NYC Bike Sharing System Analysis

Shiying Tu, Lelei Zhang, Muyao Chen

Problem definition:

With an increasing popularity, bike sharing systems are getting deployed in more and more cities around the world, it's a cost effective way of promoting a sustainable lifestyle in urban areas and it plays an important role in the transportation network and provides alternative transport methods in cities with heavy traffic and improves public health.

Despite the advantages and popularity, there are several issues arise during planning and operation of bike sharing system. For instance, the distribution of bikes, there is a great demand at specific locations and at specific times, we encountered problems such as bike shortages during peak hours. There are also some stations may be fully loaded at certain hours, making it difficult for people to return their bikes when they reach their destination. We identified which stations are shortage or overload and we plan to remove bikes from fully occupied stations and refills in stations where they are needed. This is another way to rebalance during peak commuting hours.

The bike sharing system also poses a set of challenges to be addressed and solved. Some of the challenges like estimating the demand at a station, it depends on several factors like nearby locations, passenger counts, trip distances, travel time and total payment amount etc. We approach the issue of inferring actual travel needs and use these amounts to better plan for system usage.

Methodology:

The first problem is to analyze the popularity of NYC bike sharing system preparing, which could give us a big picture overall. With the help of Apache Spark Dataframe, we are able to extract trip counts for each station out of 6 GB data in around 1 minute, which proves the great power of Spark.

We tackle the problem of maintaining system balance during peak rush-hour usage as well as rebalancing from nearby stations to prepare the system for rush-hour usage. In order to find out which area is overload and whether the bike storage at each location satisfy users needs, we use Pyspark to calculate the number of bikes in and out of the station for each day corresponding with the station capacities to determine the storage station. Then we extracted some useful information and made some interpretations to find the five nearest stations. Once we had the nearest stations, we were able to move bikes from those overload areas to shortage areas.

From the second question, we found some shortage stations doesn't have supply in their shortage hours. So we addressed our third question: Which locations have the potential to attract people to take bikes instead of taxi. Firstly, we calculate the average distance travelled by bike, then based on that distance, we filter out the 30 million taxi trips that took longer distance than the average, the second step is to select the trips with low velocity. We then use Pyspark clustering algorithm K-means in MLlib to train the k-means model, cluster the data points based on input coordinates as features to analyze the taxi data and discover the best placement of bikes to facilitate usage.

Problems:

Though we focus on processing big data on PySpark in this project, front-end UI can provide a better visualization about populated results. We had a hard time deciding the tools used for UI. The drawback is HTML + CSS does not have the dataframe data structure, which is the main structure from the backend. Another technique is Plotly + Dash, which is a brand-new technique to us and can use Pandas dataframe to populate graph. Within 3 days, each of us learnt Plotly + Dash and wrote a small demo in both HTML/CSS + Javascript version and Plotly + Dash version. Finally we decided to use Plotly+Dash for our front-end implementation.

Second problem, we found that some stations will have negative bike storage. According to citibike monthly rebalancing plan, no individual station outage shall continue for longer than 4 hours [1]. For simplicity, we assume that all bike stations will be restored to the number of available bikes at 12:00 am.

To find the five nearest stations, we cannot calculate the Euclidean distance between two points on the Earth as we need to consider the coordinate degree of the Earth's convergence. Instead, we use the Haversine Formula expressed in terms of a two-argument inverse tangent function to calculate the circular distance between two points on the earth.

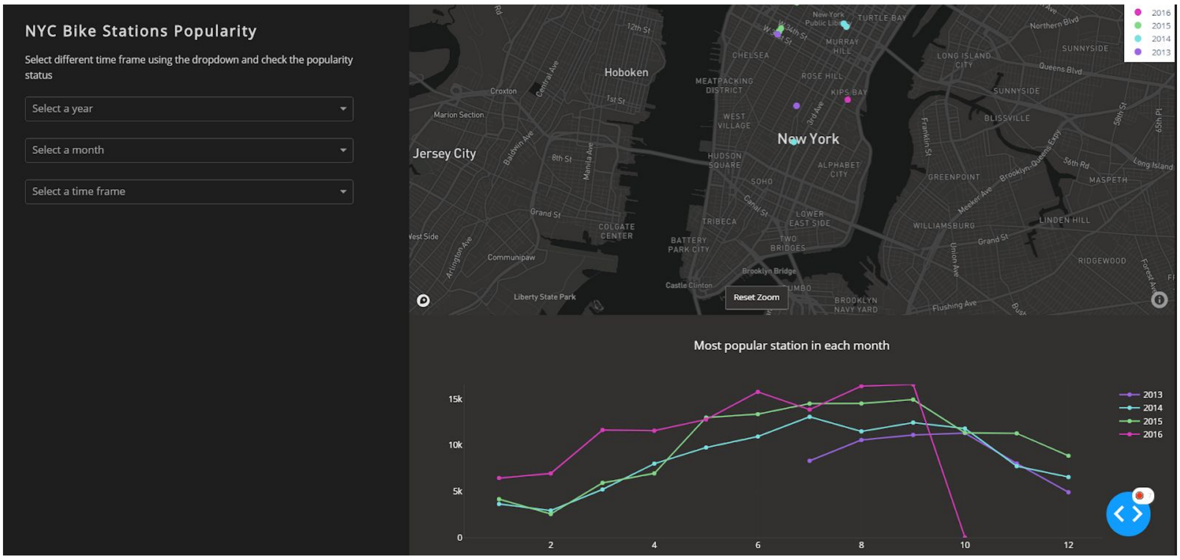
We find that the the official provided stations information are not consistent with the real travel data. For example, We found some users start or end their trips at station 'E42 st & Vanderb'. However the station is not listed in the stations_df table, which means we are not able to track its capacity and number of bikes available. To solve this, We wrote a script to update the stations_df table, and assumed that those newly added stations has 0 bikes available and 0 capacity.

Result

Popular Stations: Seen from the map, all popular stations are all located midtown part of Manhattan. Pershing Square North was the most popular one regarding yearly bike movement. In 2016, the total bike movement for this station was 111,897, which means every 2 minutes, there will be a bike borrowed or returned in this station.

However, most of the borrow-return events in Pershing Square North station are happened during the morning session. For the afternoon, the most popular stations were seasonally different,West st & Chambre St could be more popular than Pershing Square North during May to July. Lafayette & E 8 st were most popular during October to December. while in January and February, the Lexington Ave & E 24 st can be more active than lafayette & E 8.

Basically, we can summarize a pattern that during the winter (December to February), the top 10 most popular stations basically avoid all tourist attractions, they separated around Union Square, East Village, and tribeca area.



NYC Bike Stations Popularity

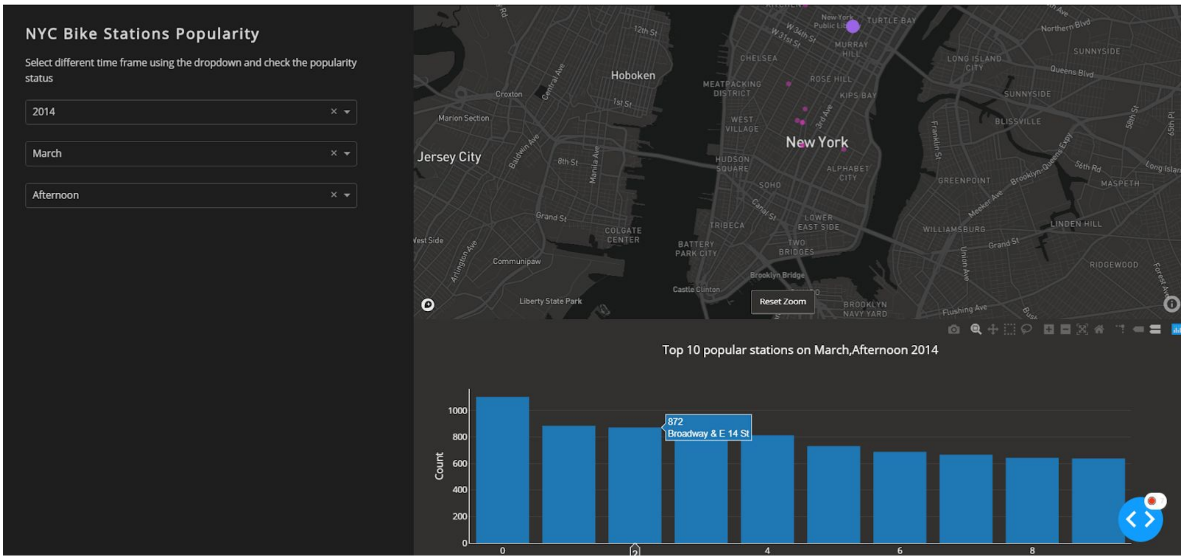
Select different time frame using the dropdown and check the popularity status

Select a year

Select a month

Select a time frame

Most popular station in each month



NYC Bike Stations Popularity

Select different time frame using the dropdown and check the popularity status

2014

March

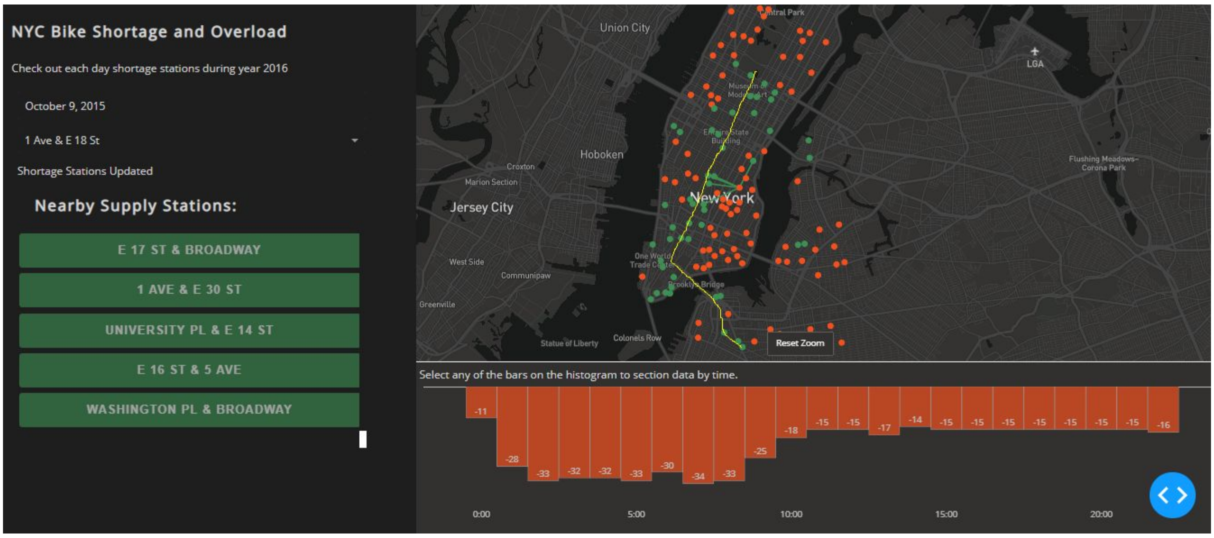
Afternoon

Top 10 popular stations on March,Afternoon 2014

872
Broadway & E 14 St

Bike Movement: There are more shortage stations than overflow stations. Basically the locations of overflowed stations separate around the metro line N in Midtown Manhattan and lower Manhattan, including Koreatown, Soho, and Wall street.

While overflowed stations locate cross along the centerline of Manhattan Island, shortage stations are closer to the edge of Island. Those stations are usually located in lower East side, East Village, GreenWich Village, Upper West Side, Upper East side, Downtown Brooklyn, and downtown Williamsburg.


NYC Bike Shortage and Overload

Check out each day shortage stations during year 2016

October 9, 2015

1 Ave & E 18 St

Shortage Stations Updated

Nearby Supply Stations:

- E 17 ST & BROADWAY
- 1 AVE & E 30 ST
- UNIVERSITY PL & E 14 ST
- E 16 ST & 5 AVE
- WASHINGTON PL & BROADWAY

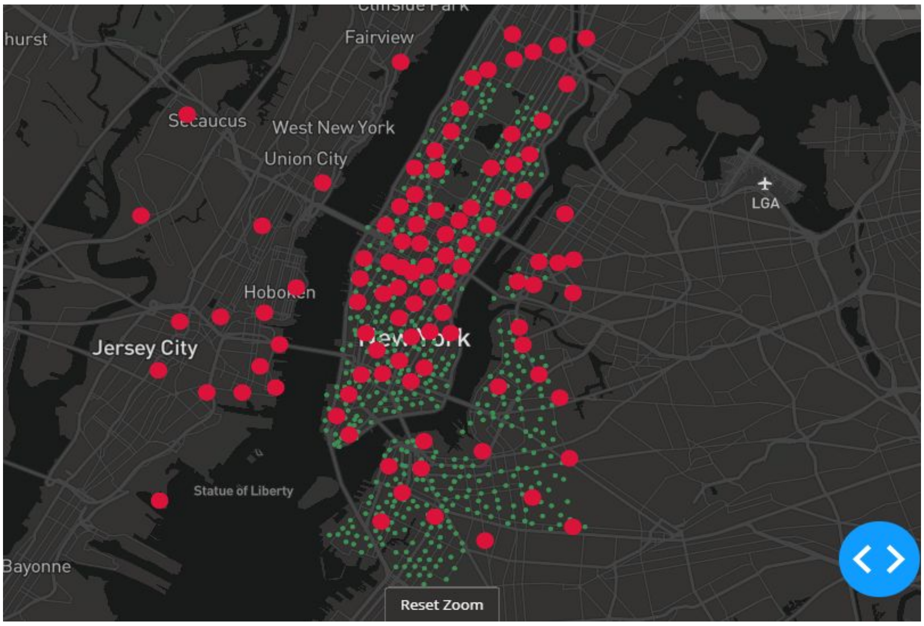
Select any of the bars on the histogram to section data by time.

(Orange data points are shortage stations, Greens are overflowed, yellow line is the N line metro)

From the daily bike movement of those shortage station, most of the bikes were borrowed from 3:00 am to 9:00 am in the morning. If those bikes are located close to the centerline of Manhattan Island, they are more likely to be refilled during 9:00 am to 12:00 pm or after 15:00 pm in the afternoon.

We tried to find some nearby stations to quickly refill those shortage ones when they are empty. However, most supply stations are more than 1~2 km far from those and some of them are not able to find any supply all day.

New Stations Suggestion: Since there are more demand than supply , we made a list of potential bike stations we can add to New York City. In addition to increase the density of bike stations in those shortage areas, we also recommend to add stations in New Jersey and Harlem. The details of the newly added stations are shown in the following map.



We learned the strengths and weaknesses of each technique. We processed the data by four layers: the first layer is to pull the data from Google Cloud BigQuery. In second layer, we process the tables by Apache Spark Dataframe. Third layer is to support last frontend layer using Python Pandas. By implementing each layer, we compared techniques we have studied in previous assignments, and finally decide to use Apache Spark Dataframe to implement second layer. However, Spark is less flexible as it passing results to frontend, so we use pandas to take the outputs of the spark and answer the queries sent from Dash By Plotly layer.

For implementation, try to avoid hard coding. For example, the address of the read file should always use relative address, otherwise it may cause error when other team members run it on their own laptop. Moreover, cache() is a good method to speedup table join. However, over-using will cause additional computation time, since spark need to re-evaluate the data evicted from the cache.

ProjectSummary:(20 points)

- Getting the data. 3
- ETL:3
- Problem: 1
- Algorithmic work: 2
- Bigness/parallelization:3
- UI: 4
- Visualization: 3
- Technologies: 4

Reference:

[1] Motivate International Inc. ,‘*Citi bike operated by motivate’*, 2016, P8.