ORIE 5530
Project: NYC Citibike

Due on Tuesday Nov 22nd, 11:59pm (EST)

Notes:

- This is a group project. Each group should contain at most three students.
- Each individual must submit the group report on Gradescope.
- You should also submit your programming code along with the report.

Data: In this project, you will work with NYC CitiBike data. You will use the dataset for the month of July 2022. Please download the dataset *"202207-citbike-tripdata.csv.zip"* from the link *https://s3.amazonaws.com/tripdata/index.html*.

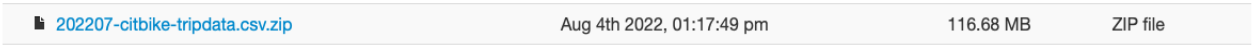

Figure 1: Screenshot of the dataset you should download

This dataset contains Citibike rides in NYC that happened in July 2022. Your dataset should contain 3,497,392 rows and 13 columns. Each row corresponds to a ride (i.e., a trip). The columns give the ride ID, the ride type (classic, electric), the start time of the ride, the end time of the ride, the start station name and ID, the end station name and ID and other info. Please find a detailed description of the dataset at the following link *https://ride.citibikenyc.com/system-data*. Note that the data has already been cleaned by CitiBike. I recommend to use Python and Pandas for this project. Here is an interesting article on how to explore this dataset using Python and Pandas *https://towardsdatascience.com/exploring-bike-share-data-3e3b2f28760c*. This article is just to get familiar with Python/Pandas, the format of the data in the article is not the same as the one you will be using.

First, I will ask you some warm-up questions to familiarize yourself with the dataset and review your probabilities material. Then, I will give you the project problem where you will use Markov Chains to study the availability of bikes in CitiBike stations. While the warm-up questions are specific, the project question is more open and flexible.

Warm-up questions:

First, clean the data by eliminating the rides that last more than 3 hours (if any). These are likely bikes that were left undocked.

1. Using the start time and end time, compute the duration of each ride in minutes and plot the histogram of ride durations.

2. What is the expected ride duration (i.e., the average ride duration)? What is the empirical variance of ride duration? What is the probability that a ride duration is greater than 20 min?

3. What is the probability that a ride duration is greater than 20 min conditioning on the fact that the user is a CitiBike member? Note that the last column gives whether the ride is for a casual client or a CitiBike member.

4. Suppose that the duration of some ride is more than 25min. What is the probability that this ride belongs to a CitiBike member?

5. What is the expected ride duration of an electric bike? What is the expected ride duration of a classic bike?

6. Suppose that the duration of some ride is less than 10min. What is the probability that this ride uses an electric bike? What is the probability that this ride uses a classic bike? Comment on the results.

Project: We would like to estimate the steady-state of the number of available bikes in a CitiBike station using Markov chains. In other words, suppose for example there are 20 bikes available in a station near Central Park early in the morning of Monday, our goal would be to come up with a simple estimation of the number of available bikes in that station at the end of the morning.

For that purpose, choose three CitiBike stations in NYC. Make sure to choose popular stations where there is a lot of movements (i.e., bike coming in/biking going out very frequently). Split the day into two blocks of time, like morning/evening and focus only on weekdays. Discretize each block in periods of 5 or 10 minutes (whatever makes most sense for you). So the difference between two consecutive time steps would be 5 or 10 min. Model the number of available bikes in a station as a Discrete Markov Chain. The states of the Markov Chain are all the integers between 0 and the capacity of the stations (i.e., number of docks). Using the full month of data estimate the transition probability matrix of each station. Note that you have to estimate one matrix for the morning block and another one for the evening block since the patterns are different. Use only weekdays because again patterns are different on the weekend. To estimate this transition probability, for each time period of 5 or 10 min, you should look to the rides that went from that station and the ones that arrive to that station to update the number of available bikes. Compute the stationary distribution for each station (the morning one and the evening one). Comment on the results. What insights did you drive from this study?