Bike Sharing in NYC

What is the optimal number of bikes in the Citi Bike fleet, and how many bikes should typically be moved overnight to keep the system in balance? In this case we will provide an answer to these questions and a couple more. Our models will likely not yield the best answer possible, but we can certainly provide a very good start.

Grab the July 2019 (to avoid Covid effects) Citibike usage data in NYC (the file "201907-citibike-tripdata.csv.zip") from the Citibike website https://www.citibikenyc.com/system-data. Complete the following steps:

1. The station information is in the file "station_information.json" supplied separately. The file contains station id and station capacity and some other information. However, some stations shown in the travel data cannot be found in the station information json file. This may be because the station information json file is up-to-date and some old stations might have been removed. Thus, please remove from the travel history all the trips whose start station or stop station cannot be found in the station information json file, or is NaN.
2. Suppose that the rides that are initiated **at each Station** i follow a Poisson process in time, with a rate function that is constant in each hour of the day on weekdays. (Ignore weekends.) In other words, at Station i, there is a constant arrival rate $\mu_0(i)$ in place from 5am to 6am on weekdays (Monday, Tuesday, Wednesday, Thursday and Friday), a potentially different constant value $\mu_1(i)$ from 6am to 7am on weekdays, . . ., a potentially different constant value $\mu_{18}(i)$ from 11pm to midnight on weekdays. Using the Citibike data you obtain from Step 1, estimate these 19 numbers for each station i, where $\mu_t(i)$ represents the average number of rides initiated per hour in the tth hour of a weekday from Station i. Ignore the variation between weekdays that you saw in HW 2. Also ignore censoring. We'll assume that the number of rides from midnight to 5am is negligible and ignore them.
3. Estimate, for each hour t the transition probabilities $p_t(i, j)$ = Pr(destination station is j | a ride leaves Station i in hour t). (This is just the fraction of rides beginning in Station i in time t that go to Station j). Effectively, you are generating 19 matrices, one for each hour, where each matrix is a transition matrix so is nonnegative and has row sums equal to 1. If no rides originate from Station i in hour t then just set $p_t(i, i) = 1$ and $p_t(i, j)=0$ for j not equal to i.
4. Compute, for each hour t, the arrival rate of riders who are returning bikes to Station i for all stations i. In other words, at Station i, there is a constant arrival rate of bikers returning bikes $\lambda_0(i)$ in place from 5am to 6am on weekdays (Monday, Tuesday, Wednesday, Thursday and Friday), a potentially different constant value $\lambda_1(i)$ from 6am to 7am on weekdays, . . ., a potentially different constant value $\lambda_{18}(i)$ from 11pm to midnight on weekdays. Compute these 19 numbers for each station i, where $\lambda_t(i)$ represents the average number of riders returning bikes per hour in the tth hour of the week from Station i. You can get these values from Parts 2 and 3 above; don't estimate them from the data.

Now, for each station in the system, compute the optimal number of bikes to place at the station by 5am to minimize outages throughout a typical weekday. Also compute the net

flow of bikes over a typical weekday at all stations (this entails computing one number for each station, and is essentially how many bikes would need to be moved per station per day to keep the system in balance).

Now you are ready to provide answers to the following questions, graphically where appropriate.

1. What is your prediction for the optimal number of bikes in the fleet?
2. How many bikes need to be moved overnight assuming ideal positioning each morning from #1 above?
3. What is your predicted number of outages over a typical weekday, assuming ideal positioning each morning? Compare your answer to the average number of rides taken in a typical weekday.
4. Suppose you are trying to set the point values for Bike Angels at 2pm on a Tuesday. These point values depend on the number of bikes in the racks at 2pm. Since we don't have the data feed, suppose that all stations are half full at 2pm (rounding down if a station has an odd number of bikes). Use your model to estimate the value of one more bike or one fewer bike at 2pm at a station in reducing outages till the end of the day, and determine which of "adding a bike" or "removing a bike" gives the better system improvement. Repeat this calculation for all stations. Show your answer graphically, by giving a single value for each station, coloring the value black if changing the number of bikes makes no difference, coloring the value blue if adding a bike will most reduce outages and coloring the value red if removing a bike will most reduce outages. If placing text on the map is too cluttered or not easy to do, instead color the stations to indicate the value. Discuss how you could use the values you get to set the points in the Bike Angels system.
5. Discuss the 4 biggest limitations of your model and computation. For each limitation, what is the weakness and is there a way to avoid it?

Please prepare the usual 6-or-fewer page writeup of your case, along with a pdf of your code as an appendix (that doesn't count towards the 6-page limit).