Data Mining and Visualization Analysis of shared bikes ——In the Case of Citi bike

Can Yang[1,a,*], Xuemei Li[2,b]

[1]School of Economics and Management, Beijing Jiaotong University, China

[2]School of Economics and Management, Beijing Jiaotong University, China

[a]17120628@bjtu.edu.cn, [b]xmli@bjtu.edu.cn

Keywords: Shared bikes, Data mining, Cycling data, Visualization.

Abstract. As a new kind of sharing economy, shared bikes solve the last kilometer problem of people's travel. This paper analyzes the riding characteristics of shared bike users in New York by using the massive data of shared bikes. Based on the cycling data of Citi Bike, the R language is used to extract, clean and analyze the data, and the analysis results are presented in the form of visualized graphs through data visualization technology. Research shows that the male ratio of Citi bike is high, and users are mainly between 20 and 50 years old; Cycling characteristics are characterized by large differences between workdays and rest days, with morning and evening peaks, and a short period of 0 to 10 minutes; In addition, there is a phenomenon of agglomeration in cycling sites, mostly in areas with high traffic such as railway stations, bus stations and scenic spots.

1. Introduction

With the rapid development of the sharing economy, shared bikes have risen rapidly in the field of travel. Shared bikes are in line with the concept of low-carbon travel, which can effectively solve road congestion and environmental pollution problems, reduce the use of private cars and public transportation pressure. Moreover, the cost of riding a shared bike is relatively low, so it has received widespread attention from the public and has achieved rapid development. To a certain extent, the emergence of shared bikes has alleviated the traffic pressure of the city. However, due to the lack of a sound management and supervision system, there are many problems after the launch, such as parking confusion and unreasonable vehicle scheduling.

In the fast-developing information society, the explosive growth of data has gradually become the norm. The era of big data has arrived, and big data has become the focus of attention in various industries after cloud computing and the Internet of Things. In fact, the development of shared bikes is also inseparable from the application of big data, such as positioning data. At the same time, the massive riding data generated by the shared bikes can plan the best path for vehicle operation, which promotes the development of shared bikes in a more standardized direction.

2. Literature references

There are many researches on shared bikes at home and abroad. The research focuses on the following three aspects: The first is the theoretical study, Y. Wang discussed the development status, problems and countermeasures of shared bicycles [1], H. Q. Lai and D. Sheng used the OFO shared bike as an example to explore the business model of shared economic enterprises [2]; The second is the study on time-space distribution and scheduling optimization, some researchers used the BP neural network to predict the initial amount of shared bikes, established an integer programming model and solved the expected number of dispatches [3], Q. M. Tao carried out cluster analysis on the areas of pick-up and returning peak hours, and divided the areas into three categories [4]; The third is the study of travel decision, some studies used the form of questionnaire to establish a LOGIT model for shared bike travel, and analyzed the riding characteristics by means of SPSS tool [5]. Among them, there are few studies on the riding characteristics of shared bicycle, and most of them are based on

Copyright © 2018, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

questionnaire data rather than massive riding data for analysis. Therefore, the contribution of this paper is to mine data on massive cycling data, find out the riding characteristics of shared bike users and provide decision support for enterprise scheduling management.

3. Data source and processing

The research object of this paper is Citi bike in New York City. In 2013, Citi bike was officially launched. In the first phase of the operation, a total of 6,000 units were put into operation. After 10 months of operation, Citi bike accumulated more than 6 million orders and 400,000 members, making it the largest shared bicycle project in the United States.

3.1 Data source

The research data is mainly derived from the open data provided by the Citi bike shared cycling platform, which covers the use of Citi bike shared bike users in New York City from May 1, 2016 to May 31, 2016, including a total of 19,488 data.

Table 1. Variable definitions.

Serial number	Name
1	Continuous riding time
2	Start time
3	End time
4	Start station ID
5	Start station name
6	Start station latitude and longitude
7	End station ID
8	End station name
9	End station latitude and longitude
10	User ID
11	User type
12	Date of birth
13	Gender

3.2 Data preprocessing

Data preprocessing refers to the necessary processing such as reviewing, screening, and sorting before collecting or grouping collected data. Data preprocessing methods include data cleaning, data integration and data conversion. These data processing techniques improve the quality of data mining and reduce the time required for actual mining.

When performing data preprocessing, the R language built-in algorithm is used to clean the data, delete the duplicate data, and process the data with missing values. At the same time, after testing, it is found that the abnormal riding time contains abnormal values, which will seriously affect the effect of subsequent data mining, so delete the records with abnormal values. After data preprocessing, 17,646 complete and reliable riding data were obtained.

3.3 Data mining

Data mining generally refers to the process of searching for information hidden in it from massive data by algorithm. The main task of data mining is to extract and analyze the Citi bike riding data, and to mine the key features in the riding data. The methods used include statistical data analysis methods, complex network analysis methods, data classification and mining methods. Firstly, perform statistical analysis on the user characteristics of Citi bike, such as user gender, age structure. Secondly, the riding characteristics of Citi bike are mined and analyzed from the perspective of time and space, such as riding time, riding concentration time and riding concentrated area. Finally, this paper provides vehicle scheduling optimization recommendations for the acquired user characteristics and riding characteristics.

4. Data visualization analysis

4.1 User feature analysis

Firstly, analyze the male-female ratio of Citi bike shared bike users in May 2016. As can be seen from Figure 1, in May 2016, Citi bike shared more male users than female users, with 13,896, accounting for 79%, and fewer female users, with 3,750, accounting for 21%.

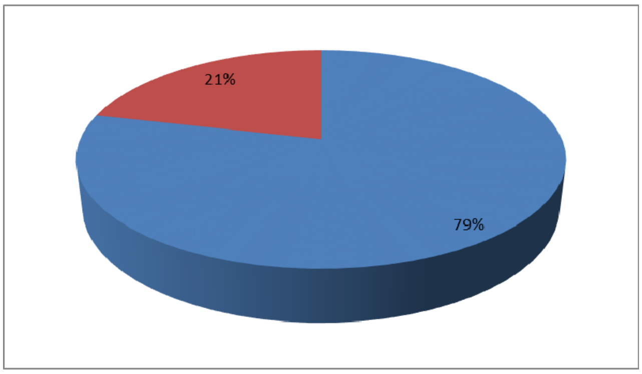

Fig. 1. Ratio of male to female Citi bike users.

Secondly, an analysis of the age structure of Citi bike users in May 2016 is shown in Figure 3-2. As can be seen from Figure 2, in May 2016, among the Citi bike users, there were 8,028 users aged 31-40, the highest proportion, followed by 41-50 years old users. And there are 27 users aged 71-80, which account for the lowest proportion. Overall, users of Citi bike are mainly concentrated in the 20-50 age group.

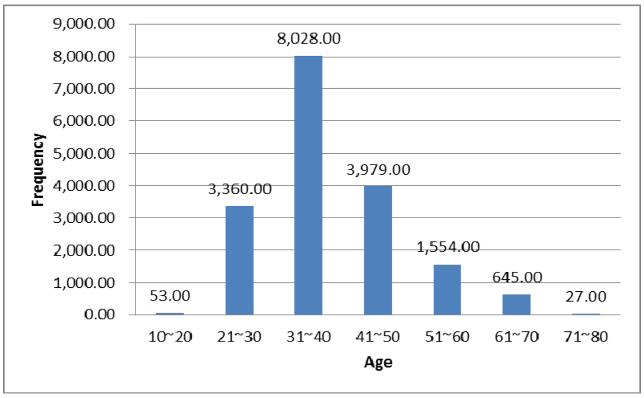

Fig. 2. Age Structure of Citi bike Users.

4.2 Cycling characteristics analysis

This paper mainly analyzes the data mining characteristics of users using Citi bike in May 2016 from the perspective of time and space.

4.2.1 Time characteristics of riding behavior

(1) Weekly characteristics of riding behavior. Because bicycle travel is greatly affected by the weather, in order to reduce the impact of weather conditions on the research results, we study the weekly characteristics of the riding behavior in the week when the weather conditions are mainly sunny and the air condition is excellent. The specific date is from May 9th (Monday) to May 15th (Sunday). The frequency of Citi bike users is counted every 3 hours, and the time distribution curve of the user's riding demand in one week is obtained.

As can be seen from Figure 3, the usage of Citi bike is significantly different on weekdays and rest days. The working day contains a large number of commuting riding behaviors, which form the phenomenon of early peak riding, in which the early peak formation has a short duration and the number of riding surges; The evening peaks contain a large number of life riding behaviors, which are characterized by a slower growth in usage frequency and a slower demand; At noon, the lunch behavior of some users triggered a certain close-range riding behavior. There are many life riding behaviors on the rest day, the riding time is relatively balanced, and there is no obvious early peak phenomenon. The casual riding behavior of the user in the evening formed the peak of the frequency of use of Citi bike. It can be seen from the time distribution of the user's riding behavior that the maximum demand for Citi bike is determined by the frequency of use of the morning and evening peak hours of the working day.

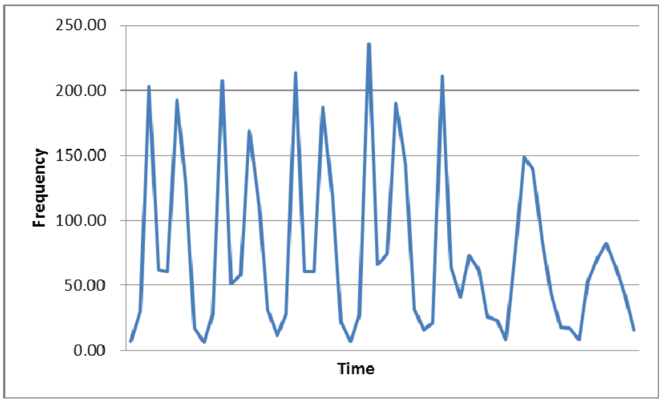

Fig. 3. Citi bike usage frequency from May 9th to May 15th.

(2) Daily characteristics of riding behavior. In order to more accurately describe the peak demand of cycling in New York City residents, determine the amount of shared bicycles, and study the selection of the full-time analysis on May 12th with high riding demand. As can be seen from Figure 4, the frequency of users using Citi bike is lower from 0:00 to 6:00 in the morning, the maximum value is no more than 20 times, and users who use Citi bike start to increase at 7:00 in the morning. The peak of the morning peak appeared around 8:00, and the peak of the evening peak appeared around 18:00. Therefore, this study determines 7:00-9:00 as the early peak period, and 17:00-19:00 as the late peak period. The riding in these two periods is mostly commuting riding behavior. From 9:00 to 16:00, the frequency of use of Citi bike has been moderated, the demand has not changed much, and there has not been a surge or a sharp drop. The riding during this period is mostly a life riding behavior. The frequency of shared bicycles at 13:00 has increased slightly, but it is caused by the noon eating behavior, and it is often expressed as short-distance riding. The frequency of use from 20:00 to 24:00 is small and shows a gradual decline. In summary, the morning and evening peak time periods are important test periods for the supply and demand pressure of the riding facilities. The amount of Citi bike should be increased to ease the morning peak commuting pressure.

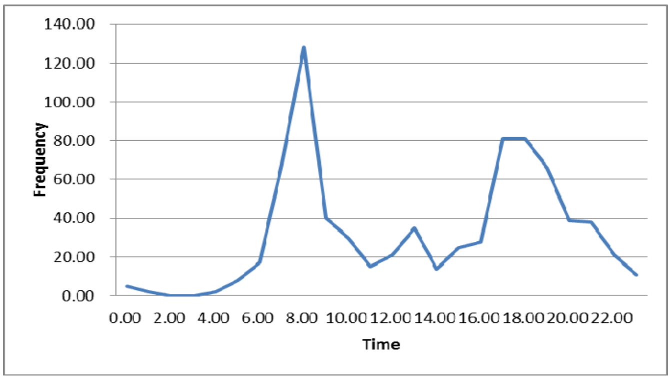

Fig. 4. Citi bike usage frequency on May 12th.

(3) Characteristics of riding duration. As shown in Figure 5, more than 90% of the rides are controlled within one hour. The duration of the Citi bike is concentrated in 0-10 minutes, mostly for short-distance riding, which is consistent with the characteristics of shared bikes.

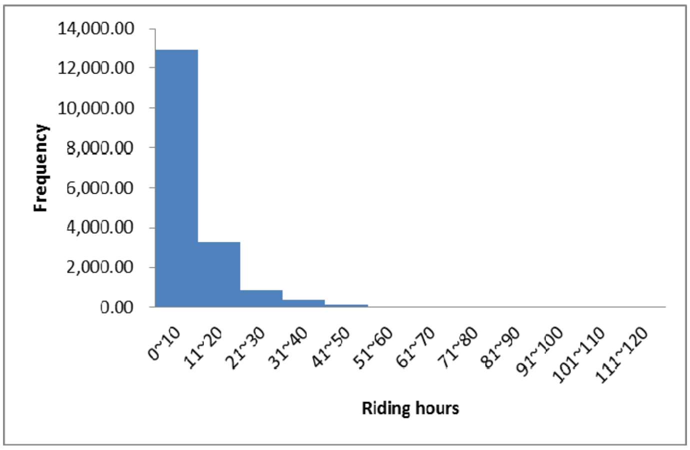

Fig. 5. User's cycling duration in May 2016.

4.2.2 Spatial characteristics of riding behavior

The spatial distribution of cycling requirements is an important basis for the planning of shared bike parking facilities. This paper analyzes the data from the starting point of the user's riding in May 2016, and obtains five areas where the starting point is densely packed, as shown in Table 2. At the same time, get the Citi bike appearance frequency of all starting points.

As you can see from Figure 6, the five stations that Citi bike shares the most in bicycle demand are Grove St PATH, Sip Ave, Exchange Place, Hamilton Park, and Newport PATH. Among them, Grove St PATH and Newport PATH are train stations, Exchange Place is a bus stop, and Hamilton Park is a famous park attraction in New York City. These locations are characterized by high traffic, so there are more users who use Citi bike to share bicycles. In summary, there will be a gathering area around the railway station, bus station, and scenic spots. The capacity allocation of shared bicycle parking facilities around these stations is also the focus of planning. Therefore, Citi bike enterprises should increase the amount of shared bicycles in these areas where aggregation occurs, and rationally dispatch vehicles in the morning and evening rush hours to ensure that users can have vehicles available at anytime and anywhere, improve service quality, and thus enhance the core competitiveness of enterprises.

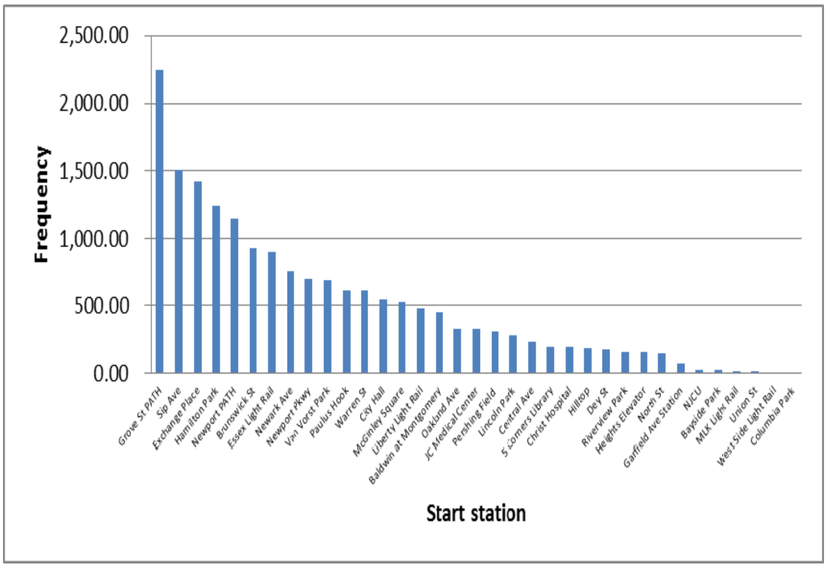

Fig. 6. Occurrence frequency of Citi bike start station in May 2016.

Table 2. Five Riding Starts with the Most Frequent Frequency.

Serial number	Start station Name	Frequency
1	Grove St PATH	2241
2	Sip Ave	1502
3	Exchange Place	1417
4	Hamilton Park	1241

5	Newport PATH	1145

5. Summary

Studies have shown that users who use Citi bike are mainly male, and mainly young and middle-aged people between the ages of 20 and 50. From the time point of view, the riding behavior is characterized by a large difference between working days and rest days in a week. There are high peaks in the morning and evening in one day, from 7:00 to 9:00 in the morning peak period, and from 17:00 to 19:00 in the evening peak period. The duration is concentrated in 0 to 10 minutes. From the perspective of space, the riding area is concentrated in the crowded areas such as railway stations, bus stations, and scenic spots. Enterprises can rationally dispatch and manage vehicles according to user characteristics and riding characteristics to improve service quality. The shortcoming of this paper is that the data has limitations, which is not enough to accurately reflect the use characteristics of Citi bike. In the future, more data will be selected for further study.

Acknowledgement

I am very grateful to my mentor, Miss Li has given me a lot of encouragement and guidance on the topic selection and writing.

References

[1] Y. Wang, Development status, profit model and main issues of shared bicycles, *Times Finance*, vol.30, pp. 249-250, 2017.

[2] H. Q. Lai and D. Sheng, Analysis of the business model in the era of sharing economy: Taking the example of OFO sharing bicycles, *Chinese Business Theory*, vol.11, pp. 3-4, 2017.

[3] L. Y. Zhou, X. J. Chang, and B. Tian, Optimization of shared bicycle dispatching based on BP neural network, *China Strategic Emerging Industry*, vol.24, pp. 51, 2017.

[4] Q. M. Tao, Time-space distribution model of shared bicycles based on cluster analysis, *China Strategic Emerging Industry*, vol.4, pp. 61, 2018.

[5] L. N. Qi and F. Li, Analysis of trip bike characteristics and travel behavior of shared bicycles, *Traffic Information and Security*, vol.35, pp. 93-100+114, 2017.