The Feature of the Most Popular Station of Bike Share System in New York

by

Bo Sun

May/2020

Abstract. The bike system has developed a few years around the world. In big cities, the share bikes provide convenience for tourists and citizens. Recently other share systems for transport are also occur in the city, include auto share system. Thus, to investigate the features for station selection is very important. The most popular spot is around restaurants. The second feature is the gym area.

1 Introduction

In this section, the background and problems will be introduced.

1.1 Background

Bicycle-sharing system provides the self service for people with shared bike and lower or free fee in a show trip. Normally customers use them for the one-way trip or round trip. So one problem is unbalance available bikes. And bike share system is a extension of public transport sector. For instance, After transfer the subway, users of bike share system reach their destination with bike. It is very convenient that users don't need to park their bikes and don't need to concern about theft. However, users can't be easy to find a bike in busy time. The customer don't want to walk more than 100 meters. The station or dock for bike is very important.

1.2 Problem

As we have described above, customers are not willing to get a bike with more than 100 meters or 3 min walks away. For company of bike share system, they must enhance the efficiency of bike uses at best number of bikes. If they just increase the number of bikes and docks, the cost is also increasing. If the density of docks is very low, users don't walk too long to get it. So company need to find an optimal number of docks and locations. So which spot is popular for customers is a problem for company. The new spot can refer the existed docks.

2 Data Describe and Data Analysis

In this section the data source and data descriptive analysis will be introduced.

2.1 Data Source

This is data link: Trip Data of New York. The data set includes the following information:

Table 1: Bike Trip Data of New York City

Data	**Data Description**
Trip Duration	seconds
Start Time and Date	
Stop Time and Date	
Start Station Name	
End Station Name	
Station ID	
Station Lat/Long	
Bike ID	
User Type	Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member
Gender	Zero=unknown; 1=male; 2=female
Year of Birth	

The original data sets are collected from 2013 to 2020. In this project I selected the data from 01/2016 to 12/2019. Because the number of the docks and bikes are changed significantly and these data can imply the actual features of docks spot selection.

2.2 Data Preprocessing

Three steps are implemented in this processing section. After data processing, all data sets merge in one larger data set.
Data is downloaded from web site. The trip duration and station name, id has missing data. Without this value we can't research the problem efficient. And these columns are not good to use the average value methods to fill. So for these rows with missing trip duration, station id and extreme data will delete from the data sets. According to the data web site, data with the trip duration is shorter than 60 seconds can be the test data from company or the customer check if the bike is locked or not. so these data will delete from data.The extreme large trip duration over 6 days is dropped in the data.

Because the majority of extreme data is customer data. This may be they forgot to lock bikes properly.
The second step is changing the type of data. The data type of start station id and end station id, gender, user type, year of birth is float or int, I convert these as string. The type of start time and end time are different in different year. The type of start time and end time is string, I convert it to date type.
The last step is to calculate the year of data, the month of data, and weekday of day. For instance, 2016-01-01 is at year 2016, the 1st. month ,the 4st day in a week. I will just use the start time as the data time, for the trip duration is longer one day, I just use the first time. So add three columns year, month and weekday and drop columns start time and end time.
After the three steps in this processing section, there are 68300047 rows in the data set.

2.3 Data Analysis

The number of bike stations increased from 654 to 938 in 4 years in New York.
The change of trip duration and the number of using are corresponding to human habits and weather. In the summer time ,the trip duration reaches the peak, about 2 hours and in the winter time the average trip duration is less than 1 hours. Figure 1 shows this result. Figure 2 shows The use of bike is shown a season cycle. The number of count is doubled from 2016 to 2019.

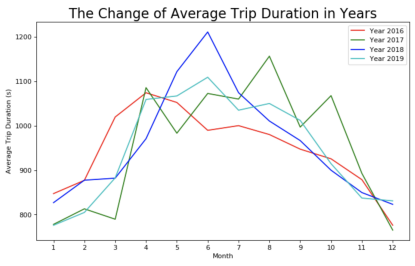

Figure 1: Trip Duration

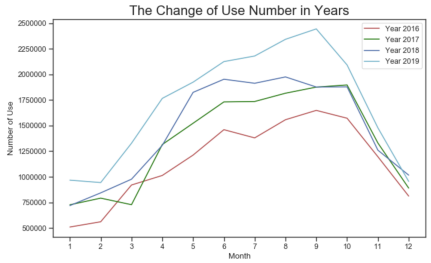

Figure 2: The number of Use

For gender, the unknown user has a significant higher trip duration. Female clients use the bike longer than male. Annual Member show no significant change in a week. However, the result of short time customer show that the user in type of customer drive a bike more in the weekend. This seems to be tourists or for sightseeing. Figure 4 shows the result.

Figure 3: The violin plot of user type in a week

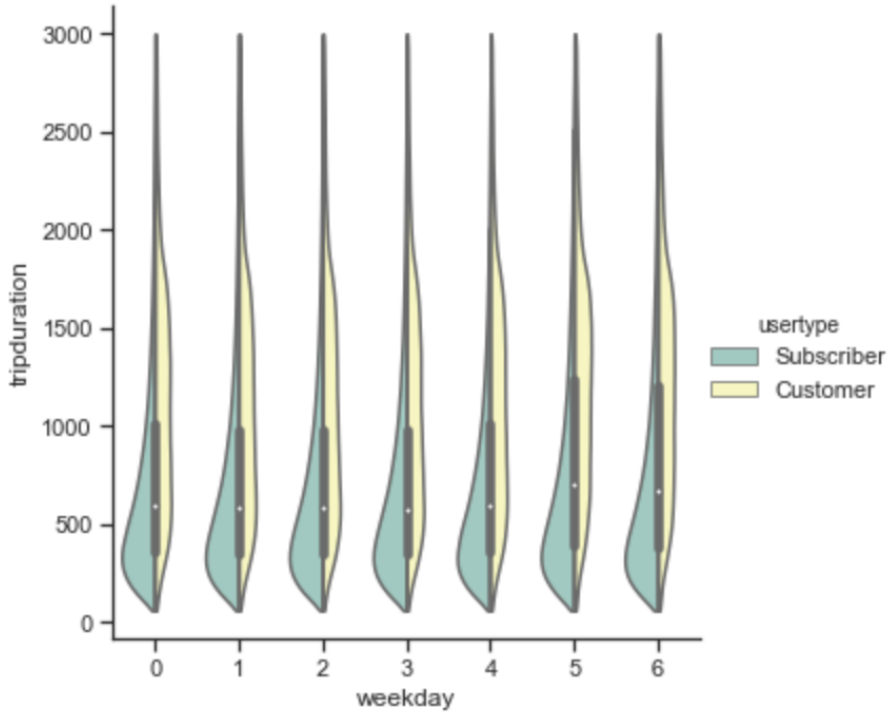

Figure 4: The density of docks in New York

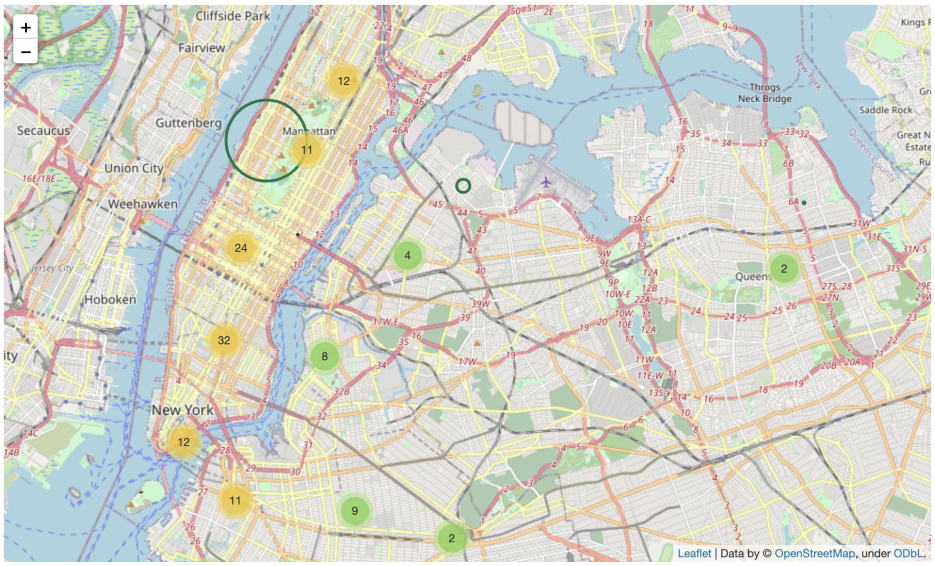

Figure 4 shows the density of docks in New York city. In Manhattan there

are the larger number of docks. Pershing Square North is the most popular start station. Around this dock there are time square, central park and Giant central station. The top 10 trip is shown in Table 2. Around the first trip is restaurants and bars. The second trip is central park sightseeing trip.

Table 2: The Top 10 Trip

	Start Station Name	**End Station Name**
0	E 7 St & Avenue A	Cooper Square & Astor Pl
1	Central Park S & 6 Ave	Central Park S & 6 Ave
2	Central Park S & 6 Ave	5 Ave & E 88 St
3	North Moore St & Greenwich St	Vesey Pl & River Terrace
4	West Drive & Prospect Park West	West Drive & Prospect Park West
5	Vesey Pl & River Terrace	North Moore St & Greenwich St
6	12 Ave & W 40 St	West St & Chambers St
7	Pershing Square North	E 24 St & Park Ave S
8	McGuinness Blvd & Eagle St	Vernon Blvd & 50 Ave
9	Soissons Landing	Soissons Landing

3 Modelling

The bike station should be located nearby house or near a metro station? Which features should docks have is the main problem for selection docks. To investigate the features around bike docks, we can solve this problem using k-means.

The k-means algorithm searches for a pre-determined number of clusters within an unlabeled multidimensional data set. It accomplishes this using a simple conception of what the optimal clustering looks like.

In this section I use cluster method to solve the problem. Using foursquare to get the venues with 200 meter of top 40 docks in Manhattan, we can know which character is important for different group of docks.

So I select the top 40 docks in the Manhattan. Because Manhattan is a popular area and a larger number of docks in New York. Figure 5 show the 4 cluster. The most important feature is around the restaurants and coffer shop. Nearby the food shop the bikes can be used efficient. For cluster 1, the gym is an feature. The public transport is important for this type of customer.

Figure 5: The Segment Analysis

4 Conclusion

The sharing-bike is used seasonally. Although the increase of station is just about 150, the number of bike is doubled in 2019. And The food area is the spot the potential customer come. Other areas like bus, mertro line are also important. Residential area does not show a must for docks. The result could be the efficiency of bike use. In this area citizens could be only use it in the morning or afternoon. One problem of this cluster is imbalance.