Citi-Bike Trip Analysis Visualization

Sankalp Gupta

17302431
guptasa@tcd.ie

Abstract

This document provides insights into the CitiBike bike sharing systems through visualization of the data. The aspects of the visualization and the idioms used have been explained in the report. Section 1 provides an introduction on the data used for visualization. Section 2 provides a background of the task performed and how it helps in answering questions and provide insights related to bike sharing systems. Section 3 discusses about the approach, the idioms and the interaction operators used in the visualization. The last section lists the references.

1. Introduction

Many cities around the world have started adopting bike sharing systems. However, they are facing a lot of issues regarding redistribution of bikes, over-crowding of bikes and unavailability of bikes at designated bike-sharing stations. I have used the dataset of Citibike bike sharing system which are operating in the New York city. It is an open dataset publicly available on Citibike website for analysis, development and visualization. The data which I have visualized is a recent dataset having trip details of December,2017. The below table shows the features used in the dataset for visualization [cit18]:

Table 1: *CitiBike Dataset Fields*

Features Used	
tripduration	end station id
starttime	end station name
stoptime	end station latitude
start station id	end station longitude
start station name	bikeid
start station latitude	birth year
start station longitude	gender

2. Task

The task is to visualize the dataset in such a way that it gives meaningful insights of the bike-sharing systems. The number of trips for every hour of the day for a period of one month is visualized. This enables us to understand and visualize the user riding pattern, visualize the peak and non-peak hours of the bike usage.

The station coordinates are mapped on the OpenStreetMap view visualizing the original locations on the world map. There are total of 752 distinct stations from where the bikes can be picked and parked after use. It is important to know the specific stations having more traffic and stations having near to zero or very less trip rides. This analysis can help in better strategizing and positioning of new stations and also removal of stations which are not required. The same applies for the end stations, the information of number of trips finishing at an end station will help us analyze the parking availability of the bike. Also, it is beneficial to visualize the end stations where the trip ends. This makes us understand the most frequent route taken and possible end stations from a particular start station. Using the demographic data, we can visualize the ratio of male and female users and the age range of the current bike users. This data can be used by bike-sharing companies to focus on a particular user segment to increase the number of users.

The following questions can be answered through this visualization

1. Stations having Maximum No of Start trips.
2. Stations having Maximum No of End Trips.
3. Time or days of the month, the most number of trips are taken.
4. Total number of trips taken for a given range of start date and end date.
5. Which is the most frequently travelled route.
6. What is the ratio of male and female users using bike sharing systems.

The answer to the following questions can help improve the redistribution of shared bikes. It can help the bike vendors better strategize the number of bikes required at each station and improve the services provided. For example, in which areas, locations and at what period, the more number of trips are taken on. The increase in user base and the number of trips over several months can be studied to help in forecasting of user usage.

2.1. Results

There were 8,89,967 number of trips of Citibikes taken during the month of December,2017 in the city of New York. There were 11,119 bikes used which were placed at 752 stations around the city of New York. It can be observed that most of the trips were taken between 08:00 AM - 10:00 AM and 17:00 PM - 19:00 PM. It can also be observed that there were less number of rides during the last week of December probably due to Christmas holiday week. Pershing Square North was the most busiest station with 9,801 trips starting and 9,773 trips terminating. The route most frequently used was from Pershing Square North to Broadway and W 32 Street (323 trips). Males are the dominant users (6,51,548) of the service, approximately 3 times than that of the female riders (1,94,802).

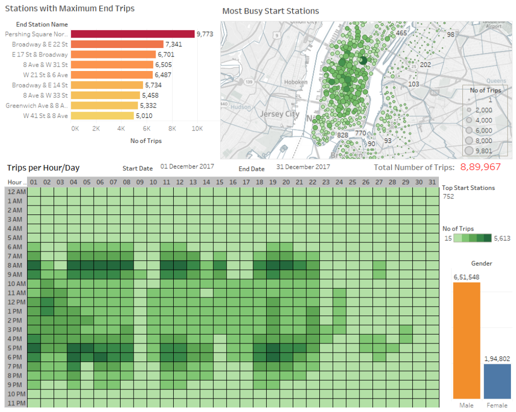

Figure 1: *CitiBike Visualization*

Further in detail results for a particular hour, or a date range or for few top stations can be obtained through various filters, interactions and selection methods incorporated in the visualization.

3. Approach

I am doing my dissertation on CitiBike sharing systems and this was one of the reasons for choosing this dataset. I have used Tableau tool [Tab18] to visualize this dataset. I was always interested to learn Tableau as it is widely used tool in the industry for interactive data visualization and reporting.

Initially, a derived variable Number of Trips was defined by calculating the count of Starttime. The Starttime (Day) was plotted against Starttime (Hour) with the above derived variable Number of Trips as a colour heat-map encoder. Darker the shade, greater the number of trips taken. A user control filtering method is provided to mention the start date and the end date of the trip which reduces the size and the content of the data according to the trips taken between that range period. [Mun14] The user can also select a particular region of interest i.e. on selecting a specific hour and day of the month in the visualization, the other parameters and layouts gets interactively modified according to the selection. On deselecting, it returns to its normal state.

A map is visualized with actual locations of the bike sharing stations. The user can navigate, zoom in zoom out, pan for appropriate scaling and viewing. The user has the freedom to visualize the top busiest start stations by accessing the user control slider for Top Start Stations. The number of trips for the selected stations gets calculated dynamically and also the list of top 9 end stations gets updated at the same time. Also, the heat-map of the trip timings gets revised as only the trips from the start stations will be considered on this selection. No of Trips are color encoded as well as with the size for improved understanding. Darker and bigger the size of the circle, greater is the number of trips for that particular station.

On selecting a particular start station, the user can view the top 9 end station names and the number of end trips associated with it. If a user wants to see all the start stations associated with the trips of a particular end station, the user can select that particular end station and all the start stations on the map gets updated dynamically.

Gender Analysis is given through a bar chart wherein the Male and Female are color encoded to differentiate them. As all the sheets are linked and connected together, the data gets dynamically filtered and updated accordingly giving better insights for each and every possible scenario.

4. Conclusion

From the above results, it can be said that visualizing the data provides a better sense and more information of the data. Visualization helps in easy comprehension of the data without which data is difficult to comprehend. The use of various interactive operands, filters, animation, dynamic user controls, idioms makes visualization come to life and provide hidden and necessary business insights.

4.1. Improvements Possible

1. Animation of the path travelled by each bike from start to end station.
2. The current visualization incorporates trips for a single month. The analysis and user riding pattern in different months can be studied if the dataset for entire year or various years is considered. The change and increase in number of trips and users along the years can be animated.
3. Weather data could have been combined with the current trip dataset to analyze the impact of the weather on Bike ridership.

References

[cit18] *Citibike Website*. 2018. URL: https://www.citibikenyc.com/system-data.

[Mun14] MUNZNER T.: *Visualization analysis and design*. CRC press, 2014.

[Tab18] *Tableau*. 2018. URL: https://www.tableau.com/.