John Jostes

6/18/20

Tableau Homework - Citi Bike Data Analysis

Data Exploration

I began this assignment by deciding which subset of the Citi Bike data I wanted to analyze. For the sake of reducing the file size of my aggregated data, I stuck to the winter months of December 2018 - February 2019 and summer months of June 2019 - August 2019. These files were concatenated into a single database using pandas. The columns ‘birth year’ and ‘gender’ were dropped, as I had no intention of looking for trends relating to that data, and to further reduce the size of my output file.

The final database had 9.6 million entries, and was saved as a csv file. I felt that for the scope of this assignment that provided a large enough sampling of the data, but small enough to be easier to import/export.

Dashboard 1 - Map Request

A condition of the assignment was to provide a map detailing the requested data from a ‘city official.’ I chose to create a dynamic map showing the popularity of both start stations and end stations for the selected winter and summer months. Using the toolbox provided by Tableau’s Pages Shelf, one can select for the given month. The need to select by year was also stipulated, but, seeing as only 1 of the 6 months is from a different year, I chose to leave it to only the month selection.

To show popularity, station locations are marked with circles that increase in size and decrease in color value (appearing darker). These, and subsequent maps, use the count of stations within the data set as its Measure. I.e. the more a station is associated with unique bike trips by users, the more that station is in use. Within this dashboard, static lists of the top start/end stations across all 6 months are also presented.

As one would expect, user activity increases from winter to summer months, while the overall distribution of station popularity remains the same, even between start and end stations. Those that are popular tend to correspond to major NYC landmarks and points of mass transit, but could also be thought of as nodes that most efficiently reduces the amount of walking to a destination for the greater number of destinations. Sort of like cow paths in a university quad.

Stations and corresponding landmark and/or transit (italics):

Pershing Square North - ***Grand Central Station***

E 17th St & Broadway - **Union Square Greenmarket**

8th Ave & W 31st St - **Madison Square Garden; *34th-Penn Station* (a few blocks north)**

Broadway & E 22nd - **Flatiron Building**

Broadway & E 14th - ***14 Street-Union Square Station***

West St & Chambers St - **possibly Rockefeller Park; One World Trade Center (a few blocks south)**

Broadway & W 60th St - **Central Park; Columbus Circle**

Several other stations not in the top 10 border Central Park

W 21st St & 6th Ave - could not easily determine

12th Ave & W 40th St - ***Lincoln Tunnel***

Christopher St & Greenwich St - could not easily determine. Possible commercial area

Dashboard 2 - Summer Weekday Traffic For Subscribers

I wanted to look more narrowly at potential phenomena relating to station popularity, which might better explain what was seen in the previous dashboard. For this one, I filtered the data to include only summer months(better weather), weekdays (commuting to work), trips taken by subscribers, and the popularity (station count) for each hour in a day across all 3 months. I looked at subscribers specifically on the assumption that they would be those users biking to and from work; the annual subscription being more cost effective than the passes for customers.

Two general trends I noticed are (A) the inverse popularity of stations during the morning and evening commutes, and (B) the locations of where these trips begin and end. For this I will refer to the Station ID.

In the morning hours, starting stations 3255, 358, 514 and 519 “light up,” along with numerous smaller stations in the Lower East Side and East Village neighborhoods. The major nodes correspond to 34th-Penn Station, Grand Central Station, and Lincoln Tunnel. I interpret this as long distance commuters getting off mass transit and finishing the remainder on bike. As for the neighborhoods, these are also commuters, but with a short enough distance to comfortably get there by bike.

Ending stations group around the major commercial areas of Lower Manhattan and Downtown, which include: 3664, 426, 402, 359 and 281. There is also a lack of end station activity in the aforementioned east Manhattan neighborhoods, as that is highly residential.

In the evening, the opposite is true, and one can imagine the flow of commuters leaving their places of work and returning to either their home, or the next exchange in their commute. Throughout the day the map appears quieter - I would consider this due to those leaving work to get lunch only going within walking distance. One thing not taken into account is the evening behavior of users - e.g. whether they are biking to social events, gyms, restaurants, etc.

Possible Uses of Dashboard 2

Advertising could be targeted by both location and time, if peak number of users has yet to be met. Early morning ads could appear for those in areas likely to be leaving toward work and feature someone on their way, riding past recognizably commercial areas. Similarly, in the evening, those locations seeing a higher amount of trips taken could target potential customers with the idea of returning home or going to post-work activities. Or whatever marketing determines would best influence those in the specific behavioral mode.

It also provides a practical visualization of times when it might be best to go into the stations in order to clean, maintain, restock, etc. when activity has decreased enough that it would not be disrupting users from coming and going at the station. Similarly, one could gauge how large of a station would be needed to meet user demands.

Dashboard 3 - Summer Behavior, Customer vs. Subscriber

After looking at commuter behavior, I wanted to see if there were any noticeable trends in weekend trips made by customers - those who don’t necessarily bike regularly, but are making trips as needed for that day. These could be tourists along with New Yorkers wanting to enjoy the weekend. I wanted to justify this by first showing the inverse relationship between subscriber activity and customer activity throughout the weeks of June, July & August. The bar graph attempts to give a qualitative sense to this relationship, which can be seen best towards the second half of July and through August as the Customer bar goes up and down while the Subscriber bar goes down and up, respectively. The line chart provides a static view of this, but differs in that a moving average is used to smooth out the data and make the trend easier to see.

Dashboard 4 - Summer Weekend Traffic for Customers

For this dashboard, I will continue focusing on Manhattan. Right off the bat one can see the popularity of both Central Park and the waterfront on the West Side. There’s a good amount of activity throughout the rest of the area, but no noticeable similarity to Subscriber weekday activity.

One thing I found interesting was the increased activity along Park Ave. during the weekends in August. As it turns out, a street festival was held on the 3rd, 10th and 17th of August (Summer Streets), which closed off parts of Park Ave. and Lafayette St. to host a wide variety of activities.

Possible Uses of Dashboard 4

As with Dashboard 2, marketing decisions could be made to try and increase using bikes for getting to leisure activities (e.g. images of Central Park, the waterfront), or as leisure in itself. It can also provide an estimate for what load to expect of future events, be it fairs, parades, etc.

Files used from "https://www.citibikenyc.com/system-data"https://www.citibikenyc.com/system-data:

201812-citibike-tripdata.csv

201901-citibike-tripdata.csv

201902-citibike-tripdata.csv

201906-citibike-tripdata.csv

201907-citibike-tripdata.csv

201908-citibike-tripdata.csv