NYC Bike Data Visualization

K Iwasaki

May 18, 2017

Contents

Introduction

Citi Bike, New York City's bike share system, has released public data since July, 2013. As a cyclist, I found interesting to investigage the publicly available data. I use the data for July 2016 for my investigation. The data contains 1,380,110 columns and some variables including trip duration, start time, stop time, station latitude, and station longitude among 15 variables. R is a statistical computing language I used for this analysis. Among some packages for R, gglot2 and ggmap were very useful for visualization.

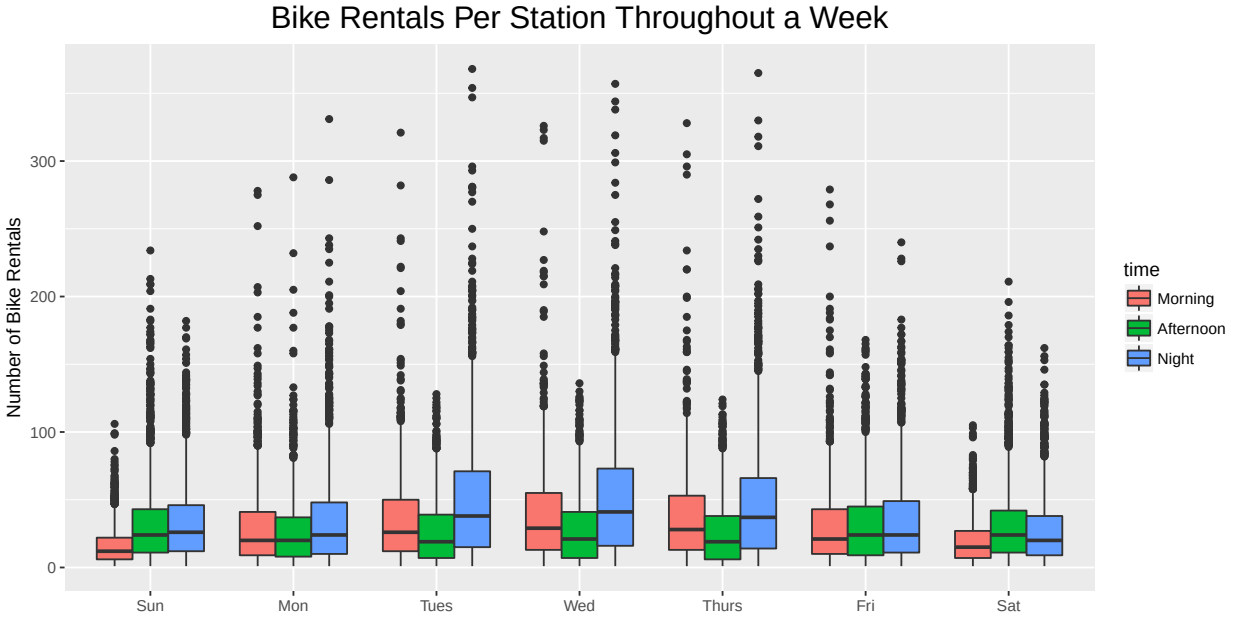
My focuses in the investigaiton are mainly around:

- Capture a big picture: find interesting usage pattern throughout a day/week, popular bike stations, and etc.
- Investigate some patterns identified: understand what is actually going on
- Think about ways to exploit the patterns in problem solving
- (Bonus:) Use a spatial visualization package ggmap for visualization

Data exploration

Big picture - understand a pattern of bike rentals at high level

I initially approached the data by looking from bike station standpoint. There are 483 bike staions in NYC and I assume there are popular stations and not so popular stations probably because of their locations. I used a rental bike service some times when I visited cities in the U.S. and outside. For me, I looked for a bike station closed to a train station or a hotel that I stayed.



Look at the chart which represents a count of bike rentals per station in each day of a week. There are **many outliers in each day**. Outliers are shown above whiskers on the boxes. Take a little bit closer look. In Monday morning, a mean value for rentals per station is 20. However, a station has the largest number of rentals actually got 278 rentals. This is more than 10 times of the mean. In addition, we can see some outliers with a value from 100 to 200. It seems it worth explore more on outliers.

Sidenote1: Since this dataset contains all transactions in the month of July 2016, each day of the week contains four or five days of the day. For example, Sunday contains transactions in Sundays 3rd, 10th, 17th, 24th, and 31st of July.

Another interesting observation is that Saturday and Sunday show a similar pattern and so do Tues, Wednesday, and Thursday. Rentals are slow in the morning on Saturday and Sunday with a narrower IQR and with relatively fewer outlieres. Rentals go up toward afternoon and night. For Tuesday, Weneday, and Thursday, rentals are higher in the morning and night, and are lower in the afternoon with fewer outliers and lower maximum.

Sidenote2: IQR, the interquatile range is a measure of variability. Quartiles divide a rank-ordered data set into four equal parts: Q1, Q2, Q3, and Q4.

- Q1 is the "middle" value in the first half of the rank-ordered data set.
- Q2 is the median value in the set.
- Q3 is the "middle" value in the second half of the rank-ordered data set.

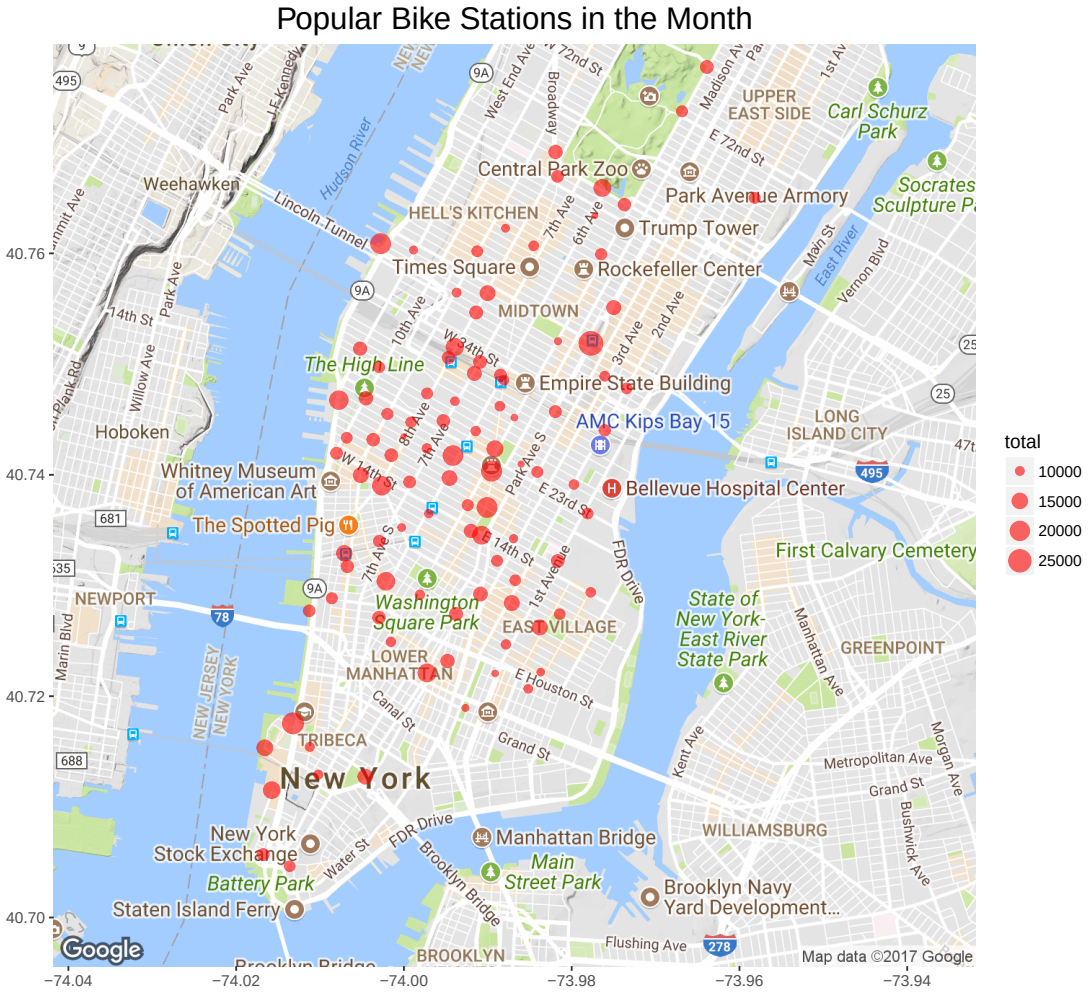
The interquartile range is equal to Q3 minus Q1.

Outliers - where are popular bike stations?

Think about outliers articulated in the previous chart. Those outliers are stations renting out a lot of bikes for customers. Customers rent a bike and return a bike. That means there are stations receiving bikes from customers. I want to investigate outliers of **bike stations not only for departing (renting) but also stations for arriving (receiving) bikes**.

I calculated a number of departing bikes and arriving bikes for each of 483 stations and picked up 100 stations with the highest number of transactions for the month. Then, I plotted them on the map as below, using a

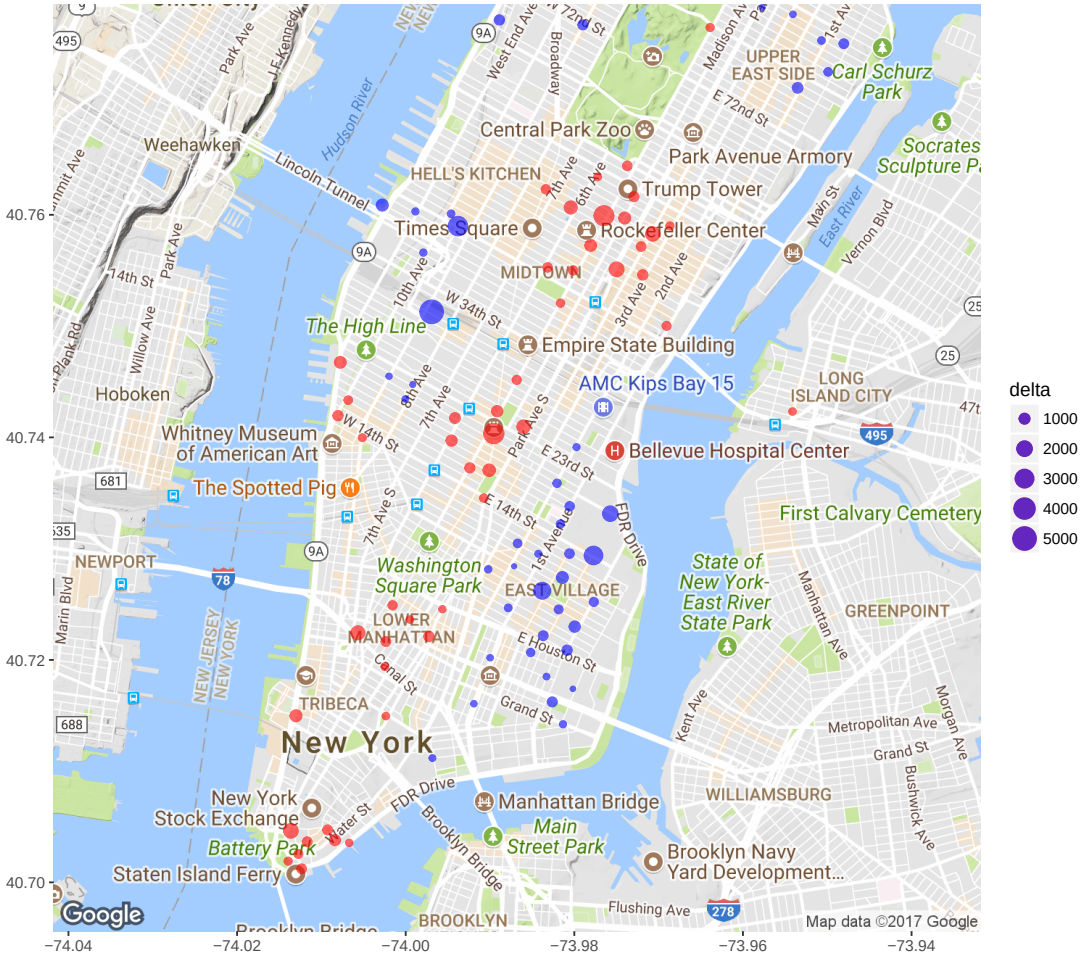
longitude and a latitude data for each bike station. The map tells a location of each stations and its volume of transactions, however, not very insightful.



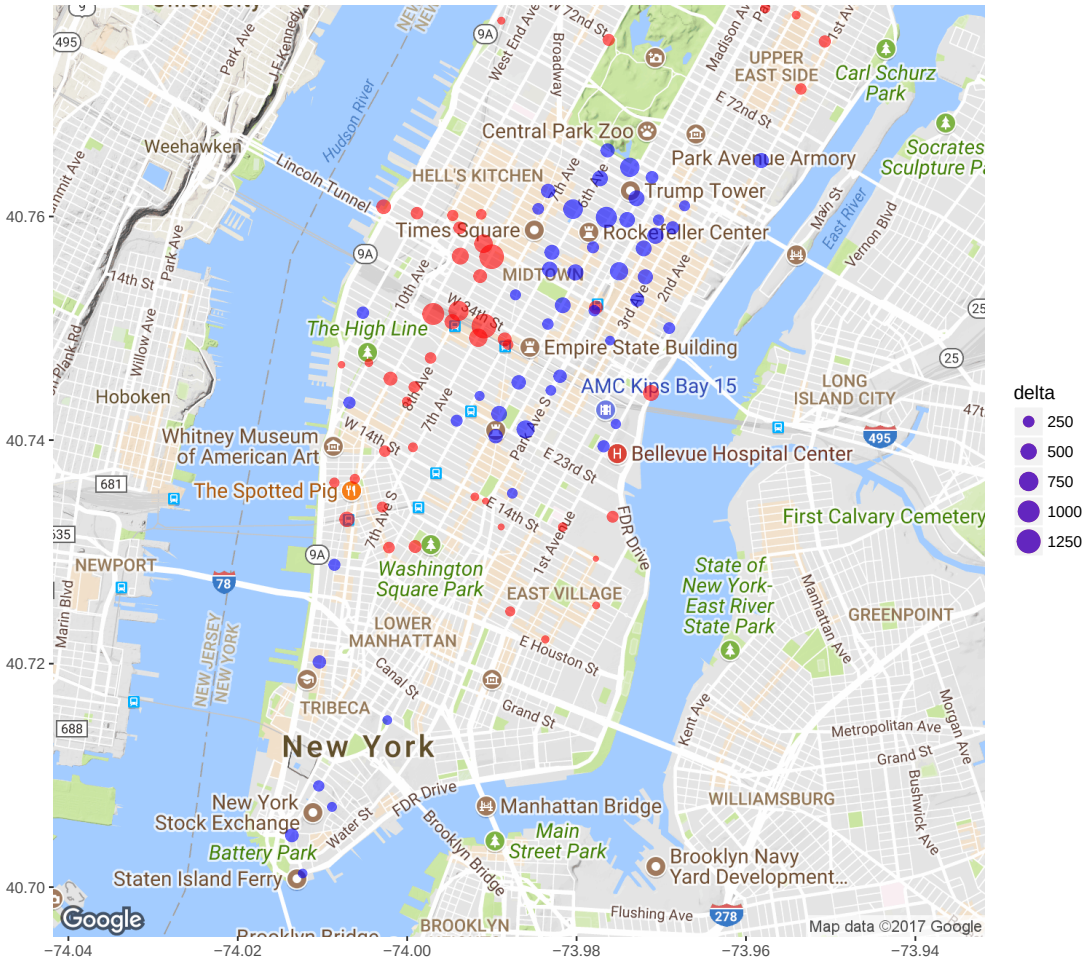
Step back and think about how to analyze popular stations further. Look at the first chart again, **number of bike rentals changes during a day** and it seems worth confirming if popular stations change during a day. For example, there might be morning bike stations where people rent bikes in the morning but people don't come after the noon. Another angle to investigate is to **classify stations into two categories: arriving stations and departing stations**. This might uncover interesting facts.

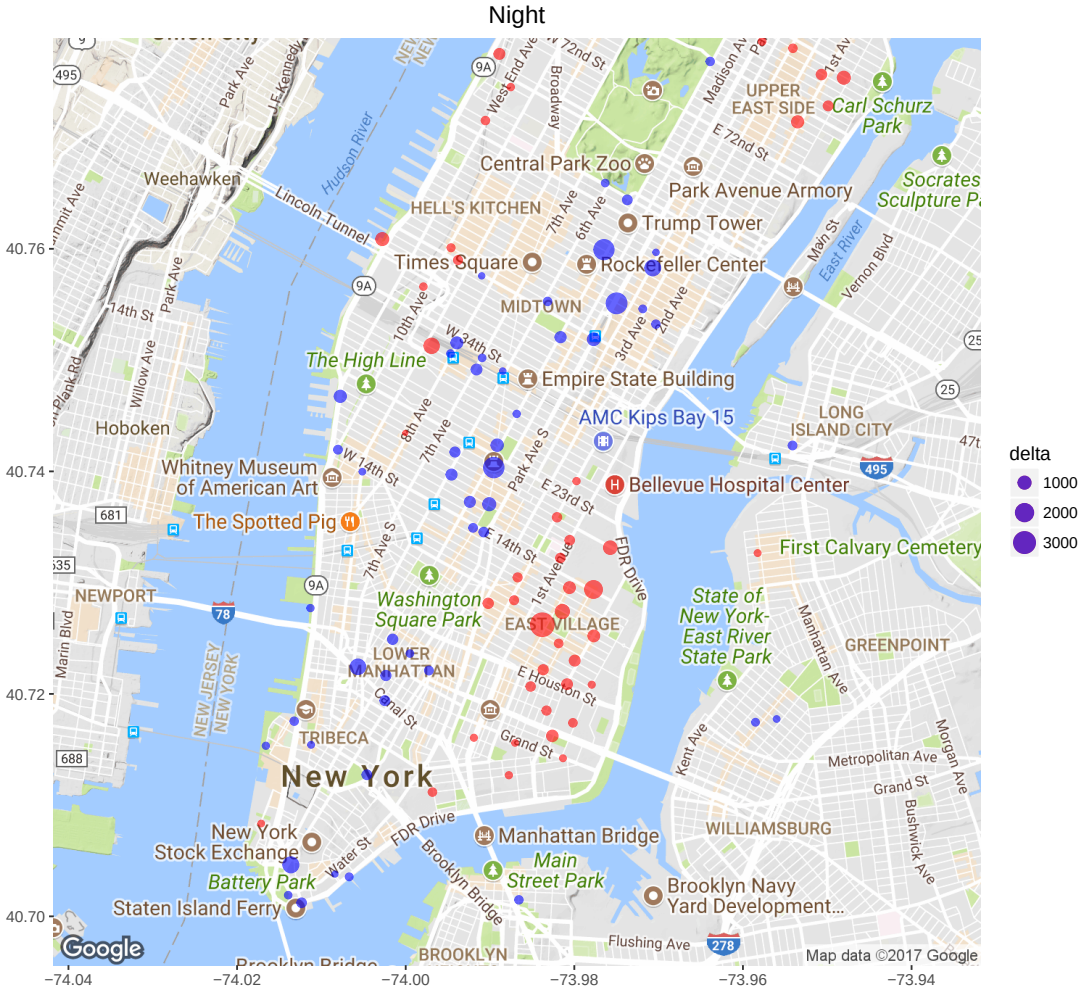
Drill down - look for interesting patterns

Morning


delta
1000
2000
3000
4000
5000
40.76
40.74
40.72
40.70
−74.04
−74.02
−74.00
−73.98
−73.96
−73.94
Weehawken
Hoboken
Central Park Zoo
Park Avenue Armory
Trump Tower
Times Square
Rockefeller Center
MIDTOWN
HELL'S KITCHEN
UPPER EAST SIDE
Carl Schurz Park
The High Line
Empire State Building
AMC Kips Bay 15
LONG ISLAND CITY
Whitney Museum of American Art
Bellevue Hospital Center
The Spotted Pig
First Calvary Cemetery
NEWPORT
Washington Square Park
EAST VILLAGE
LOWER MANHATTAN
State of New York-East River State Park
GREENPOINT
TRIBECA
New York
New York Stock Exchange
Battery Park
Staten Island Ferry
Manhattan Bridge
Main Street Park
WILLIAMSBURG
Brooklyn Navy Yard Development...
BROOKLYN
Lincoln Tunnel
Hudson River
FDR Drive
Brooklyn Bridge
Google
Map data ©2017 Google

Afternoon


delta
250
500
750
1000
1250
40.76
40.74
40.72
40.70
−74.04
−74.02
−74.00
−73.98
−73.96
−73.94
Central Park Zoo
Park Avenue Armory
Trump Tower
Rockefeller Center
Times Square
HELL'S KITCHEN
MIDTOWN
The High Line
Empire State Building
AMC Kips Bay 15
Bellevue Hospital Center
Whitney Museum of American Art
The Spotted Pig
Hoboken
Weehawken
Washington Square Park
EAST VILLAGE
LOWER MANHATTAN
TRIBECA
New York
NEWPORT
State of New York-East River State Park
First Calvary Cemetery
GREENPOINT
WILLIAMSBURG
Manhattan Bridge
Main Street Park
Brooklyn Navy Yard Development...
New York Stock Exchange
Battery Park
Staten Island Ferry
BROOKLYN
UPPER EAST SIDE
Carl Schurz Park
LONG ISLAND CITY
Map data ©2017 Google



In the three maps above, blue dots represent the top 50 popular departing stations and red dots repesent the top 50 popular arriving stations during a day. Again, a size of dots reflect a number of renting and returing a bike. The new visualizations provides us more insights than the previous one. First, arriving stations and departing stations create clusters separately. In the morning, we can observe three(or four) clusters for arriving stations and two clusters for departing stations. In the afternoon, there are two big clusters to the North. In the night, clusters appear similar locations to the ones in the morning but locations for arriving stations and locations for departing stations switched.

Sidenote3: I calculated "surplus" and "deficit" to come up with the departing stations and arriving stations. "Surplus" is calculated for each station by a number of rentals subtracted by a number of receivals. "Deficit" is an opposite of "surplus", a number of receivals subtracted by a rentals. If there is "surplus" for a station, a station is renting bikes more than receiving it. I define this type of station as an arriving station and station with "deficit" as a departing station.

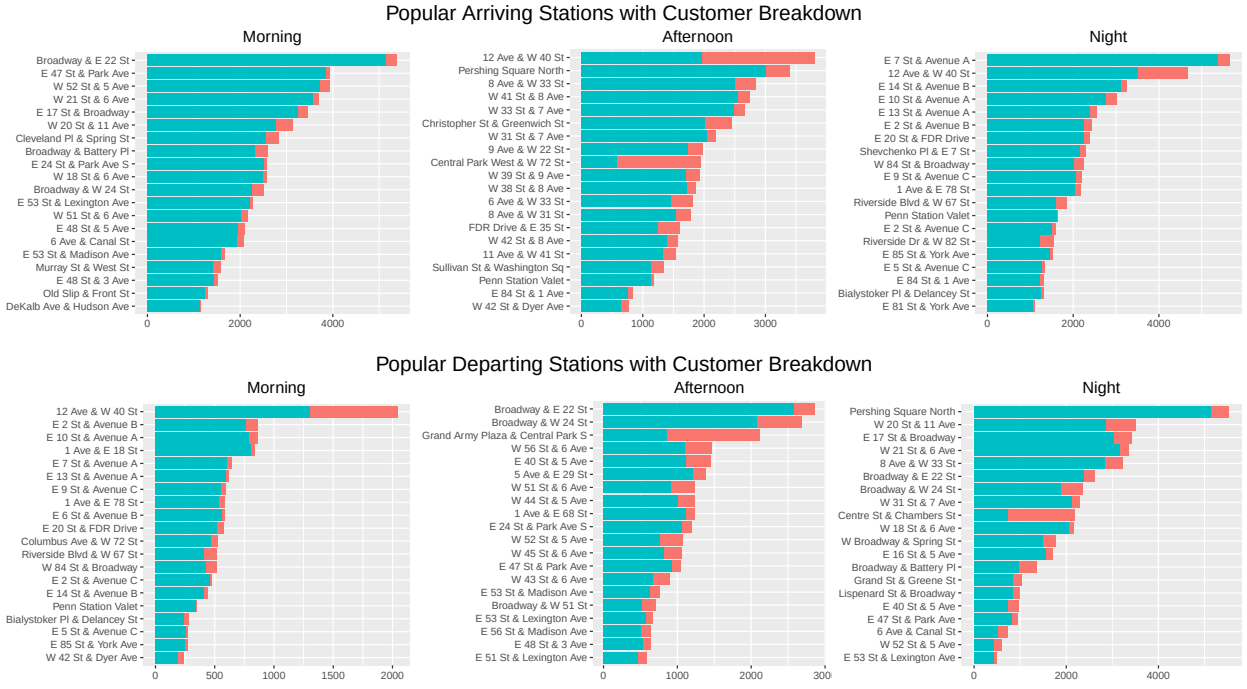
One way to interpret this pattern in the morning and in the night is that customers rent a bike at East Village, Pennsylvania Station and around (blue bubble clusters) and bike to Midtwon, Lower Manhattan, and so on (red bubble clusters). I assume this is mostly for commute because areas under blue bubble clusters have lot of offices. In the night, customers head back from their office (blue bubble clusters) to their home, their transit metro stations or others (red bubble clusters).

To validate this idea, let's look at types of customers at popular stations. As shown in below, I popularted charts for both popular arriving stations and popular departing stations with customer breakdown: subscriber (blue in the chart below) and customer (red). Most of popular stations are dominated by subscribers especially

morning and night at arriving stations. This supports my idea of bike users using a bike for their commute because subscribers are supposed to use a bike regularly and they pick up and drop off a bike at certain stations.

Let's pick some stations to see where they are located and think about what's going on at the stations.

- Broadway & E 22nd St. Station: Ranked No.1 as an arriving station in the morning and located at a center of the downtown. No wonder people dropping a bike here. It's also interesting to notice this sation appears in the lists of popular departing stations in the afternoon (No.1) and in the night (No.6).
- 12 Ave & W 40 St. Station: Ranked No.1 as a departing station in the morning and No.1 as an arriving station in the afternoon. Users are a mix of customres and subscribers unlike most of other popular stations. The station is located next to Hudson River and is close to many tourist attractions such as museums, sightseeing cruises, and so on.
- Grand Army Plaza & Central Park S Station: Ranked No.3 in the afternoon as a departing station and located next to Central Park. The station is popular for customers (one-time users). Considering time and location, I think tourists are taking a ride in Central Park from the station.



Conclusions - ideas to utilize the data for business

We found out some interesting patterns in bike rentals through visualizations and analysis. Logical next step is to squeeze values out of the findings. Before diving into some ideas to turn data into values, let's understand who is involved in this bike rental service. The service operator, subscribers, customers (one-time users), businesses on a bike route and around a station, and so on. Then, think about their potential demands and pains for the bike rental service. For this time, I foucsed on the service operator, subscribers, and customers and list up some questions to think about.

Important business questions to ask

The service operator: - How to handle fluctuating demands throughout a day at 483 bike stations? - More specificially, how to manage rental requests at popular departing stations? - Should they provide same service

to a subscriber and a customer? - How active are subscribers using the service? Who are dormant customers and why? - There are so many bikes (in fact 8143 bikes). Which bike should they check and repair?

Subscribers: - How to commute safely by bike? - Communute is not exciting. Is there anything to do to make commute fun? - What time is a best time for communite? How is a traffic at the time?

Customers: - Don't know where to go and see as a tourist. Where should I go by bike? - Where is a bike station nearby? - Why bike? What is an advantage of renting a bike over taking a Uber/Lyft?

Some questions are quite important business questions to answer and others are not so much. Let's focus on some important questions the dataset might help to answer and provide values for them. Below I summarized ideas about how to solve the selected questions by using the dataset.

Potential solutions leveraging the dataset

Inventory optimization

- Problem: popular stations for renting out bikes might suffer inventory shortages that cause opportunity loss and bad reputation for the service especially for tourists
- Solution: focus on popular departing stations. First, confirm if inventory shortages actuallly happen at the popular stations by relocating bikes from the unpopular stations to the popular stations. We might keep relocating bikes to the popular stations to understand a real demand at each of the stations. Second, once understand a real demand at the popular stations, the operator defines a relocation scheme which in essense 1) move bikes to the popular stations at the end of a day to prepare for next day morning. 2) move bikes to the popular stations during a day if necessary. Third, refine the scheme with data analysis.
- Implementation: Focusing on a few dozen or even less popular departing stations should make the analysis and the implmentation easier and faster. Trying to optimize inventories at every station is a loser's game.

Apps for a customer

- Problem: tourists don't know where to go and what to do after renting a bike
- Solution: build an app to suggest the five most popular destinations and tourist attractions around a bike route from a starting station to a ending station.
- Implementation: the app is easy to build since the current dataset has a starting station and a ending station for each user. Roadblocker is to combine a bike route information with another dataset containing information for tourist attractions.

Safe bike trip

- Problem: some potential/current users like biking but have concerns in commuting by bike because of safety concerns especially in rush hours.
- Solution: introduce an insurance service for a subscriber. First, the servive operator needs to collect an accident information to quantify frequency and materiality of a bike accident that a user encounter for a certain period of time. Second, they build and introduce an insurance package for a subscriber.
- Implementation: Not easy however the insurance can be a lucrative business. Creating an insurance package is not easy but feasible by utilizing the data. Think this as an mid-term project to tackle instead of a low-hanging fruit.

Optimization of bike maintenance

- Problem: maintenance costs might be very high to keep running 8143 bikes.

- Solution: identify and prioritize which bike to check and repair by analyzing the dataset. The dataset contains “bikeid” and “tripduraiton”. With these datapoints, it is easy to analyze which bike has been used for how long. Assumption here is the longer time it is used for, the more likely it requires maintenance. Pick highly utilized bikes for maitenance. If we add maintenance history on each bike in the dataset, we can also identify optimal maitaintenance cycle for a bike. This helps avoid unecessary check-up and maintenance for a bike.
- Implementation: easy to implement with the set of information they have already collected.