Final Capstone Proposal

Introduction – New York City has the highest population per capita for a city in the United States. In result of this, many citizens of the city need to use different modes of transportation to get to and from work as opposed to driving like many do in other parts of the country.

- What is the problem you are attempting to solve?
    - Predicting the usage of bike travel based on the NY City borough they pick up and drop off while using the latitude/longitude values and borough data columns from the NY census dataset.
        - Variables used: Boroughs/ Pick-up & Drop Off/ Latitude and Longitude/Men/Women/Gender/Station Lat/long/ Bike ID/ User Type

- How is your solution valuable?
    - Knowing which borough should have more bikes at different times of day so that they do not run of bikes for rental.
        - Milestones:
            - Successfully merge the two datasets
            - Build and run successful models to predict goals
            - Design visuals that help explain data
            - Build presentation
        - Project Goal: To determine bike use from borough to borough and when the height of rental is for each of these areas.

- What is your data source and how will you access it?

Please review the attached website – merging these two data sets:

https://www.kaggle.com/muonneutrino/new-york-city-census-data#nyc_census_tracts.csv

https://www.citibikenyc.com/system-data

Using jupyter notebooks to

- What techniques from the course do you anticipate using:
    - Clustering
    - Random Forest
    - ARIMA- modeling trend in bike usage over time

- What do you anticipate to be the biggest challenge you'll face?

It'll be exciting and challenging to see how my models predict the trend of Citi bike rentals in the different boroughs of New York City. But some of my challenges are-

- Understanding my data and lining up the start and stop stations with the latitude/longitude of the boroughs from the two different data sets.
- Recognizing why a specific model is working well or the other way
- Deciding visualizations to convey my message in the best possible way
    - Using two different data sets and converging them could present problems if the data doesn't align up well, so it may require further data cleaning and organization to make sure that the models run appropriately and predict accurately.