Part 1: Planning the Dashboard for Research Questions

Research Questions:

1. What are the most popular bike stations in New York? (bar chart)
2. During which months are the most trips taken, and is there a correlation with weather?
3. What are the most popular routes?
4. Does temperature affect bike usage?

1. What are the most popular bike stations in New York?

- To answer this question, we need to analyze the 'start' and the 'end' station columns in the CitiBike trip data. By counting how often each station appears as either the start or end point of a trip, we can identify the most frequently used stations.

- We can visualize the data using bar charts or maps (heatmaps) to show the top stations in New York.

2. During which months are the most trips taken, and is there a correlation with weather?

- We will group the CitiBike data by month and count the number of trips per month. To examine the correlation with weather, we will use temperature data from NOAA for the same time period. By merging this weather data with the trip data based on the date, we can compare the number of trips in each month with the average temperature for that month. This can help determine if weather conditions influence the number of bike trips taken.

- We can show it using line or bar plot.

3. What are the most popular routes?

- Popular routes can be identified by looking at the combination of start and end stations. By counting how often each pair of stations appears, we can identify the most frequently used routes in the CitiBike network.

- This can be visualized using flow maps or a simple ranking of the most common station pairs. Or Edge-Weight Network (networkx)

4. Does temperature affect bike usage?

- To analyze the relationship between temperature and bike usage, we will need to compare temperature data (e.g., average temperature per day) with the number of bike trips taken on that day. We can calculate the correlation between daily temperature and daily bike usage. If there's a significant correlation, we can further explore how different temperature ranges (e.g., cold vs warm weather) affect bike usage patterns.

- Scatter Plot, Correlation Heatmap.

Data Collection and Setup:

To collect data for the above research questions, the following steps are required:

1. CitiBike Trip Data:
We will need the CitiBike trip data for the year 2022 that includes columns like:
- `start station`
- `end station`
- `starttime` (timestamp of when the trip starts)
- `endtime` (timestamp of when the trip ends)
- Other columns related to bike trip details.

2. Weather Data from NOAA:
We will need historical weather data from NOAA (National Oceanic and Atmospheric Administration). Specifically, we will look for daily average temperature data (TAVG) for New York City (or a specific station like LaGuardia). The dataset should include:
- `date`
- `temperature` (in Celsius or Fahrenheit)

3. Merging Data:
The CitiBike trip data will be merged with the weather data based on the date to enable analysis of how weather affects bike usage. We'll ensure both datasets have a consistent date format to merge them correctly.

Tools and Libraries:

- **Pandas :** For data manipulation and merging.
- **Matplotlib/Seaborn/NetworkX/Folium/Plotly :** For data visualization (e.g., bar charts, line graphs, and heatmaps).
- **Requests :** For fetching weather data via the NOAA API.

In the following steps, we will perform data cleaning, data merging, and visualization to answer these research questions.