**Motivation**

As Heinz students, studying public policy and information systems, we are interested in how we can apply data science to public problems. We are interested in public transit and commuter patterns through cities. When we saw the expansive data set from New York City’s bike-share system, we were interested in applying machine learning techniques to the data.

**Overview**

For our project, we aim to perform a clustering analysis of the hotspots where the program should invest more docks to keep up with demand. We also will try to predict ridership patterns based on time of day so the city can better understand the usage patterns.

**Domain & Learning Type**

We plan on applying clustering, classification, and regression techniques for this project. As we discussed in the lecture, “A Computer Program is said to learn from experience with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with E.” For this project, our program will learn from experience how to cluster, classify, and predict (the tasks). We will measure performance based on the different tasks. For example, we will compare the accuracy and precision of the classifications, and we will assess the RMSE for the length of time prediction. Our goal is to find a feasible model that can explain the data and provide business insight. Our objective for the performance of the models is, to take classification as an example, to achieve over 50% of accuracy and other performance metrics as the minimum standard and as high as possible.

**Data Set**

1. **Citi Bike rideshare data (https://ride.citibikenyc.com/system-data)**

* The data is stored monthly (total of 37 CSV files)
* Each month is a separate CSV file
* Rows: around 50000 for each CSV file (Sum of about 2 million rows)
* Columns: 13
* The data itself is cleaned. It is also expandable to look at the distance between the start station and the end station.

**Features in the dataset**

* Ride ID
* Rideable type (Classic Bike or Electric Bike)
* Started at (MM/DD/YY H:MM:SS)
* Ended at (MM/DD/YY H:MM:SS)
* Start station name
* Start station ID
* End station name
* End station ID
* Start latitude
* Start longitude
* End latitude
* End Longitude
* Member or casual ride (Causal or Member)

**Related Work**

1. Citi Bike Struggles to Keep Up With New Yorkers’ Love of Cycling (<https://www.nytimes.com/2021/12/02/nyregion/citi-bike-parking-docking-station.html>)

In this article, the author describes the growth of the New York City bike share post-pandemic era and the pain spot of the system. There are demand problems that do not meet the increasing popularity of rideshare and also distribution problems although the Citi bike rideshare system has already deployed hundreds of thousands of bikes, there are situations where users are not able to find bikes at destined bike stops. This implies that a study about the usage of rides to solve the redistribution problem might be urged.

1. Survey Says: City Voters Support Public Money Invested in Bike Share (<https://nyc.streetsblog.org/2021/05/11/survey-says-city-voters-support-public-money-invested-in-bike-share/>)

This article talks about the significance and public support for increasing subsidies and public funding for the New York City rideshare. New York City is the most crowded city in the U.S., and if having as many cars per household as other cities, the traffic situation would become worse. As the article mentioned, “we need to learn more and more into — opening up public spaces, getting out of our cars, focusing more on public transportation. This is the way of the future, unquestionably” This also leaves a question for public transportation developers - how to spend the money most efficiently so that can best serve the public needs of New York City citizens?

1. New York 25x25 (https://nyc25x25.org/)

This website is the main page of the advocacy, the New York 25x25 plan. This plan asks mayoral candidates to commit to dedicating 25 percent of the space now designated for vehicles — including 19,000 miles of roads and three million on-street parking spaces — as space for people by 2025 so that New Yorkers might have room to recover from (and thrive after) the COVID-19 pandemic, which has highlighted the city’s lack of equity when it comes to active transportation and green space.

**Questions**

1. Based on clustering analysis, where should the city invest more docking stations across the city?
2. Based on predictive analysis, can we predict the length of trip and time of day?
3. Based on the time spent, start station, end station, and distance, can we classify the user type and bike type?

**Possible findings and Implications**

Our clustering analysis could provide the data to drive the strategy of where to expand the program and add additional docking stations. Perhaps we can also better predict ridership patterns to address the pain points that Citi listed in the above NYT article: “Increased ridership and unpredictable commuting patterns during the pandemic have made it hard to distribute bikes and docks evenly, Citi Bike has said. The bike-share company and the city Department of Transportation are planning to add 8,000 more docks and 4,000 more bikes by the end of next year, largely in Manhattan, but also in parts of Queens and Brooklyn.”