**Assignment 1:**

In this assignment, we'll tackle the first-ever open data challenge hosted by the Metropolitan Transportation Authority (MTA) in New York City (NYC). Our goal is to analyze NYC subway ridership data using a graph-based representation approach. One interesting problem statement from the available dataset is: Determine the most congested subway lines and riding hours patterns to optimize the number of subways along certain routes.

We'll be working with two primary datasets:

1. **MTA Subway Stations**: This comprehensive dataset contains information about all subway and Staten Island Railway stations in NYC, including:
   * Location details (latitude, longitude, etc.)
   * Station Master Reference Number (MRN)
   * Complex MRN
   * GTFS Stop ID
   * Services offered (e.g., express trains, local trains, etc.)
   * Type of station (e.g., underground, elevated, etc.)
   * Whether the station is located within Manhattan's Central Business District (CBD)
   * ADA accessibility status

**Dataset Link:** <https://github.com/animeshbchowdhury/mta_subway_station/tree/main>

**Detailed description**: [MTA Subway Stations | State of New York (ny.gov)](https://data.ny.gov/Transportation/MTA-Subway-Stations/39hk-dx4f/about_data)

1. **MTA Subway Origin-Destination Ridership Estimate**: This dataset provides an estimate of subway travel patterns based on scaled-up OMNY and MetroCard return swipe data for a subset of ridership in January 2024 (approximately 9.1 million ridership). The dataset includes:
   * Estimated passenger volumes for all populated origin-destination (O-D) pairs
   * Aggregated by day of the week and hour of day
   * Information about the origin and destination subway complexes, including names, IDs, and approximate latitude and longitude

**Link:** <https://drive.google.com/drive/folders/1fV47SWGv5_AFPR_gRfvK1ra1LfSFCgOw?usp=sharing>

**Detailed description**: [MTA Subway Origin-Destination Ridership Estimate: 2024 | State of New York (ny.gov)](https://data.ny.gov/Transportation/MTA-Subway-Origin-Destination-Ridership-Estimate-2/jsu2-fbtj/about_data)

**How should you approach solving this problem?**

To create a graph-based representation of the data, you may start considering the following:

* **Nodes**: MTA Subway Stations
* **Edges**: Source and destination pairs of Ridership
* **Edge types**: Directed? weighted? Self-loops?
* **Node attributes:** What are the relevant information from the MTA Subway Stations dataset, which will help you do the analysis: Station type, CBD status, and ADA accessibility status, Train routes, GPS locations, Borough information?
* **Edge attributes:** What are the relevant information from the MTA Subway Origin-Destination Ridership Estimate dataset will be of help? Passenger volume, day of the week, and hour of day?

Once you’re able to represent this data, as part of graph-based representation, you are expected to develop queries on the graph-based representation which should return the following:

1. What are the top 5 origin sub-way stations from where most riders took subway ride:
   1. Across each borough (Manhattan, Brooklyn, Queens, Bronx, State Island)
   2. On Monday, Tuesday, Wednesday (combined)
   3. On Saturday and Sunday (combined)
   4. Between 1am-5am across all days and boroughs
   5. Between 6am-9am across all days and boroughs
2. What are the top 5 destination sub-way stations from where most riders took subway ride:
   1. Across each borough (Manhattan, Brooklyn, Queens, Bronx, State Island)
   2. On Thursday and Friday (combined)
   3. On Saturday only
   4. Between 12am-5am across all days and boroughs
   5. Between 6pm-9pm across all days and boroughs
3. What are the top 10 congested source-destination sub-way stations pair on:

* Monday between 1pm-2pm?
* On Queens borough, on Fridays between 6pm-9pm?
* On Brooklyn borough, Ridership between 1am-5am?
* Source is Brooklyn, Destination is Manhattan, Monday-Thursday 6am-7am?
* Source is Bronx, Destination is Manhattan, Monday-Thursday 6am-7am?
* Source is Staten Island, Destination is Manhattan, Monday-Thursday 6am-7am?

**Assignment guidelines:**

**Group:** We encourage you to solve this problem in a group of two.

**Libraries to be used for developing this project**: You may choose to utilize one of the following libraries to bring your graph-based model:

1. **NetworkX**: A widely-used and highly versatile library for creating and analyzing complex networks. NetworkX offers a vast array of tools and algorithms for graph manipulation, traversal, and analysis.
2. **Python-Graph-Tool**: Another powerful library for creating and manipulating graphs, offering advanced features such as graph filtering, clustering, and community detection.
3. **Other libraries**: Feel free to explore other Python graph libraries, such as Pygraphviz, Graphviz, or igraph, depending on your specific needs and preferences.

**Coding Assistance and API Integration**

If you require additional support or want to leverage the capabilities of coding Large Language Models (LLMs), you're encouraged to use them to develop small building functions and utilities as needed. Additionally, you can tap into the power of NetworkX's API to access its extensive range of graph-related functionality.

**Key Focus Areas**

While implementing your graph-based representation, keep in mind that this assignment is primarily focused on demonstrating your ability to:

1. **Break down complex problems into manageable components**: Effectively decompose the real-world problem of analyzing NYC subway ridership data into a graph-based representation.
2. **Develop a robust framework**: Create a solid foundation for your graph-based model, incorporating necessary data structures, algorithms, and techniques to achieve your goals.

**Further interesting readings:**

**Perks if you’re successfully complete this assignment/project:**

Participate with your submission in MTA Open Data Challenge ([MTA Open Data Challenge](https://new.mta.info/article/mta-open-data-challenge))

**Some cool visualizations of MTA data analytics:**

[The Subway Origin-Destination Ridership dataset (mta.info)](https://new.mta.info/article/introducing-subway-origin-destination-ridership-dataset)

**Can you do better than “Google” subway connection recommendation for NYC?**

[Where is everybody on the subway going? (mta.info)](https://new.mta.info/article/where-everybody-subway-going)

**Cool project proposals:**

[Open data discoveries for MTA enthusiasts](https://new.mta.info/article/open-data-discoveries-mta-enthusiasts)

**Pandas notebook for graph representation using networkx:**

[cs224w-slides-to-code/notebooks/01-introduction-machine-learning-for-graphs.ipynb at main · mnslarcher/cs224w-slides-to-code (github.com)](https://github.com/mnslarcher/cs224w-slides-to-code/blob/main/notebooks/01-introduction-machine-learning-for-graphs.ipynb)