1. Downloaded January 2024 green\_tripdata\_2024-01.parquet file from the given link <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
2. Downloaded and Installed Microsoft SQL Server on premise version.
3. I used one of the online tools to convert the parquet file into CSV file.
4. Below is the site which I used to convert the parquet file into CSV file. <https://dataconverter.io/view/parquet>
5. After initial analysis of the raw data where I found the pickup date time value, which is in UTC time zone so I transformed the date time into actual date format using below excel functions.

Day **=LEFT(B2, 3)**

Weekday **=LEFT(B2, 3)**

Date **=MID(B2, 9, 2)**

Month **=CHOOSE(MATCH(MID(B2, 5, 3), {"Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sep","Oct","Nov","Dec"}, 0), 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)**

Year **=MID(B2, 25, 4)**

Trip\_Date **=TEXT(DATE(Y2, X2, W2), "DD/MM/YYYY")**

1. Below are the few observations from the dataset.

* there are some null/blank values
* some negative tip amount
* 0 passenger trips with tip amount

**Note:** I have assumed the negative values as positive and performed the sum. I have not ignored the tip amount with 0 passengers for the sum.

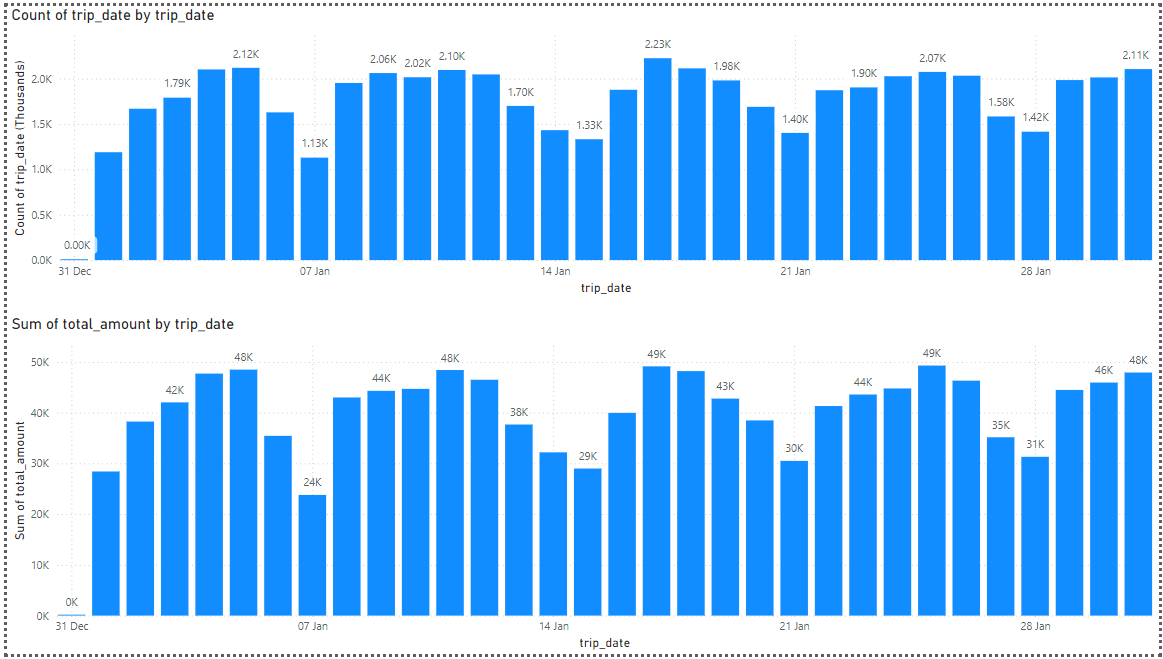
1. After the modifications I have loaded the modified dataset into SQL server using import flat file option.
2. Later I did some research to find the zone, Borough, service zone values for New York City Taxis
3. During this course, I came across this page <https://www.kaggle.com/code/aguado/nyc-yellow-green-cabin-analysis/input> where taxi zone lookup values are listed for New York City Taxis.
4. Assuming the values for Zone, Service Zones, and Borough remain constant over time, I have used these values for my task. The provided data associates these values with the year 2022, while my dataset pertains to 2024.
5. After these assumptions, I have loaded those details into SQL server.
6. Created the star schema design: trips\_fact as Fact Table, date\_dim as Date Dimension table, location\_dim as Location Dimension table by using SQL queries.
7. Inserted the raw data into created fact and dimensions tables using SQL queries.
8. Performed Data analysis operations using SQL queries for the below:

* Calculate the total number of trips per day.
* Identify the top 5 zones with the highest total fare amount.
* Calculate the average trip distance by borough.
* Determine the most common pickup and dropoff locations.
* Calculate the total tip amount per passenger count.

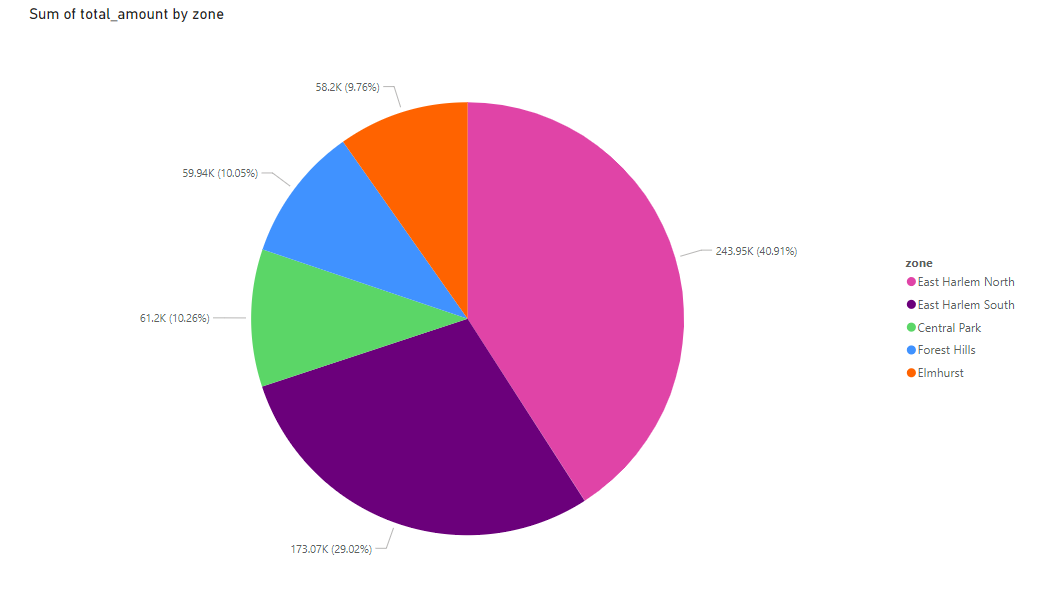
1. I attached all the developed scripts in the GitHub repository, below is the link: <https://github.com/JosephVikas/Nestech-DE-Task>

Power BI Graphs

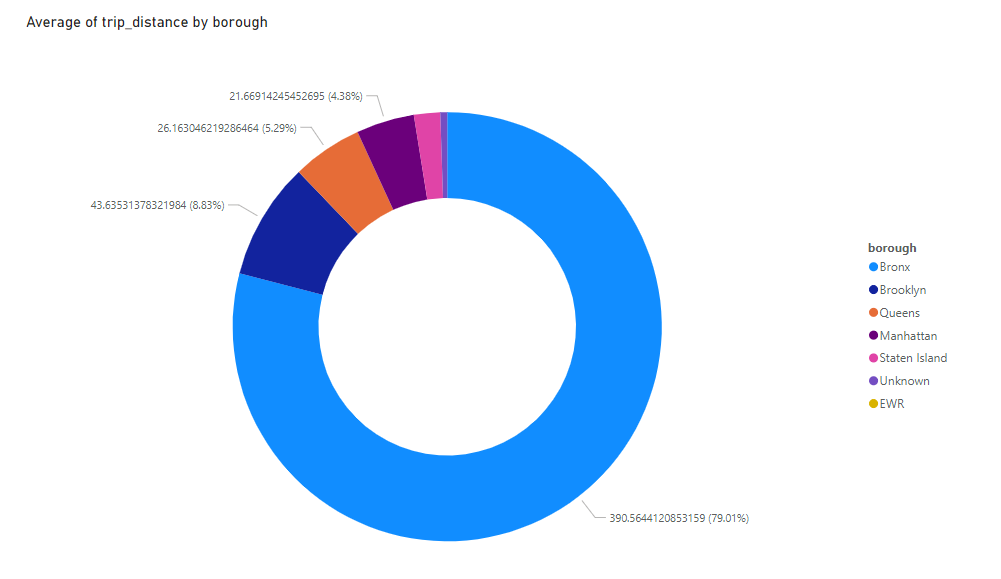
Total Number of Trips per Day and Total amount per Day



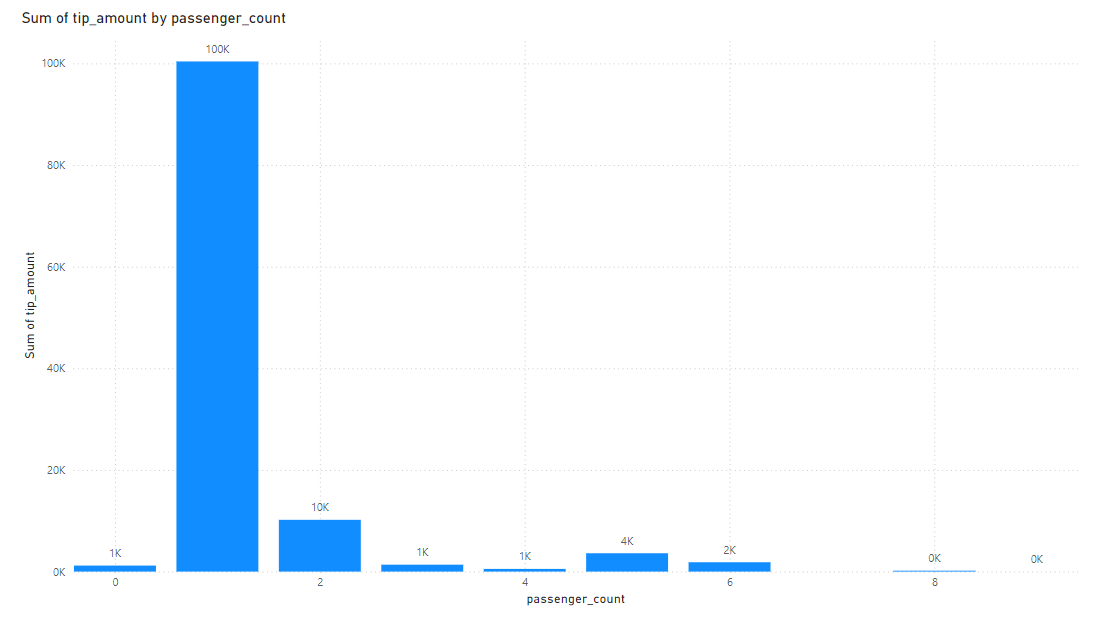
Top 5 Zones with highest total amount



Average Trip distance by Borough



Total tip amount by passenger count



Common pickup and drop off locations

