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")**

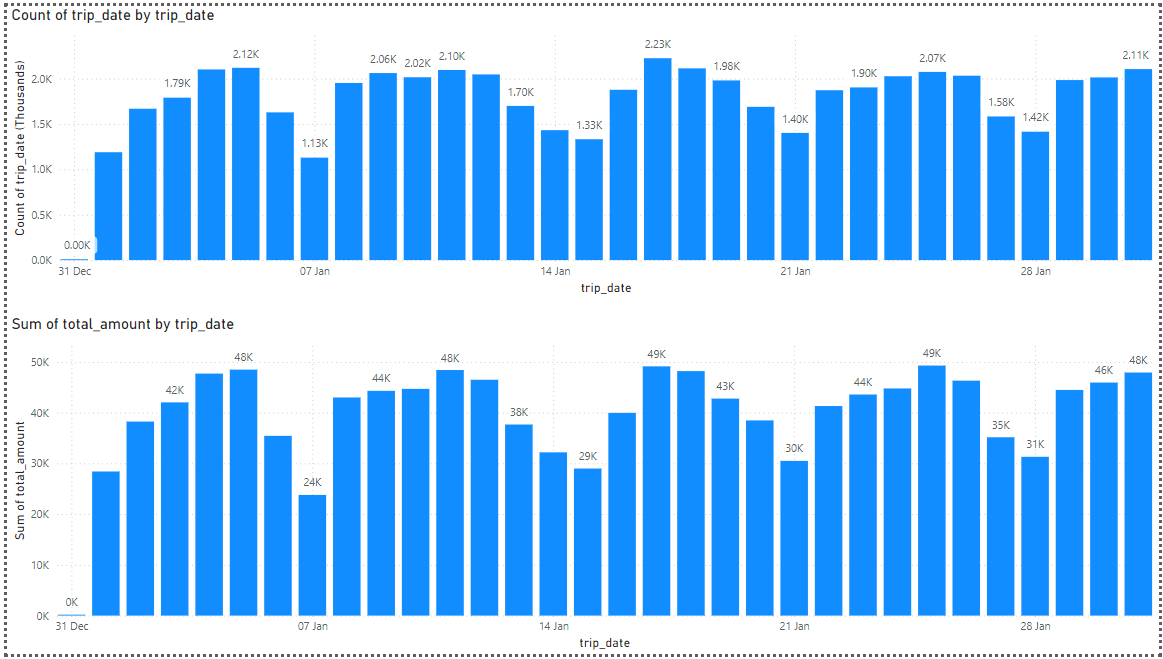
1. Below are the few observations from the dataset.

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

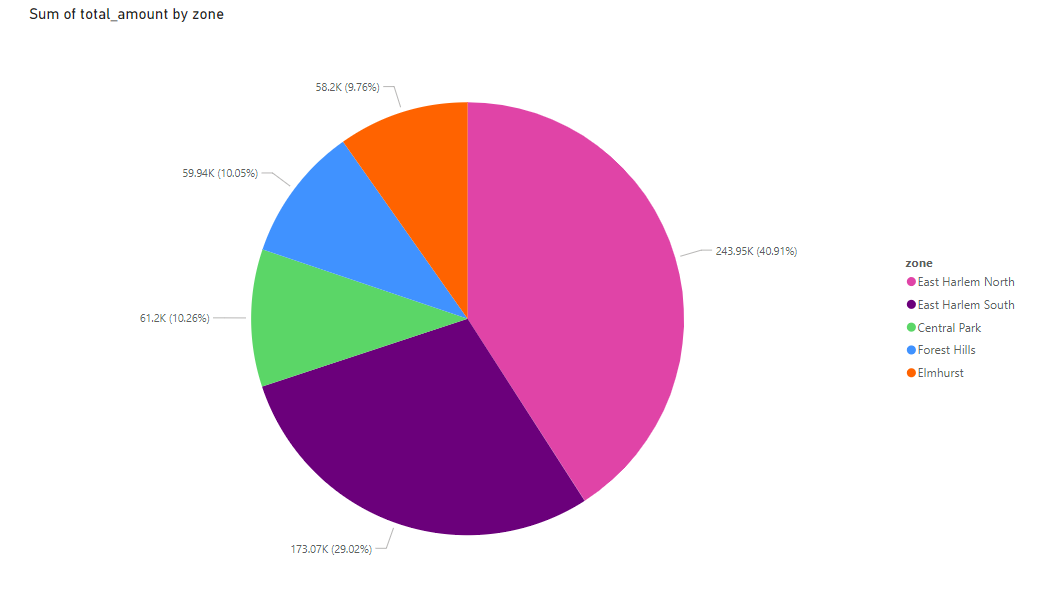
1. After the modifications I have loaded the modified dataset into SQL server using import flat file option.
2. Later I did some research and found the zone, borough, service zone values for New York City Taxis and I loaded those details into SQL server.
3. Created the tables trips\_fact, date\_dim, location\_dim using SQL queries as per the instructions.
4. Inserted the data into required tables using SQL queries.
5. Performed Data analysis operations using SQL queries.
6. I attached all the developed scripts in the github repository.

Power BI Graphs

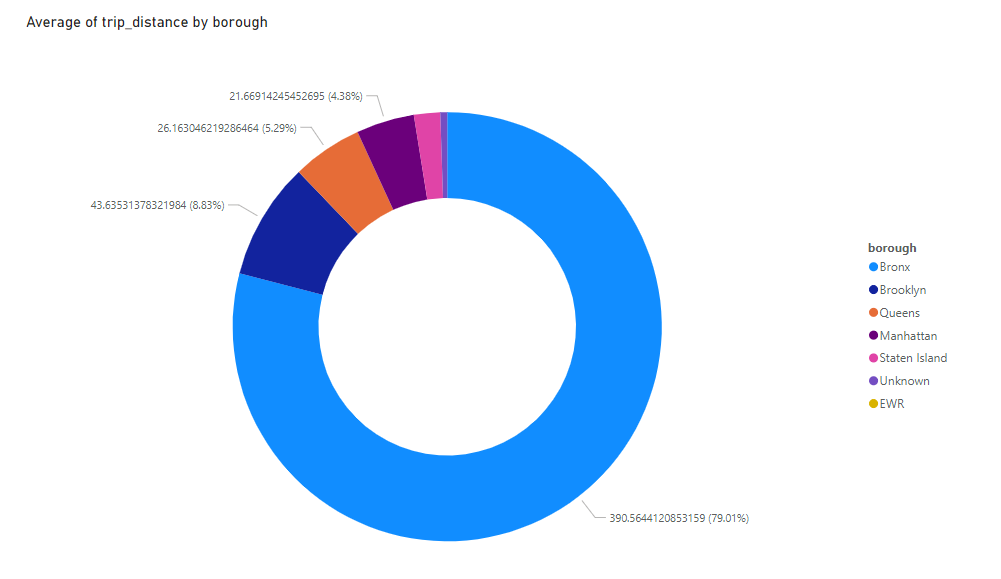
Total Number of Trips per Day and Total amount per Day



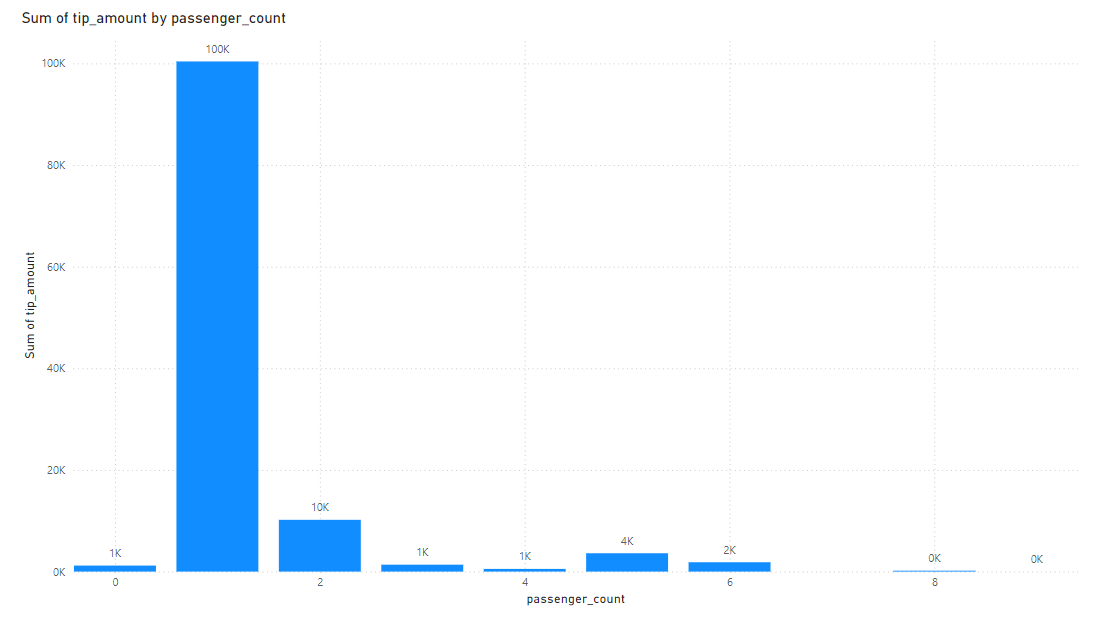
Top 5 Zones with highest total amount



Average Trip distance by Borough



Total tip amount by passenger count



Common pickup and drop off locations

