CSC - 501

REPORT (ASSIGNMENT – 2)

SPATIO-TEMPORAL DATA

SUBMITTED BY - GROUP B

Kapoor, Nikita	
Malik, Mona	V00935224
Pattamatta, Suseela	
Vemula, Sannath Reddy	V00949217

Submitted To - Prof. Sean Chester

schester@uvic.ca

SECTION - 1

Data Modelling

- **ER diagram**

We are majorly using two tables as the core base for our Data Analysis. The tables are taxi_details with 17 columns (taxi-sample.csv) and the given shape file. Shape file is a geodataframe, within which queries can be performed for filtering and projecting the required rows.
The given data includes inconsistent dirty data for which the information is unknown. For instance, information RateCodeId=99 is not mentioned in the data dictionary of yellow taxi trip records. Such data like VendorID =4, future dates, PULocationID or DULocationID > 263 has been removed from the table as a part of data cleaning which reduced the row count from 1048576 to 1008970 (cleaned about 4% of inconsistent data).

To make the data model flexible and handy while working on the dataset to create insights, the 17column table is divided into 1 Fact table and 2 Dimension tables. The fact table, of 14 columns, is obtained by adding an extra column named ID and by dropping 4 columns - tpep_pickup_datetime, tpep_dropoff_datetime, PULocationID, DOLocationID. Fact table is related to Dimension tables by column name ID. One of the dimension table is Date_time table which has 9 columns – ID, PU_Date, PU_hour, PU_minute, PU_AM_PM, DO_Date, DO_hour, DO_minute, DO_AM_PM. Another Dimension table is Loc_CRS which has latitudes and longitudes of all the LocationID's fetched from shapefile. The below ER diagram explains the whole model and the relation between tables.
The main intension of this model is to make the temporal and spatial data readily available as individual components (eg. time is divided into hours and minutes) which avoids the use of multiple lines of code to extract a required part of temporal/spatial data while working on visualizations. For instance, extracting year out of datetime column while plotting a graph would add steps like casting of data types, operations on strings which are also considered redundant if the same part of temporal data is being used in multiple plots.

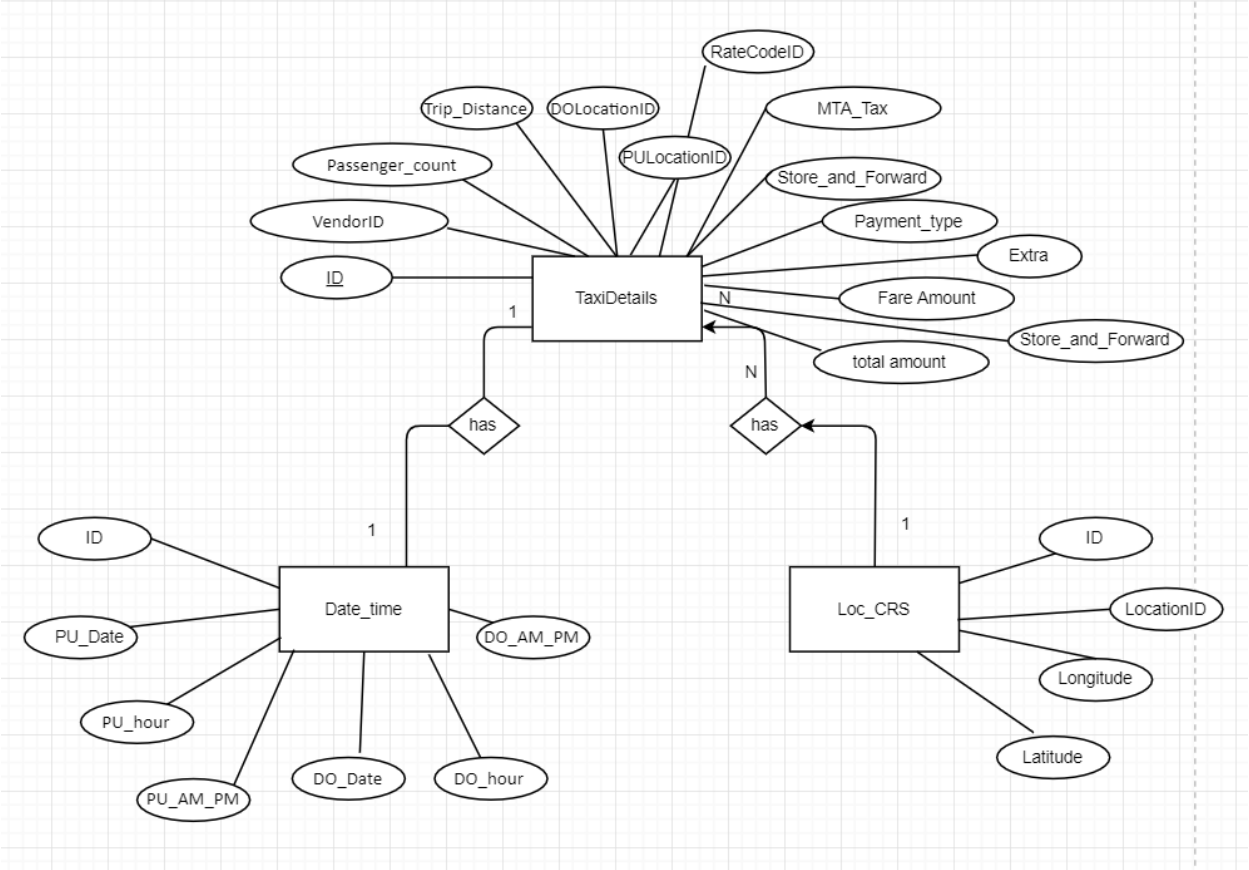


Table Structure of the ER DIAGRAM

Fact Table:

create table if not exists taxiDetails as select rowid as ID,VendorID, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount from taxi_details

Dimension Tables:

create table if not exists Date_time as select rowid as ID,
substr(tpep_pickup_datetime,1,10) as PU_Date,
substr(tpep_pickup_datetime,12,2) as PU_hour,
substr(tpep_pickup_datetime,15,2) as PU_minute,
substr(tpep_pickup_datetime,21,2) as PU_AM_PM,

```
substr(tpep_dropoff_datetime,1,10) as DO_Date,
substr(tpep_dropoff_datetime,12,2) as DO_hour,
substr(tpep_dropoff_datetime,15,2) as DO_minute,
substr(tpep_dropoff_datetime,21,2) as DO_AM_PM     from taxi_details")
```

For the creation of database, we have used postgres which resulted in better performance and execution time when compared to sqlalchemy. In our previous assignment, sqlalchemy outclassed other libraries in terms of execution time for read files and loading them as tables into database. But for this assignment, creation of a table in database from csv file along with data cleaning, sqlalchemy took around 1 minute. Whereas, postgres was successful in achieving the task within 20seconds.

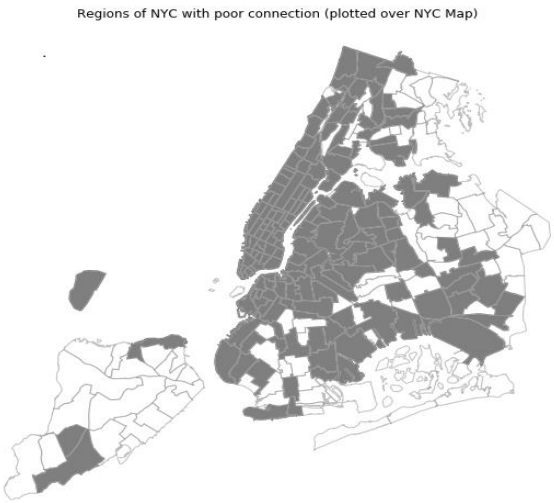
SECTION - 2

Visualization

INSIGHT #1 : **Connectivity**

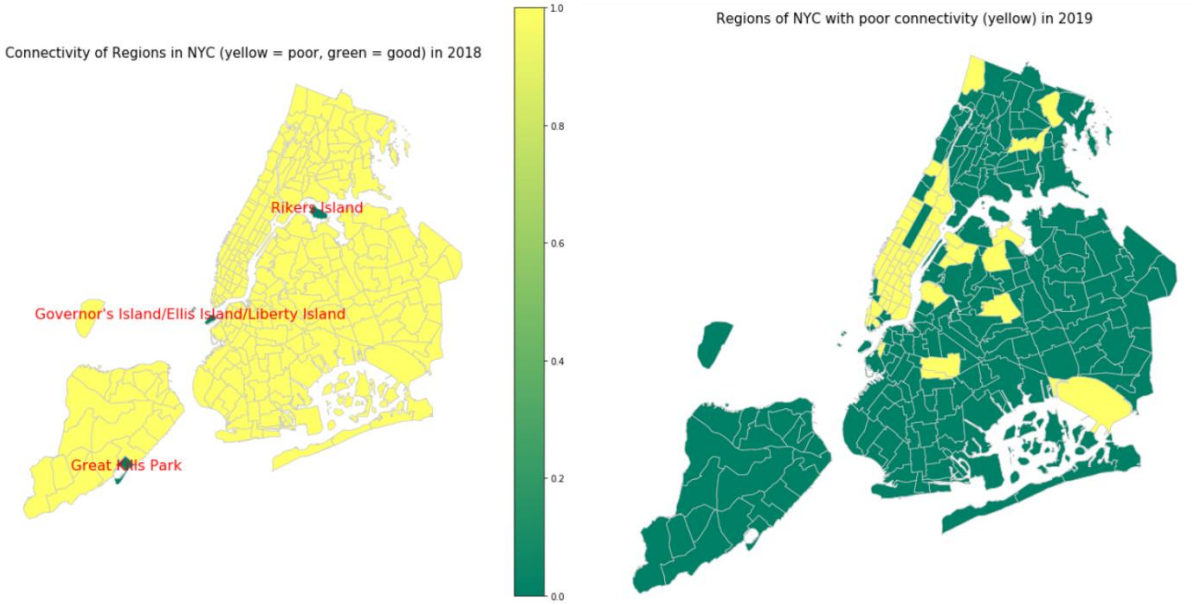
1.1 Regions where the connection between the vehicle and the server is poor

The desired results were achieved by extracting the values of Store_and_forward_flag from TaxiDetails table. DOLocationID of the record where Store_and_forward_flag=Y is considered as the location with poor connectivity. The below plot shows the poor connectivity regions for the entire dataset (which was loaded into table after cleaning).



1.2 Poor Connectivity regions during 2018 and 2019:

The below plots show the same feature for the years 2018 and 2019(yellow color represents poor connectivity regions),

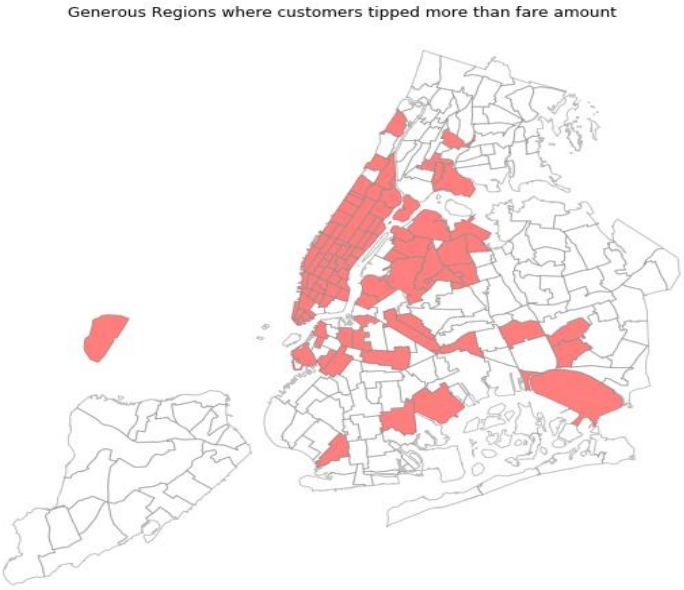


Insight – The above visualization shows that the number of regions with poor connectivity decreased in 2019 when compared to 2018 which implies that the technology being used to receive data from the vehicle to server has been enhanced by yellow taxi to improve the connectivity.

INSIGHT #2: Generous Regions (where given tip amount > fare amount)

2.1 All regions where tip amount is greater than fare amount:

DOLocationID, Tip_amount and Fare_amount columns of taxidetails are used for filtering the required locations with tipamount>fareamount which resulted in 94 distinct locations.
(Red color represents the generous regions.)



2.2 Top ten generous regions:

The below plot gives the top 10 regions of the feature,

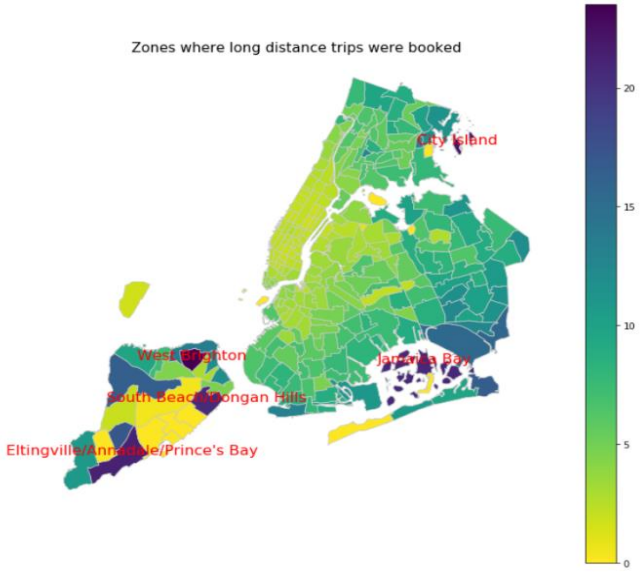
Insight – We can see that most of the regions in Manhattan, according to the data, are considered as generous and even the top ten regions which are plotted belong to Manhattan. The above visualizations conclude that, drivers of yellow taxi can gain more out of tips by working at these top ten generous locations.

INSIGHT #3: Long Distance trips

3.1 Regions from where Long-Distance trips have been booked:

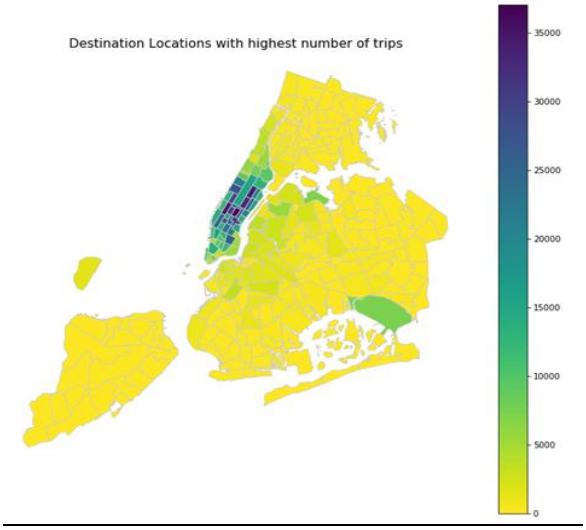
PULocationID and Trip_distance column of the table was picked to plot this feature and average of the Trip_distance grouped by PULocationID was considered.

The below plot shows the source locations in ascending order of the trip distances. Yellow colored locations represent the source locations with least trip distance and dark blue colored locations represent the source locations with highest trip distance.



Insight – The top regions of the feature plotted are a part of islands of NYC.
People living in Islands of NYV prefer to take a taxi for long-distance trips which leads to the assumption that within Islands, to travel longer distances, NYC has limited conveyance services or the available services for conveyance takes more time than a taxi because of which people prefer taxi over public transport.

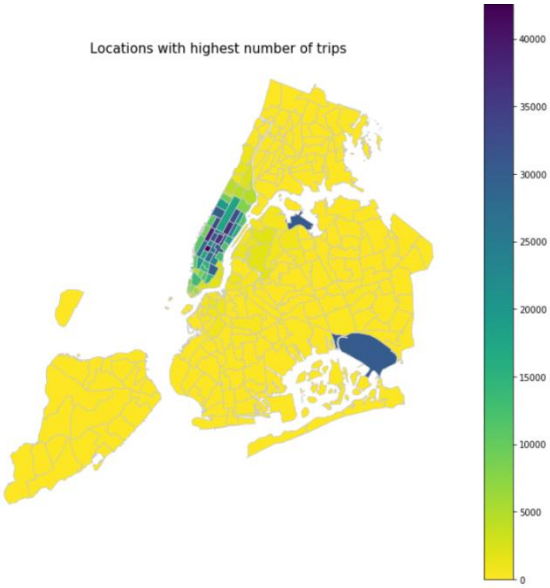
3.2 Regions to where Long-Distance trips have been booked:



Insight – Destination locations are taken into consideration for the above plot, which shows green and blue shade in Manhattan. Apart from Manhattan, Airports like John F Kennedy and Laguardin are colored green which are next to regions Manhattan.

INSIGHT #4: **Highest Number of trips**

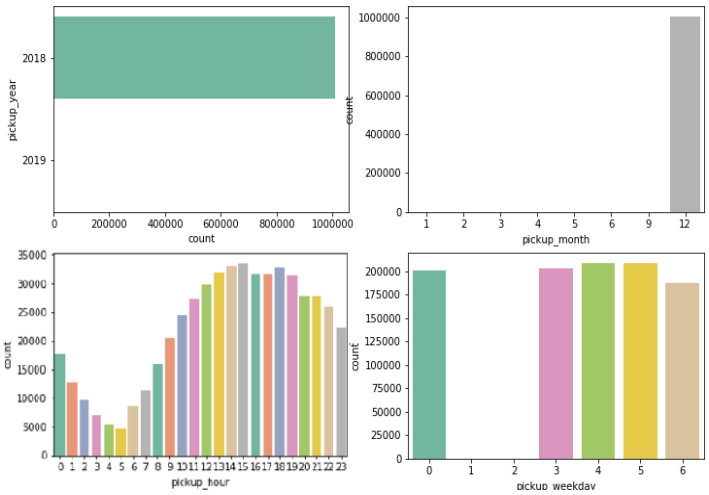
For this feature, count of PULocationID was considered.

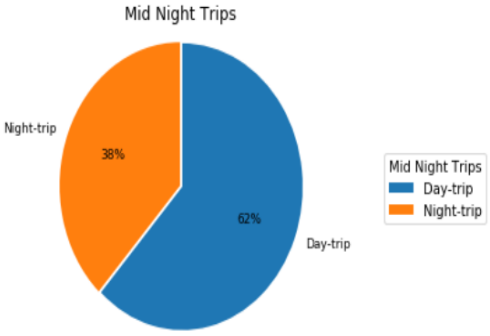


Insight – Upon analysing the above map, Manhattan has the highest number of trips compared to other regions. Laguardin and John F Kennedy Airport are also the locations with a greater number of trips.

INSIGHT #5: Highest Number of trips

Below are the graphs based on temporal data of the given dataset. Them main columns fetched to plot the graphs are PU_date, PU_Hour, PU_Minute, PU_AM_PM of Date_Time table.





Insights –

The above graphs show that

- most of the trips of the dataset belong to year 2018,
- December is the month with the greatest number of trips,
- Graph showing Pickup_hour for the entire day has pattern which says that the number of trips of the day start increasing around 11am and reaches peak at 2pm. It then drops slightly over the odd hours.
- Monday and Tuesday are found out to be the days of the week with least number of trips whereas the other days have almost same number trips.
- The Pie chart shows the percentage of trips during days and percentage of trips during nights

SECTION – 3

Algorithmic Considerations:

Data being the initial point for reporting the analyses, we started with loading the given csv into database as a table. We used a library called **SQL Alchemy** in order to load the data from csv file into the db. It uses data chunking as taught in the class in order to process larger data sets. In order to create database and load 1million records, SQL Alchemy approach took around 1 minute. In previous assignment, performance using sqlalchemy proved to be too good for 27million records data. But for 1 million data, along with data cleaning and data loading into table postgres performed much better than using sqlalchemy alone. Execution time using postgres,

```
Wall time: 20 s in normalizing data
Wall time: 16.8 s in inserting data to database
```

Library used for postgres: **psycopg2**

Since the given data include spatial data, libraries in python like folium, geopandas and JavaScript libraries like d3.js, chart.js for interactive data visualization were initally considered.
As the assignment progressed, geopandas was found to be appropriate considering the extent of the library and to work with shape files which is supported by datatype called geodataframe.

Below mentioned listed are execution times of the graphs,

Regions with poor connectivity in 2018
919ms in data collecting
396ms in plotting graph

Regions with poor connectivity in 2019
418ms in data collecting
205ms in plotting graph

Locations where trips booked a longer distance
602ms in data collecting
261ms in plotting graph

Highest number of trips in zones
626ms in data collecting
216ms in plotting graph

Generous Regions where customers tipped more than fare amount
Graph-1 Wall time: 1.89 s
Graph-2 Wall time: 741ms

System specifications:
Processor - intel i7 7th gen CPU 2.7GHz - 2.9GHz 64-bit OSx64 based processor
RAM - 8 gb
Tool/software - Anaconda version 2019.07
Language – Python 3.7

Using Tableau will minimize the time consumptions for operations like joins, data transformations and even results in much better performance. Also considering that the data could scale up in future, visualising the data through the tool would be more reliable for accurate results without considering any additional computations which should be done if any programming language like python is used.

SECTION - 4

Relationship to Spatial Temporal Challenges:

1) Issues with spatial data:

The shape file has data of 263 locations of NYC. When attempted to get the latitude and longitudes for all the locat ions in shape file using LinearRing, an error occurred stating that ‘MultiPolygon’ object has no attribute 'exterior'. To overcome this error BBox concept of shapely library was implemented which retrieved coordinates of both poly gons and multi-polygons. But this implementation created redundant/inconsistent coordinates,

	LocationID	longitude	latitude
101	100	9.873556e+05	213800.923426
102	101	1.063366e+06	210296.728371
103	102	1.018557e+06	195300.116634
104	103	9.717312e+05	190599.495169
105	103	9.729665e+05	193824.073186
106	103	9.789282e+05	189983.282593
107	103	9.717312e+05	190599.495169
108	103	9.729665e+05	193824.073186
109	103	9.789282e+05	189983.282593
110	103	9.717312e+05	190599.495169
111	103	9.729665e+05	193824.073186
112	103	9.789282e+05	189983.282593
113	106	9.869716e+05	184562.609078
114	107	9.886797e+05	207908.949316
115	108	9.890224e+05	153835.351552
116	109	9.423589e+05	139362.989826
117	110	9.489495e+05	137238.924021
118	111	9.866870e+05	176779.198074
119	112	9.986127e+05	204798.780373
120	113	9.856447e+05	206046.020318
121	114	9.850308e+05	204700.765412
122	115	9.600075e+05	165214.984444

270	262	999443.4972422048	222247.38159526885
271	263	997818.2714909017	223006.18593746424

271 records for 263 locations

Finally, the issue was solved by using an online tool mygeodata.cloud which takes .dbf, .shx and .shp file as input and generates a csv file as output with all latitude and longitude of the locations.

2) Issues with the given date format:

Initially while working on the temporal attribute of the dataset, date format which is given in the 2^20 subset is MM/DD/YYYY. This created a minor issue when attempted to use sqlite3 as the expected/preferred date format is YYYY-MM-DD. For this we had to convert the date into the suitable format if sqlite3 was to be used for data extraction. But then implementing postgres solved the problem with the format which has to_char() function.

3) Visualizing quick Insights Challenge:

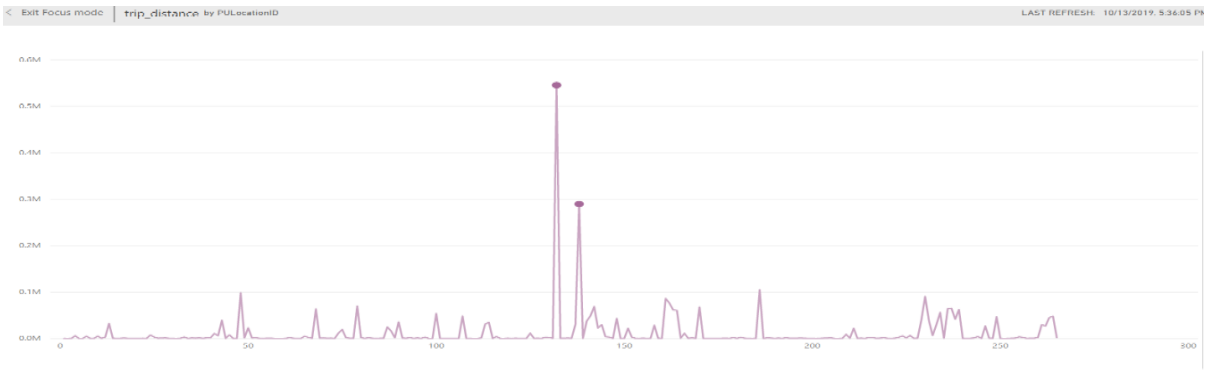
One of the major challenges of spatial-temporal data is to find interesting data patterns from the given data. In the initial phase we used Tableau and Power BI taking idea from the research paper for the insights.

Microsoft Power BI and Tableau are few of the tools that addressed and solved all the above-mentioned challenges while visualizing the spatio-temporal data.

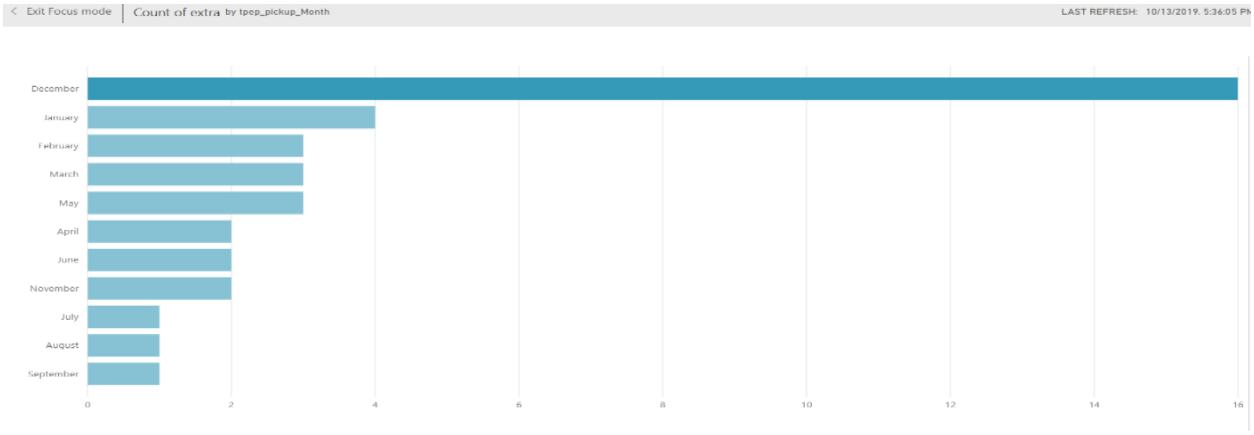
POWER BI

Power BI is a cloud-based business analytics service by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities. We used Power Bi to generate some quick Insights related to the New York Taxi dataset and come across some interesting insights listed below: -

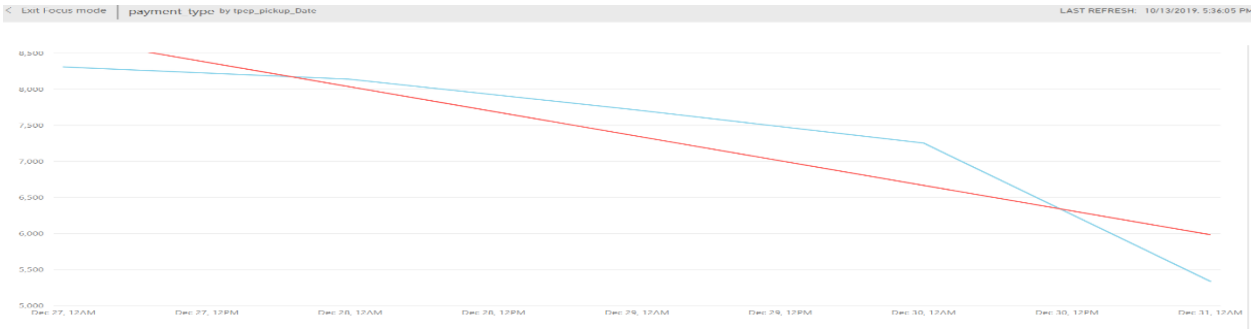
- Trip_distance by PULocationID. This plot shows that trip distance has **outliers**.



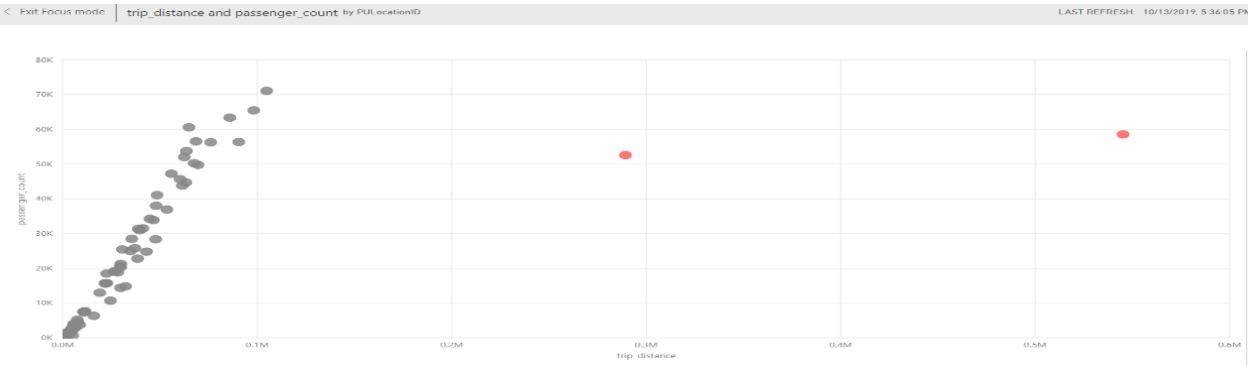
- Count of Extra by **TPEP_PICKUP_MONTH**. This plot shows that December has noticeably more rides.



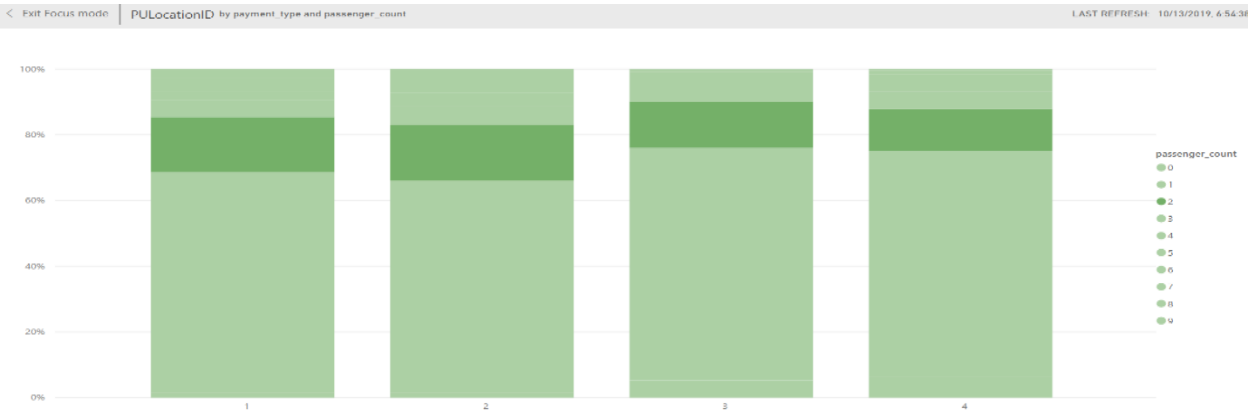
- Payment_type by **tpep_pickup_date**. This plot shows that payment type is trending downwards for RateCodeID 2 and tpep_dropoff_month 'December'.



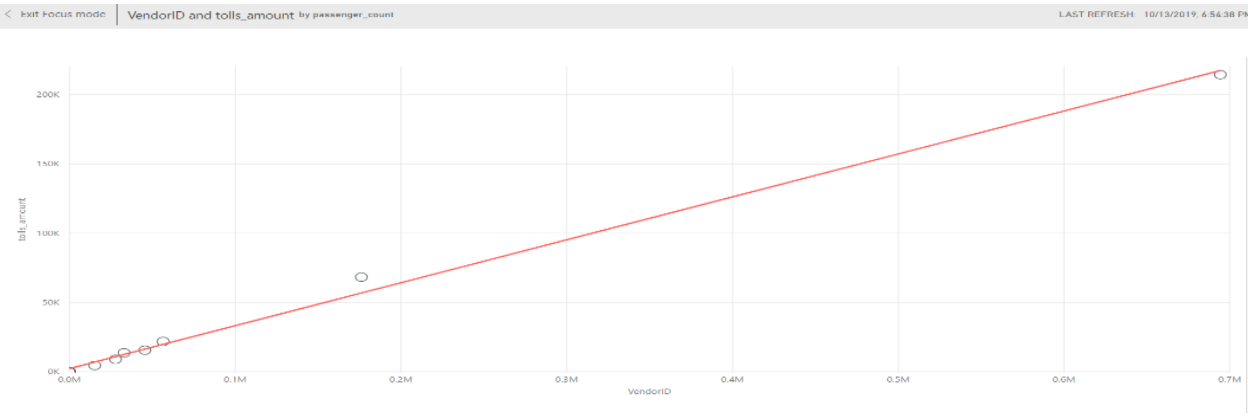
- Trip_distance and passenger_count by **PULocationID**. This plot shows that trip distance and passenger_count are correlated by PULocationID with outliers at 132 and 138.



- **PULocationID** by payment_type and passenger_count. This shows that Payment_type 2 has a relatively steady percent of total PULocationID



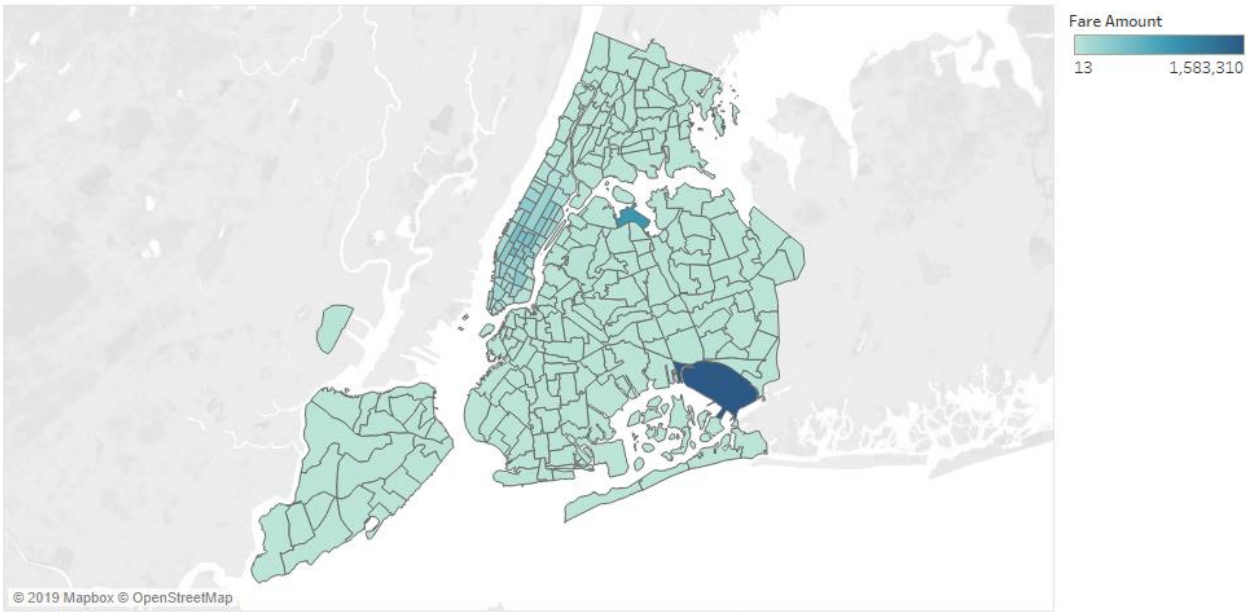
- **VendorID** and **tolls_amount**. This shows that there is a correlation between VendorID and tolls_amount



Tableau

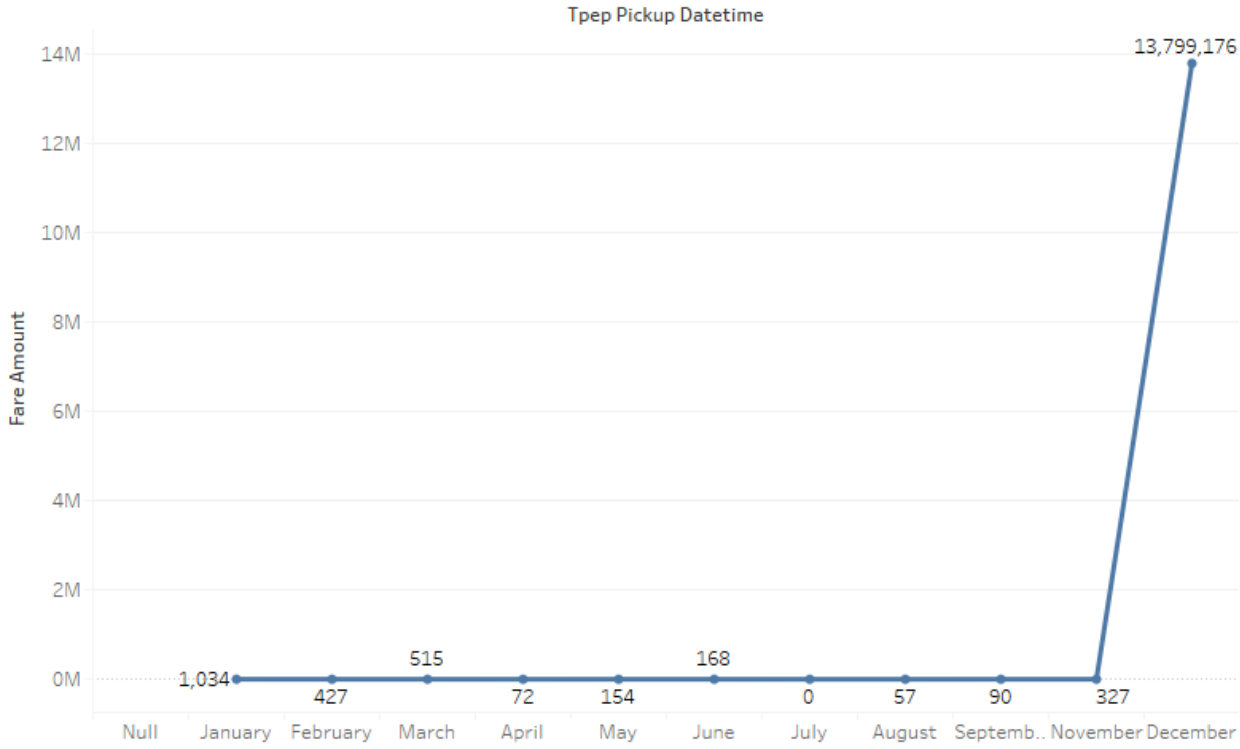
Initially we used tableau to understand the data and geospatial insights. Tableau is a great tool available for data modelling visualizing dataframes and geodataframes. We used Tableau Prep builder for data manipulation and cleaning. Further an outer join was performed on the shape file and the provided CSV file on (**PULocationID=LocationID)** condition for the desired visualization results. Some of them are listed below

Sheet 1



Map based on Longitude (generated) and Latitude (generated). Color shows sum of Fare Amount. Details are shown for Zone.

Sheet 2



The trend of sum of Fare Amount for Tpep Pickup Datetime Month. The marks are labeled by sum of Fare Amount.