Project Report
CS661: BIG DATA VISUAL ANALYTICS
2022-2023 Semester II
Project Title: NYC Taxi Trip Analysis

Team members:
Himanshu PS, 190379, hpshetty@iitk.ac.in
Bashaboyna Vasavi, 190228, vasavi@iitk.ac.in
K.Vyjayanthi Reddy, 190433, vyreddy@iitk.ac.in
IIT Kanpur

1 Introduction:

New York City is a bustling metropolis with millions of residents and visitors. One of the most common modes of transportation in the city is taxis, which provide a convenient way to navigate the city's busy streets. With millions of taxi trips taking place every year, it can be challenging to make sense of the data generated by these journeys. This is where visual analytics comes in. By using powerful tools to analyze and visualize taxi trip data, we can gain insights into patterns and trends that would otherwise be difficult to discern. From peak travel times to popular destinations, visual analytics can help us understand the intricacies of taxi travel in New York City. In this article, we will explore the world of NYC taxi trip visual analytics, looking at the tools and techniques used to analyse this fascinating data set.

2 Tasks:

1. Provide an detailed overview of number of trips in each zone of NYC on selected day(s) and chosen hours.
2. Analysis of number of rides started from each zone and their destination.Also ,get an estimate of trip total cost and trip duration for a selected start point and end point.
3. Get the Fare, Tip, Total Amounts of taxi rides across the months of 2010.
4. Understand the preferred mode of payment by users,get an idea about the famous places and destinations in NYC

3 Proposed Solution:

1. **Data Collection:** The NYC Taxi and Limousine Commission is the official data provider of New York City taxi trips .We've chosen the 2010 dataset and the data was provided for each month of 2010 stored in the PARQUET format.
2. **Choosing Tech Stack:** We used Python which provides versatility and a large ecosystem of libraries that make it an ideal language for data analysis tasks. As the dataset was very huge and memory consuming ,using Pandas was time consuming to process and work on the data.On the other hand,Vaex allows users to work with huge datasets that cannot fit into memory. It uses memory-mapping and lazy

evaluation to efficiently handle large datasets, making it an ideal tool.So, Vaex was used to work with this massive dataset.while Dash allows developers to create interactive web applications entirely in Python,we've used Python,Dash and Vaex together in creating interactive visualizations.

3. **Cleaning Data set:** Firstly, to work with Vaex, the PARQUET data files were converted into hdf5 format. The data had some serious outliers and were removed after a careful analysis.The parameters to clean the data were chosen based on the various histograms, heatmaps and line graphs generated for the different features of the dataset.We created a few columns like trip duration, trip speed, pickup hour, pickup day which were derived from the existing columns for quick and easy access .

4. **Data Visualisation:** We used the following features of data set:
'pickup longitude', 'pickup latitude', 'dropoff longitude', 'dropoff latitude', 'total amount', 'trip duration min', 'trip speed mph', 'pickup hour', 'pickup day','dropoff borough', 'dropoff zone', 'pickup borough', 'pickup zone' .
A Geographical map would be generated to get an overview of number of pickup in each of NYC.This Geographical map could be further customized on choosing the Vendor(CMT or VTS), desired timings and day(s) of week .To get an understanding about the outflow of taxis from a zone, a Sankey diagram would be a perfect choice which helps in understanding the popular destinations from the chosen zone, get an entire overview of drop off points from the selected pickup point.
A HeatMap of NYC based on number of trips with latitude and longitude, upon that allowing the user to select the desired pickup and drop off point based on the location coordinates and providing histograms of cost and trip count of trips belonging to the selected start and end points , also giving an estimation of trip cost and trip duration would give a better analysis of the trip.
The details about the month wise average of total amount,fare amount, tip amount, trip distance and trip duration could be provided in a line graph The number of trips in each borough(Queens ,Staten Island, Brooklyn, Bronx,Manhattan,EWR) on each day of week could be compared and understood from a line graphs.Finally , to comprehend the preferred payment type and total transaction amount on each day of week a bar graph is used.

4 Results:

The results of various visual techniques implemented are uploaded in the below google drive link.
https://drive.google.com/drive/folders/1bmhkruHNiUig3T8WBhXSX1lXaejZJ7GV?usp=sharing

5 Conclusion:

This visual analysis of NYC Taxi trip helps in easy understanding of entire data.Based on a user interest, this could help them in many ways – know the famous places of NYC,figure out the peak traffic hours of day, get the cost estimate of a trip, know the availability of cab on desired day and hour of a taxi vendor and so on.

6 Link to source code:

We've uploaded our code for this project in the below repository
https://github.com/pshimanshu/CS661-NYC-Taxi-DataVis

References

- https://dash.plotly.com/tutorial/
- https://dash.gallery/Portal/
- https://dash.plotly.com/dash-core-components/
- https://www.machinelearningplus.com/python/vaex/
- https://plotly.com/python/sankey-diagram/