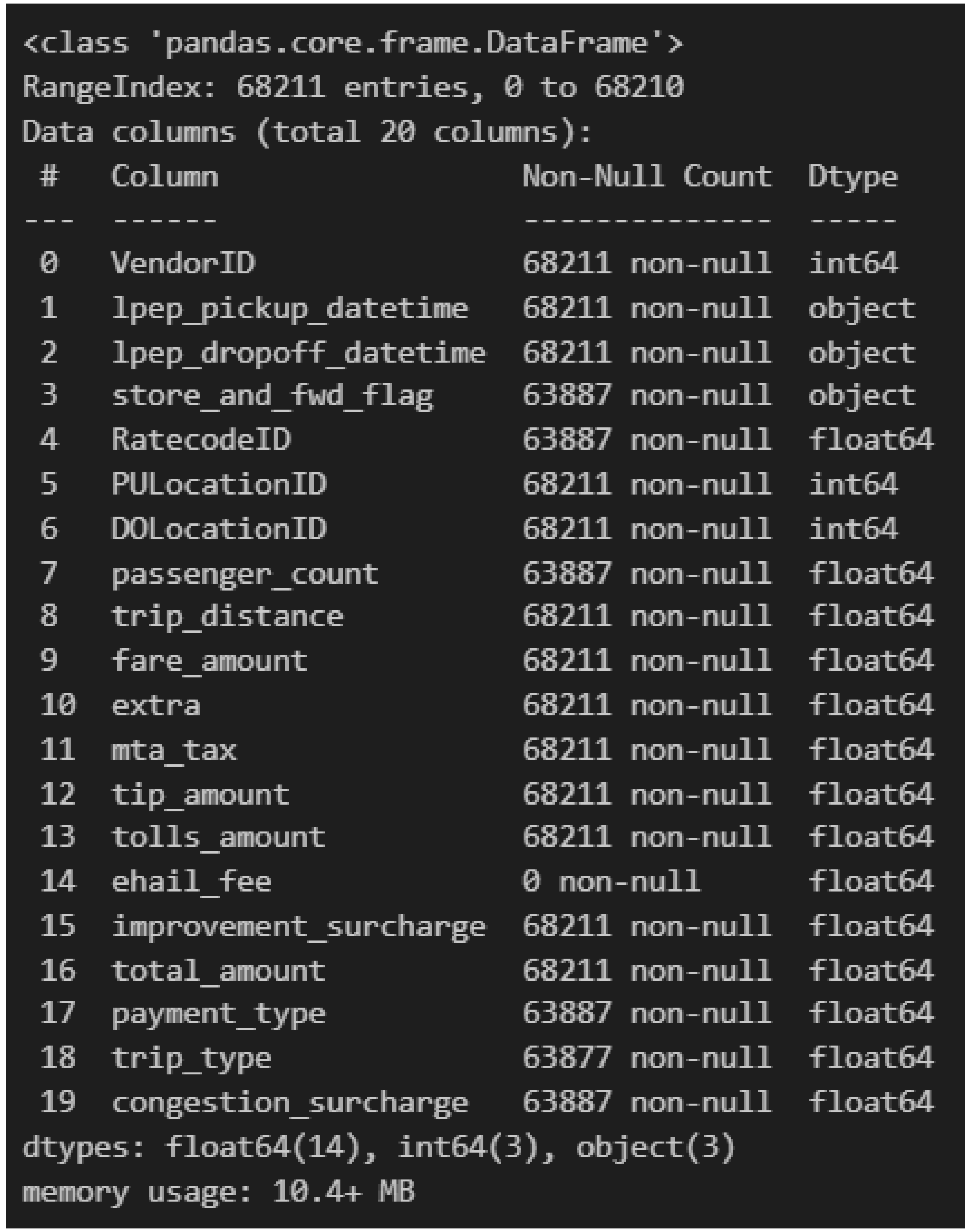


Feature	Description
VendorID	A code indicating the LPEP provider that provided the record. - 1 = Creative Mobile Technologies, LLC. - 2 = VeriFone Inc.
lpep_pickup_datetime	The date and time when the meter was engaged.
lpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles as reported by the taximeter.
PULocationID	TLC Taxi Zone in which the taximeter was engaged.
DOLocationID	TLC Taxi Zone in which the taximeter was disengaged.
RateCodeID	The final rate code in effect at the end of the trip. - 1 = Standard rate - 2 = JFK - 3 = Newark - 4 = Nassau or Westchester - 5 = Negotiated fare - 6 = Group ride
Store_and_fwd_flag	Indicates whether the trip record was held in vehicle memory before sending to the vendor due to lack of server connection. - Y = Store and forward trip - N = Not a store and forward trip
Payment_type	Code for how the passenger paid for the trip. - 1 = Credit card - 2 = Cash - 3 = No charge - 4 = Dispute - 5 = Unknown - 6 = Voided trip



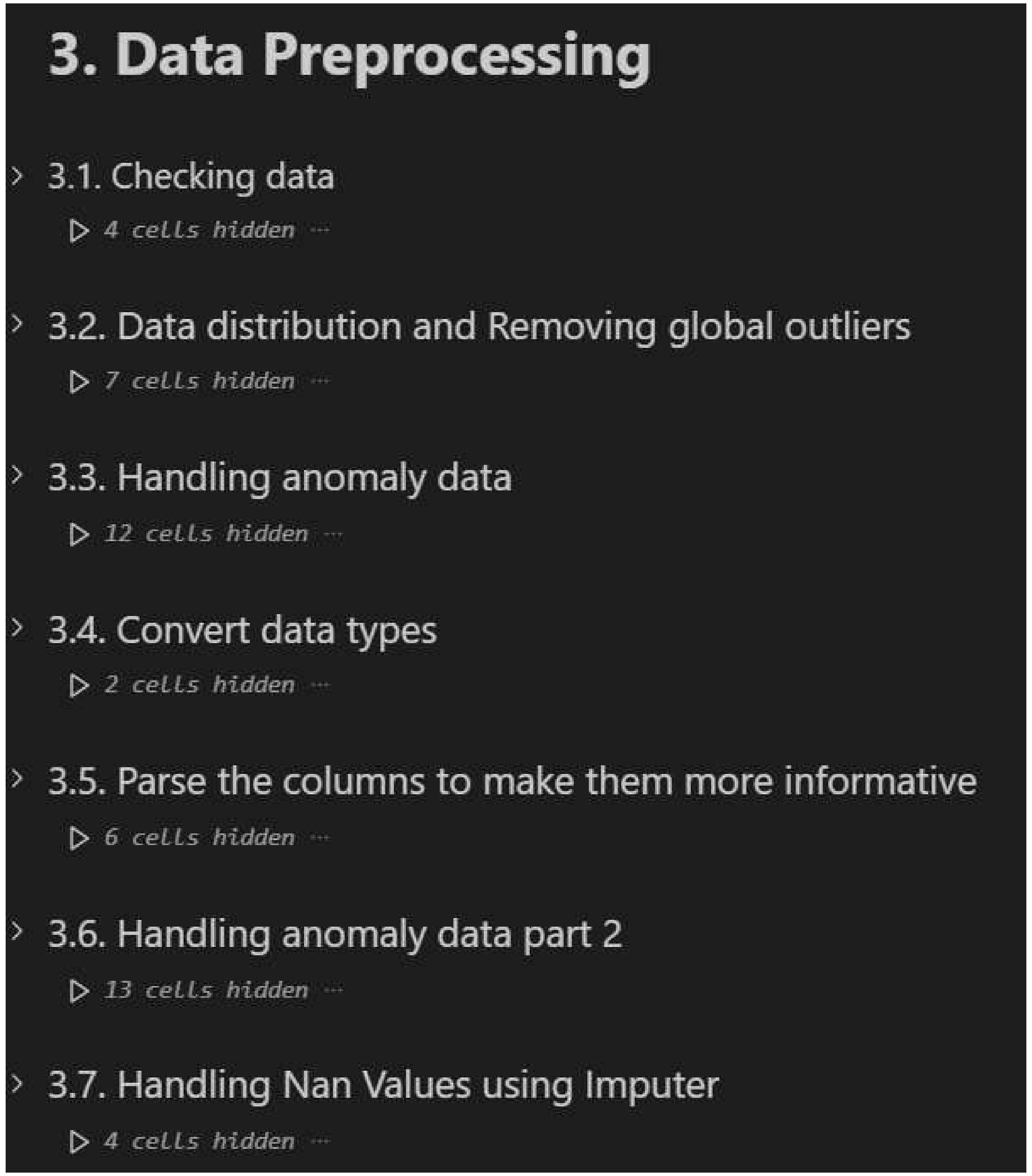
Fare_amount	The time-and-distance fare calculated by the meter, including extras and surcharges like $0.50 and $1 rush hour and overnight charges.
MTA_tax	$0.50 MTA tax automatically triggered based on the metered rate in use.
Improvement_surcharge	$0.30 improvement surcharge assessed on hailed trips at flag drop; introduced in 2015.
Tip_amount	Automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	The total amount of all tolls paid during the trip.
Total_amount	The total amount charged to passengers, excluding cash tips.
Trip_type	Code indicating whether the trip was a street-hail or a dispatch. Automatically assigned but can be modified by the driver. - 1 = Street-hail - 2 = Dispatch

DataFrame Information



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68211 entries, 0 to 68210
Data columns (total 20 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   VendorID               68211 non-null  int64  
 1   lpep_pickup_datetime   68211 non-null  object 
 2   lpep_dropoff_datetime  68211 non-null  object 
 3   store_and_fwd_flag     63887 non-null  object 
 4   RatecodeID             63887 non-null  float64
 5   PULocationID           68211 non-null  int64  
 6   DOLocationID           68211 non-null  int64  
 7   passenger_count        63887 non-null  float64
 8   trip_distance          68211 non-null  float64
 9   fare_amount            68211 non-null  float64
 10  extra                  68211 non-null  float64
 11  mta_tax                68211 non-null  float64
 12  tip_amount             68211 non-null  float64
 13  tolls_amount           68211 non-null  float64
 14  ehail_fee              0 non-null      float64
 15  improvement_surcharge  68211 non-null  float64
 16  total_amount           68211 non-null  float64
 17  payment_type           63887 non-null  float64
 18  trip_type              63877 non-null  float64
 19  congestion_surcharge   63887 non-null  float64
dtypes: float64(14), int64(3), object(3)
memory usage: 10.4+ MB
```

Data Preprocessing and EDA in Jupyter Notebook



3. Data Preprocessing

› 3.1. Checking data

▷ *4 cells hidden* ···

› 3.2. Data distribution and Removing global outliers

▷ *7 cells hidden* ···

› 3.3. Handling anomaly data

▷ *12 cells hidden* ···

› 3.4. Convert data types

▷ *2 cells hidden* ···

› 3.5. Parse the columns to make them more informative

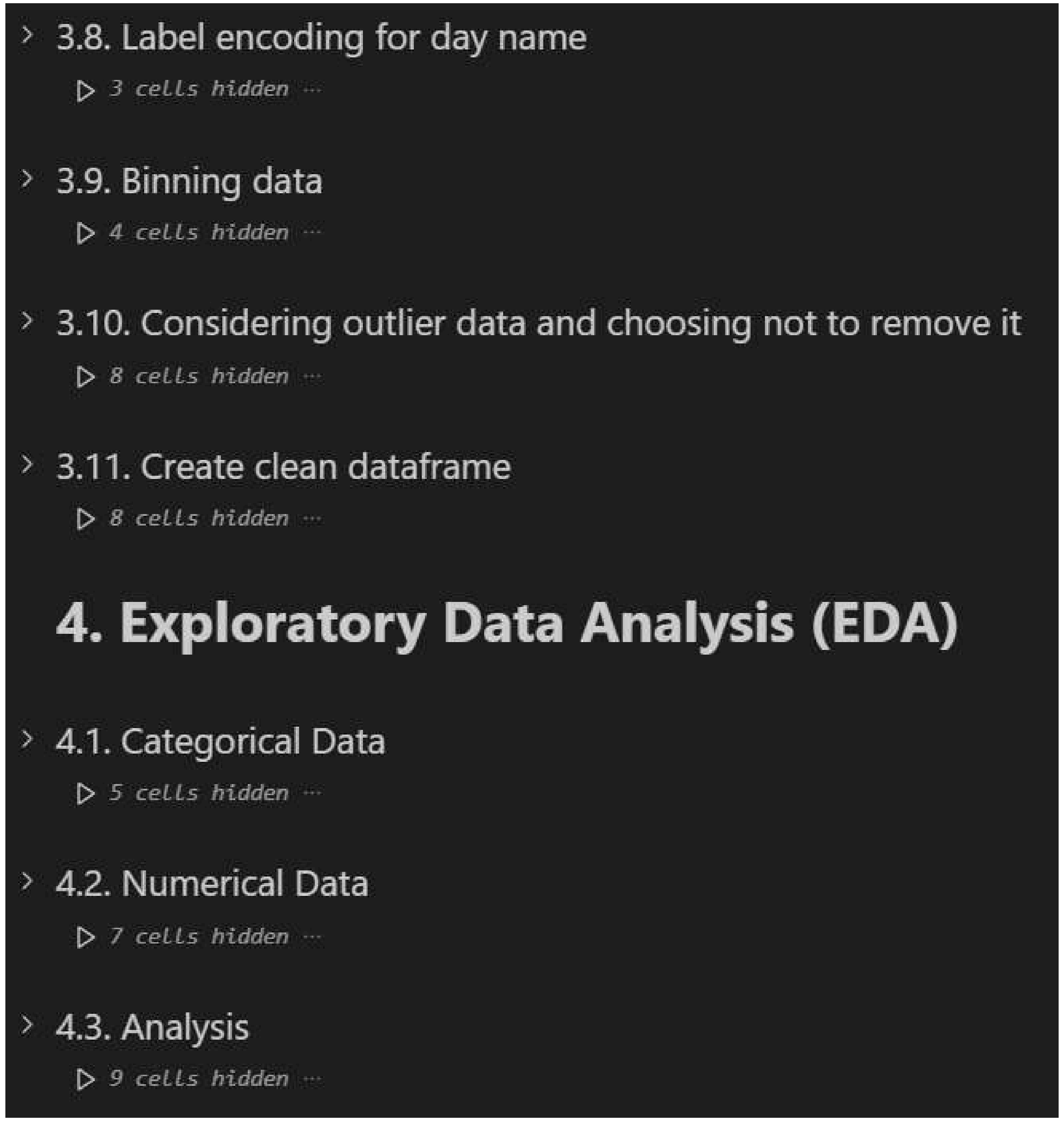
▷ *6 cells hidden* ···

› 3.6. Handling anomaly data part 2

▷ *13 cells hidden* ···

› 3.7. Handling Nan Values using Imputer

▷ *4 cells hidden* ···



› 3.8. Label encoding for day name

▷ *3 cells hidden* ···

› 3.9. Binning data

▷ *4 cells hidden* ···

› 3.10. Considering outlier data and choosing not to remove it

▷ *8 cells hidden* ···

› 3.11. Create clean dataframe

▷ *8 cells hidden* ···

4. Exploratory Data Analysis (EDA)

› 4.1. Categorical Data

▷ *5 cells hidden* ···

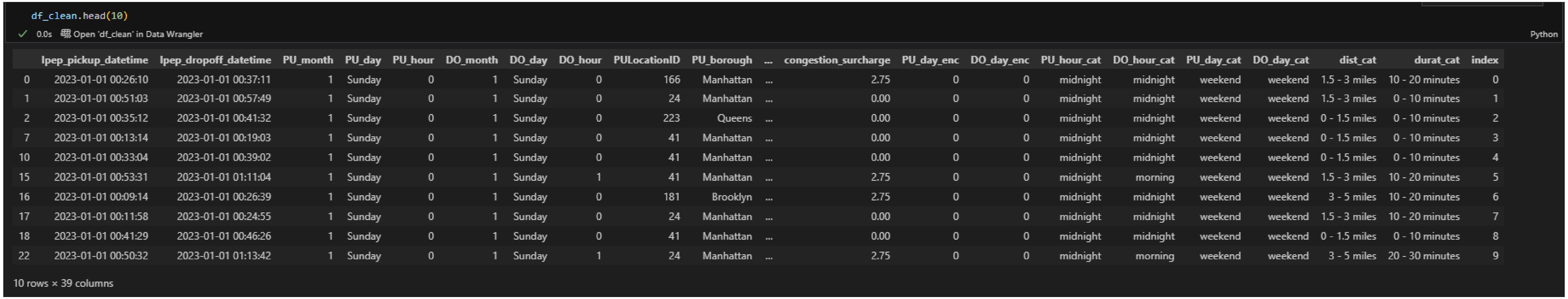
› 4.2. Numerical Data

▷ *7 cells hidden* ···

› 4.3. Analysis

▷ *9 cells hidden* ···

Clean DataFrame



```
df_clean.head(10)
```

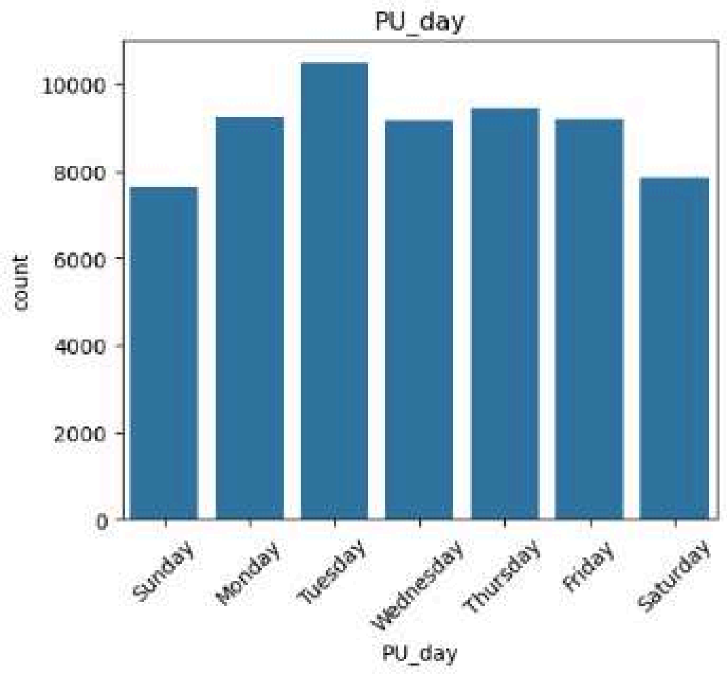
✓ 0.0s Open 'df_clean' in Data Wrangler

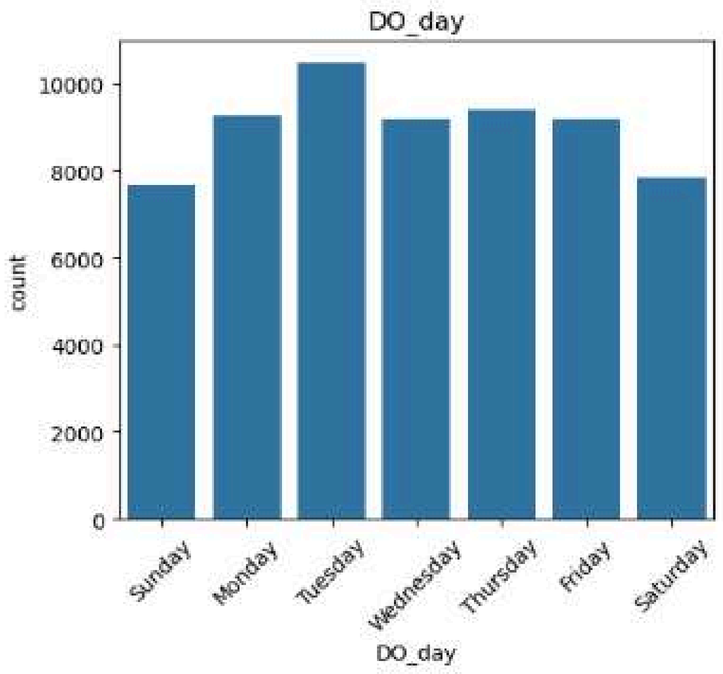
	lpep_pickup_datetime	lpep_dropoff_datetime	PU_month	PU_day	PU_hour	DO_month	DO_day	DO_hour	PULocationID	PU_borough	...	congestion_surcharge	PU_day_enc	DO_day_enc	PU_hour_cat	DO_hour_cat	PU_day_cat	DO_day_cat	dist_cat	durat_cat	index
0	2023-01-01 00:26:10	2023-01-01 00:37:11	1	Sunday	0	1	Sunday	0	166	Manhattan	...	2.75	0	0	midnight	midnight	weekend	weekend	1.5 - 3 miles	10 - 20 minutes	0
1	2023-01-01 00:51:03	2023-01-01 00:57:49	1	Sunday	0	1	Sunday	0	24	Manhattan	...	0.00	0	0	midnight	midnight	weekend	weekend	1.5 - 3 miles	0 - 10 minutes	1
2	2023-01-01 00:35:12	2023-01-01 00:41:32	1	Sunday	0	1	Sunday	0	223	Queens	...	0.00	0	0	midnight	midnight	weekend	weekend	0 - 1.5 miles	0 - 10 minutes	2
7	2023-01-01 00:13:14	2023-01-01 00:19:03	1	Sunday	0	1	Sunday	0	41	Manhattan	...	0.00	0	0	midnight	midnight	weekend	weekend	0 - 1.5 miles	0 - 10 minutes	3
10	2023-01-01 00:33:04	2023-01-01 00:39:02	1	Sunday	0	1	Sunday	0	41	Manhattan	...	0.00	0	0	midnight	midnight	weekend	weekend	0 - 1.5 miles	0 - 10 minutes	4
15	2023-01-01 00:53:31	2023-01-01 01:11:04	1	Sunday	0	1	Sunday	1	41	Manhattan	...	2.75	0	0	midnight	morning	weekend	weekend	1.5 - 3 miles	10 - 20 minutes	5
16	2023-01-01 00:09:14	2023-01-01 00:26:39	1	Sunday	0	1	Sunday	0	181	Brooklyn	...	2.75	0	0	midnight	midnight	weekend	weekend	3 - 5 miles	10 - 20 minutes	6
17	2023-01-01 00:11:58	2023-01-01 00:24:55	1	Sunday	0	1	Sunday	0	24	Manhattan	...	0.00	0	0	midnight	midnight	weekend	weekend	1.5 - 3 miles	10 - 20 minutes	7
18	2023-01-01 00:41:29	2023-01-01 00:46:26	1	Sunday	0	1	Sunday	0	41	Manhattan	...	0.00	0	0	midnight	midnight	weekend	weekend	0 - 1.5 miles	0 - 10 minutes	8
22	2023-01-01 00:50:32	2023-01-01 01:13:42	1	Sunday	0	1	Sunday	1	24	Manhattan	...	2.75	0	0	midnight	morning	weekend	weekend	3 - 5 miles	20 - 30 minutes	9

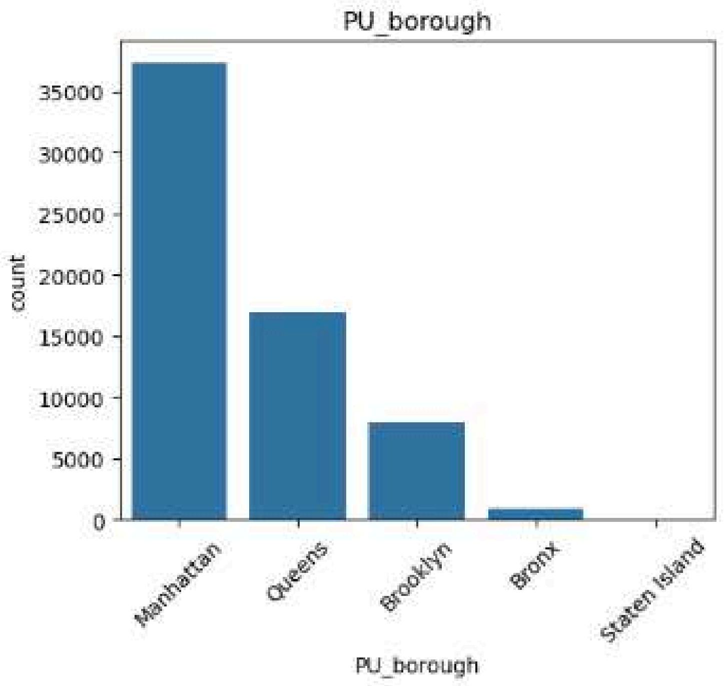
10 rows × 39 columns

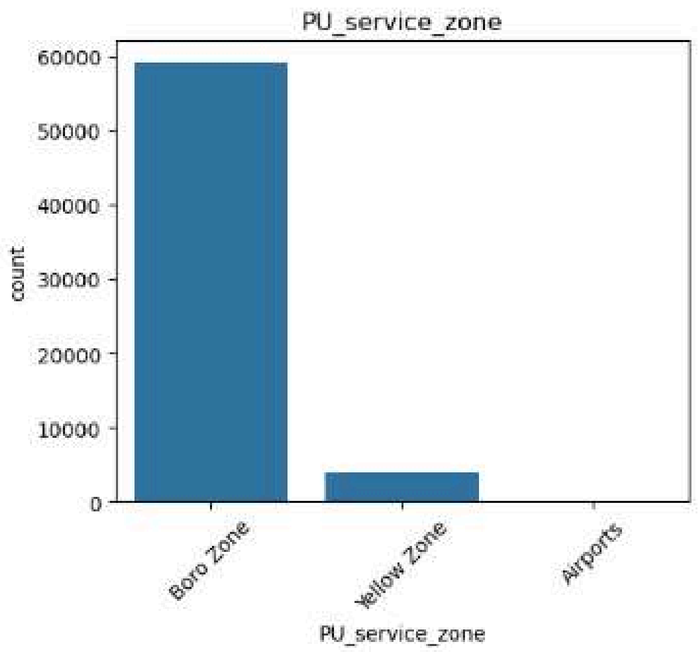
3. EXPLORATORY DATA ANALYSIS (EDA)

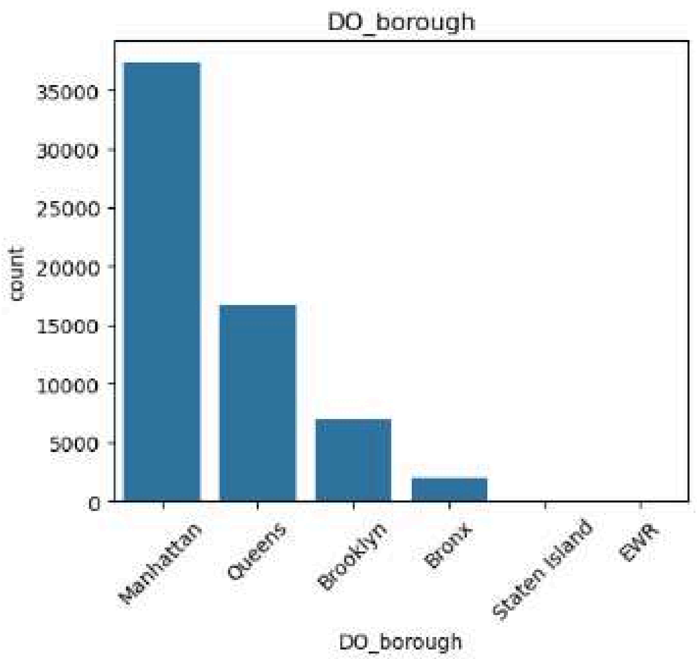
Categorical Data Distribution

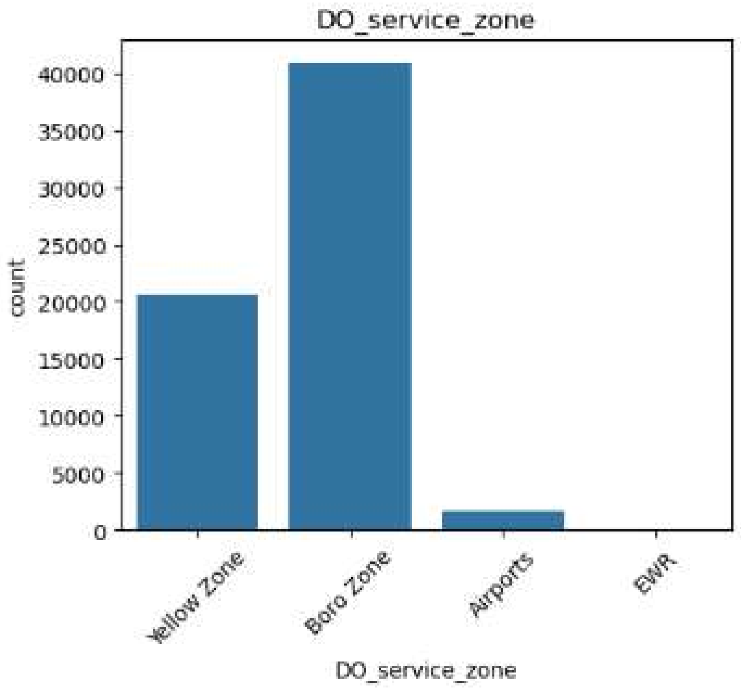




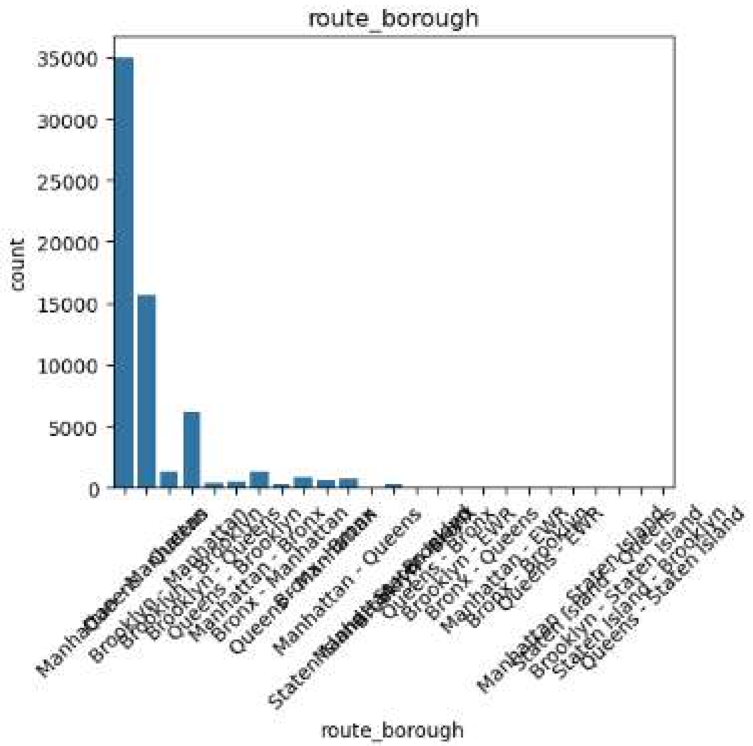


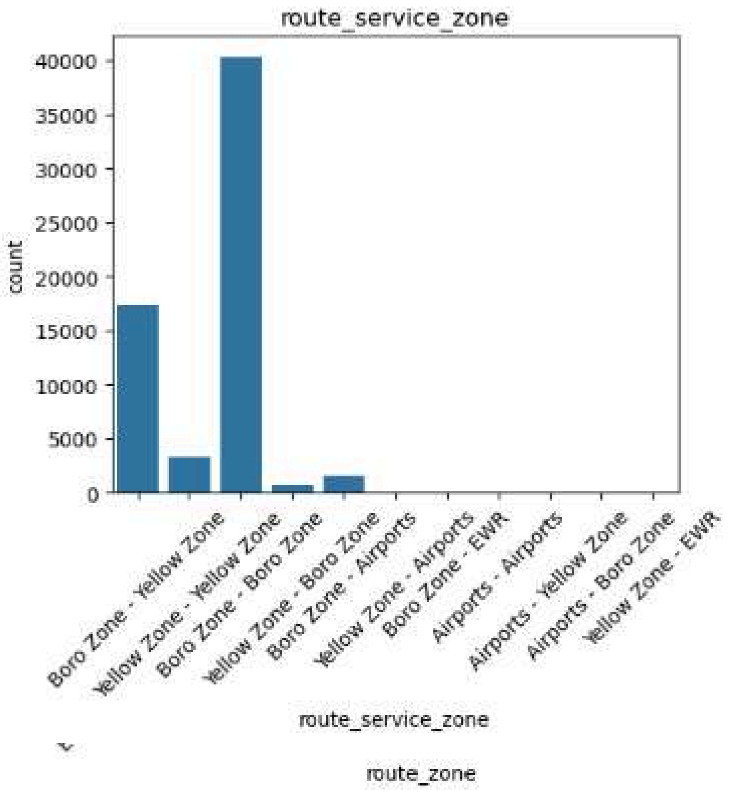


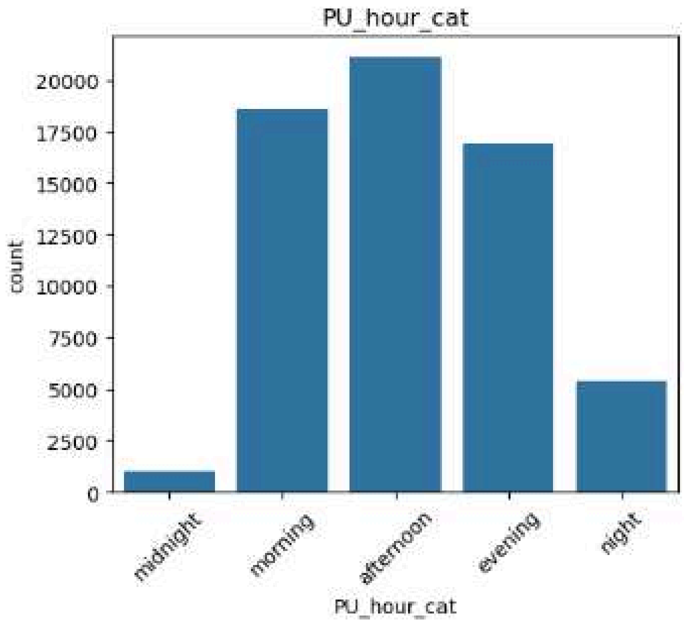


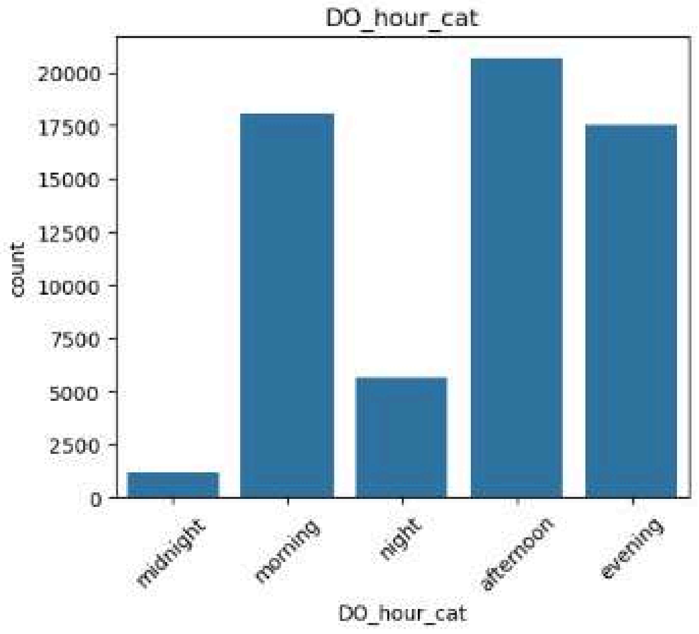


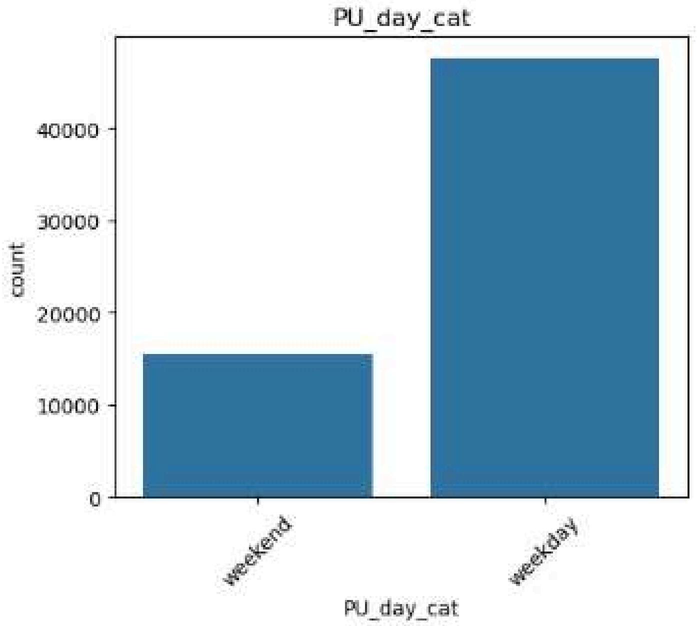
Categorical Data Distribution

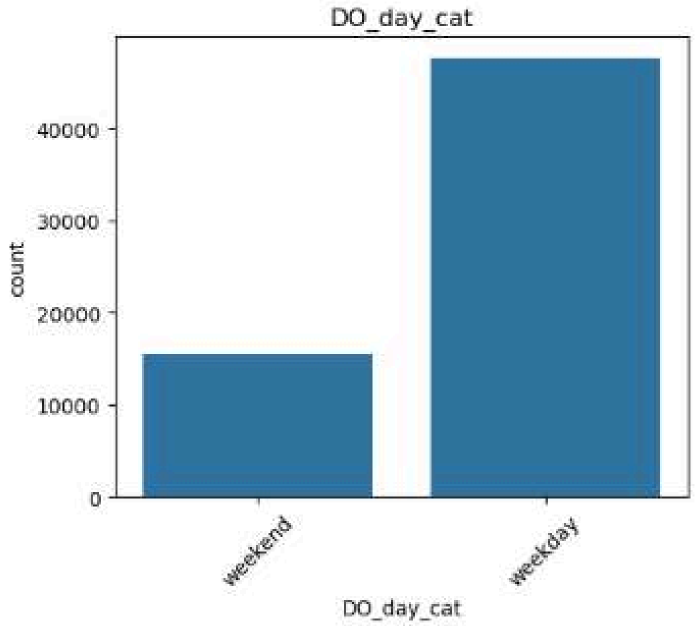


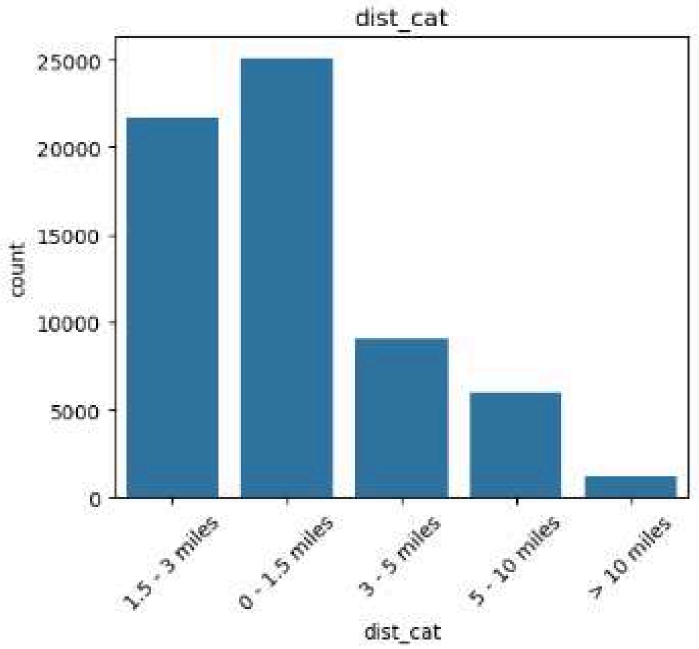




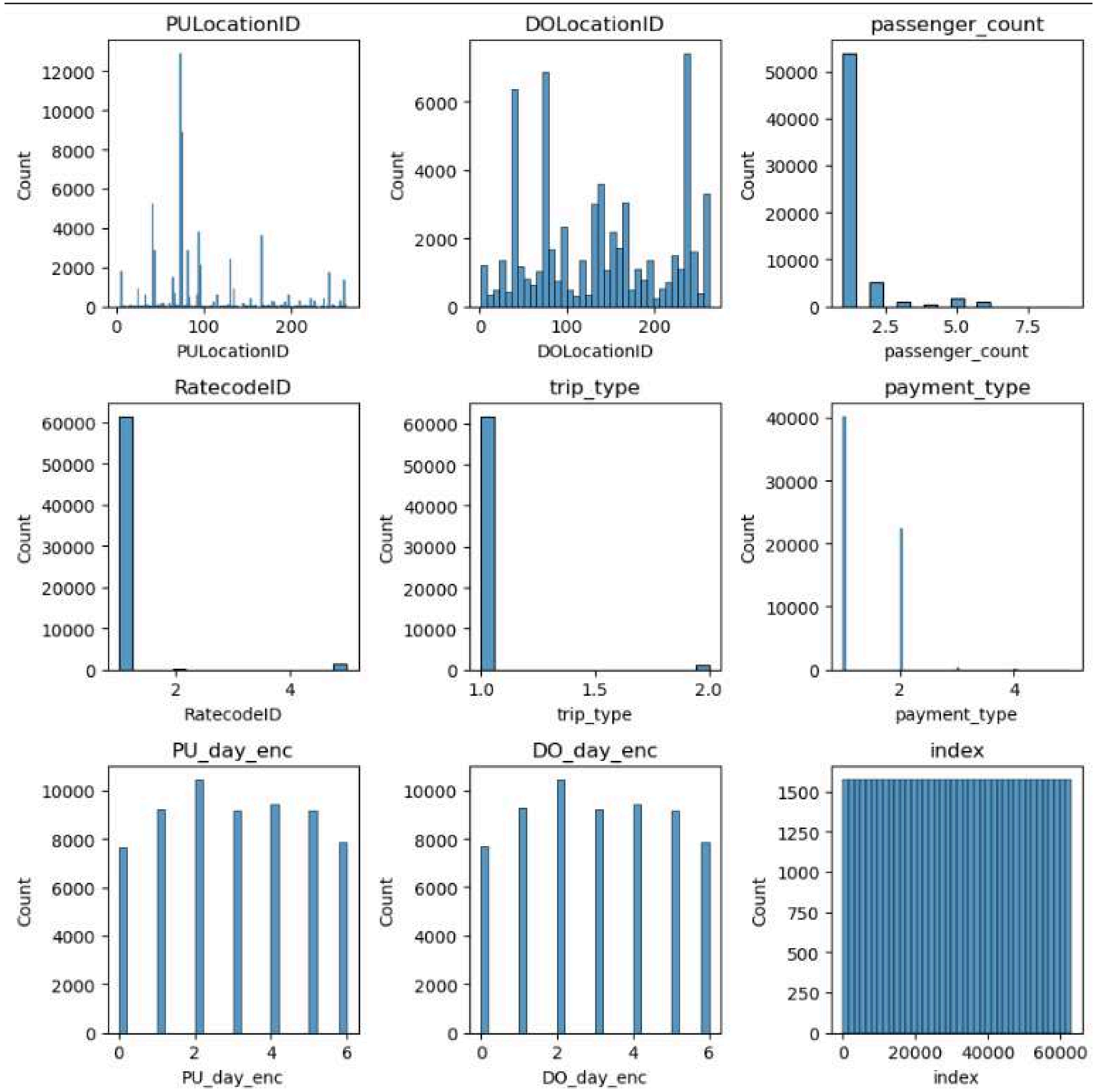




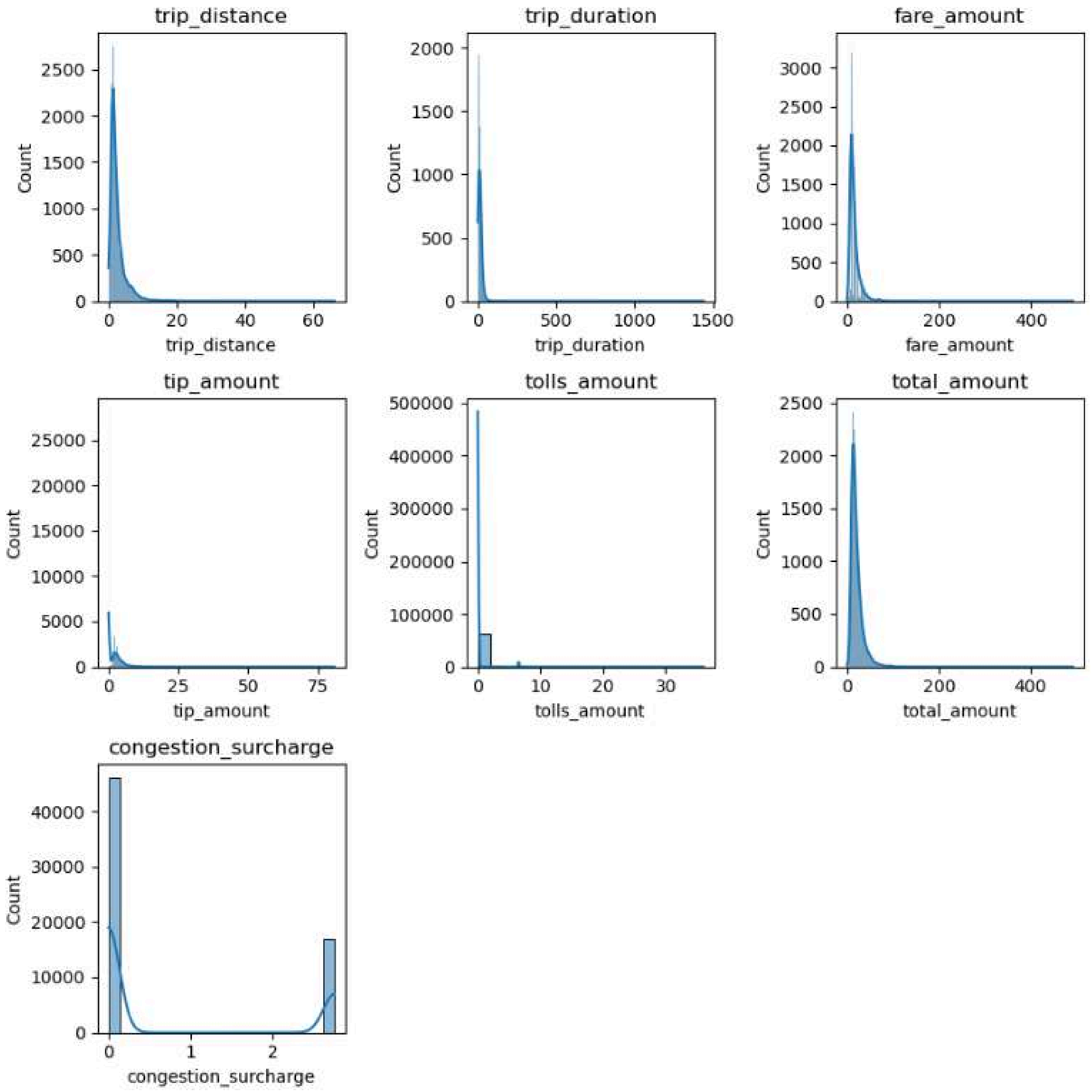




Numerical Data Distribution



Numerical Data Distribution



4. PROJECT OBJECTIVES

Objectives

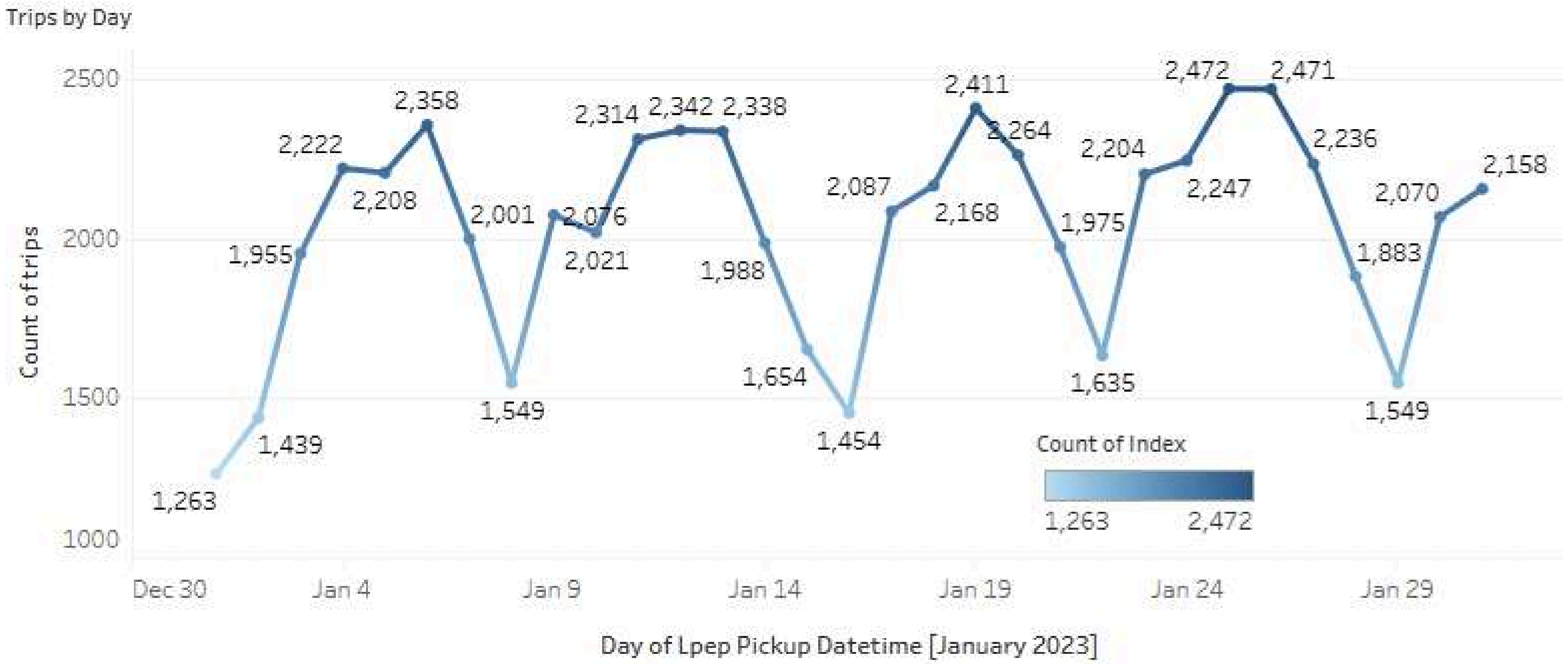
Identifying strategic locations and times to increase the company's revenue.

Approach

Analyzing customer preferences in using New York taxi services.

5. TABLEAU VISUALIZATION AND STATISTIC

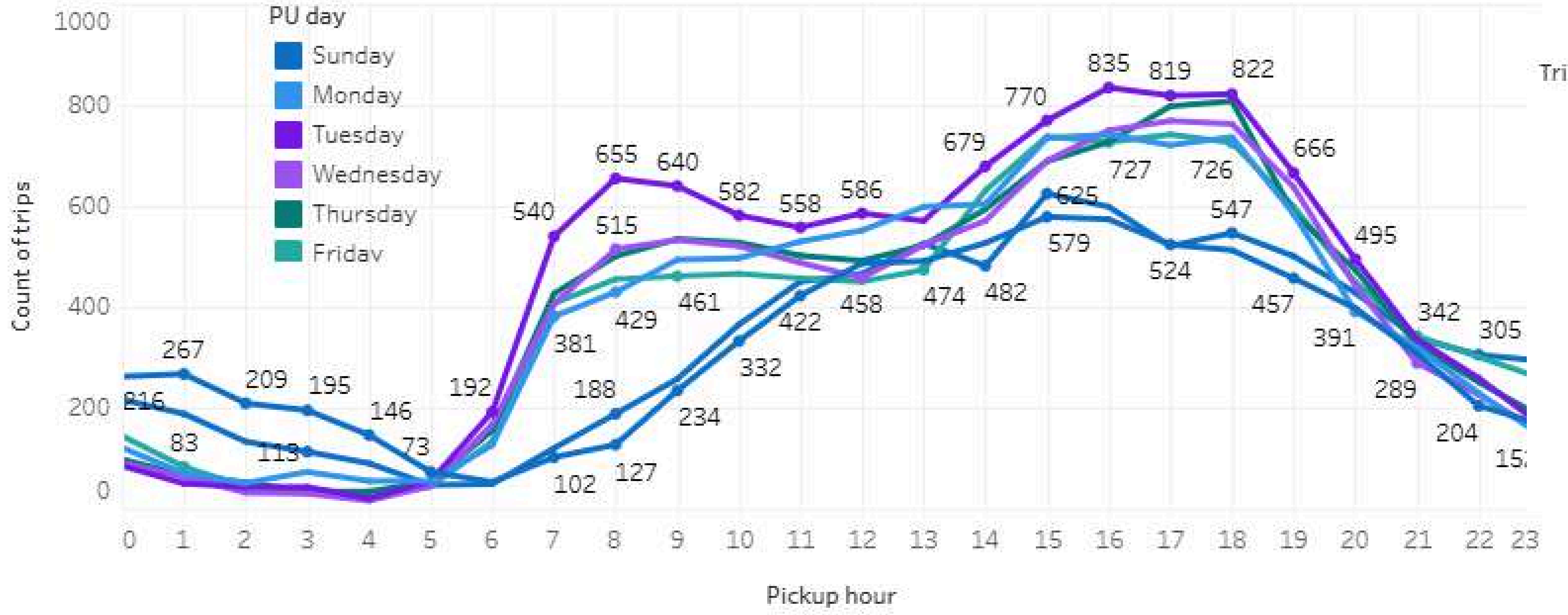
1. Trips by Time Analysis



The number of taxi trips fluctuates, forming a wave-like pattern. The peaks occur on Thursdays (January 12, 19, 26, 2023) and Fridays (January 6, 13, 20, 27, 2023), while the troughs fall on Sundays (January 1, 8, 22, 29, 2023).

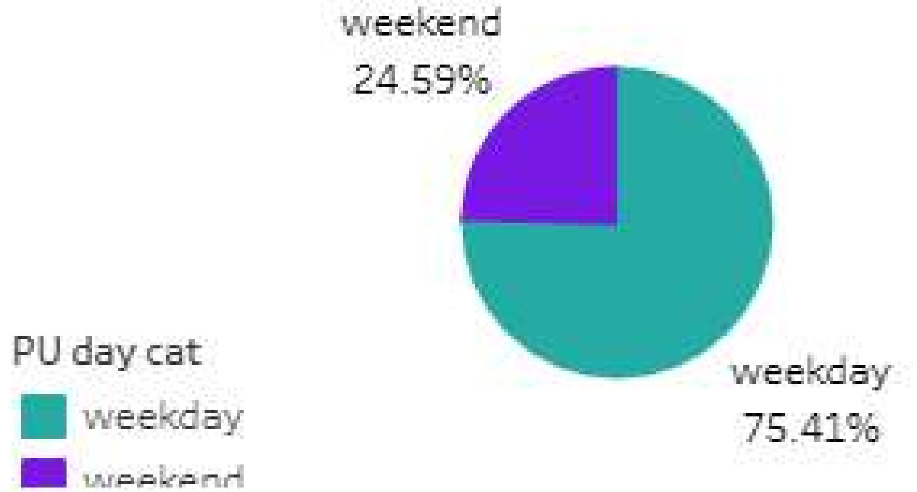
1. **Trips by Time Analysis**

Trips by Hour for each Day

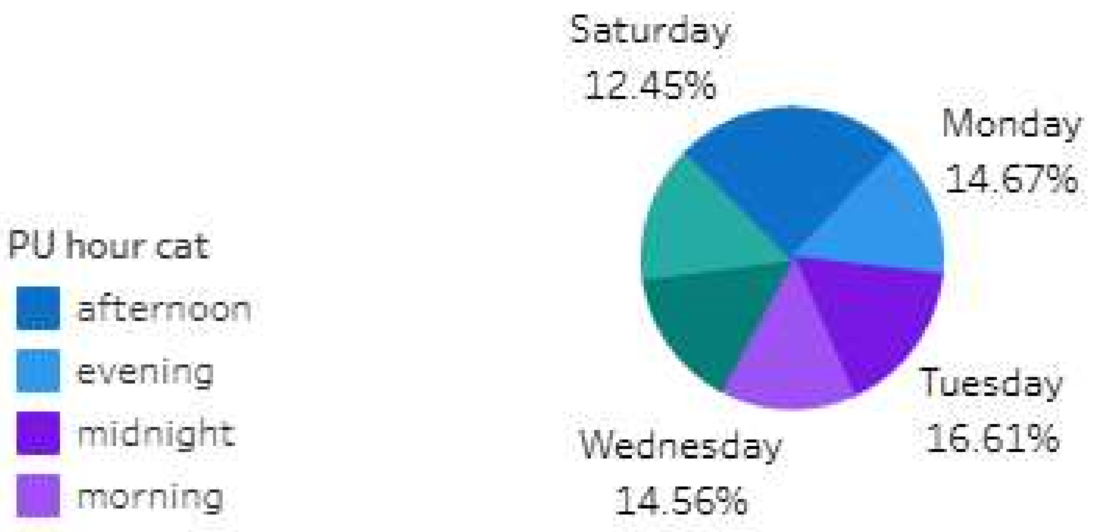


Cumulatively, the highest number of trips occurs on **Tuesdays**. **The peak hours are from 7-9 a.m. and 3-5 p.m.**, or during the morning and afternoon. This suggests that most trips take place during commute times, when people are traveling to and from work or school on weekdays.

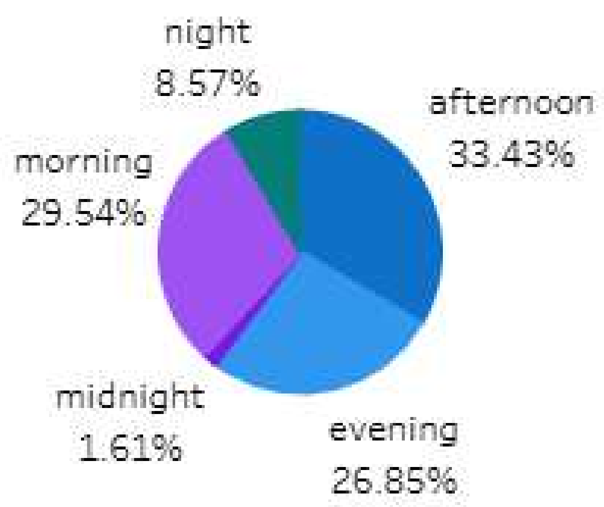
Trips by Day Type



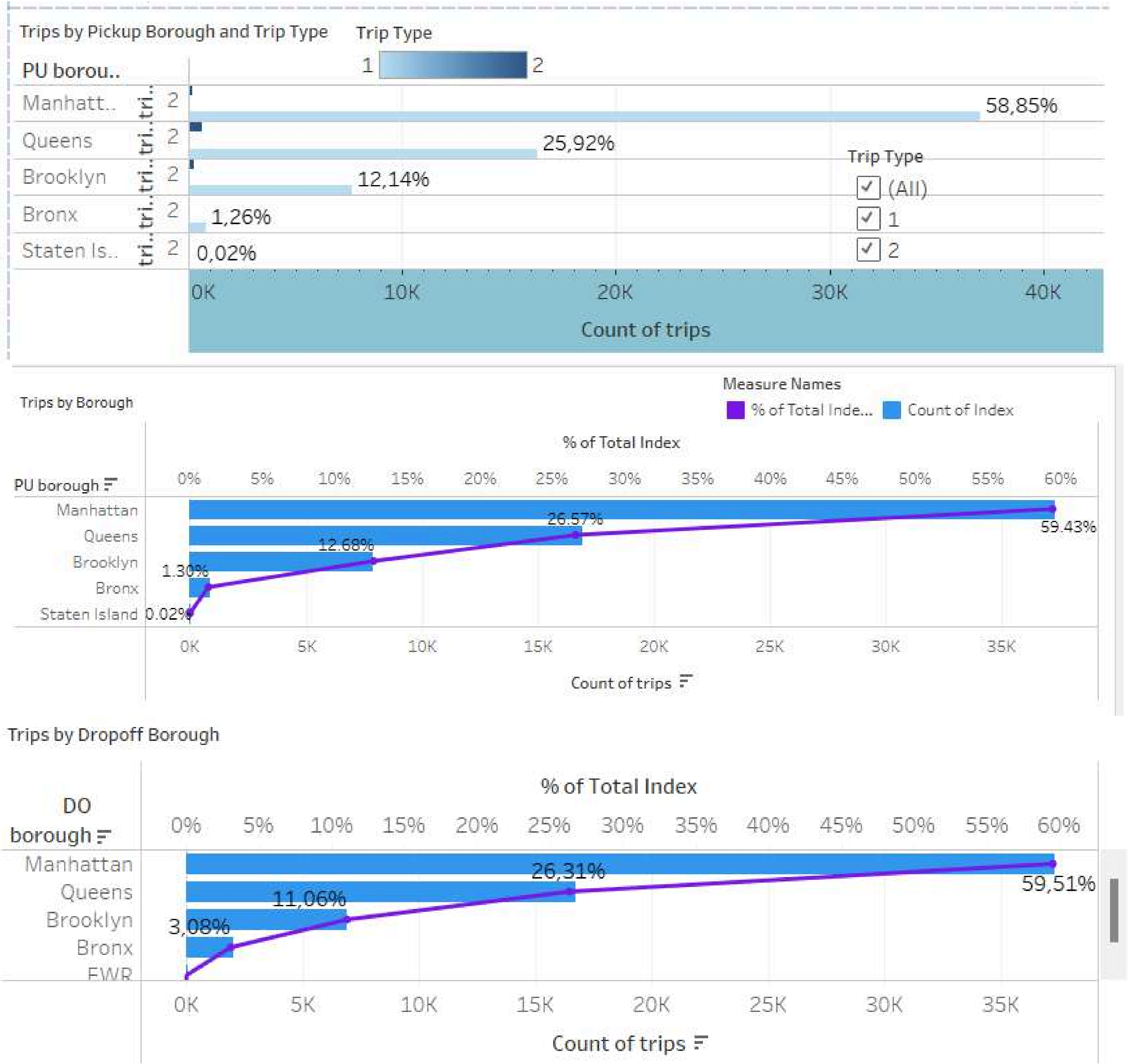
Trips by Day of the Week



Trips by Time Category



2. Trips by Location Analysis



Trip Type

☑ (All)

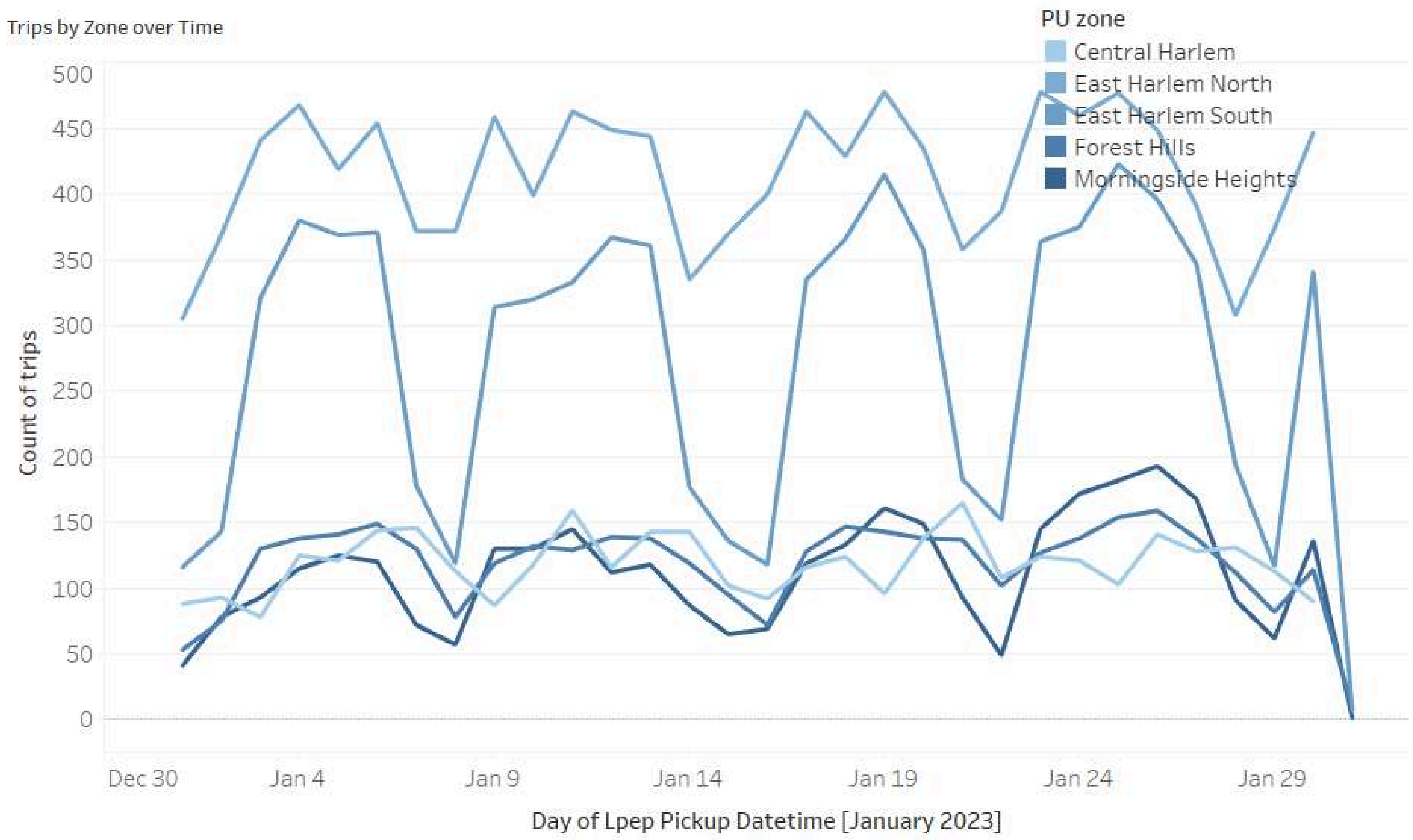
☑ 1

☑ 2

Manhattan is the borough with the highest number of pickups, indicating that the majority of taxi users are in this area.

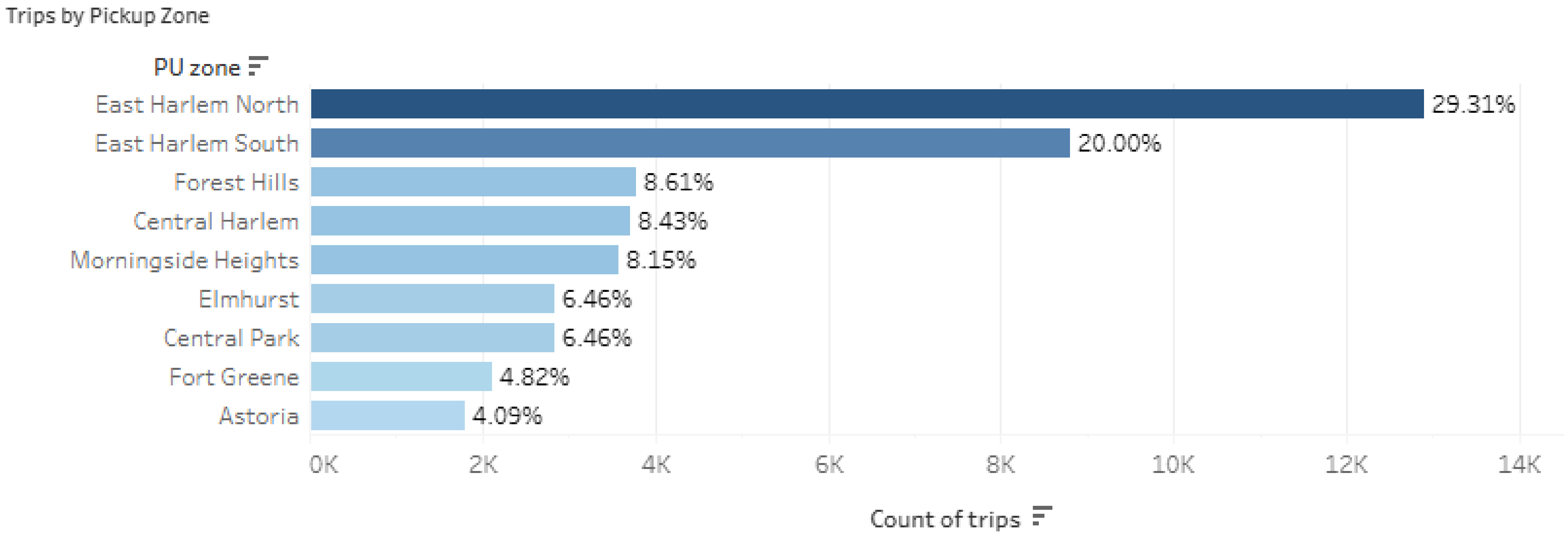
More than half (58.85%) of the trips in Manhattan are also **street hail** rides, which aligns with Manhattan's dense population and high traffic.

2. Trips by Location Analysis



Trips by Zone over Time

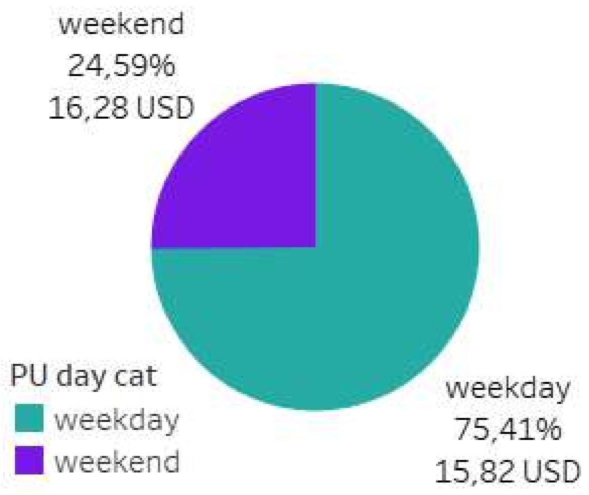
The zones with the highest number of pickups are East Harlem North and East Harlem South, both of which are part of the Manhattan borough.

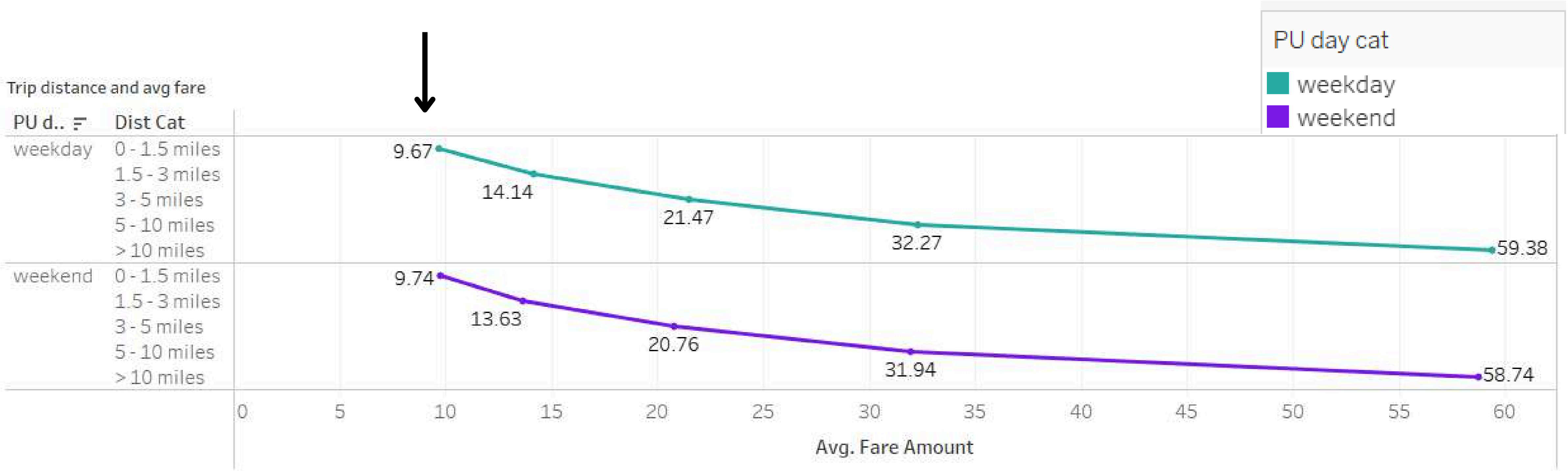


Trips by Pickup Zone

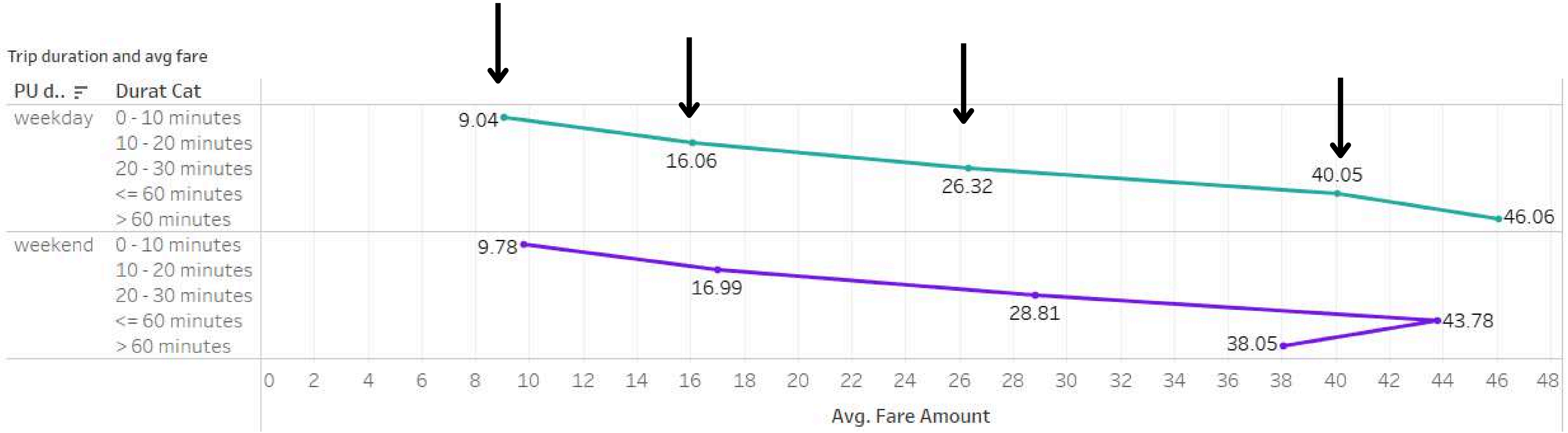
3. Fare amount and Revenue by Time

The average fare on weekdays is lower ($15.82) despite the fact that the **majority of trips** (75.41%) occur on these days.

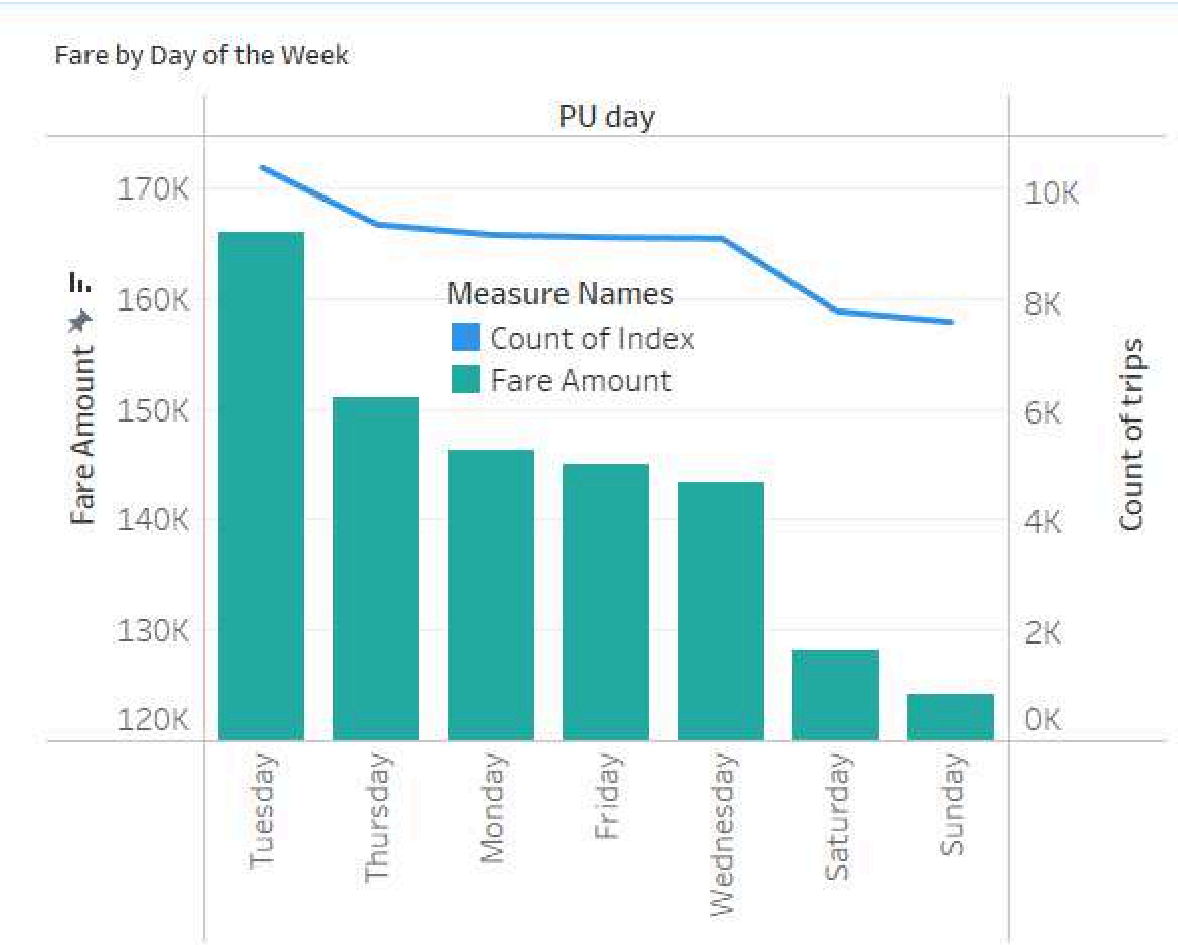




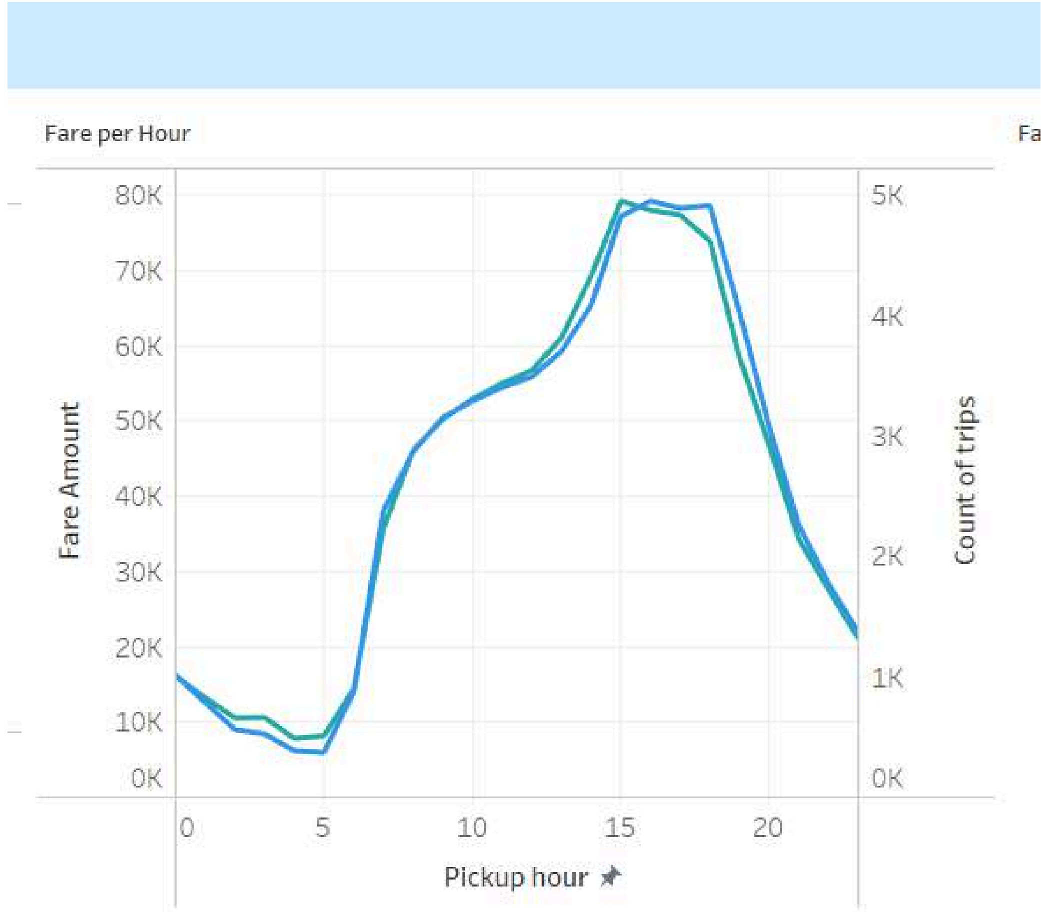
For the **same distance**, fares are **lower on weekends**, while for trips of the **same duration**, fares are **lower on weekdays.**



3. Fare amount and Revenue by Time

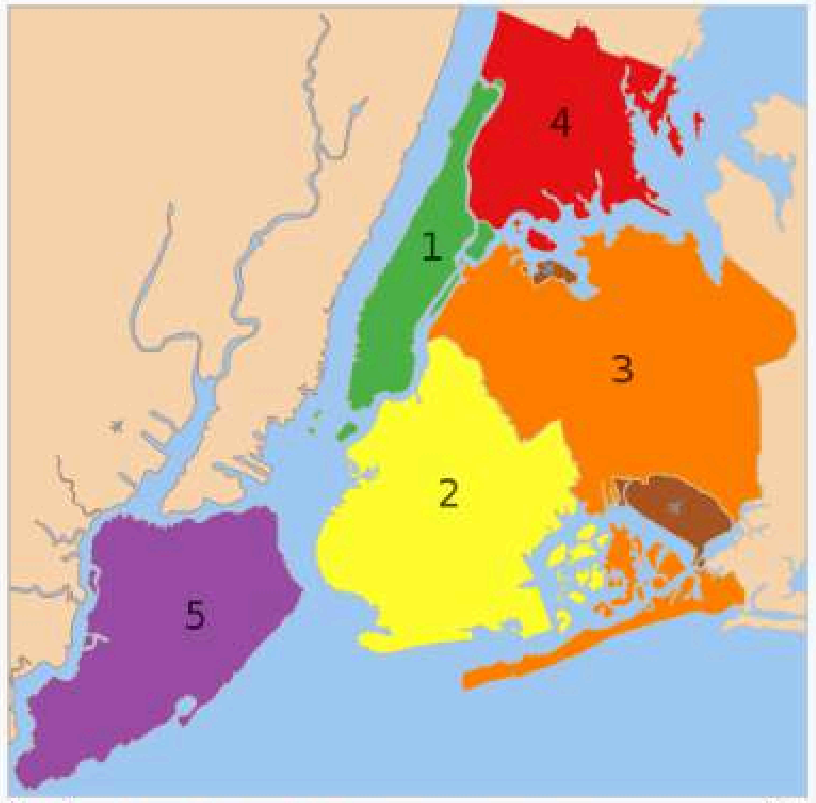


Since **Tuesday** has the highest number of trips, it also aligns with generating the **most revenue** compared to other days.



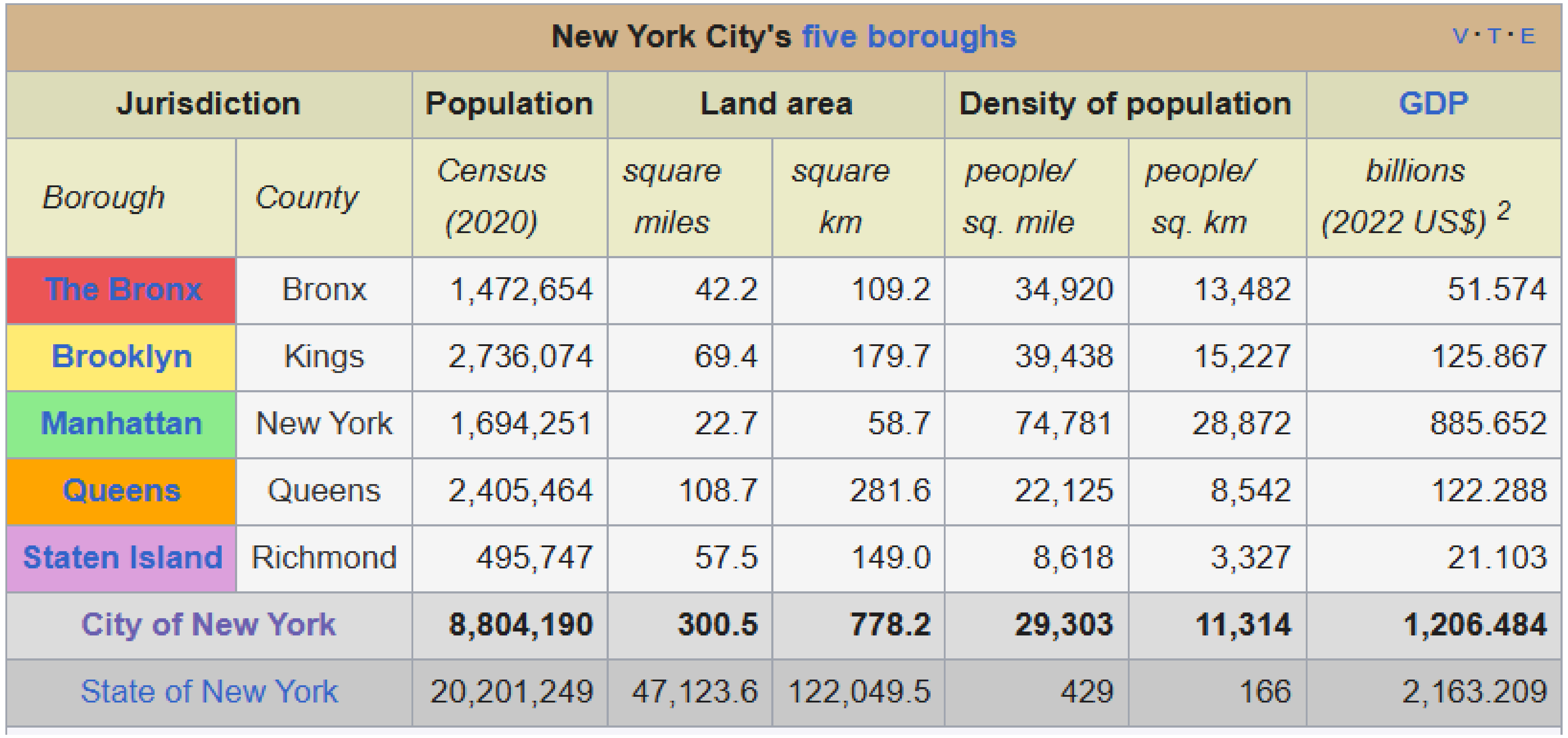
The peak hour that generates the **highest revenue is 3 p.m.**

3. Fare amount and Revenue by Location



🟩 1. Manhattan (New York County)
🟨 2. Brooklyn (Kings County)
🟧 3. Queens (Queens County)
🟥 4. The Bronx (Bronx County)
🟪 5. Staten Island (Richmond County)

Note: JFK and LGA airports are both located in Queens (marked by brown).



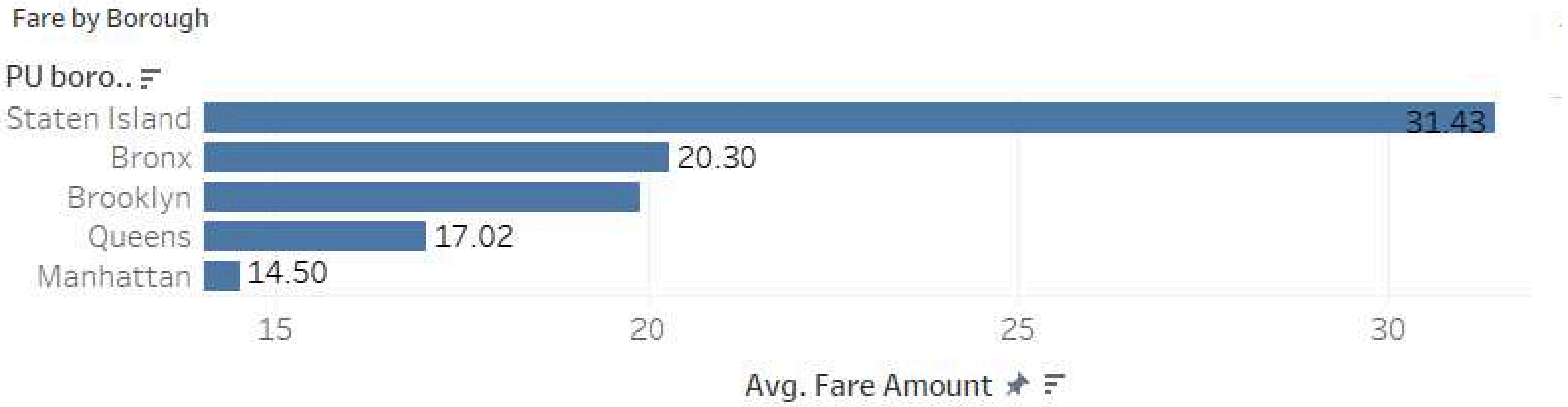
New York City's five boroughs V · T · E

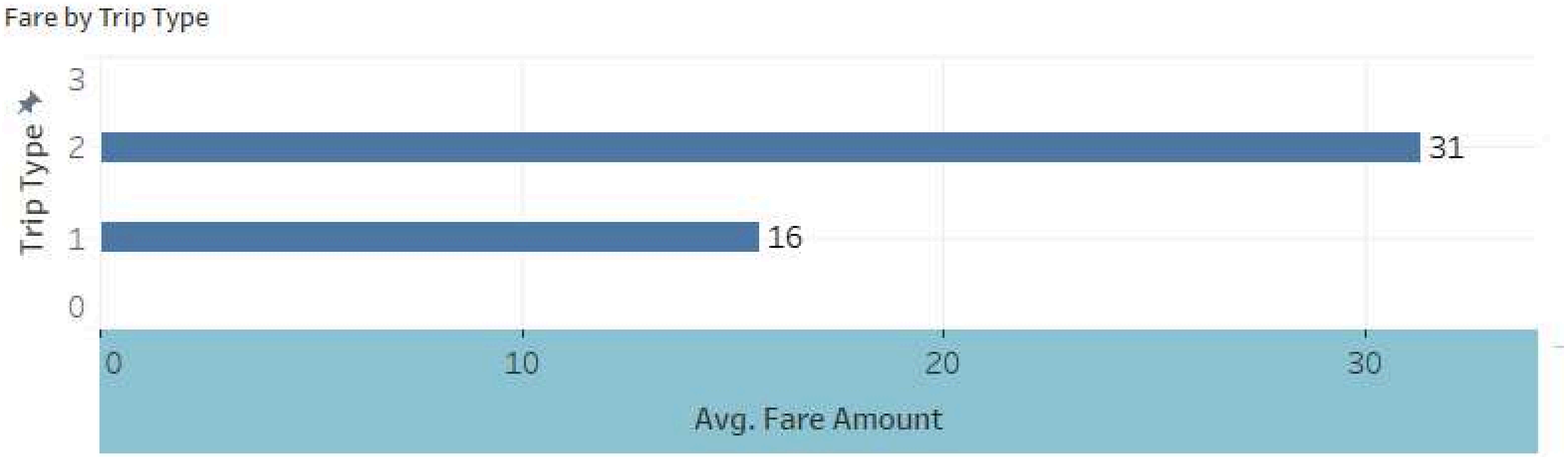
Jurisdiction		Population	Land area		Density of population		GDP
Borough	*County*	*Census (2020)*	*square miles*	*square km*	*people/ sq. mile*	*people/ sq. km*	*billions (2022 US$)* [2]
The Bronx	Bronx	1,472,654	42.2	109.2	34,920	13,482	51.574
Brooklyn	Kings	2,736,074	69.4	179.7	39,438	15,227	125.867
Manhattan	New York	1,694,251	22.7	58.7	74,781	28,872	885.652
Queens	Queens	2,405,464	108.7	281.6	22,125	8,542	122.288
Staten Island	Richmond	495,747	57.5	149.0	8,618	3,327	21.103
City of New York		**8,804,190**	**300.5**	**778.2**	**29,303**	**11,314**	**1,206.484**
State of New York		20,201,249	47,123.6	122,049.5	429	166	2,163.209

Brooklyn has the largest population among the boroughs, while **Manhattan has the highest population density as well as the highest GDP per capita**. Thus, for these two reasons, Manhattan is the borough with the highest number of taxi trips.

source : https://en.wikipedia.org/wiki/Demographics_of_New_York_City#cite_note-40 (2020-2022)

3. Fare amount and Revenue by Location



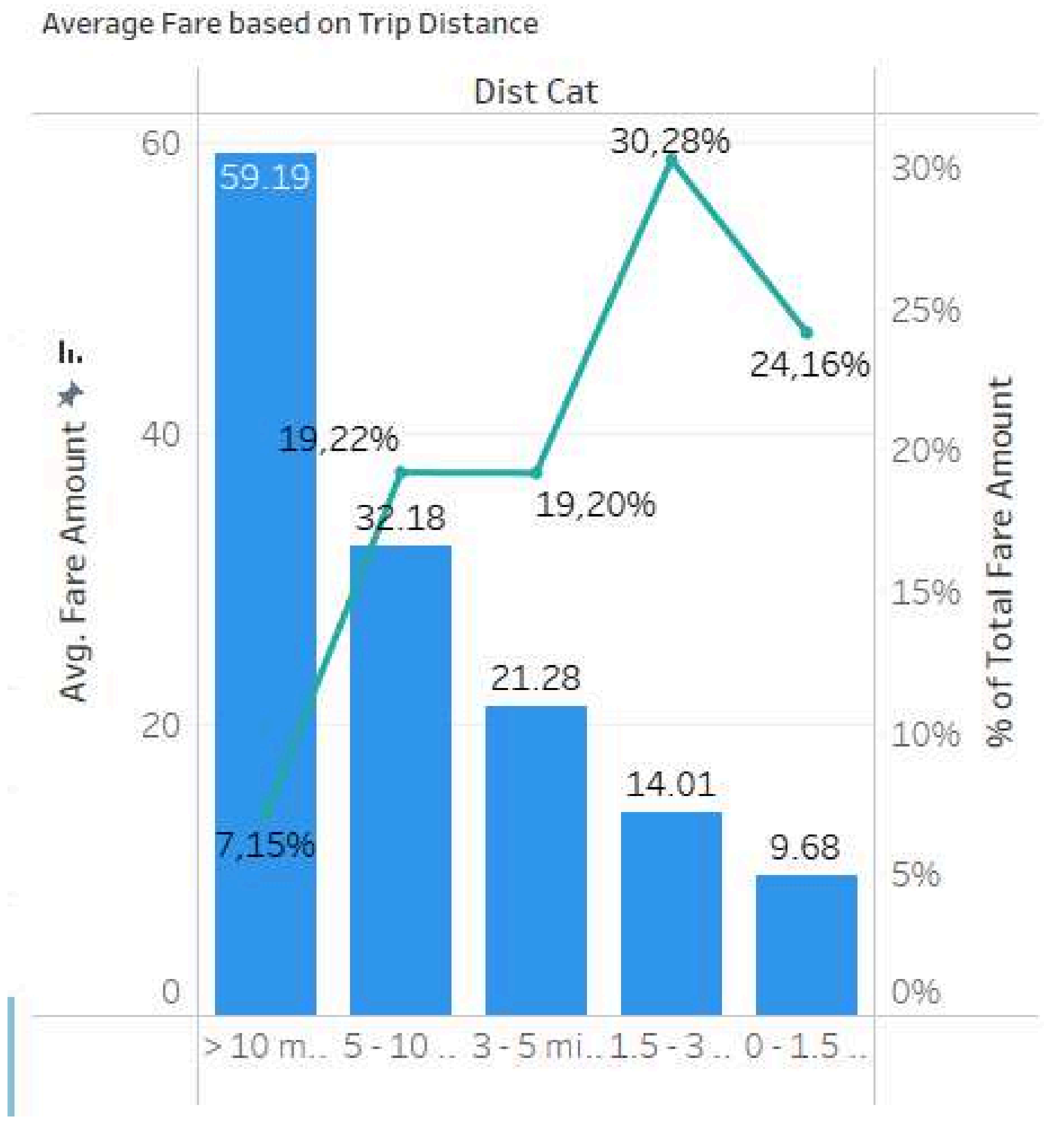


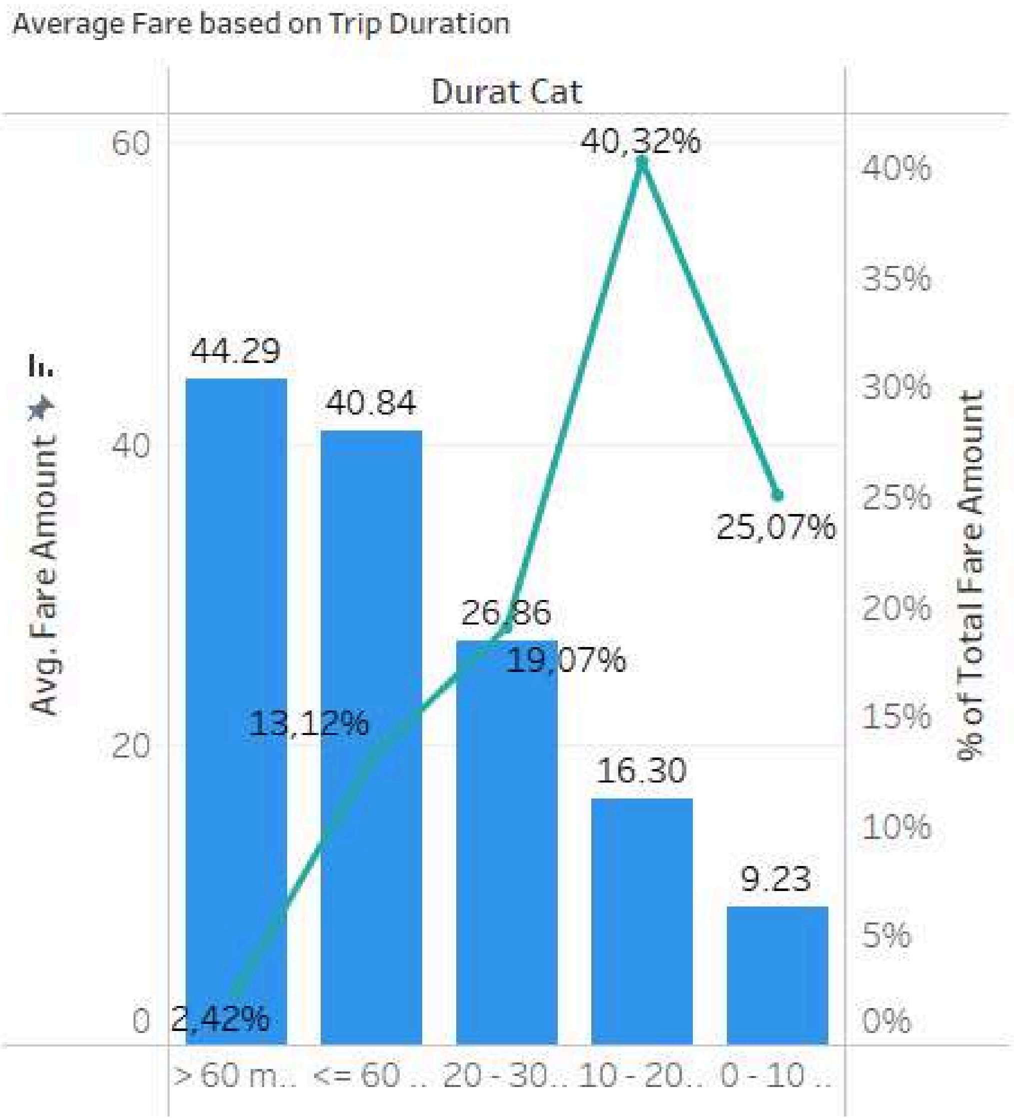
The **average fare for trips in Staten Island is the highest**, although the **number of trips** in this borough is the **lowest** compared to other boroughs.

In contrast, the **average fare in Manhattan is the lowest,** but it has the **highest** number of trips.

Additionally, **dispatch trips** have a **higher average fare** compared to street hail trips.

3. Fare amount and Revenue by Trip Distance and Trip Duration

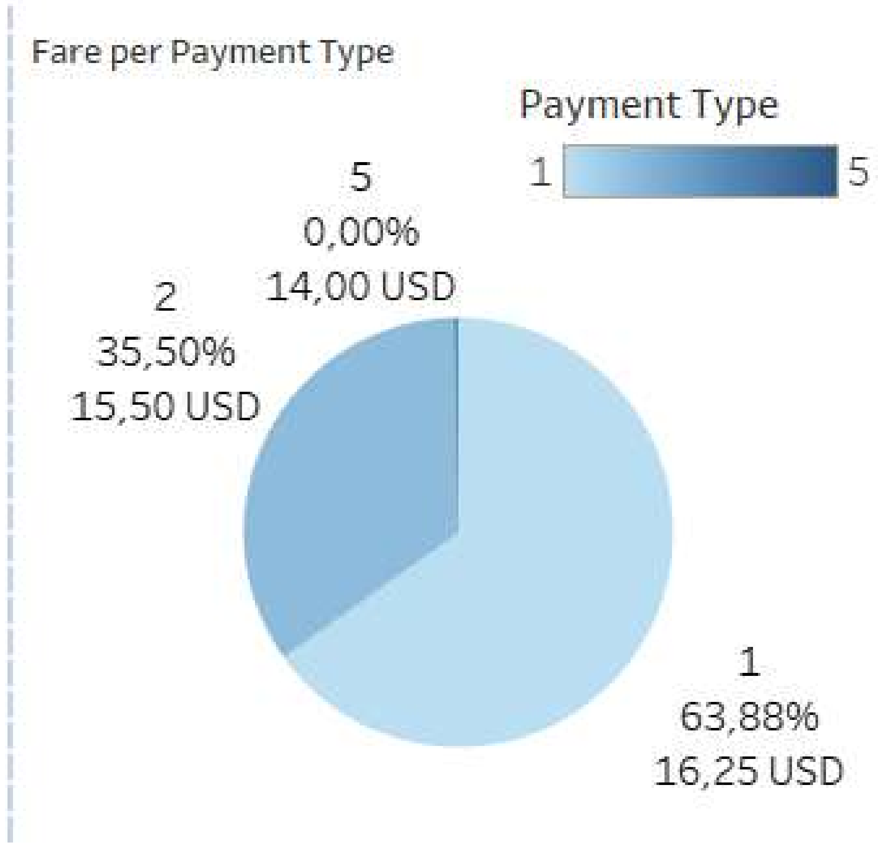




The highest average fares occur for the longest trips (>10 miles) and the longest durations (>60 minutes).

However, **most passengers travel distances of 1.5-3 miles and durations of 10-20 minutes,** both of which have the **second-lowest average fares.**

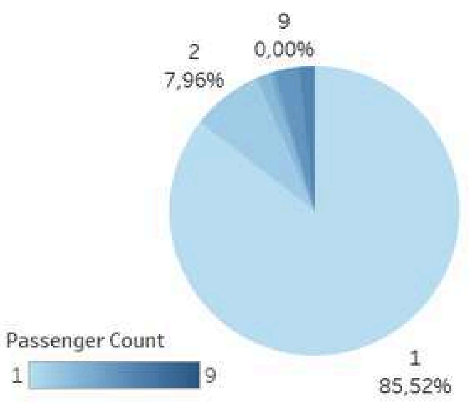
4. Passenger Preferences



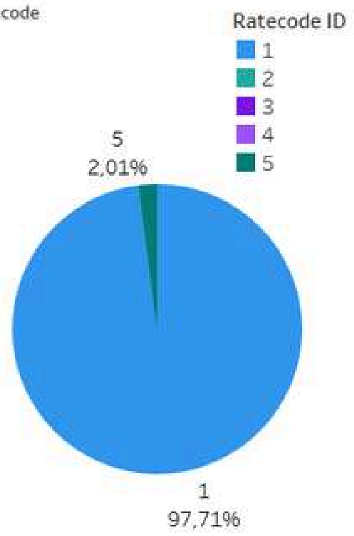
The **credit card payment method is the most preferred,** with an average fare of $16.25 paid via credit card, whereas cash payments have a lower average fare.

4. Passenger Preferences

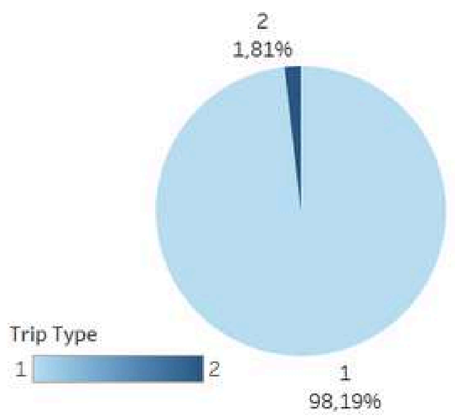
Proportion of Passenger Count



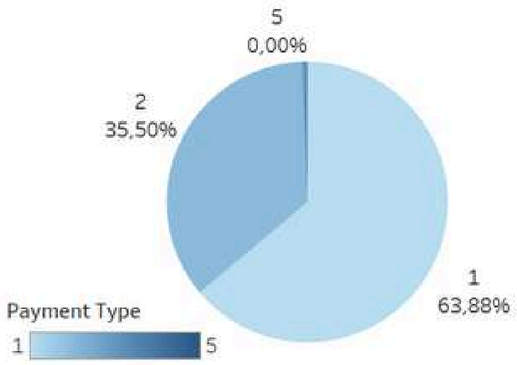
Proportion of Ratecode



Proportion of Trip Type



Proportion of Payment Type

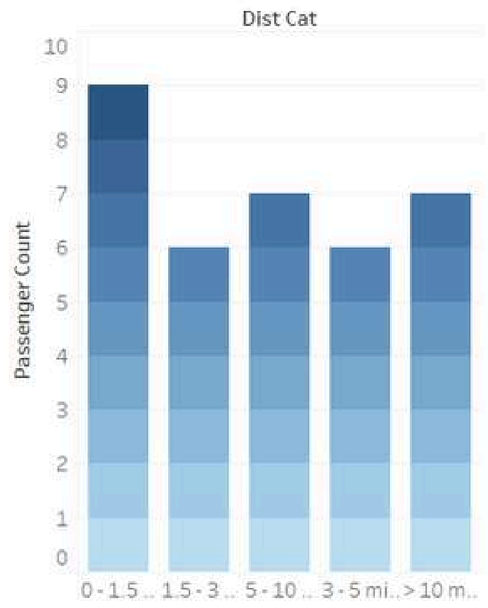


The majority of passengers ride taxis alone, use the standard fare rate, access taxi services through street hail, and make payments with credit cards.

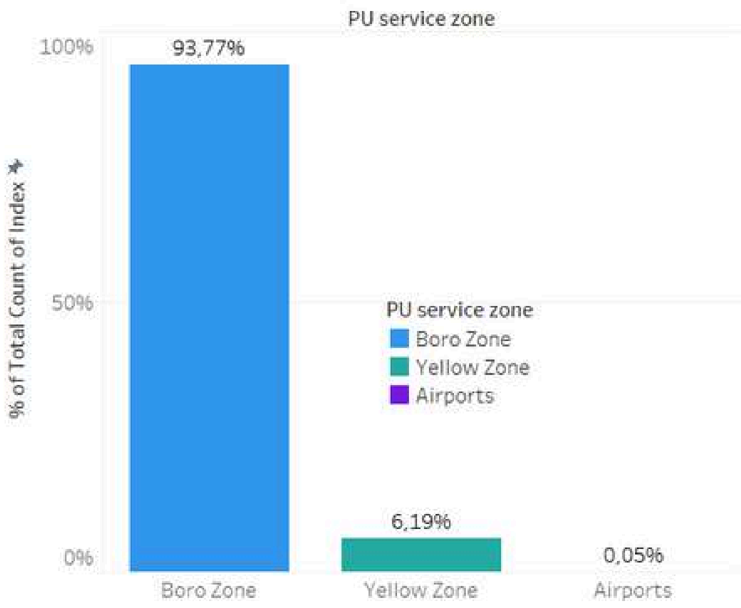
The most frequently used service zone for taxi users is the Boro Zone.

The majority of trips take place within the Manhattan area.

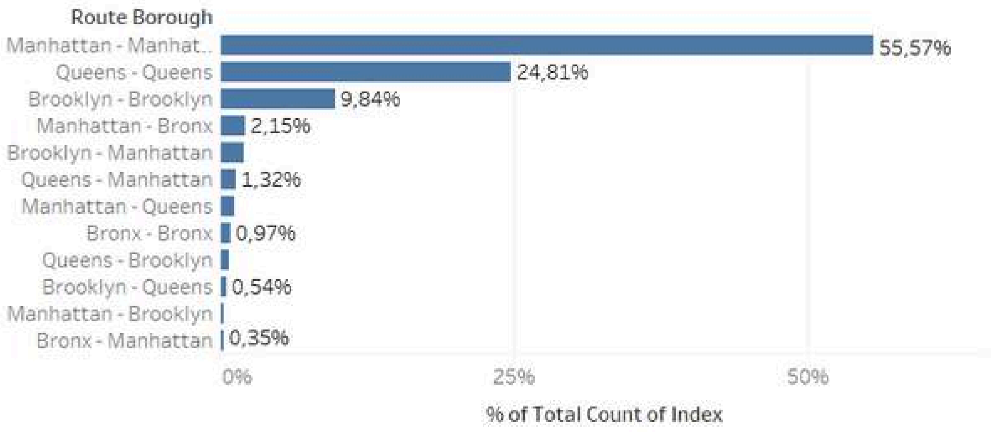
Proportion by Distance Category and Passenger Count



Proportion of Service Zone



Proportion of Borough Route

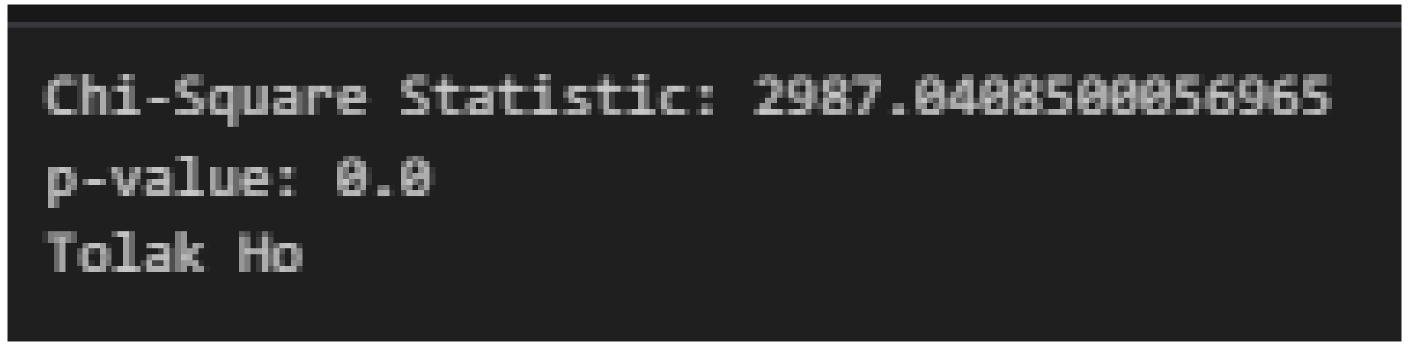


5. STATISTIC

1. Trip Revenue by Pickup Location Zones : statistic test

Statement H0 : The distribution of high-revenue trips no varies significantly across different pickup zones.

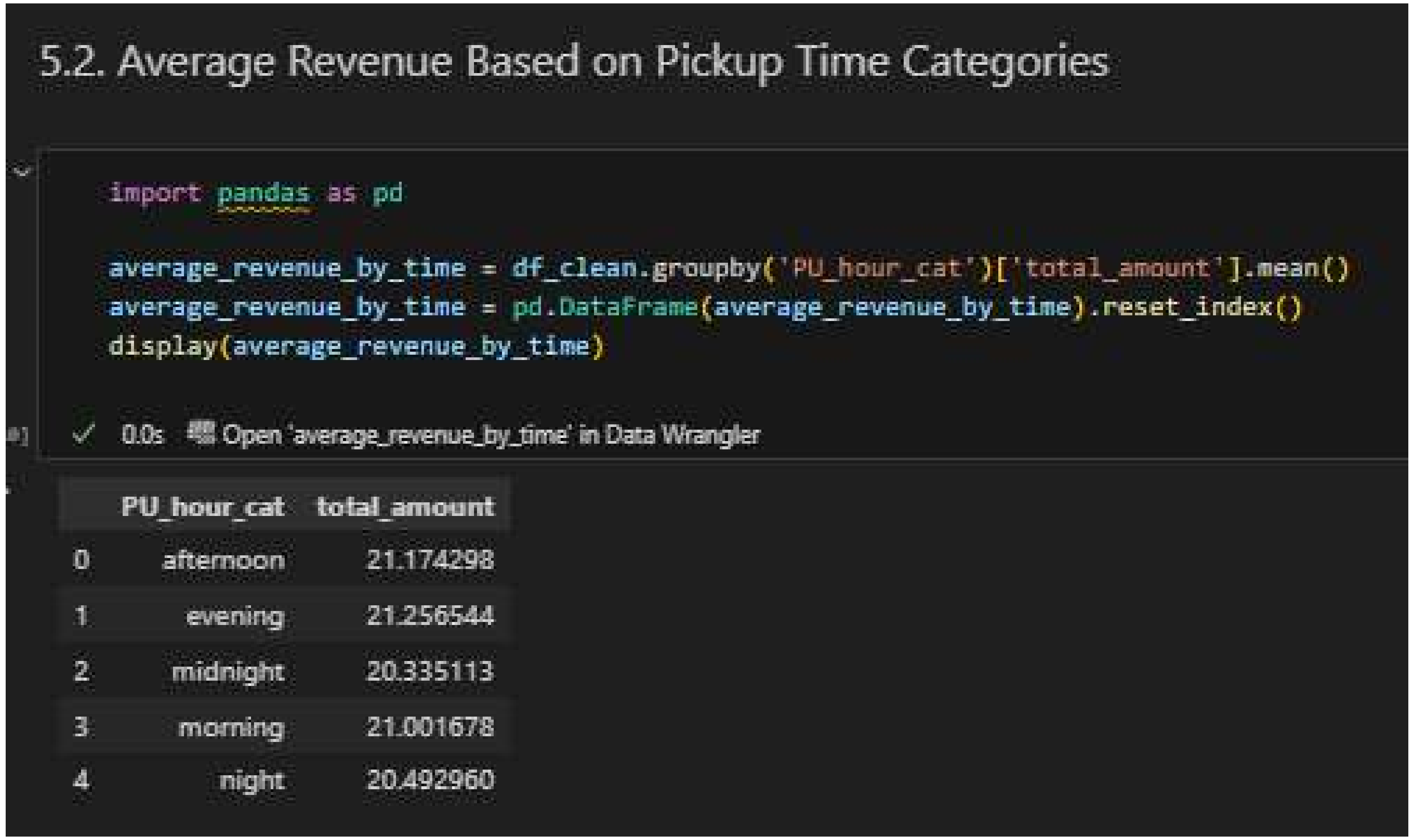
Hypothesis Test: Chi-Square test of independence.



```
Chi-Square Statistic: 2987.0408500056965
p-value: 0.0
Tolak Ho
```

- There is sufficient evidence to reject H0.
- There is a significant relationship between pickup zone (PU_zone) and high revenue (high_revenue category based on total_amount). Certain pickup zones tend to generate higher-revenue trips.

2. Average Revenue Based on Pickup Time Categories



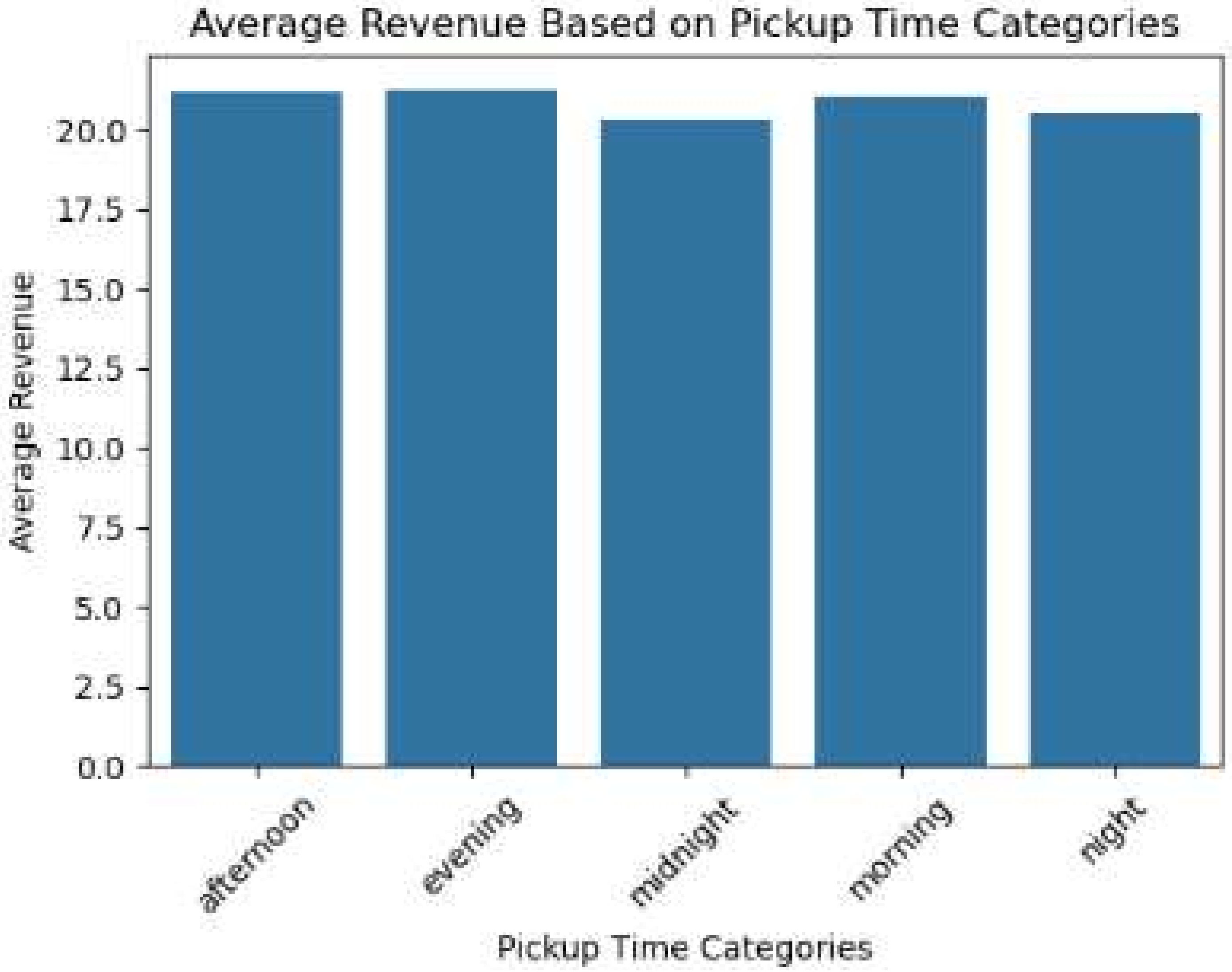
5.2. Average Revenue Based on Pickup Time Categories

```python
import pandas as pd

average_revenue_by_time = df_clean.groupby('PU_hour_cat')['total_amount'].mean()
average_revenue_by_time = pd.DataFrame(average_revenue_by_time).reset_index()
display(average_revenue_by_time)
```

✓ 0.0s Open 'average_revenue_by_time' in Data Wrangler

	PU_hour_cat	total_amount
0	afternoon	21.174298
1	evening	21.256544
2	midnight	20.335113
3	morning	21.001678
4	night	20.492960

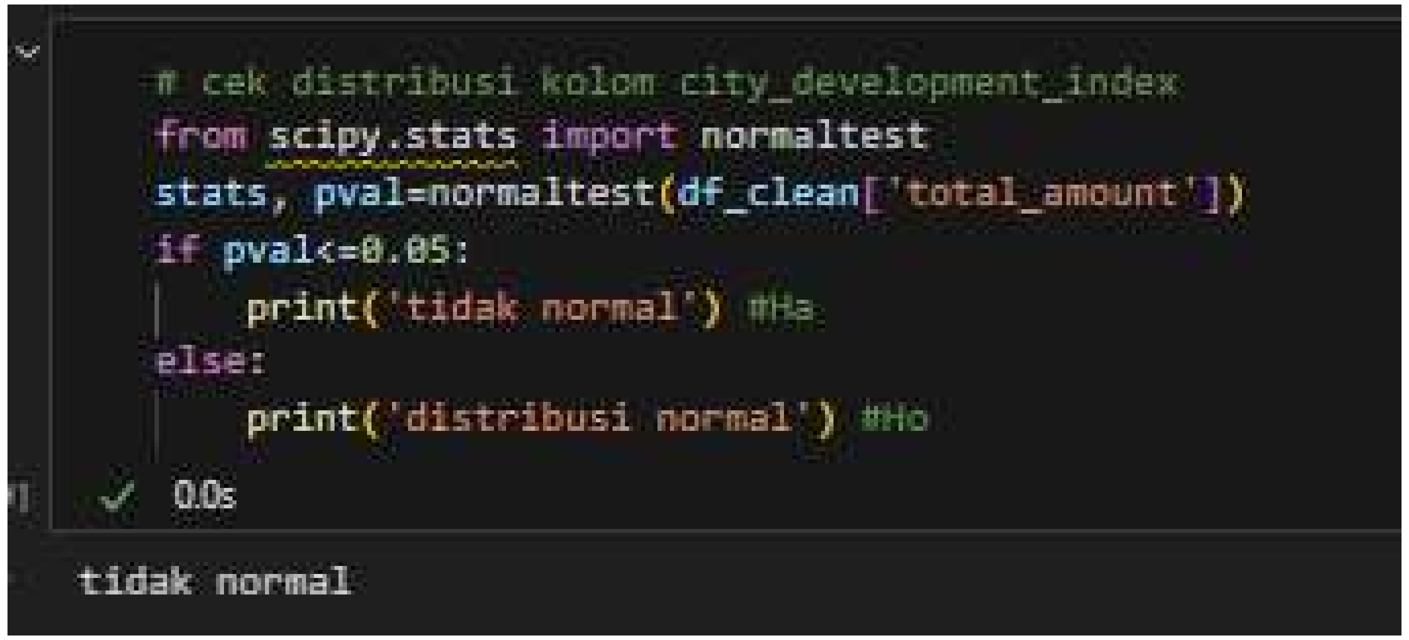


Here, the average total amount per time category is displayed in a DataFrame table and bar plot. The goal is to determine whether the total amount varies across time categories.

2. Average Revenue Based on Pickup Time Categories : statistic test

Statement H0 : The average revenue per trip varies no significantly different between time categories

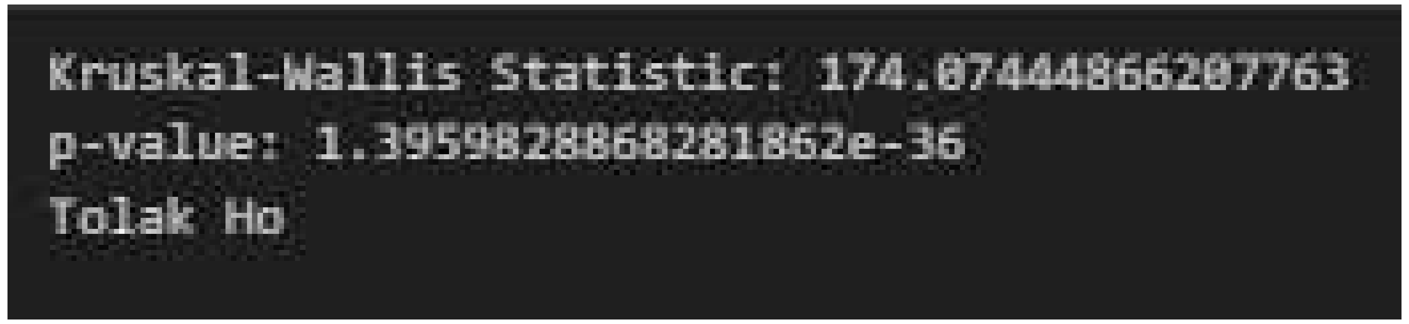
Hypothesis Test : Kruskal-Wallis test.



```
# cek distribusi kolom city_development_index
from scipy.stats import normaltest
stats, pval=normaltest(df_clean['total_amount'])
if pval<=0.05:
    print('tidak normal') #Ha
else:
    print('distribusi normal') #Ho
```

✓ 0.0s

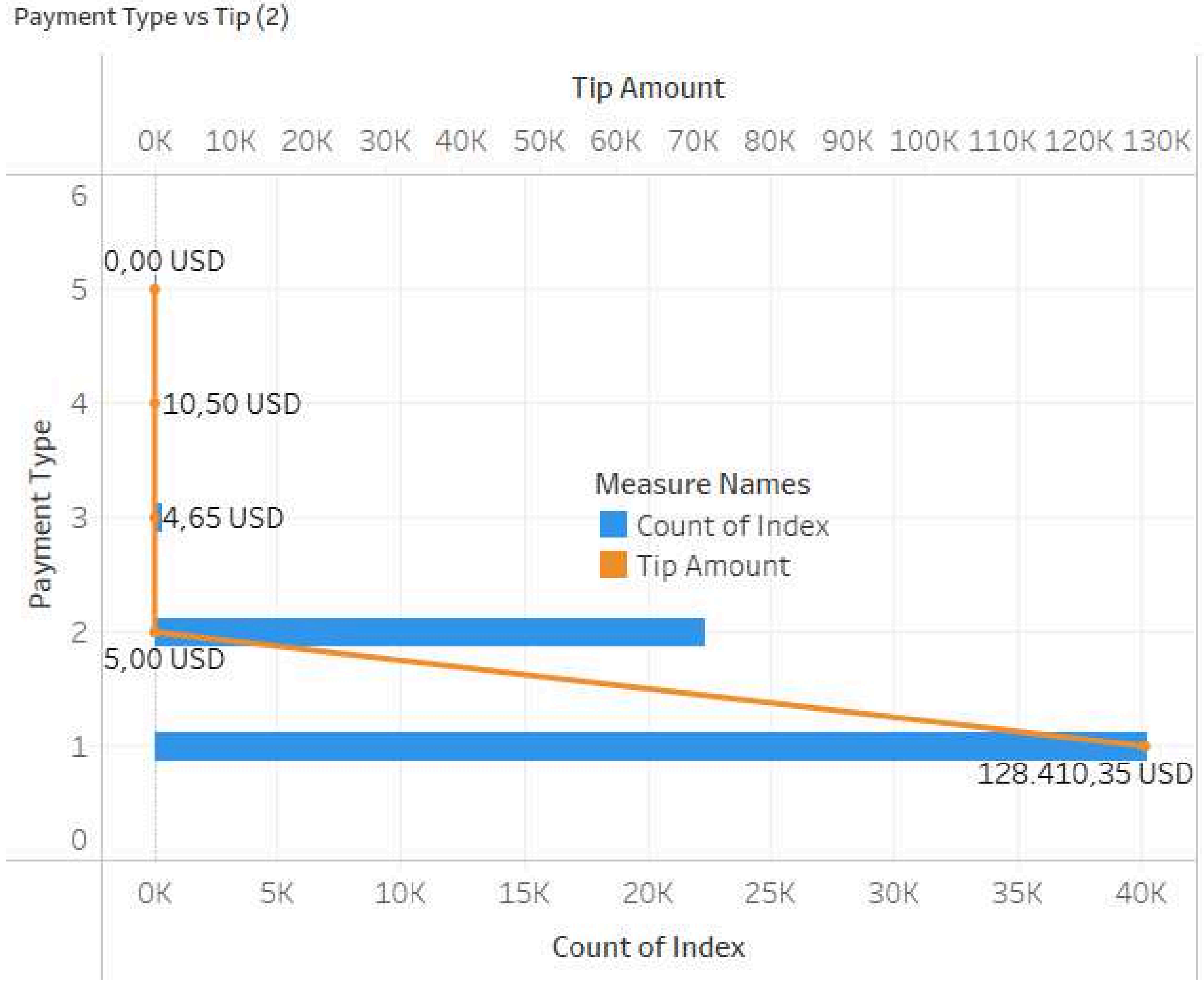
```
tidak normal
```

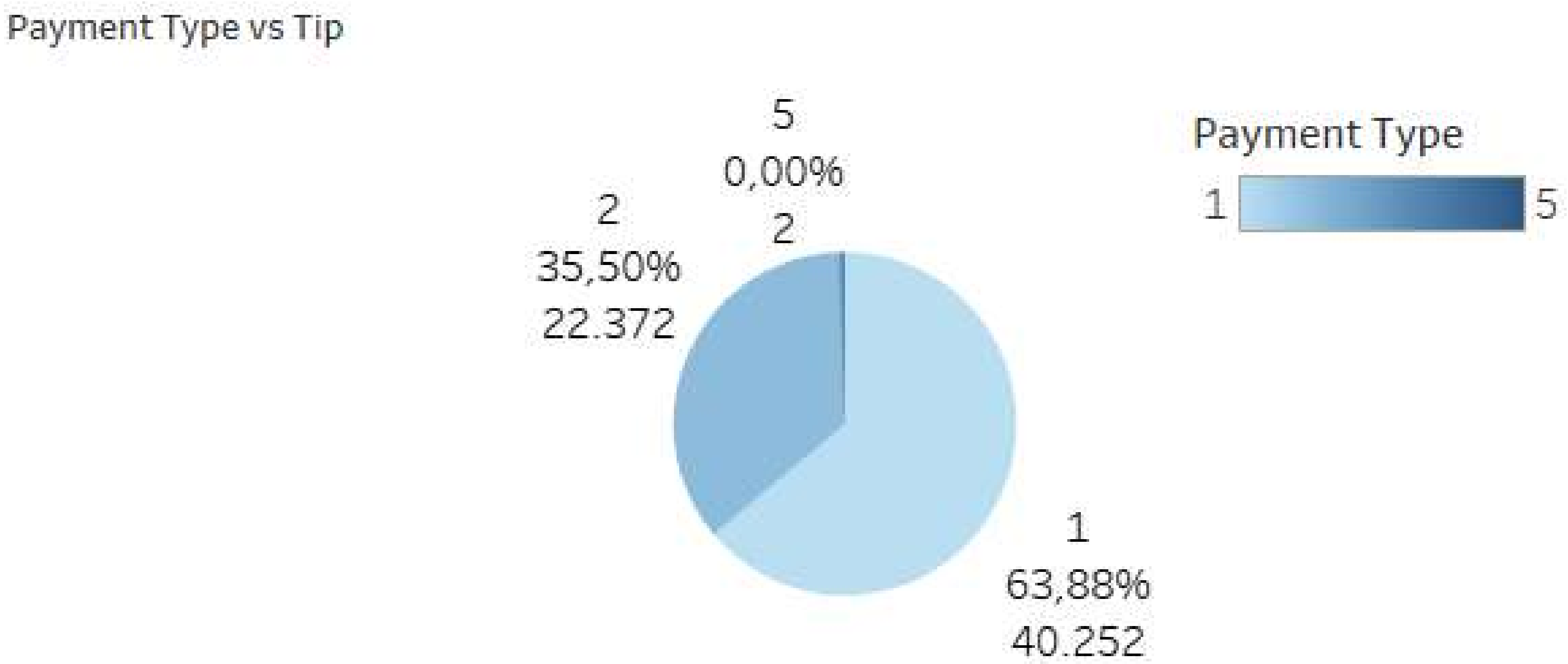


```
Kruskal-Wallis Statistic: 174.07444866207763
p-value: 1.3959828868281862e-36
Tolak Ho
```

- There is enough evidence to reject H0.
- There is a significant difference in the total number of trips (total_amount) received by taxis across different time categories such as midnight, morning, afternoon, evening, and night.

3. Payment Type and Tip Proportion



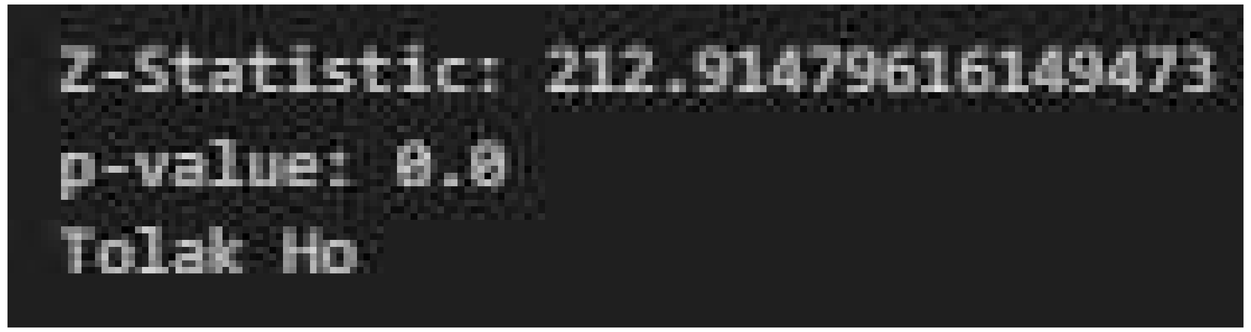


The total tip amount from credit card payments is $128.41, while the tip amount from cash payments is only $5.

3. Payment Type and Tip Proportion : statistic test

Statement H0   : The proportion of trips with tips no differs significantly between cash and non-cash payment types

Hypothesis Test : Proportion Z-test



```
Z-Statistic: 212.91479616149473
p-value: 0.0
Tolak Ho
```

- a Z value of 212.91 indicates that the difference between the tested proportion and the expected proportion is far beyond the critical value (typically ±1.96 for a two-tailed test at a 0.05 significance level).

- The p-value suggests that the difference observed in the proportions is highly likely to be real and not due to random fluctuations.

- There is enough evidence to reject H0.

- The proportion of trips with tips differs significantly between cash and non-cash payment types. This means there is a highly significant difference in tipping behavior between customers who pay with cash and those who pay with non-cash methods.

6. CONCLUSION AND RECOMMENDATION

Summary from Tableau [link]

1. The highest number of trips occurs between 7-9 AM and 3-6 PM.
2. The day with the highest number of trips is Tuesday, accounting for 16.61% of total trips.
3. The majority of trips occur in the afternoon (33.43%), followed by morning (29.54%), with the least occurring at midnight (1.61%).
4. The zone with the highest number of pick-ups is East Harlem North (29.31%), followed by East Harlem South (20%).
5. The borough with the highest number of pick-ups is Manhattan (59.48%), and the least is Staten Island (0.02%).
6. The borough with the highest number of drop-offs is Manhattan (59.5%), and the least is Staten Island (0.02%).
7. The majority of passengers use the street hail method to hail a taxi.
8. The average fare amount is lower on weekdays (15.82 USD) compared to weekends (16.28 USD), despite weekdays having a higher percentage of users (75.41%).
9. The highest fare amount occurs on Tuesdays and in the afternoon, while the lowest occurs at midnight.
10. Passengers using credit cards are more likely to tip, with the total tip amount given by credit card users being 128 USD.

Summary from Tableau [link]

11. Credit cards are more often used with an average fare of $16.25.
12. The highest average fare is in the Staten Island pickup area ($31.43), while the lowest is in Manhattan ($14.50).
13. Dispatch rides are more expensive than street hail rides, with an average fare of $31 compared to $16.
14. Trips over 10 miles have the highest average fare but contribute the smallest percentage (7.15%) to the total fare data. Trips between 1.5 - 3 miles contribute the largest percentage (30.28%).
15. Trips over 60 minutes have the highest average fare but contribute the smallest percentage (2.42%) to the total fare data. Trips between 10-20 minutes make up the largest percentage (40.32%).
16. Most passengers use taxis for a single rider (85.52%), standard rate for rate code (97.71%), street hail ride type (98.19%), and credit card payment (63.88%).
17. The most used pickup service zone is the Boro Zone (93.77%).
18. The most frequent route taken by taxi users is within Manhattan (Manhattan-Manhattan) at 55.57%.
19. Passengers who use credit cards are more likely to tip.
20. The total tip amount paid by credit card users is $128.

Recommendation

- **Optimize Taxi Operations During Peak Hours:**

Since the highest number of trips occur between **7-9 AM and 3-6 PM**, it’s essential to ensure there is an **adequate number of taxis available** during these peak periods to meet the demand and reduce waiting times for passengers.

- **Target Marketing on Tuesdays:**

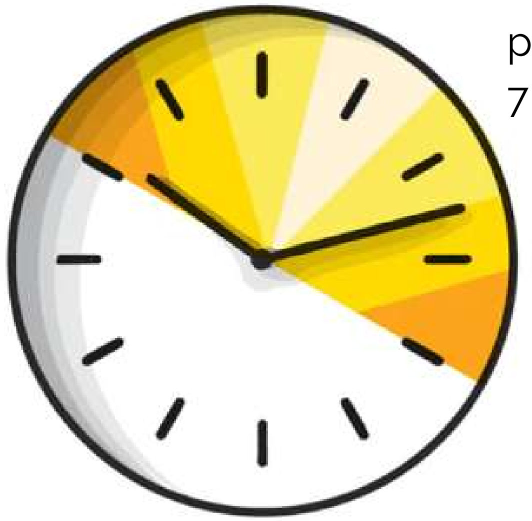
Given that Tuesdays account for the highest percentage of trips (16.61%), **targeted marketing campaigns or promotional offers on this day** could help further increase ridership and revenue.

KHUSUS HARI INI SAJA!!!

Diskon

20%*

example promo on Tuesday



peak hours
7-9 AM and 3-6 PM

Recommendation

- **Increase Availability of Taxis During the Afternoon and Morning:**

The afternoon (12-17 AM with 33,43%) and morning (1-12 AM with 29,54%) periods make up the largest share of trips, so **ensuring that more taxis are operating during these times** can **improve service quality** and **meet passenger demand.**

- **Focus on East Harlem Zones:**

With East Harlem North and East Harlem South having the highest pick-up percentages (29.31% and 20%), **optimizing operations and increasing taxi availability in these zones** could help improve service efficiency.

Recommendation

- **Strengthen Manhattan Coverage:**

Manhattan is the primary borough for both pick-ups and drop-offs (59.48% and 59.5%), so maintaining a **strong presence of taxis in this area** should be a priority to accommodate the majority of passengers.

- **Leverage Street Hail for Marketing:**

Since most passengers use street hail, efforts could be made to promote the benefits of hailing a taxi through **street signs** or **local advertising** to further increase this usage.

The two recommendations above are related, as street hailing is generally more common in urban areas like Manhattan, which serves as a city center

Recommendation

- **Tailor Services for Weekdays and Weekends:**

Given that fare amounts are generally lower on weekdays, yet **weekday** ridership is much higher, **different pricing strategies could be considered**, such as **offering discounts or loyalty programs on weekdays to encourage more passengers.** For **weekends,** the company could implement strategies that **emphasize premium service or unique experiences,** as weekend fares are generally higher. For example:

- Offer Premium or Exclusive Weekend Services, to justify the higher fare amounts and attract weekend passengers seeking a unique experience.
- Weekend Promotions for Group can encourage group or family bookings by offering discounts for multiple passengers

Recommendation

- **Address Midnight Demand:**

Although trips at midnight make up only a small percentage (1.61%), these trips could still be targeted by **offering special incentive**s to passengers during late-night hours, ensuring availability even during off-peak periods. Providing special incentives during this off-peak time could help **build loyalty with these passengers** and ensure they have reliable service options when demand is typically lower.

- **Encourage Credit Card Payments:**

Since passengers using credit cards are more likely to tip, further **promoting cashless payments** could not only **improve customer convenience** but also **increase the total tip amount received.**

Recommendation

- **Focus on Shorter Trip Distances:**

As the majority of fare amounts come from trips of 1.5 - 3 miles, taxi services should consider **offering promotions or ensuring availability for these shorter trips to** maximize the volume of trips within this range.

File Project

Github Link

https://github.com/huwaidanur/CapstoneProject2-DataScience-Purwadhika/

Tableau Link

https://public.tableau.com/app/profile/huwaida.nur.asysyifa.mufarrida/viz/NewYorkTLCTripRecord/Story1

Youtube Vide Link

https://youtu.be/xyn4WiX7kjQ

THANK YOU

● FOR YOUR NICE ATTENTION