Citi Bike Station Activities Analysis Visualization

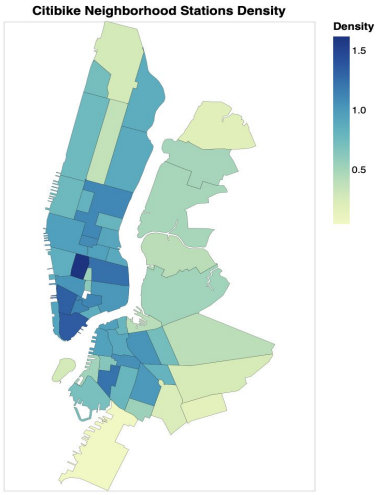
Group 6: Angelica Hernandez, Connie Wu, Hanh Hoang, Michael Ousseinov

Abstract

Citi Bike is New York's largest bike share system with 12000 bikes, 750 stations spreading across Manhattan, Brooklyn and Queens. The large scale of the systems poses a challenge in management which involve many optimization problems including the pressing issue of rebalancing bicycles among stations as well as the issue with the performance of individual stations. With this visualization, we explore the activities at each station and stations in each neighborhood as a whole. The insights are designed to create an intuitive understanding on the scheme and specific trends for each station.

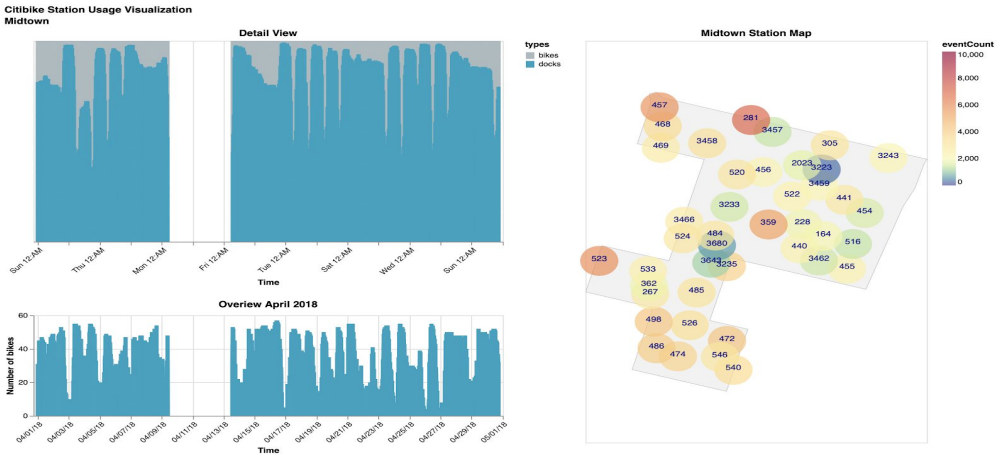
Introduction to our tool

First there is an overview of the neighborhoods.



(Figure 1)

By clicking on a neighborhood, user is then redirected to a detailed view of that neighborhood where there is a map of the stations on the right and bar charts of number of bikes of individual station on the left.



(Figure 2)

Research question

Traditionally, Citi Bike has to manually moving bicycles around by trucks which takes 45 minutes in average and usually is done at night. It now also has a Bike Angels program which rewards users traveling in reverse traffic from a full station to an empty station. There are also valet stations where employees drop off bicycles during rush hour. To further optimize the issue surrounding the lack of bicycles when the demand is high in certain stations, we want to build a tool to give hints to the following questions:

1. Where are the hot pick-up/drop-off stations during rush hours and off-peak hours?
2. The shortest less active station around a popular station during rush hours that Bike Angels can use?
3. How are the performance of stations comparing to each other in a neighborhood?

Method description

Data

For this project we mainly used two CSV files that contained information about CitiBike usage throughout New York City. The first CSV file contained all information about every bike trip that was taken in NYC during the month of April in 2018. This CSV was mainly used to extract explicit data such as location of bike stations, the name of the bike station, ID's of bikes etc. The second source of data that was used was a CSV file that contained information about how many bikes were at a station at any given time during the month of april. This was used explicitly for one of our data visualizations.

Data Processing

Event Counting

An important part of our data processing was counting the events at every station. An event is defined as a bike of a certain bikeID leaving a certain station and going to another station. An event is only counted for the station for which the bike exits. The purpose of counting the events that occur at a station is to be able to visualize the performance of each station. Stations are meant to be compared against other stations. Stations with more events are seen as performing better than stations with less events.

Neighborhood Filtering

Another important part of the data processing was filtering out all data extracted from the CSV's by neighborhood. This allowed us to showcase our data visualizations in a more local sense being able to derive implicit information about areas by filtering the data by neighborhood. The neighborhoods were extracted by looking at bike station coordinates and mapping them to a particular neighborhood.

Bikes At Station Counting

The bikes at station counting was done through parsing the CSV file that contained data about how many bikes were at any station at a given time. This CSV was convenient for the needs of the project as all it needed to be was formatted and then visualized in Vega.

Neighborhood of each stations

In order to filter stations into each neighborhood detail view we need to know which neighborhood a station belongs to. We achieve this by spatial join the stations data table and the neighborhood data table on the geometry (latitude and longitude)

Station density of a neighborhood

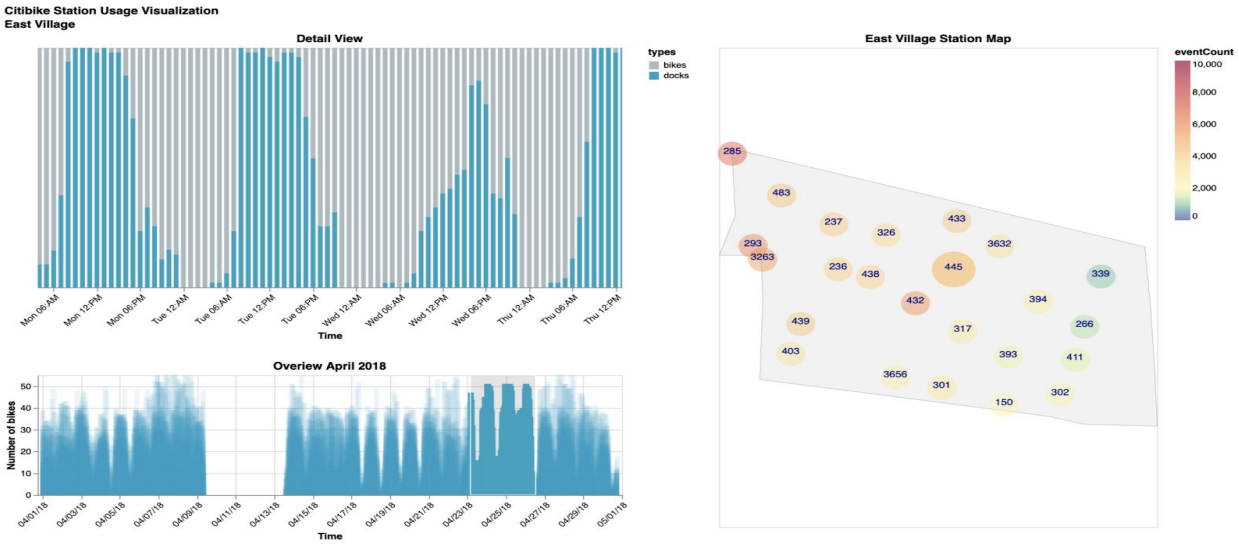
Citibike provides more bike stations in higher demand area. By that fact we can color each neighborhood according to the number of bike stations within it. But in order to make it comparable between neighborhood, we need to normalize the number of stations by the area of each neighborhood. The area of a neighborhood is calculated as area a polygon using. The density of a neighborhood is then simply by counting the number of stations in that neighborhood and divide by the area.

How the tool works

For the neighborhood overview visualization, each neighborhood is colored using a sequential colormap accordingly to its density in the number of bike stations. The density is calculated by the number of stations divided by the area of the neighborhood that those stations are located in. This color channel helps users to easily identify and navigate to a specific neighborhood of interest.

(Figure 3)

With the neighborhood detail visualization, there is a map of stations' location within the neighborhood on the right. Each station is represented by a point and a number which is the station ID. Each station is colored by the number of activities that has happened during the entire period of time (the month of April). The diverging colormap is chosen to demonstrate the stations with either very high or very low activity. When the user clicks on a station, the size channel is used, creating a pop-up effect to indicate the chosen station versus the rest.



(Figure 4)

On the bottom left of the neighborhood detail visualization is the overview bar chart of number of bikes at a station over time. The time is label in month/date/year format. Users can choose a specific time interval by brushing on the bar chart. The opacity channel is used to show the contrast between the chosen time interval and the rest.

The filtered time is then shown on the top left stacked bar chart which acts as a zoom in for the overview, and the time axis is further labeled into time of the day and day of the week. The blue color bars represent the number of docks and the gray bars represent the number of bikes.

Result

For example, in Figure 4 we inspect station 445 in East Village. By selecting a time interval from Monday 6PM to Thursday 12PM we can clearly see there's a trend where bikes are taken out around morning rush hours and return around evening rush hours. Given the fact that this is a residential area and there's no nearby subway station around this bike station we can safely infer that the people in this neighborhood use bikes this station to go to work in the morning and come back in the evening. We could also note that this station have a very low number of bikes, making it a hot pickup station during morning rush hours 8AM to 10PM.

Our contributions

There has been many tools built to visualize the unbalanced in bikes at each stations but there has not been many that focus on comparing the performance between stations. What our tool can do is to make it easier to identify under performed stations within a neighborhood. Combining with other tools could help in decision making of adding or removing stations in a neighborhood.

Conclusion

Currently our vis colored the stations on the number of events, either pick-up or drop-off. But we're also interested in seeing the unbalance in the amount of pick-up and the amount of drop-off at each station. So we plan on deriving the bike unbalance and having options where user can choose to color the station on eventCounts or bikesUnbalanced.

Even though isolating each neighborhood in the detail view helps with reducing visual noise, we lose the information of nearby stations from neighboring neighborhoods. This information might be helpful for route planning of Bike Angels. In the future, we plan on including neighboring stations within a certain radius on the neighborhood map as well.

We also think that it is useful to include the subway stations on the neighborhood map and the information about each neighborhood, whether it is more commercial or residential area or both.

We have taken note that it would be more clear for the user if there is a a notification system for missing data in certain neighborhoods. We also plan on further creating a dynamic set of data that allows more months that can be chosen from - up to the most recent data provided. This data visualization is a stepping stone for researchers to further delve into the enterprise of bike sharing foresight.