**Abstract**

In this short study, I explored the travel patterns of CitiBike in NYC using one month of data from September 1 to 30, 2017. Based on quantitative and visual analysis, I identified stations that are subject to significant bike stock overloaded and under-supplied when re-balancing is not in effect. This insight helps to quantify the significance of having a robust re-balancing operation to support consistent service delivery. This is an important issue based on current studies [1]. However, I’d like to understand more about the dynamics specific to NYC. As I drilled down to more details, I discovered various distinct recurring travelling patterns that can help to support operation and expansion planning.

**Project Motivation and Questions**

Deploying, rebalancing, and servicing bikes are crucial parts of bike sharing business. Based on a common observation, bikes tend to accumulate in certain locations while deplete rapidly in others due to popular travel patterns. I wonder how important is bike rebalancing. However, the insight of bike balancing operation is not available to the public. I am curious to understand the dynamics of how bikes distribute across the city throughout the day. If I am the business owner of a bike sharing company, I’d be concerned about aligning the rebalancing capacity with station expansion in order to maintain stable services. If I am a dispatch operator who is responsible for balancing bikes, I'd be interested to know the following:

1. What are the most popular departing and destination stations?
2. What are the stations with the most and least inventory of bikes by Sept 30, 2017, if bike re-balancing is not in effect?
3. Which areas may experience the most risk in overflowing and under-supplied bikes?
4. What are the travelling pattern based on the most popular stations?

**Data**

I used the September 2017 CitiBike data to support this analysis. There is a total of 1,878,098 trips and 694 unique bike stations across all NYC boroughs from this month of data. To enhance my analysis, I integrated two additional datasets: NYC Subway Entrance and Zip boundary shape files.

Standard data cleaning, such as de-duplication and bad data removal were applied. For example, some stations have longitude and latitude of (0,0).

**Method**

The main data process is to calculate hourly bike balance of each station in order to understanding bike accumulation and depletion.

The *Bike Balance* at the beginning of hour *i* is defined as: Balancei = Balancei-1 + Net flowi-1

The *Net flow* at hour *i* is defined as: Net Flowi = # of Departurei – # of Arrivali

A wide data frame was created with the following key columns based on this approach:

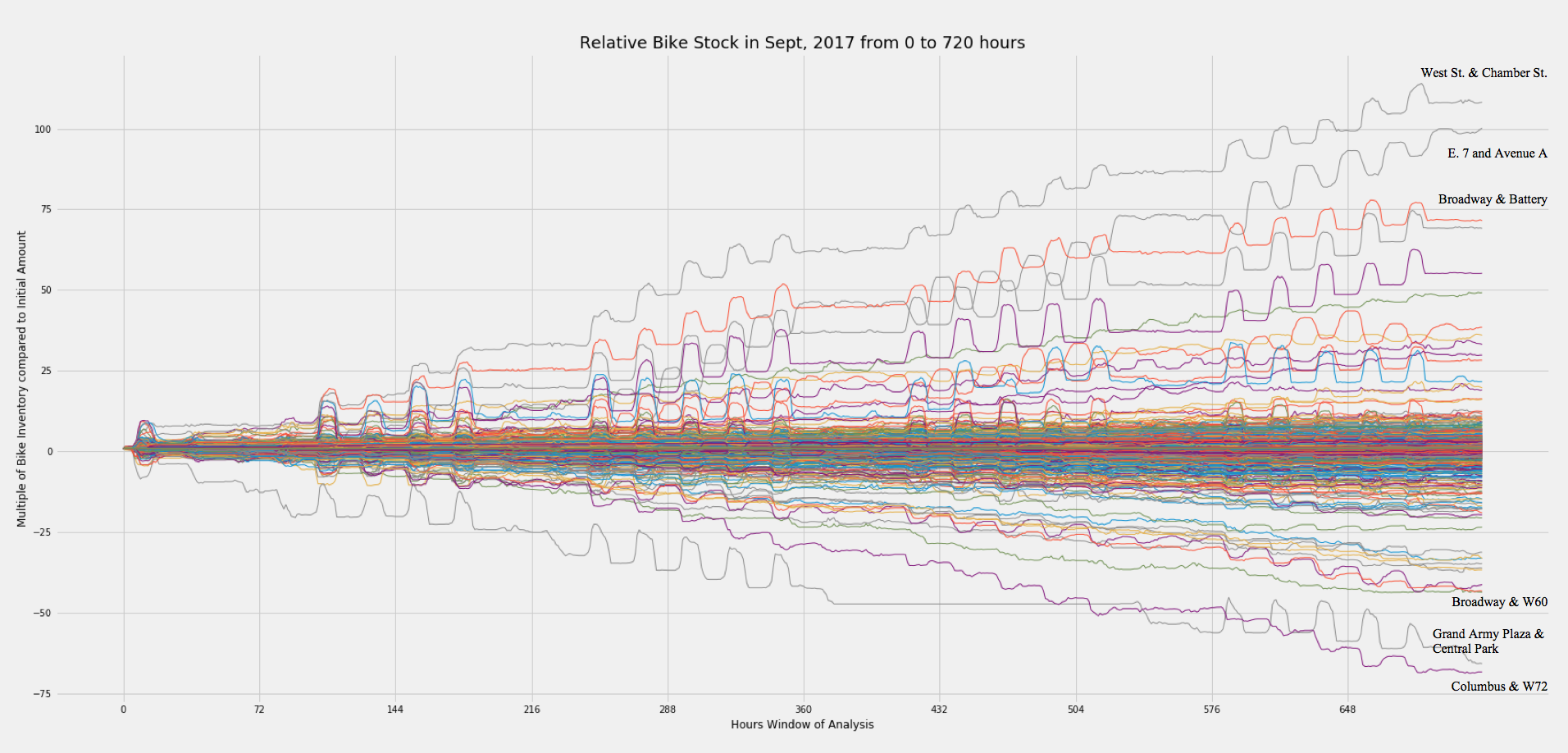
* Station ID and locaiton information
* Hourly Departure from 0 – 23 hour for everyday from Sept. 1 – 30, 2017
* Hourly Arrival from 0 – 23 hour for everyday from Sept. 1 – 30, 2017
* Hourly Net Flow from 0 – 23 hour for everyday from Sept. 1 – 30, 2017
* Hourly Balance from 0 – 23 hour for everyday from Sept. 1 – 30, 2017
* Hourly Change Multiple relative to 12 am, Sept1, 2017

In order to complete this calculation, I made an assumption that there are 50 bikes at each station at 12 am, September 1, 2017. This is a simplified assumption of the reality because certain popular stations will have more bikes. However, I think it is reasonable and conservative because the assumed quantity is aligned with the largest station in the city near Time Square.

Having this dataset allowed me to perform Time Series visualization and quantify how bikes accumulate and deplete at a granular level. It also allowed me to discover common areas where people depart and arrive, which helped me to infer general geo-clusters of travelling patterns.

Additional datasets with aggregated counts of trips by station and routes (starting and end stations) were created to support additional analysis.

**Results & Discussion**



**Figure 1:** A time series plot of bike inventory multiplier compare to initial inventory of 50 bikes at 12am, September 1, 2017 without re-balancing. The stations with the most and least bikes are labeled.

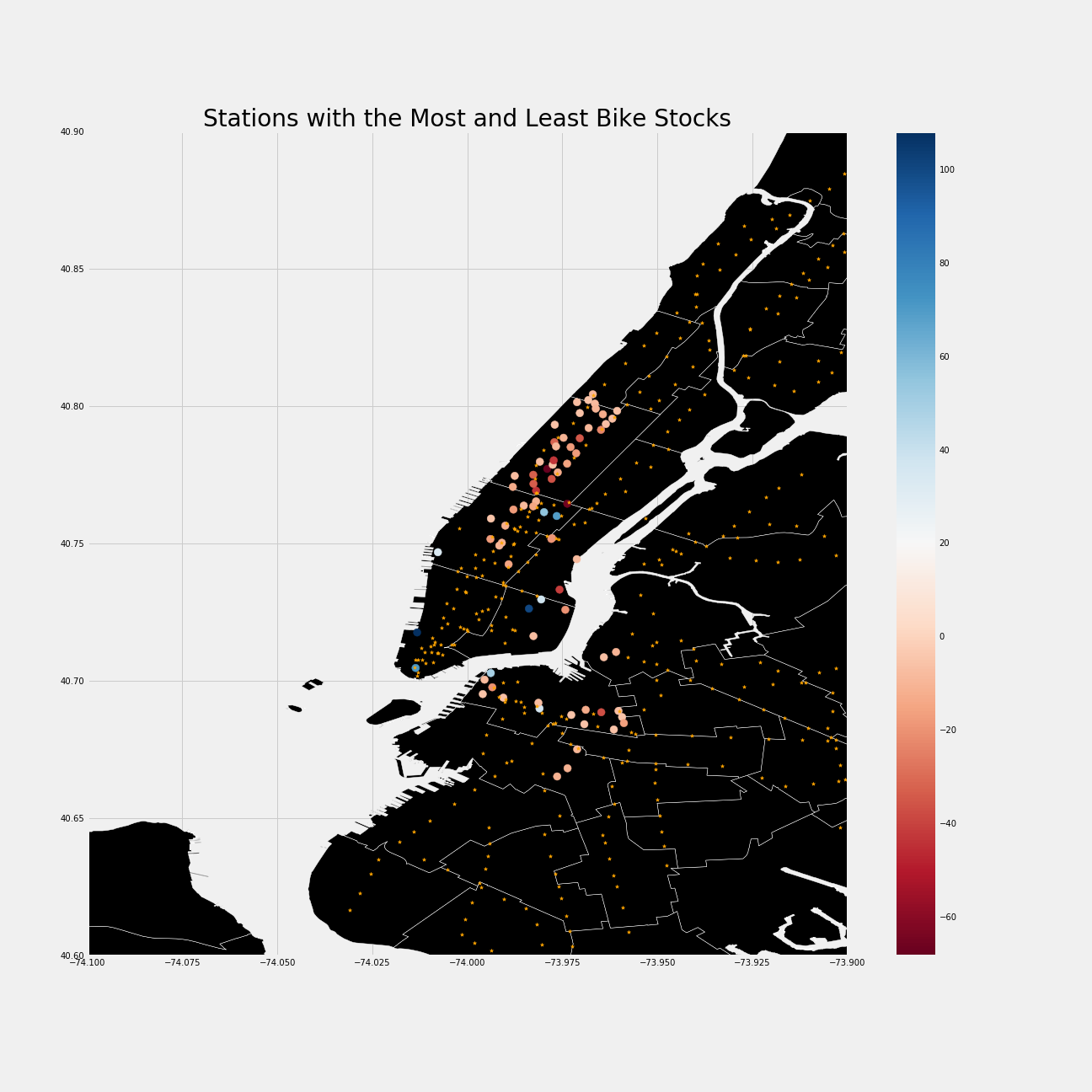
This visual analysis confirmed and quantify various observations. First, some stations are more popular as origination while others are common destinations. Over a period of one month, the top originations with the most bike accumulation is the stations at West St. and Chamber St. The most common departure station is the station at Columbus & W72.



**Figure 2:** Google Map screen shots of West St. & Chamber St. and Columbus & W72

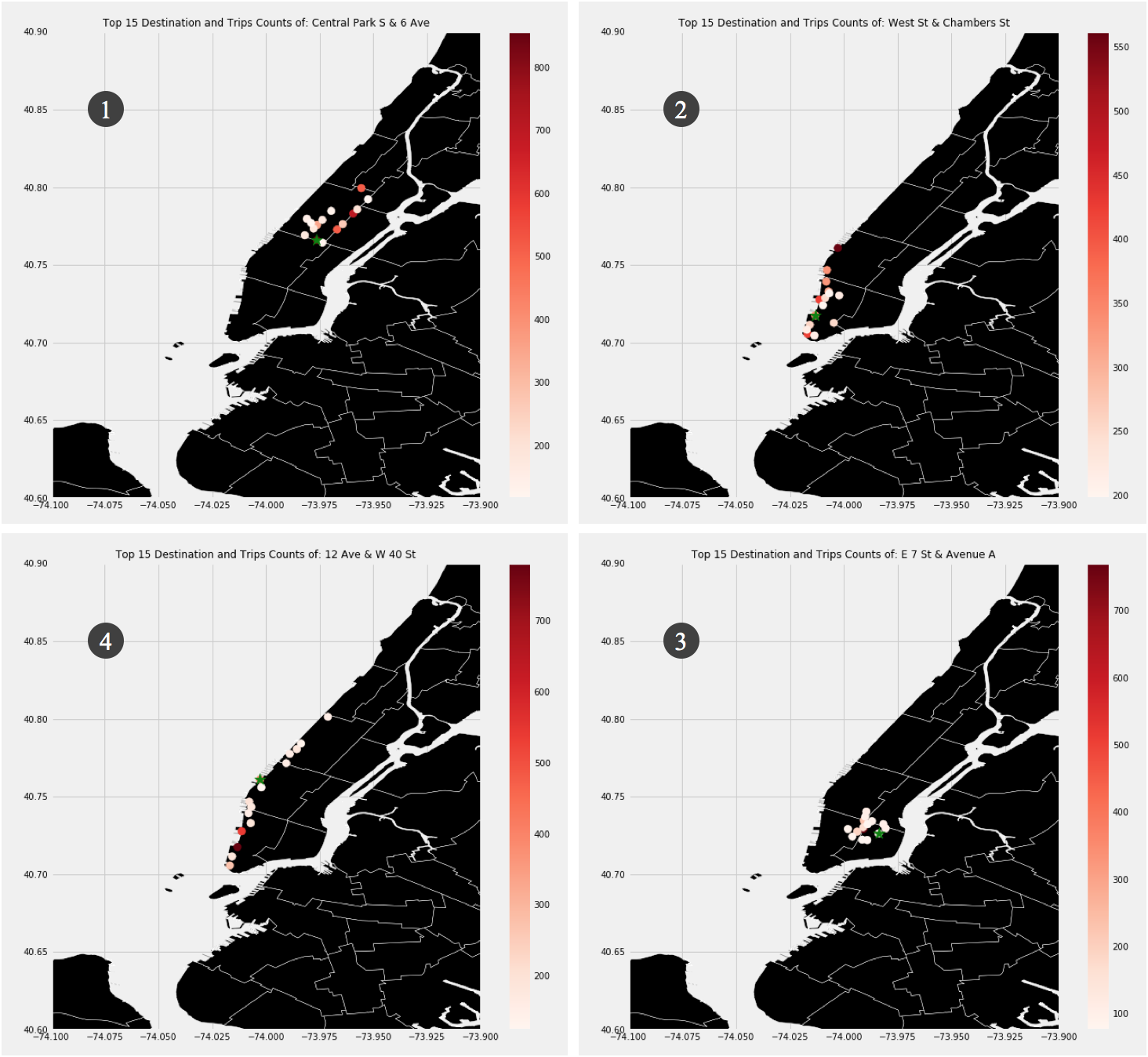
Secondly, this analysis helps to quantify the importance of bike re-balancing. If each station has 50 bikes on Sept 1, 2017, in extreme cases, some stations will accumulate over 3,000 to 5,000 bikes (60 to 100x the original). This causes significant overflowing of bikes. Bikers will have to find a different station nearby to park, otherwise the bikes will simply block the walkways. In contrast, some stations will be need 2,000 to 3,400 bikes to fulfill rider demand.

Last but not the least, the periodic spikes of bike inventory suggest that some of the stations are frequently used by commenters. There are large departure and arrival trips around morning and afternoon rush hours.



**Figure 3:** The most high risk areas with overflowing (blue) and under-supplied (red) bike inventories. Small orange dots are subway stations.

This analysis demonstrates that areas near Upper West side have a high risk of running out of bikes. Core commercial and tourist areas, such as lower Manhattan, south edge of Central Park, East Village, Dumbo, and downtown Brooklyn, tends to have overflowing problems. The overlaying with subway stations locations also suggests that some of the most popular stations are slightly remote, but not too far, from the subway.



**Figure 4:** This visual shows the 15 most common recurring routes of the top four stations. The green star is the starting station. The red dots are the destinations with color scaled by the number of occurrences in September 2017.

This analysis demonstrates distinctive traveling neighbourhoods (corresponding to the numbering labels):

1. Start from the tip of Central Park and tour around
2. Start near financial district near Manhattan towards lower west side
3. Start from east village toward midtown Manhattan
4. Start from middle west Manhattan to lower and upper west sides.

These insights can be used to support decision in operation and expansion planning. For example, where and when should Citibike allocate staffs to conduct re-balancing. There are specific cycles and locations pattern based on the analysis. In addition, some travelling pattern in relation to subway and core city areas can be generalized (although many are city specific) and applied to new city expansion for services similar to Citibike. Bike sharing services are booming in Asia at the moment.

**Areas to be Improved**

There are two specific visualization improvements to be made. First, the geo-maps are slightly distorted when ranges of x- and y-axis were specified. Although the relative positions of different geo-elements (e.g. stations, areas, and subway stations) remain intact because they were re-scaled together, it can be distracting if the overall geo-map is not aligned to what readers are accustomed to see (e.g. Google Map).

Secondly, I’d like to align zero to the white color in the color bar in Figure 3, which is aligned to light red currently. This may help to avoid the misinterpretation of due to the scale. Performing normalization can be a solution, but the true counts will be lost. Further investigation is needed.

**Additional Work**

I steered away from using unsupervised clustering techniques to find common departure and arrival, and travel pattern because descriptive analysis can be more direct and explainable. However, using clustering techniques can help to discover hidden patterns that are not obvious. I can use them as inspiration to form scientific hypothesis to further studies.

In a transportation planning context, it will be interesting to conduct hypothesis tests on whether bike sharing helps to bridge the last mile travelling, encourage multi-modal commuting, and complement MTA services. Some of the insights from this short study provided positive hints, but a more rigorous study is required to reach any conclusion.

In addition, I am also curious to develop a modelling solution – as a practical product - to help support work force planning. An idea can be using Monte Carlo simulation with time series forecasting to understand the best and worst-case scenarios for rebalancing for upcoming hours.

**Bibliography**

[[1] https://www.citylab.com/transportation/2014/08/balancing-bike-share-stations-has-become-a-serious-scientific-endeavor/379188/](https://www.citylab.com/transportation/2014/08/balancing-bike-share-stations-has-become-a-serious-scientific-endeavor/379188/)

**Data source**:

CitiBike - <https://www.citibikenyc.com/system-data>

NYC Area: https://data.cityofnewyork.us/api/geospatial/cwiz-gcty?method=export&format=Shapefile

NYC Subway Station Locations: https://data.cityofnewyork.us/api/geospatial/arq3-7z49?method=export&format=Shapefile