Capital bike share network analysis
By using python

By:
Yixin Liu
EMSE 6992
December 6, 2017

Introduction

Bike share is a newly-developing travel mode, especially for the trip under 2 miles in a metropolitan like DC. The living pace in DC is fast, and the traffic is slow, riding a bike can take people to the place he or she wants in a convenient, time-saving and economical way. Fast increasing demand for bike share brings a huge market need, and the cost of operating a bike share company is affordable for a mid-size enterprise. The competition in bike share market is intense.

As a capital bike share annual member, I ride for approximately 900 miles a year, and there is a widespread issue that dock is full at the destination or there is no bike at the start point. So, that inspire me want to dig into the data find where and which time docks are under massive demand compare with at the same time, where and how many bikes and docks are in leisure; how many bikes need for repair compared with bikes on duty.

Problem Statement

For a bike share company, there are 2 key points to raise the market share. Lower the operation cost and give a good experience to current users.

Currently, the issue is there are empty docks in peak hours for users can't find a bike (start point), and there is less corral than the need for storing excess bikes in attractive docks, users can't find empty docks to return the bike (end-point).

1. Are the location of docks and corral reasonable?
2. What is the percentage of people taking a round trip?
3. Which days in which area should build more corrals for returning bikes?

There are 5 types of users, annual membership, 5-day pass, 3-day pass, 1-day pass and single trip. The charging method now is membership fee plus using time charge.

4 Compare with the daily riders, is there a significant difference in riding behavior (riding time, distance, frequency) of registered members?

Technical Specifications

Python Version: 3.6

Environment: macOS High Sierra, Jupyter notebook

Module: pandas, numpy, geopy, matplotlib

Data source

Row data:

Collect from capital bike share website, public database.

- Duration – Duration of trip
- Start Date – Includes start date and time
- End Date – Includes end date and time
- Start Station – Includes starting station name and number
- End Station – Includes ending station name and number
- Bike Number – Includes ID number of bike used for the trip
- Member Type – Indicates whether user was a "registered" member (Annual Member, 30-Day Member or Day Key Member) or a "casual" rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

First- step analysis:

For lower the operation cost:

1 Status of bike: Leisure, using/ total number of bike on duty

2 Usage rate: running bike/ total number of bike on duty (recognize by operation hours)

3 Leisure rate: Leisure bike/ total number of bike on duty at each dock

4 Dock popular level: normalize the usage rate of all docks.

For user experience evaluation:

1 Average riding distance (per day, distinguish by user type: register or causal)

2 Average riding time (per day, distinguish by user type: register or causal)

3 Average riding frequency (per day, distinguish by user type: register or causal)

Analysis

By observing the original dataset, there are 10 columns which include:

Duration	Start date	End date	Start station number	Start station	End station number	End station	Bike number	Member Type

Fig.1 feature name

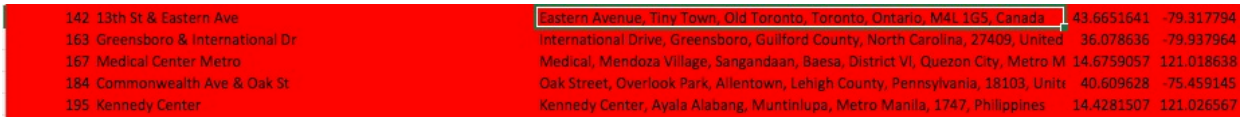
The first step is to get the geometric information of all stations to process following analysis like average riding distance, popular riding area.

After laundering the dataset, cleaning Nans value and set the order, we got 458 locations of docks. Then we use the geopy package to get the latitude and longitude of all stations.

index	address	latitude	longitude
0	19th & E Street NW	38.896015	-77.049911
1	California St & Florida Ave NW	38.917761	-77.040620
2	Jefferson Memorial	38.881406	-77.036551
3	Army Navy Dr & S Nash St	38.861744	-77.068061
4	16th & K St NW	38.902578	-77.055190

Fig.2 sample geo info

However, there are 176 addresses are located not in the area we want, apparently.



142	13th St & Eastern Ave	Eastern Avenue, Tiny Town, Old Toronto, Toronto, Ontario, M4L 1G5, Canada	43.6651641	-79.317794
163	Greensboro & International Dr	International Drive, Greensboro, Guilford County, North Carolina, 27409, United	36.078636	-79.937964
167	Medical Center Metro	Medical, Mendoza Village, Sangandaan, Baesa, District VI, Quezon City, Metro M	14.6759057	121.018638
184	Commonwealth Ave & Oak St	Oak Street, Overlook Park, Allentown, Lehigh County, Pennsylvania, 18103, Unite	40.609628	-75.459145
195	Kennedy Center	Kennedy Center, Ayala Alabang, Muntinlupa, Metro Manila, 1747, Philippines	14.4281507	121.026567

Fig.3 sample noise data

By setting a range of latitude and longitude value, these noise data are marked by red, and then we use web scrapping package combining with Latlog which is a google map API to precisely locate these errors in the area we want.

With the correct geometric location, we can generate a scatter plot which x-axis is the longitude and y-axis is the latitude. Here we can see the red line is Potomac river and docks are more concentrates in local DC area and along the river shore.

	index	address	latitude	longitude
0	0	19th & E Street NW	38.896015	-77.049911
1	1	California St & Florida Ave NW	38.917761	-77.040620
2	2	Jefferson Memorial	38.881406	-77.036551
3	3	Army Navy Dr & S Nash St	38.861744	-77.068061
4	4	16th & K St NW	38.902578	-77.055190

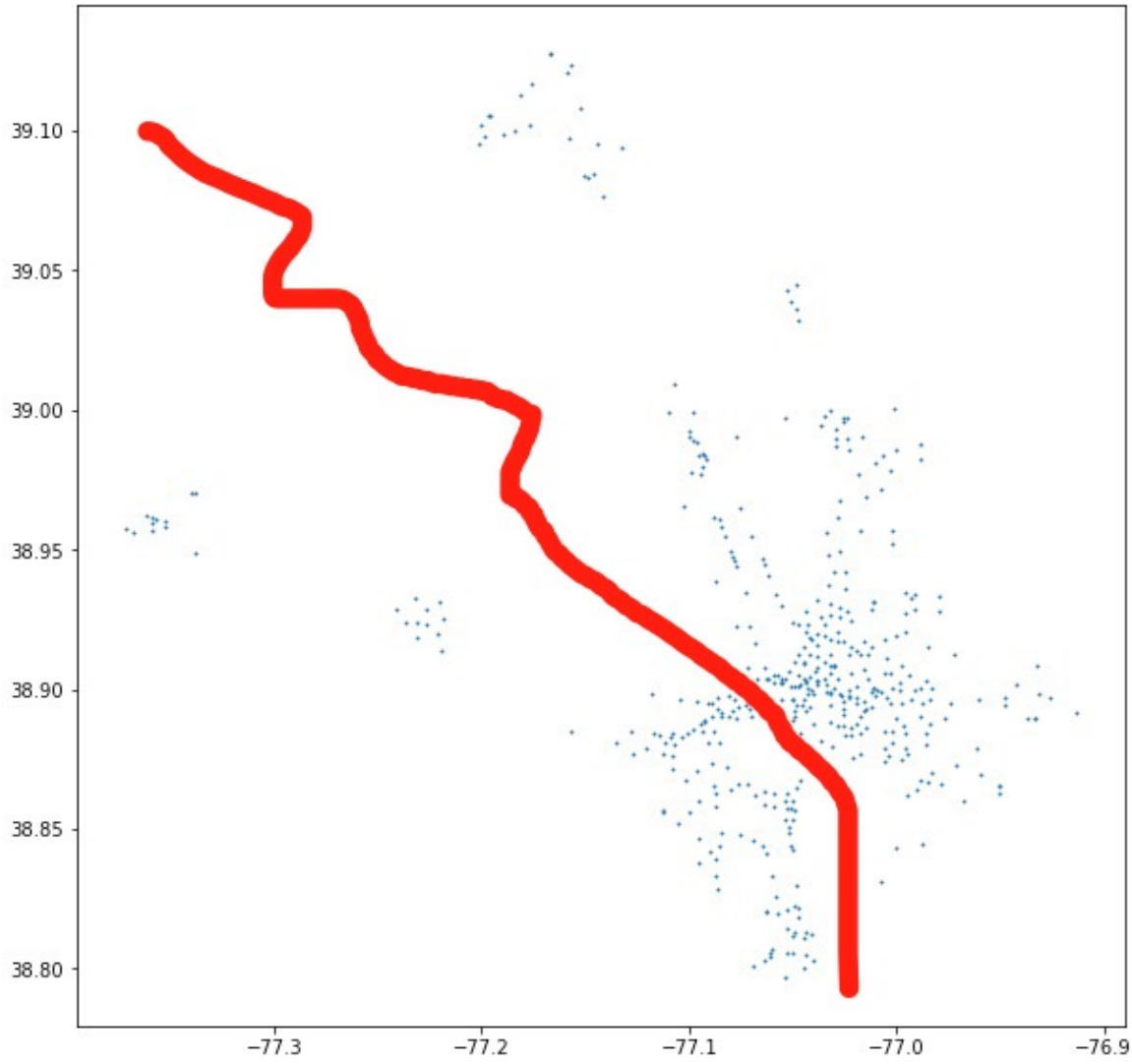


Fig.4 scatter plot of all docks

By importing the original dataset, we can use the geo information of each dock to calculate the distance between start station and end station of each trip.

But the writer faced a technical issue here with the data frame of the original dataset which is not same as the geo info set generated by the writer self. It's hard to build a for loop logic to check and import each station's geo-information from geo info dataset to each row of the original dataset without python. The geo information part has to stop at here.

For the membership diagnosis, consider of the original data is gathered within the first season of 2017. Here the writer generate 4 pie plots which are indicate the total, January, February and March.

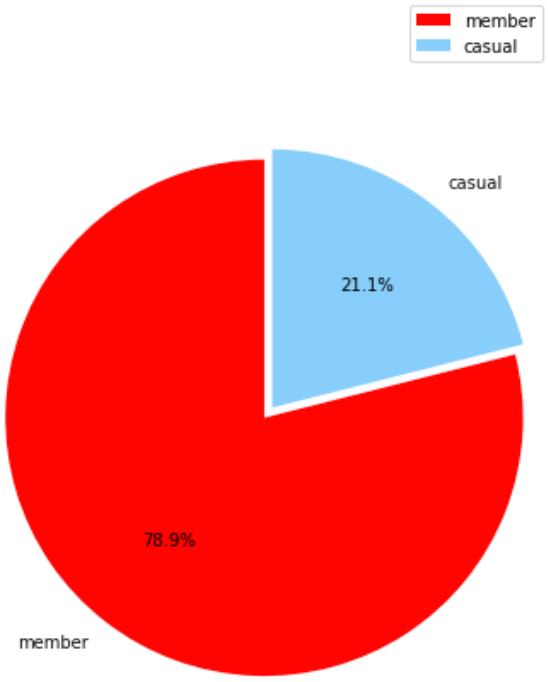


Fig.5 total membership plot

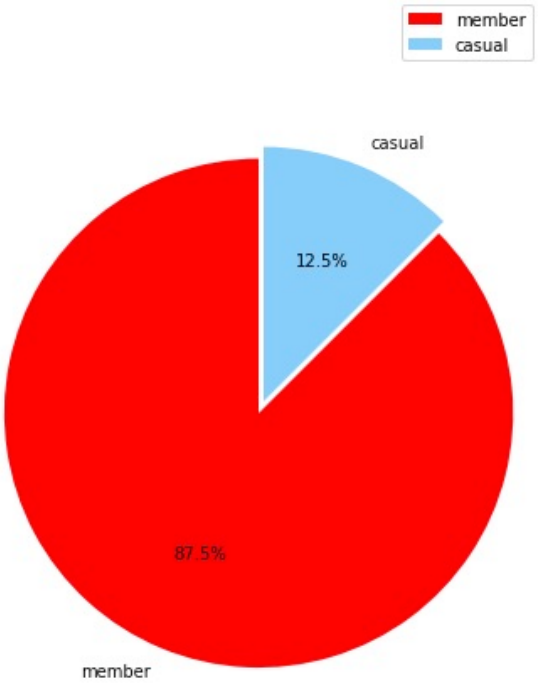


Fig.6 January membership plot

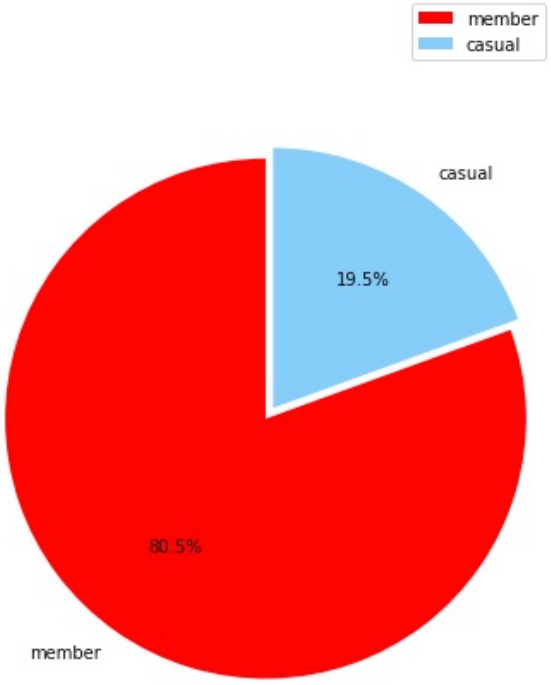


Fig.7 February membership plot

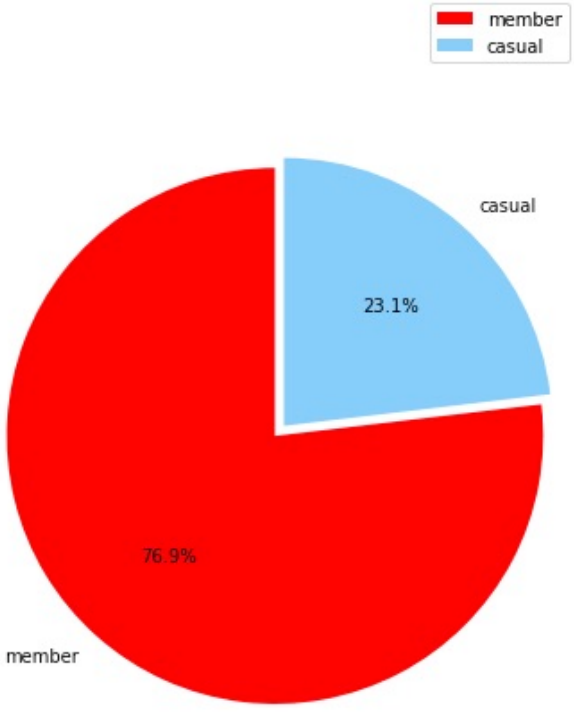


Fig.8 March membership plot

The advantage of pie plot is they can show the share of each feature we chose, here the write divide the users into 2 group. Membership user which are paid for register and causal user which are temporary users. At the last part, to evaluate which are is most popular the writer generate a bar plot to show the frequency of every dock's usage by month.

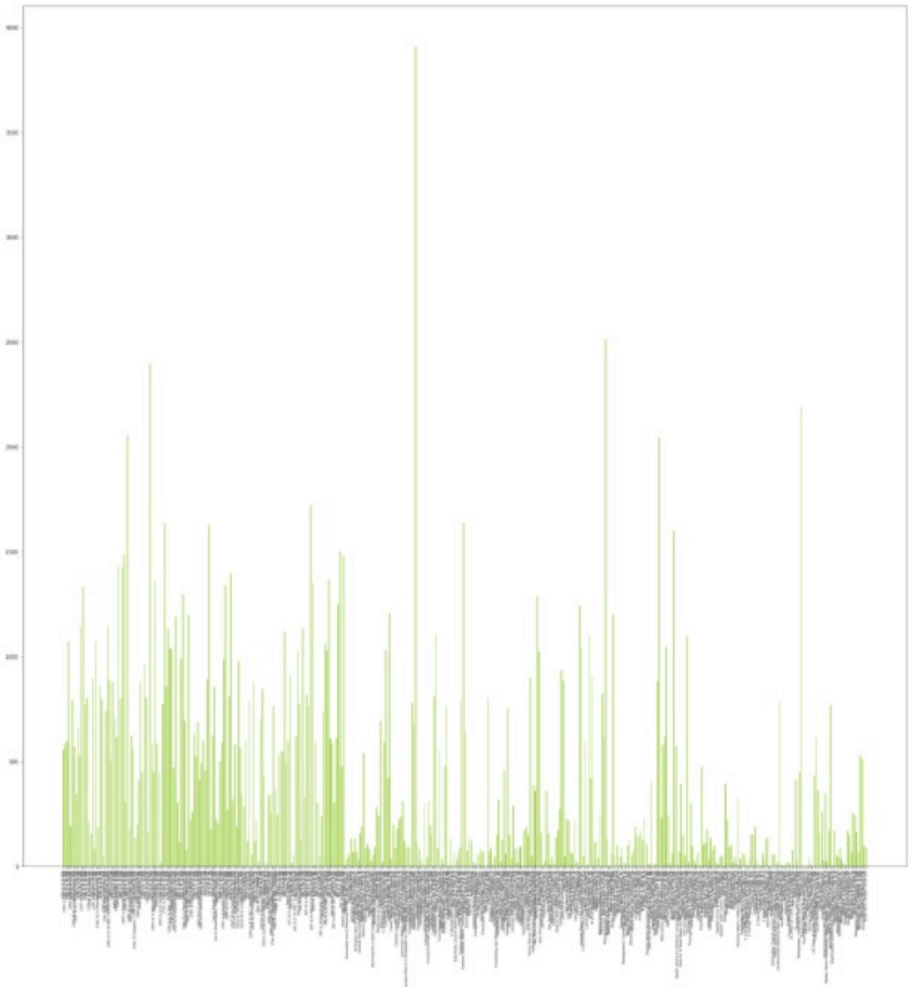


Fig.9 dock usage plot in January

Here we can see a concentrate highest usage happened in the middle which indicates docks in this area has the most usage.

Conclusion

As the scatter plot shows, combining with the map, we can get the most concentrated area of docks is located at Washington local and Bethesda, which is marked by red. Arlington and Alexandria have lots of docks, which is characterized by yellow. But docks in Arlington are sparse, and I think that is because the highway road is more concentrated at Arlington area, bikes are not the primary choice of transportation. There are several tiny concentrations of docks in Tysons, Rockville, and Ashburn, which are marked by green. These docks are giving service to the passenger of the metro. People can ride a bike to the destination after they get out from the metro station. The other area has no capital bike-share service.

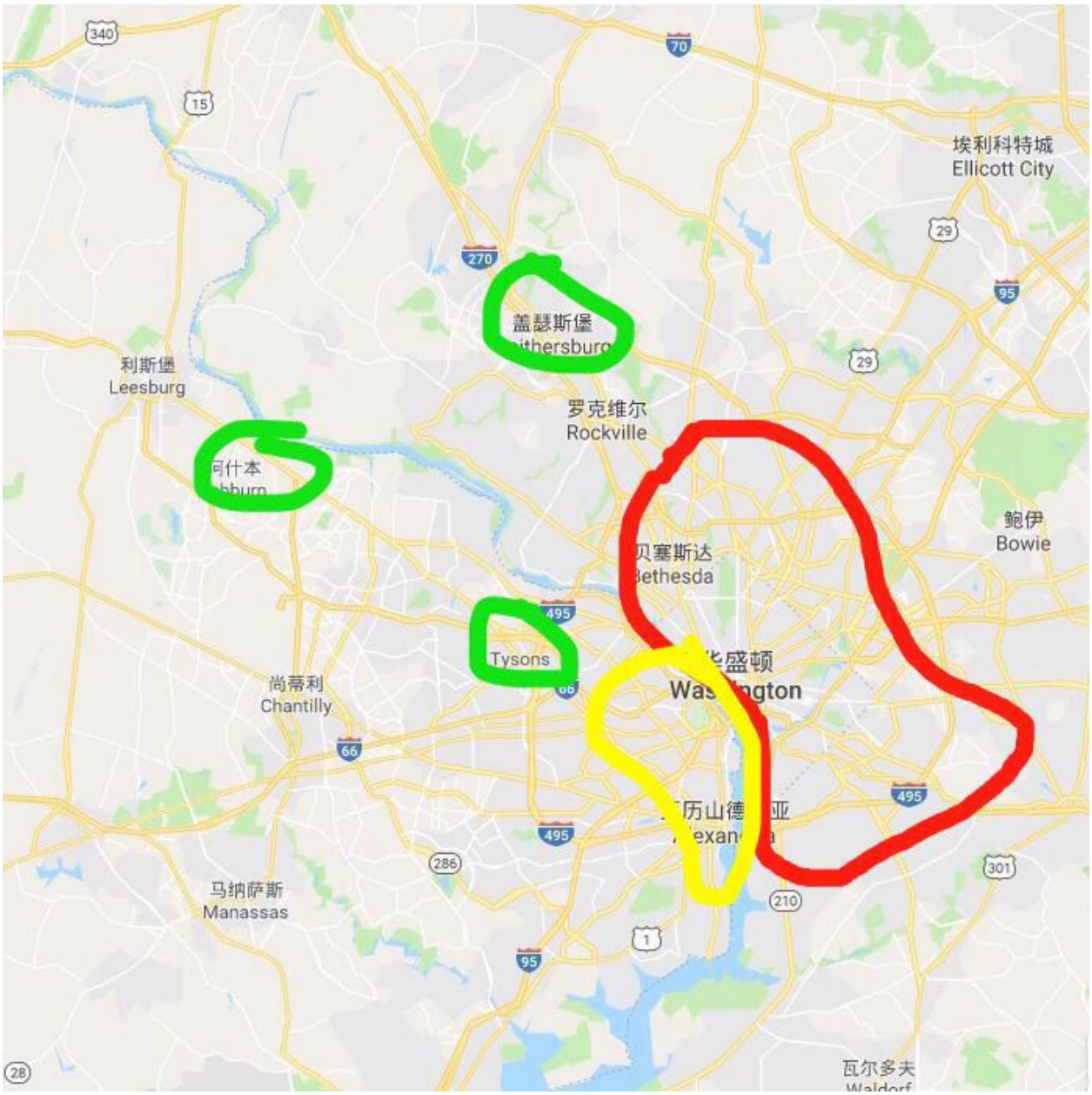


Fig. 10 dock location map

From the pie plot, we found the membership is popular in riders, but the interesting part is the percentage of membership is decreasing from January to March. I think the reason is as the temperature goes high,

more people who are not ordinary riders choose temporary using the bike, for tourism or last 1 mile trip to the destination.

The most popular area of the bike using is Georgetown waterfront to Lincoln memorial area, and I think tourist contribute a considerable volume of the bike using data.

For further analysis, I try to link the geo info with the original bike using dataset by groups. Then I can do more analysis like average riding behavior analysis in different membership, temperature.