REPORT

CENTRE FOR INFRASTRUCTURE, SUSTAINABLETRANSPORTATION AND URBAN PLANNING

Thejas N
PES UNIVERSITY

QUESTION 1:

This report analyses a dataset containing 6,867 bicycle trips over one day. The dataset contains information such as start and end times, as well as the latitude and longitude of the starting and ending depots

➢ Task 1 - Removing Zero Duration trip

The following steps were taken to solve the problem of removing zero duration trips from the bicycle-sharing system dataset

- A new function was created to find the difference between two times, which was used to calculate the duration of the trip.
- A new column named 'duration' was created, which stores the difference between the end and start times of each trip. This column was created to make it easier to identify and handle any issues related to time.
- The zero minute duration trips were segregated from the dataset, and a new dataset was created without these trips.
- The maximum and minimum duration of the trips were calculated from the new dataset, along with the percentage of circular trips.
- The runtime of the function was calculated to determine the time taken to complete the task of removing zero duration trips from the dataset.

```
Number of zero duration trip is : 47
Maximum duration of trip (in minutes): 518
Minimum duration of trip (in minutes): 1
Total number of trips corresponding to the minimum duration: 89
Percentage of total circular trips: 3.035190615835777
Runtime of the function (in seconds): 4.875218868255615
```

➢ Task 2- Feasible Pairs of trips

The following steps were taken to solve the problem of finding feasible pairs of trips:

- The original dataset was filtered to include only the trips starting between 06:00 AM and 06:00 PM.
- The number of feasible pairs was calculated by iterating through the dataset and checking if two trips can be served in succession by the same bicycle. A feasible pair is defined as two trips A and B, where the end location of trip A is the same as the start location of trip B, and the start time of trip B is greater than or equal to the end time of trip A.
- To solve this problem efficiently, a map was used to store all the ending locations(latitude and longitude) as keys and their respective values as a list of ending time.
- The dataset was then iterated again, keeping track of the start location of each trip. If the start location was already in the map and the ending time was less than the starting time, the count of feasible pairs was incremented.
- By making use of a map(as the dataset was in sorted order of starting time), the above problem was solved in linear time complexity.
- Total number of feasible pairs of trip was after filtering the dataset was found to be 44476

```
Total feasible pairs of trips: 44476
```

➢ Task 3 – Shortest Path distance

The following steps were taken to solve the problem of finding feasible pairs of trips:

- To address this problem, the dataset was reduced to only include the first hundred entries.
- The Osmnx module was used to formulate a map of the respective coordinates in the dataset.
- For each coordinate system in the map, the nearest depot was identified.

- The shortest distance between each depot and its respective coordinate system was calculated.

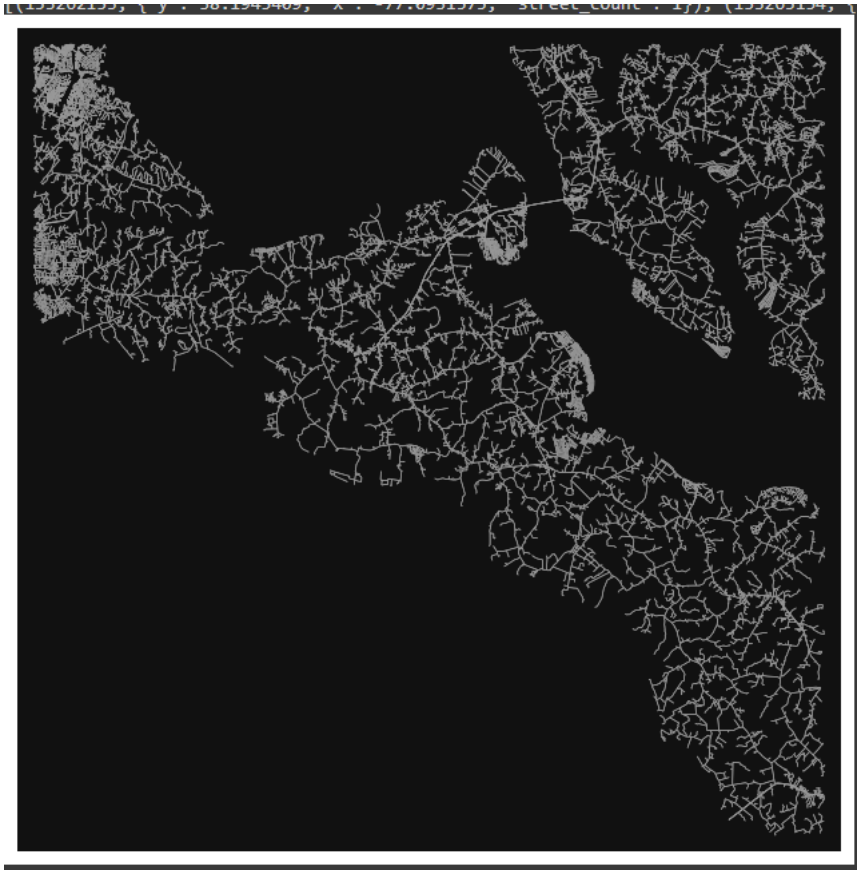
➢ Observation :

The given information can be summarized as follows:

- The number of unique depots used is 1.
- For each location system, the nearest depot is the same.
- Therefore, the shortest distance between any location and the depot is always 0.

  This suggests that all the locations are located at the same point as the depot, i.e., they are all the same location. Thus, the graph of the system would appear as a single point at the location of the depot.

  City Map View :



  Output :

```
{218522241}
Number of unique depots used: 1
Total runtime: 7.90 seconds
No feasible paths found between any pair of depots.
```

➢ Conclusion:

In conclusion, this report analysed a bicycle-sharing system dataset and answered three questions using exploratory data analysis and critical thinking. The results of the analysis provide insights into the system's usage patterns, feasibility of pairs of trips, and shortest path distances between depots. The analysis can be used to optimize the system's operations and improve its efficiency.

QUESTION 2:

As the dataset provided for this question was very large (1.3G) performing computation on this dataset was a challenge . Performing a simple linear time complexity function would take a lot of time . To tackle with this problem I made use of multiprocessing programming by creating different pools

➢ Preprocessing of dataset

Upon initial examination of the dataset, it became clear that it contained 8 columns. My immediate intuition was to group the data by the unique 'individual ids', as this would enable us to easily track and solve any issues pertaining to individual records within the dataset.

➢ Task 1

To calculate the total distance travelled by the individual . Steps taken to solve this problem :

- grouped the dataset by individual ids and make use of multiprocessing concepts to parallelize the code, resulting in faster execution times
- As the location coordinates were given, I used the Haversine method to calculate the distance travelled between 2 points
- While reading the data , computed the total distance travelled by the individual

```
[1194694.0390591281, 497596.02137196605, 1504176.1342267706, 1175265.0305609414]
```

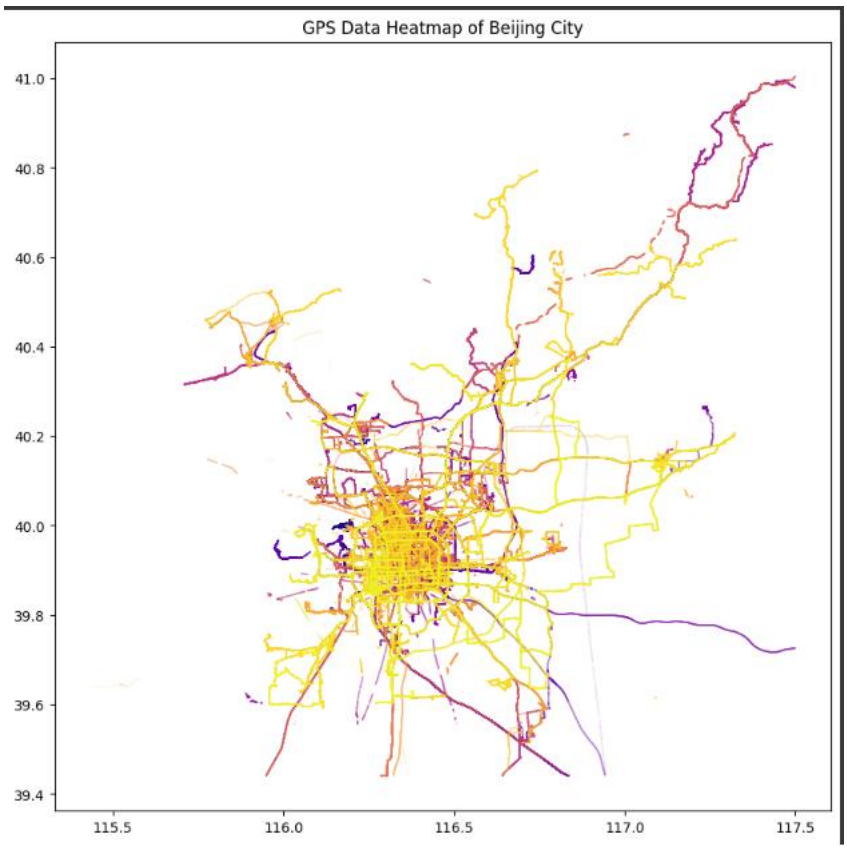
	user_id	total_distance
0	1	1.194694e+06
1	2	4.975960e+05
2	3	1.504176e+06
3	4	1.175265e+06

➢ Task 2

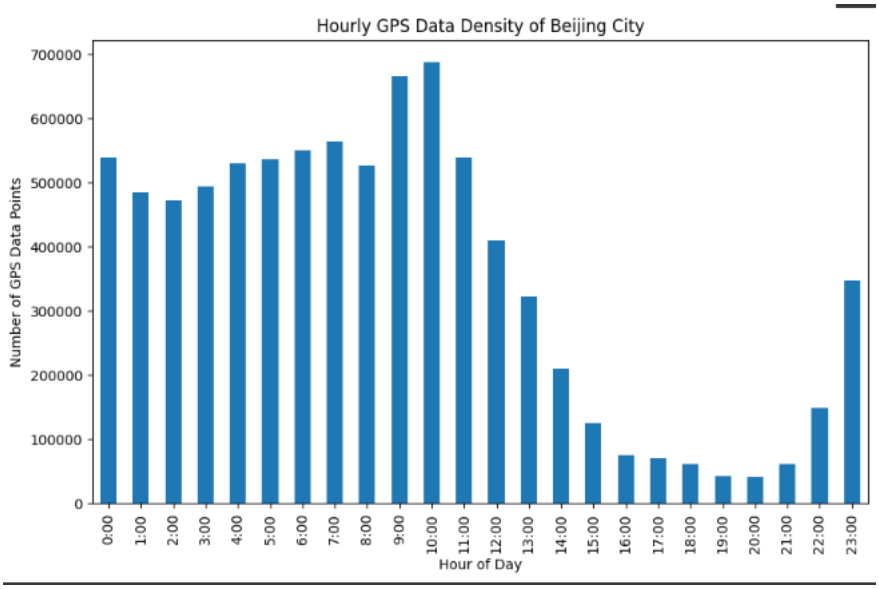
Visualisating techniques used : heatmap and histogram

At first filter out the coordinate that does not fit within the coordinates of Beijing City .

Furter plot the visualisation tools required for computation and understanding the dataset



This heatmap reveals that the area with longitude between 116.0 and 117.0, and latitude between 39.8 and 40.2, has a high population density. This valuable information can be utilized in various ways. For instance, if we need to promote a product, advertising it in the densely populated area could ensure a broader reach.



Based on the histogram, it appears that there is more movement and activity in the morning hours before noon. This could suggest that people are busier and more active during this time as they are likely commuting to their destinations.

➢ Task 3

With access to a GPS-tracking dataset that includes anonymized information such as latitude, longitude, altitude, date, and time of thousands of individuals over an extended period, the potential applications of this data are numerous. One issue that particularly interests me is traffic congestion in urban areas.

To solve this problem, I would use a methodology that combines data analysis and machine learning techniques. First, I would preprocess the data by cleaning and filtering the raw data to remove any anomalies or errors. Then, I would extract relevant features such as travel time, average speed, and distance covered to build a machine learning model.

The model would predict traffic congestion levels in different areas of the city based on historical GPS-tracking data. I would use clustering algorithms to group similar trajectories together to identify patterns of congestion. For example, if there is a high concentration of slow-moving vehicles at a particular time of day in a particular area, it may indicate

congestion. The model could then provide real-time predictions of congestion levels for different areas of the city based on current GPS-tracking data.

This information could be used by city planners and transportation agencies to optimize traffic flow and reduce congestion. For instance, if the model predicts heavy traffic in a particular area, the transportation agency could redirect traffic to an alternate route or adjust traffic signal timings to reduce congestion. Similarly, if a city planner is considering building a new road, they could use this model to identify areas of the city where new roads would be most effective in reducing congestion.

In conclusion, using GPS-tracking data to solve traffic congestion in urban areas is an exciting and challenging research problem. By combining data analysis and machine learning techniques, we can extract insights from the data that can help improve transportation infrastructure and reduce congestion.