Big Data - DS_GA_1004

Final Project

Professor: Juliana Freire

Student: Sebastian Brarda, Ali Josue Limón, Osvaldo Bulos Ramirez

NetIds: sb5518, ajl649, obr214

Date: 05/13/2016

URL:https://docs.google.com/document/d/1-oxAWLJde9kmmZ2EErLl_JX9d3yFczSbRtX-Pdh4sPQ/edit

Github URL: https://github.com/obr214/BD_FinalProject.git

1) Abstract

Citi Bike is a sharing bike system that has become very popular in NYC. It has expanded rapidly in the last years, with hundreds of stations and thousands of bikes. In this project we focused on analysing the impact of weather conditions in Citi Bike usage during 2015.

Since millions of Citi Bike trips occur every, and all of them are recorded, analysing this problem require extensive usage of Big Data tools such as Hadoop or Spark. We preprocessed the weather data from NOAA in order to make it compatible with the Citi Bike data, then joined both datasets. After that, we conducted a series of map-reduce tasks over the joined dataset in order to conduct our analysis. Finally, we created a visualization in D3.

Some interesting conclusions were drawn from the analysis:

- Women and short-term customers are more sensitive to rain than men and Subscribers.
- Rain, Temperature and Wind do not affect the distribution of people travelling by age or gender.
- Higher temperatures have higher variance and average duration trips
- There are stations that are used mainly for commuting purposes, showing extreme behaviours during rush hours.
- Rain conditions might affect these commuting stations and interfere on the expected amount of bikes at each dock station at rush hours.
- This can have a lasting effect for users that do not find bikes where they expected and ultimately diminish customer satisfaction.

2) Introduction

How Citi Bike Works

Citi Bike is New York City's bike sharing system which was Launched in 2013. Since then, thousands of bicycles and hundreds of docking stations have been added to the system, while the number of both subscription based customers and temporary customers has increased dramatically. The service is provided in Manhattan, and many parts of Brooklyn, Jersey city and Queens 24 hs, 365 days. Citi bike’s customers use the service both for commuting and for leisure.

Users can buy a 24 hs, 7 day, or annual subscription. The first two, are considered “Customers” while the annual purchasers are considered “Subscribers”. After taking a bike from a dock station, customers have 30 minutes to use it while subscribers have 45 minutes without incurring in extra fees. However, if someone uses a bike for 30 minutes, he/she can leave it in a dock for 1 minute and then take it again without charges. If a dock station is full after the 30/45 minutes so that user cannot leave the bike on time, user can request for a time extension in order to find another dock station nearby.

Users have access to an interactive map and mobile app where they can see in real time the amount of available bikes (and free spaces) in each of the dock stations of the city.

Goal of the Project

The idea of the project is to understand how different weather conditions impact on the usage behavior of city bike, which ultimately conditions the operation of the service. To accomplish this goal, we used anonymized data provided by Citi Bike itself that accounted for all the trips occured in 2015, and the weather conditions for the same year, provided by NOAA.

Since biking as a transportation service is, for obvious reasons, considerably more affected (in a negative way) by weather conditions than the subway, buses, and taxis, understanding how these weather conditions affect the service can allow Citi Bike to operate more efficiently and increase their profits.

Some of the questions that we would try to answer in the project are:

- How the different types of users are affected by the weather conditions and how to quantify these changes.
- How do weather affects usage by gender and age.
- What are the weather variables that affect usage.
- How does weather affect the operation of the service.
- What are the negative consequences of bad weather for Citi Bike and which is the lasting effect that they have.

Infrastructure

Since Citi Bike is a massive transportation service, their dataset including information about all their trips is huge. Because of that, it is important that we leverage the right tools in order to manipulate the data and conduct our analysis. More specifically:

- We used Python to clean, interpolate and fix the weather dataset.
- We used Spark to join the weather and citi bike's data.
- We used several map reducers written in Python in order to process the data and conduct the analysis. We run those map reducers in Amazon AWS.

- We used Python, Jupyter Notebooks, Matplotlib and Plot.ly to manipulate, analyze and plot the aggregated data obtained after the map-reduce processes.
- Finally, we used D3 to build an interactive visualization

The citi bike data for 2015 has information about 9,937,969 trips and weights around 3GB. After joining this dataset with weather information from 2015, the size of the data grows considerably, going from 15 to 32 attributes. Manipulating this amount of information by loading it directly into main memory is feasible, but would have made the process very slow. Furthermore, our analysis and scripts can be used to analyze any given or several years, which would result in a further increase in datasize.

3) Experimental Techniques and Methods

Weather Data Pre-Processing

We utilized weather data from NOAA. The dataset is available in the following link, while the metadata is available in this other link. The dataset contains hourly information for 2015 about temperature, wind, rain, snow, and several other variables. The time stamps for the observations are in GMT.
The Citi Bike data is open, and can be found in the Citi Bike's webpage

In order to clean the weather dataset and prepare it to join it with the Citi Bike data, we conducted the following pre-processing tasks:

- Create a new column with the timestamp in Standard Eastern Time, in order to make it comparable to the Citi Bike's data timestamps. It was important to be very careful in this task, since the NYC time changes in November and March (winter special time).

- Take a sub-sample of columns and rename them. Many columns were mostly filled with Null values, and many others were not useful for our analysis.

- Group data in order to have one value per hour. We found some cases where there was an observation at 15.51 and another at 16.00 (for example), but the first one had mainly Null values and added very little information. In order to merge very close time registries, we applied grouping and mean or mode depending on the type of variable (mean for numeric, mode for categoric). In this way, we got rid of many Null values.

- Fix the timestamps after conducting the previous process.

- Interpolated weather data in order to generate one weather row for each minute occured in 2015. We created a lineal function that took equally spaced temperature values between the temperature observations for each hour in order to generate approximate temperature values for each minute. We understand that this process is not exact, but we think it is a good approximation of the weather conditions at each particular minute. We conducted the same process for other numeric variables such as wind speed, etc.

- For rain, we created three buckets: no rain, high rain, and low rain:
    - No rain: if the cumulative mm of the one hour observation that follows that observation is 0 mm
    - Low rain: if the cumulative mm of the one hour observation that follows that observation is greater than 0 and smaller or equal than 0.049 mm
    - High rain: if the cumulative mm of the one hour observation that follows that observation is greater than 0.049 mm

  Again, we understand that this process is not perfect, but we think that for a certain hour registry, if the cumulative rain in mm of the last hour was high, then there was a high probability that is was raining for the most part of the previous hour. In practice, when we analyze hourly data, the amount of trips occurred when raining will clearly be over-estimated, since probably many minutes marked as “rain” were not minutes when it was actually raining. But we know that at least the weather was not good, and the probability that it rained at that particular minute was high.

Weather Data and Citi Bikes data Join

After conducting the pre-processing of the weather data, very little work was required to join the datasets. More specifically, we already had a weather dataset with one registry for each minute occured in 2015, and one dataset of citi bike with each trip occured in 2015. The idea was to associate each trip with the weather condition at that particular time.

In the Citi Bike data, we had two timestamp: start_time and end_time. We decided to join both datasets based on the start_time timestamp of the Citi Bike dataset, since it is the moment when the decision of starting a trip is made. In other words, if it is raining at the moment that a particular user needs to travel from one point to another, user might decide to take a taxi or the subway instead. However, if user was riding a bike at the moment that it started raining, he/she would still need to return the bike to a dock station and finish that trip.

Since the start_time column contains information about seconds, we rounded every time stamp to its nearest minute in order to conduct the join.

In order to join both datasets we used Apache Spark with the following with one master and four nodes. The Join took 11 minutes.

Map Reduce tasks for Analysis

The following Map Reduce Scripts were developed in order to aggregate the Data for analysis. The input data for all these tasks was the joined table described in the previous section.

Map Task	Reduce Task	Configuration	Charts created with output	Type of Operation
map1.py Key: each combination of rain and gender Value: 1	**reduce1.py** Sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 TIme: 2 min	1, 2 and 3	Aggregate, sum
map2.py Key: each combination of type of customer and rain. Value: 1	**reduce2.py** Sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time: 2 min	4	Aggregate, sum
map3.py Key: rain Value: Trip Duration in Seconds	**reduce3.py** Averages trip duration over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time: 2 min	5	Aggregate, average
map4.py Key: rain Value: Computes L2 distance in terms of Long and Lat	**reduce4.py** Averages distance over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time: 3 min	5	Aggregate, average

map5.py Key: Each combination of age and rain Value: 1	**reduce5.py** Splits key and sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time: 2 min	6, 7	Aggregate, sum
map6.py Key: Each combination of rain, start or end time, start or end station, period of the day (morning or afternoon) and indicator of star or end trip Value: 1	**reduce6.py** Splits key and sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 TIme: 5 min	14,15,16,17,18, 19,20,21 and 22	Selection, Aggregate, sum
map7.py Prints line of the original file if it belongs to a certain station	No Reducer, it's just a filtering operation	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 TIme: 1 min	23, 24, 25	Selection
map8.py Key: Each combination of temperature, wind, and gender Value: 1	**reduce8.py** Splits key and sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time: 2 min	8,9	Aggregate, sum
map9.py Key: Each combination of temperature, wind and age Value:1	**reduce9.py** Splits key and sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time:2 min	10,11	Aggregate, sum
map10.py Key:Each combination of temperature, wind and arrival time.	**reduce10.py** Splits key and sums values over all keys	Standard AWS config. Instance Type m3.large Number of instances 3 (1 master and 2 core nodes) Reducer tasks 2 Time: 3 min	12,13	Aggregate, sum

4) Results and Discussion

Effect of Rain on Trips

The first question to answer is which is the extent up to which rain can alter the number of trips conducted. We filtered the trips by each of the rain conditions: No Rain, Low Rain, High Rain, and got the following results:

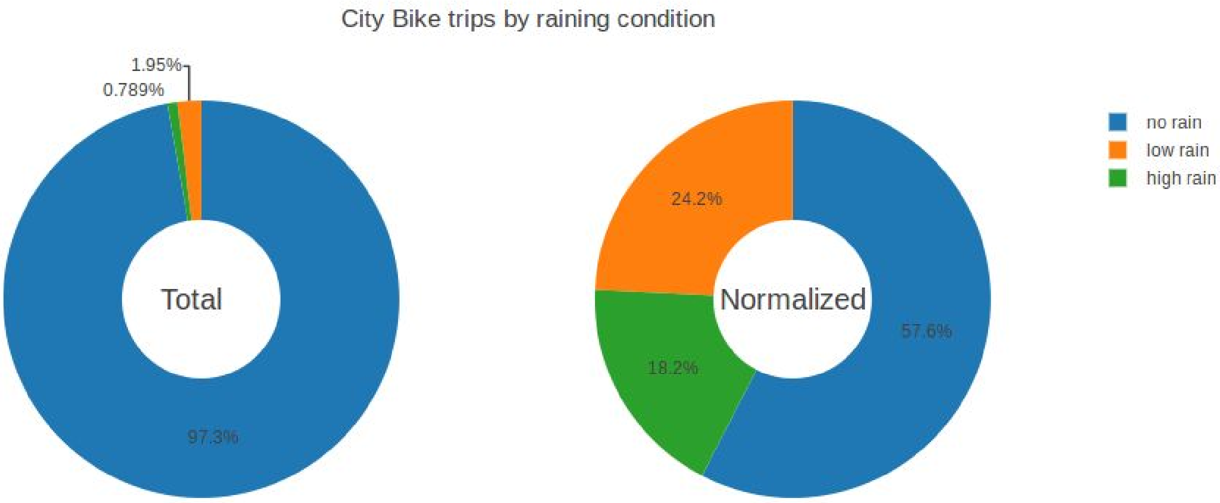


Chart 1

The chart on the left shows the number of percentual trips occurred with each of the raining conditions. However, the small amount of trips conducted with rain can be consequence of both less trips conducted when raining, or by a smaller proportion of raining minutes during the year. The second chart, is the same information but normalized by the amount of minutes in the year when each of the raining conditions occurred. This gives as a clearer perspective of the effect of rain on the average number of trips per minute.

Average trips per minute by rain condition

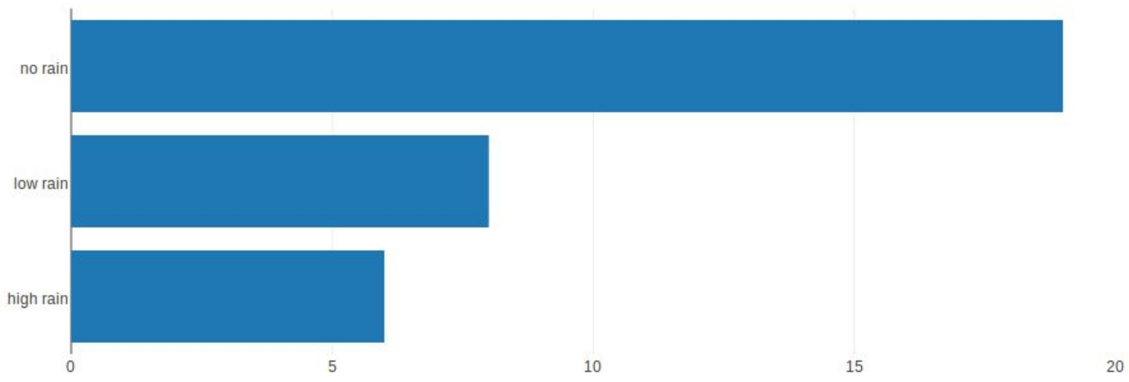


Chart 2

This second chart shows more clearly this effect: the average number of trips per minute for each of the raining conditions is less than half the number of trips when no rain. As we mentioned before, the number of trips occurred with raining conditions might be over-estimated, so these difference might be even higher in reality.

The second question to answer is whether raining conditions might affect the different types of customers in different ways.

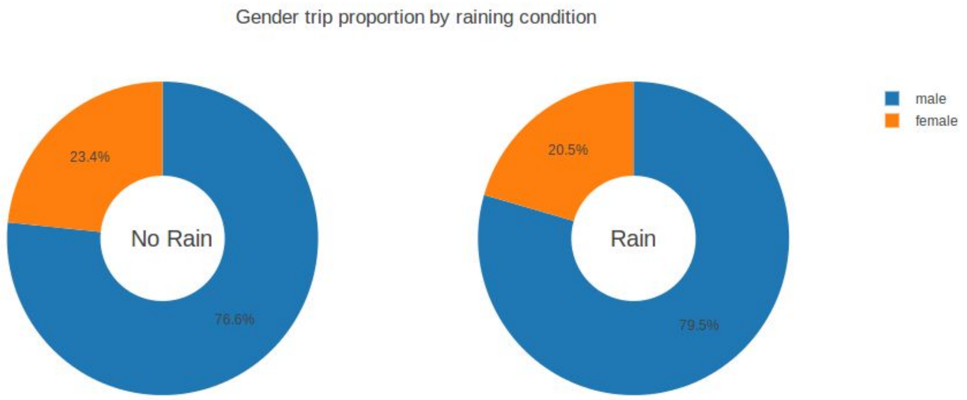


Chart 3

As it is clear from chart 3, women are more sensitive to rain than men. There is a 12.4% decrease in the proportion of women making trips when raining, which is statistically significant.

It is also important to understand the effect of rain on the different type of customers/subscribers.

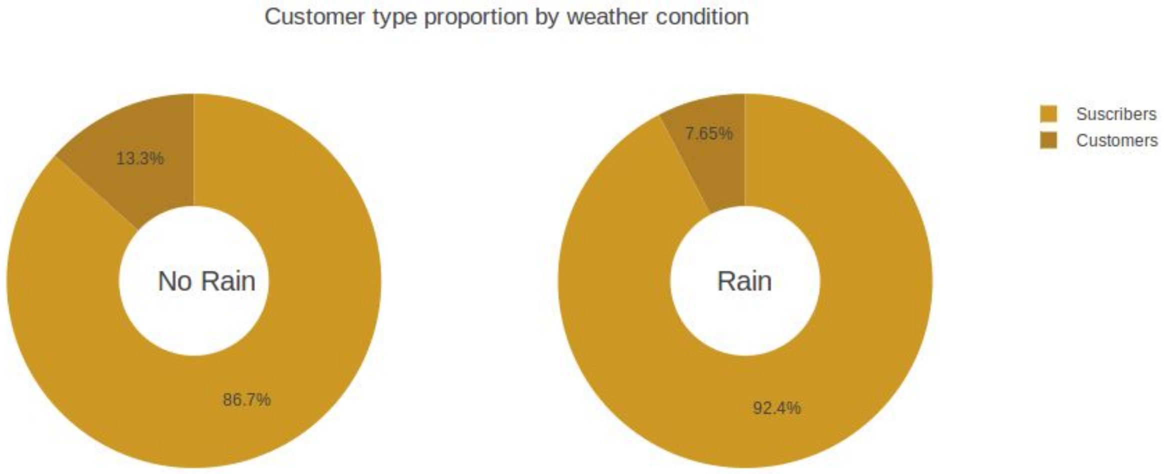


Chart 4

When raining, the proportion of trips conducted by short-term customers diminishes dramatically. This might be related to the fact that annual subscribers are more used to the service, even in bad weather conditions. On the other hand, some short term customers pay on the go, and might decide to use another transportation service.

Something interesting to understand as well is how raining conditions affect trip time and distances. The travel time in seconds was already available in the Citi Bike dataset. For the distance, we calculated the L2 norm between stations in terms of latitude/longitude. It is not exact, since users might take different routes and travel speed might be different at different parts of the city. However, we assume that it is a reasonable proxy for travel distance.

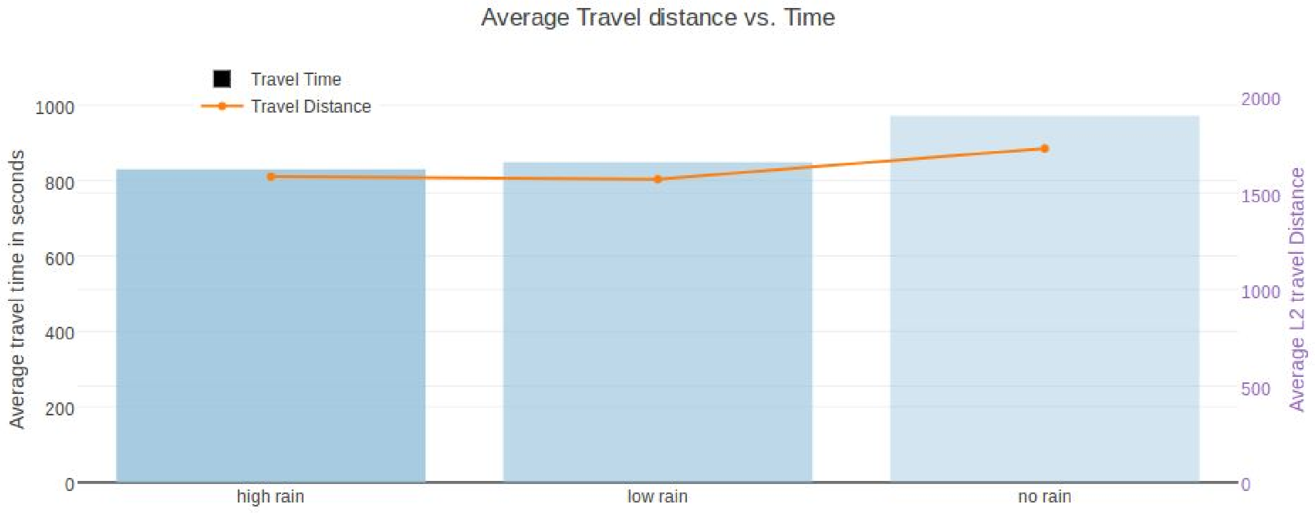


Chart 5

It was interesting to find that when raining, both the average trip time and the average travel distance diminishes. The reason might be that if it starts raining in the middle of a trip, users might decide to stop the trip earlier and leave the bike in the nearest station. Also, it is possible that when raining, the proportion of non-commuting users diminishes (we cannot distinguish them from our dataset) which might explain a smaller average trip distance since usually commuting trips are shorter.

However, the diminishment in the average travel time when raining, is higher in percentage terms than the diminishment in average trip distance. In other words, the diminishment in average travel time when raining, is only explained partially by lower average trip distances. This might mean that when raining, users not only choose to make smaller trips, but also make them faster. This might be related to the fact that if it starts raining in the middle of a trip, users might increase speed in order to hurry and get as fast as possible to the nearest dock station.

Another point that is important to understand, is how raining conditions affect users from different ages. If we plot the distribution by age of the number of trips occurred in each of the raining conditions, we get the following chart.

Distribution of Ages of users by raining Condition.

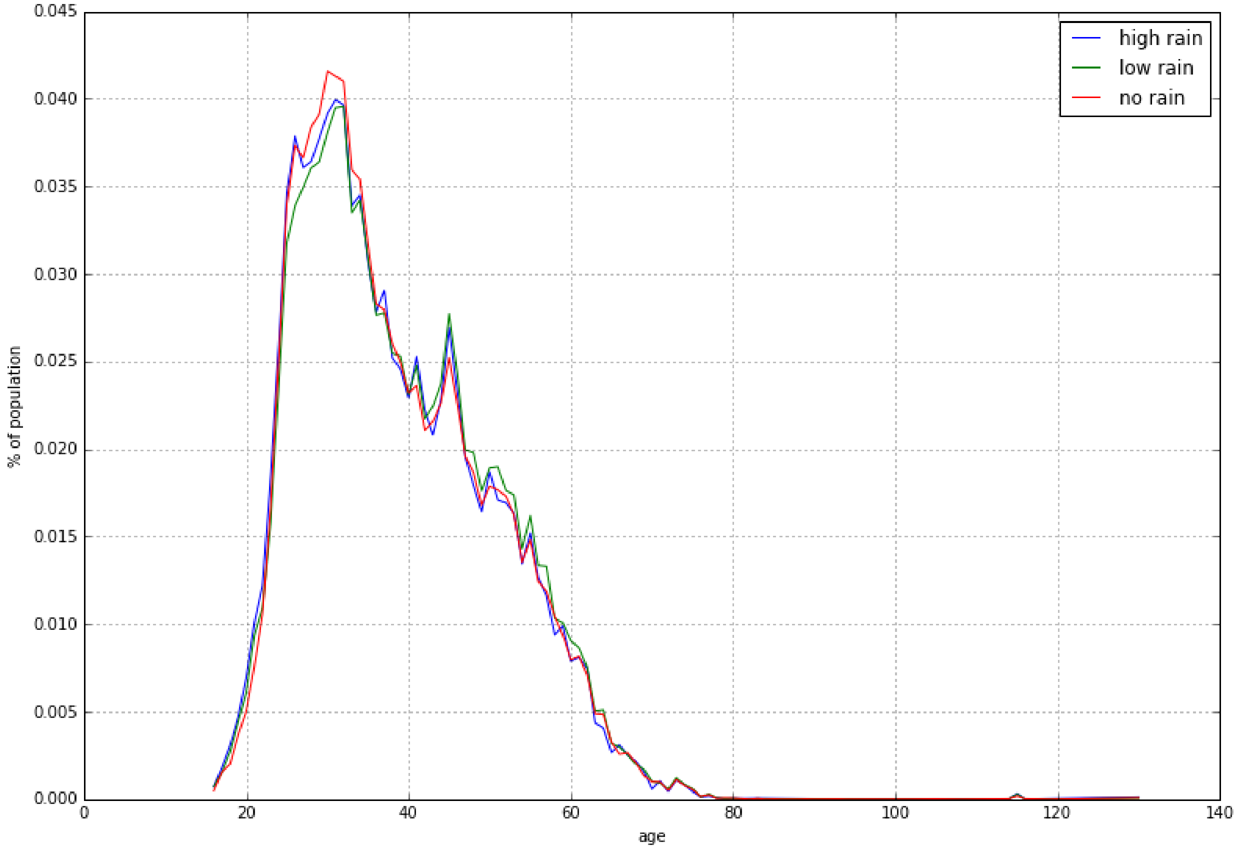


Chart 6

Other plots were made with the same information in order to identify differences:

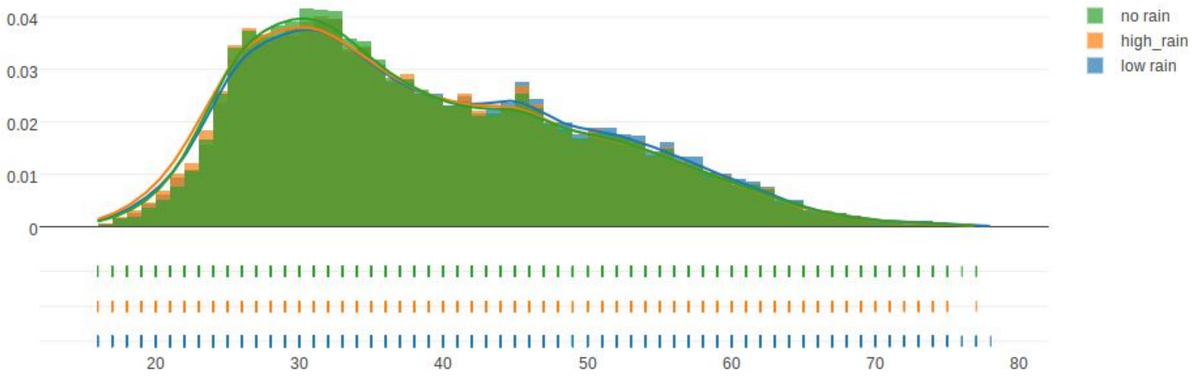
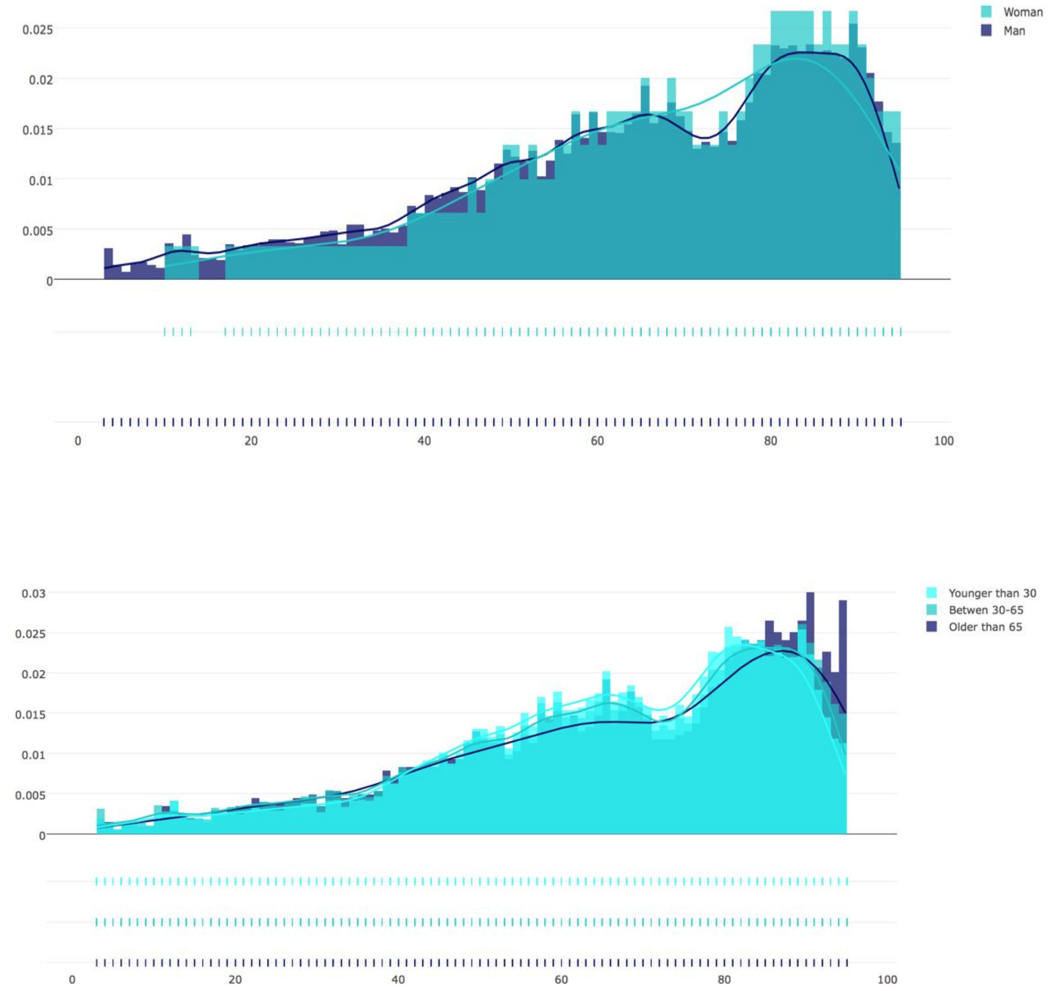


Chart 7

After reviewing the charts and conducted an ANOVA analysis, we concluded that there is no statistical difference between the distributions, so rain does not affect users from different ages in a different way.

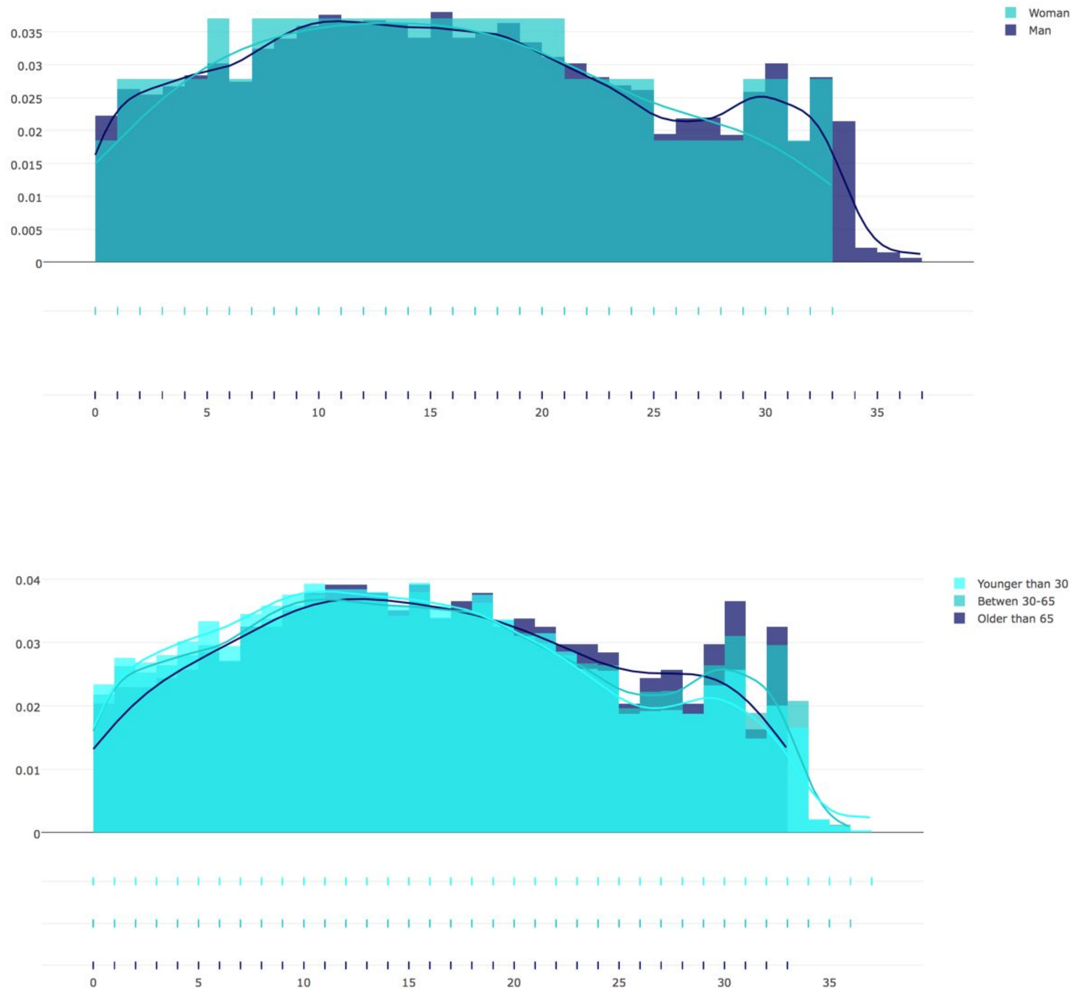
Effect of Weather and Wind on Trips

In this section we will analyze the impact of different temperatures and wind speed over the number of trips by age, gender and duration. Figures 8 and 9 show the distribution of number of trips by temperature, each color represents a gender. Similarly to the section of rain, we normalized by the number of times that each temperature appeared. It is clear that the distribution does not change between gender. As it was expected, the number of trips is significantly bigger for high temperatures. Similar conclusions are drawn if we consider Age instead of gender.



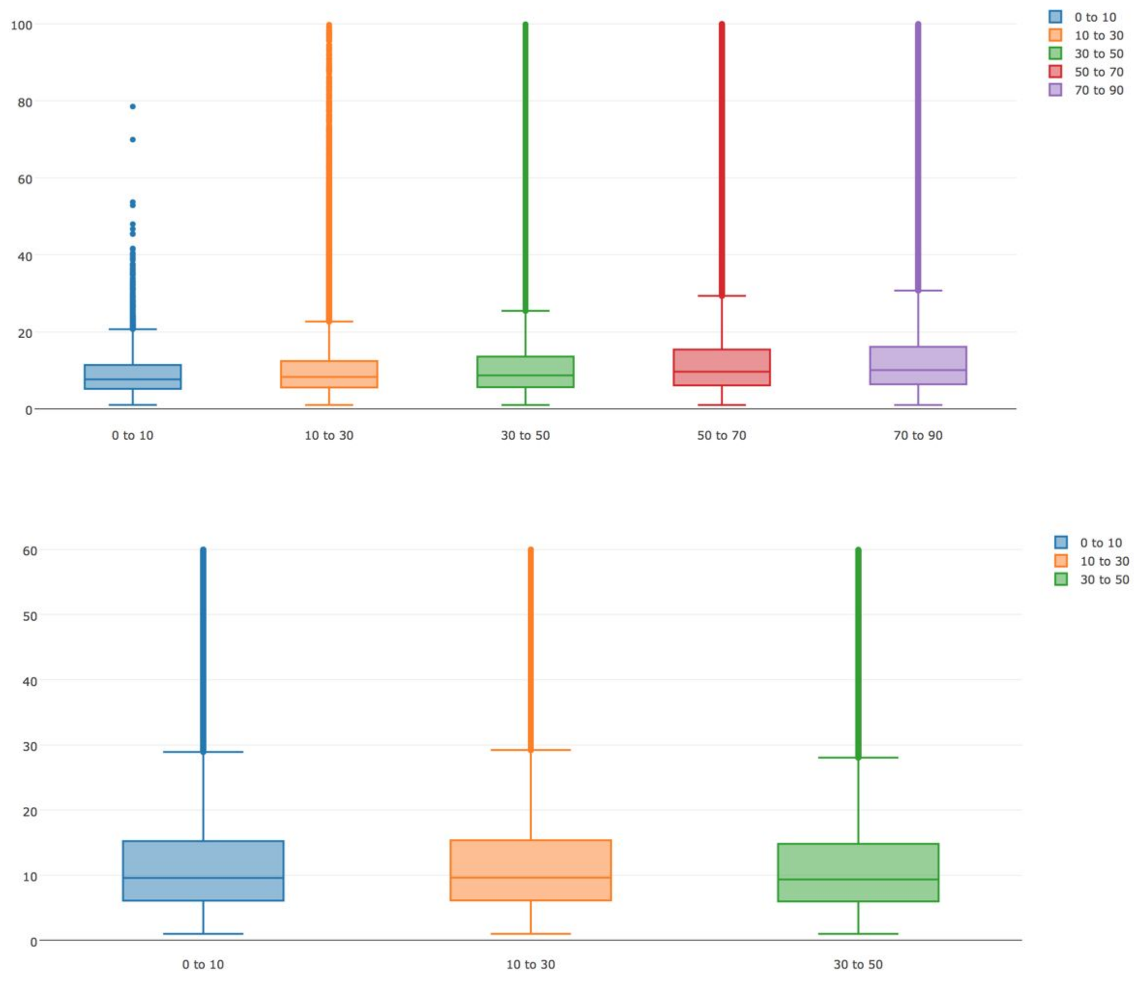
Charts 8 and 9

In terms of wind, the distribution seems to be flatter than temperature. There is no significant difference between gender and age.



Charts 10 and 11

Histograms about the trip duration are shown in figures 12, 13. Regarding temperature, it can be seen that higher values of temperature are associated with higher variance and mean in the trip duration. In the case of wind, the boxes are similar and we cannot make any deeper conclusions.



Charts 12 and 13

Analysis of variance of trips by Station

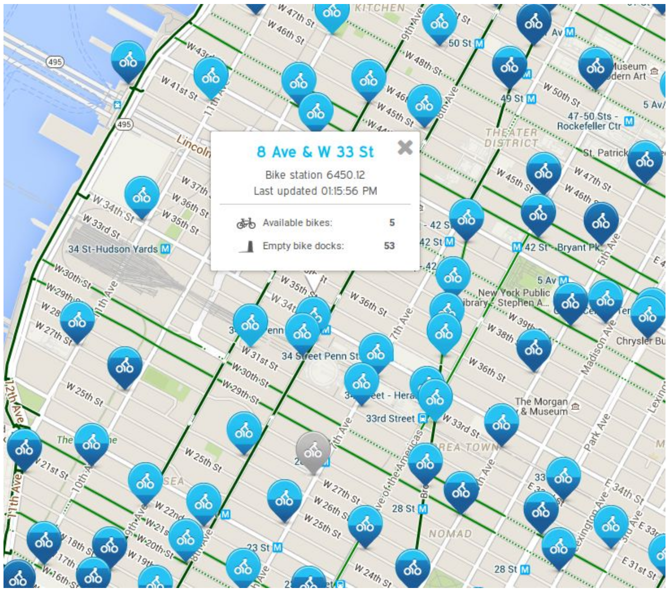
After researching about how Citi Bike works, and interviewing some frequent users, we decided to focus our analysis on how the different dock stations behave and how weather can alter that behavior.

More specifically, interviewed users were commuters (used bikes to travel to work and back home at rush hours). They mentioned that at certain times in the morning and afternoon, some dock stations appeared to be empty, and some other completely full.

In other words, some stations worked as “morning trips starters” and “afternoon trips enders”, while some other worked as “morning trips enders” and “afternoon trips starters”. For the first kind, we would expect to find them full of bikes over the night, and empty during working hours. On the contrary, for the second type we would expect to find them empty during the night, and full during working hours. For the first type, we would expect them to be located in residential areas, or next to connecting points from other travel services (for example next to the Grand Station). For the second type, we would expect them to be located in areas where people work, such as Wall Street.

After manual inspection of the City Bike webpage, this was confirmed. The following screenshots were taken today around noon.

Typical “Morning starter” station near Penn Station

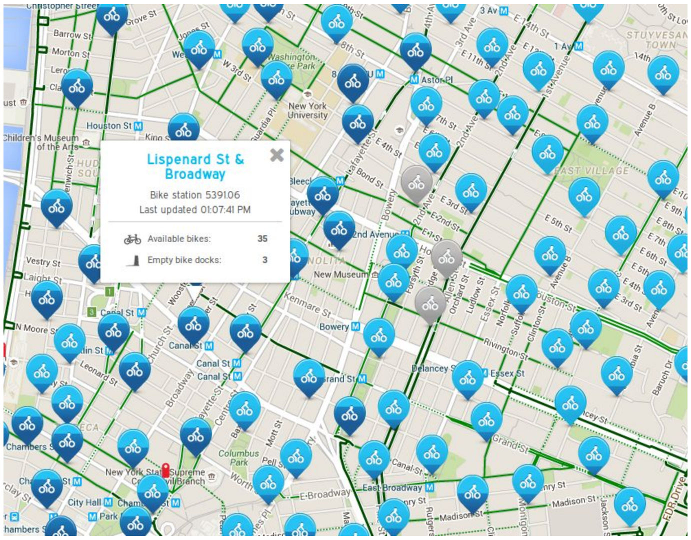


8 Ave & W 33 St

Bike station 6450.12
Last updated 01:15:56 PM

Available bikes: 5
Empty bike docks: 53

Typical “Afternoon Starter” station in Downtown



Lispenard St & Broadway

Bike station 5391.06
Last updated 01:07:41 PM

Available bikes: 35
Empty bike docks: 3

Of course that some other stations would be more intermediate without showing these extreme behaviors.

To analyse this fact, we filtered our dataset and restrict our analysis to the commute hours: morning between 7 am and 10 am, and afternoon between 5 pm and 8 pm, from Monday to Friday. Afterwards, we computed the average % of trips started and ended at each station over the total trips of that station at each of the time windows.

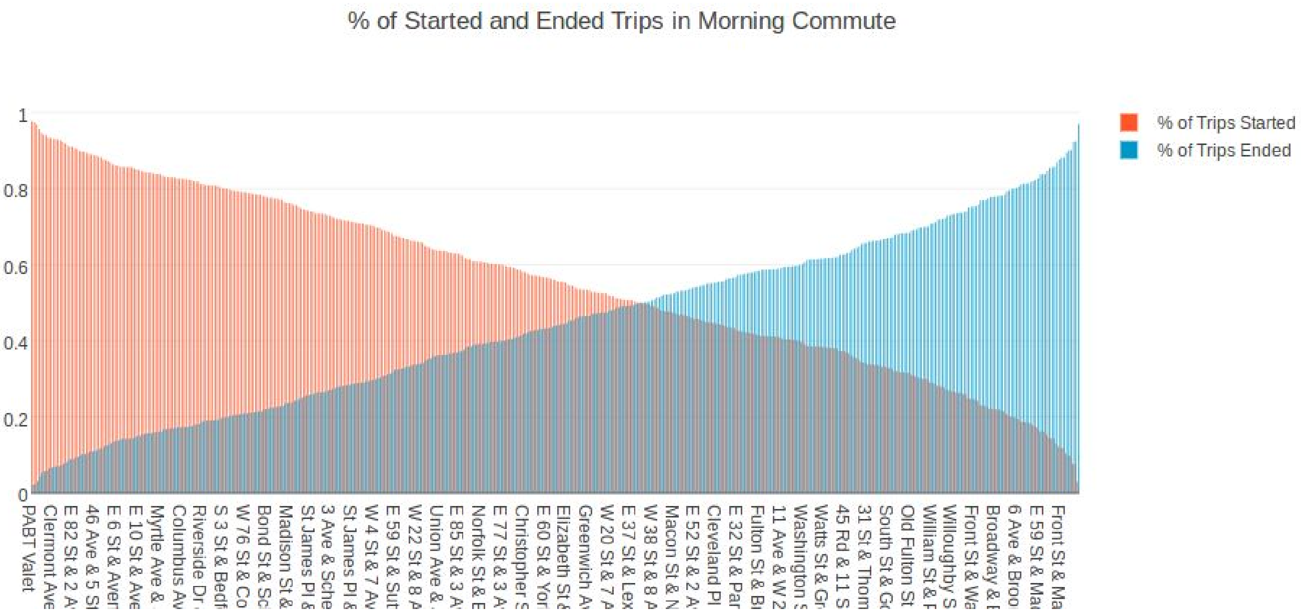


Chart 14

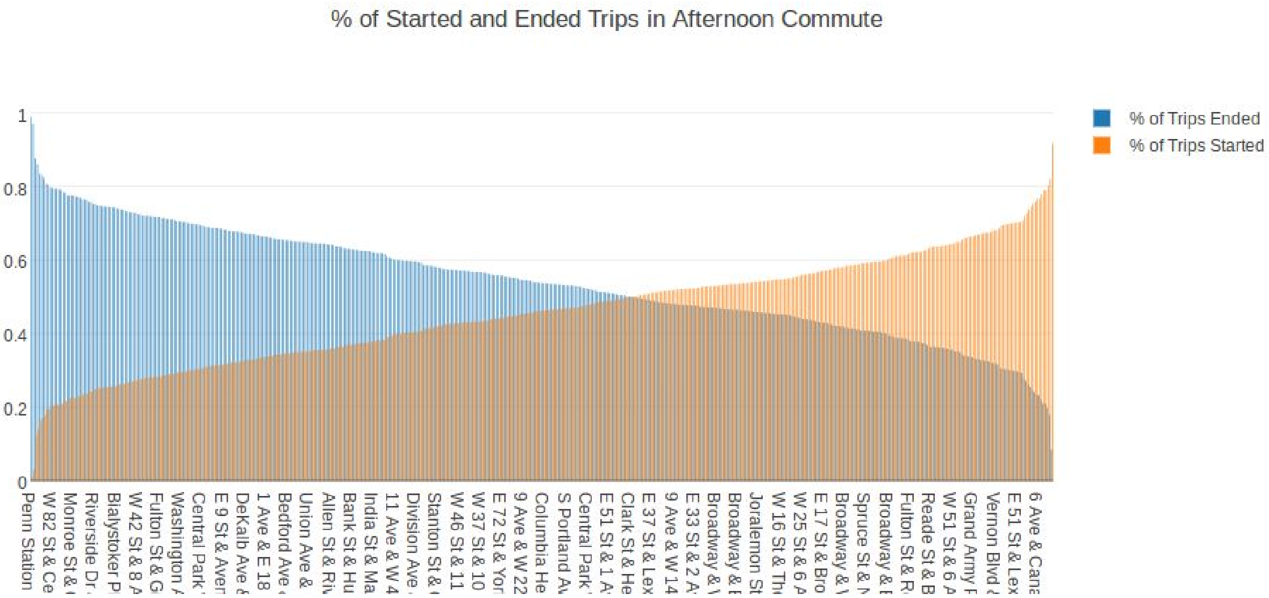


Chart 15

The X axis shows the name of the station, and the Y axis, the average % of trips that started/ended during that time window.

From the chart, it is clear that some stations have strict commuting usage. We would expect That the stations on the left of the first chart (start a lot of trips in the morning) would be on the right side of the second chart (start a lot of trips in the afternoon). And that’s exactly what happens.
More specifically, we are interested in the correlation of the % of trips started in the morning, and % of trips started in the afternoon across all stations.

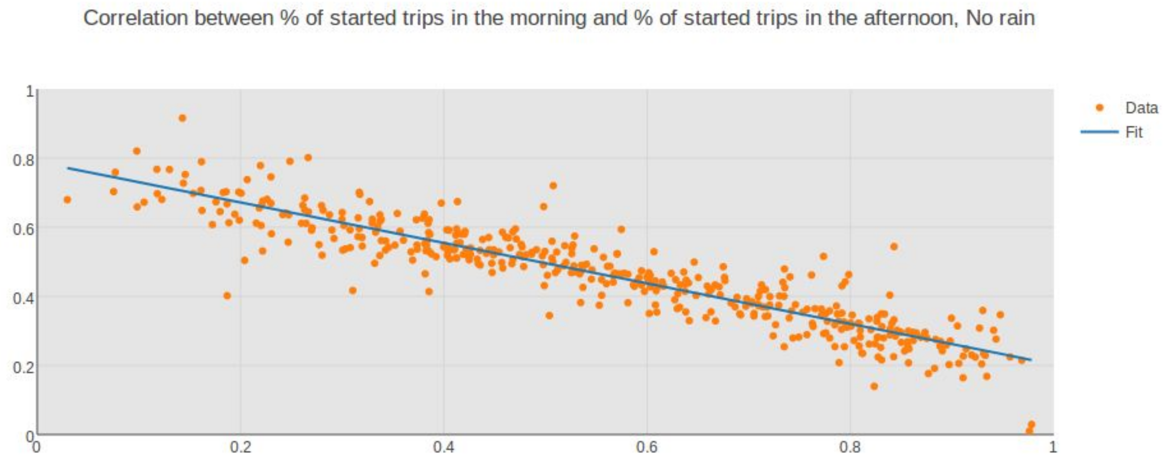


Chart 16

With an R-squared of 83.74% the relationship becomes confirmed when there is no rain. If a particular station has a high percentage of trips started in the morning, it will have a low percentage of trips started in the afternoon (or a high percentage of trips ended in the afternoon). If a Station has a small percentage of trips started in the morning (or a high percentage of trips ended in the morning), it will have a big percentage of trips started in the afternoon.

However, we cannot say the same when rain is high.

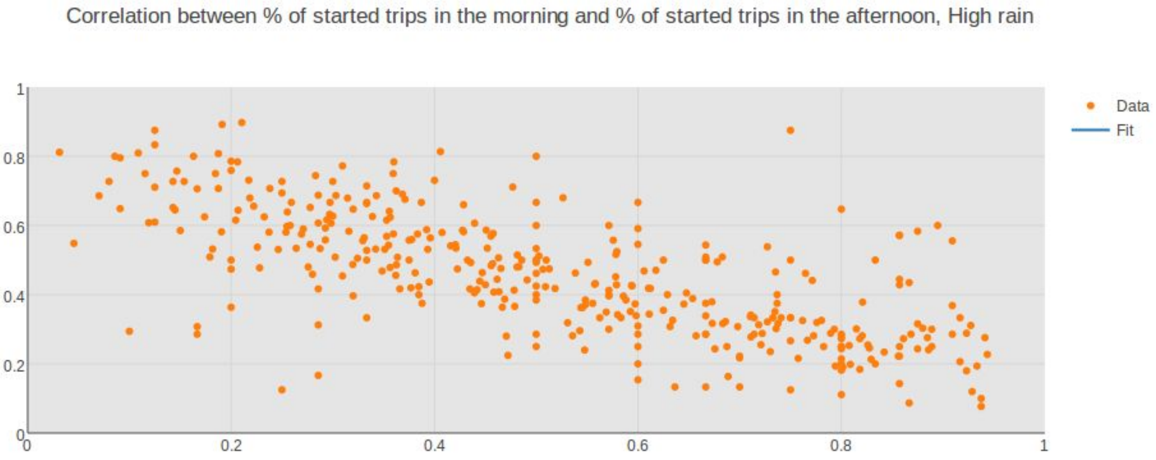
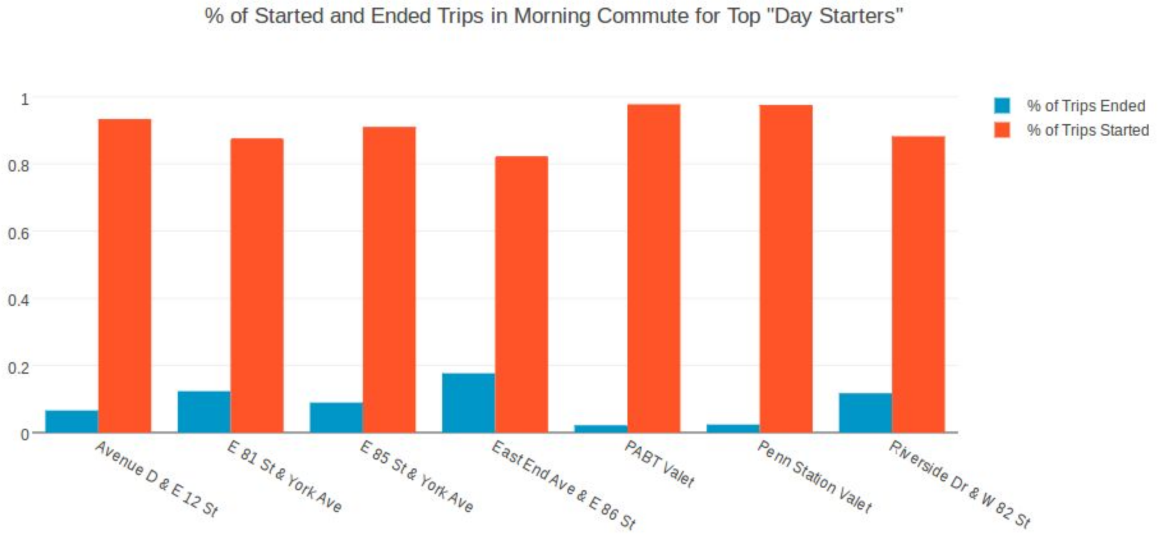
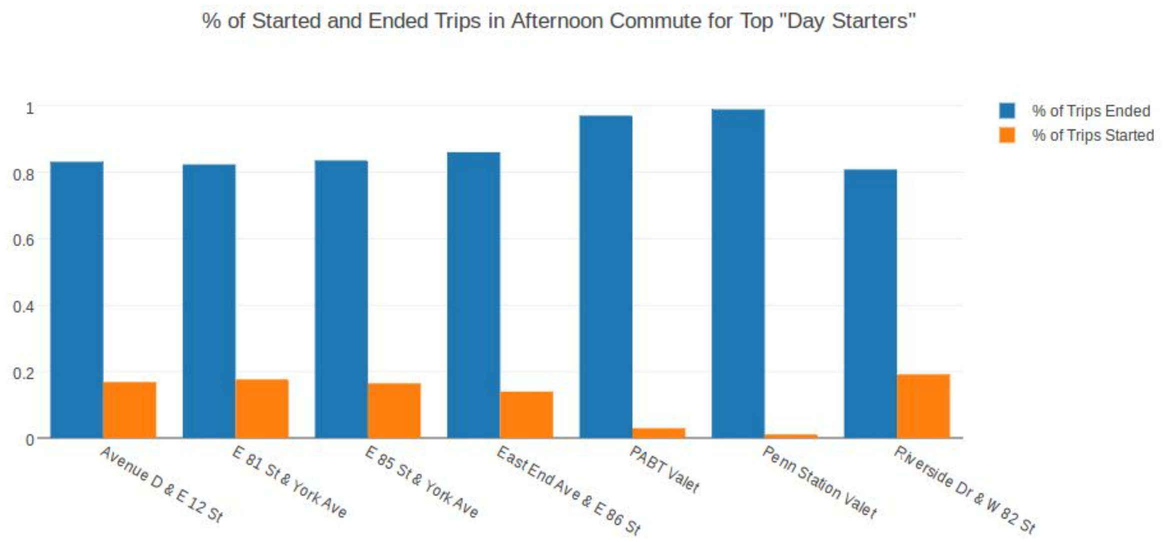


Chart 17

When rain is high, the model does not fit well, and the relationship is not as clear.

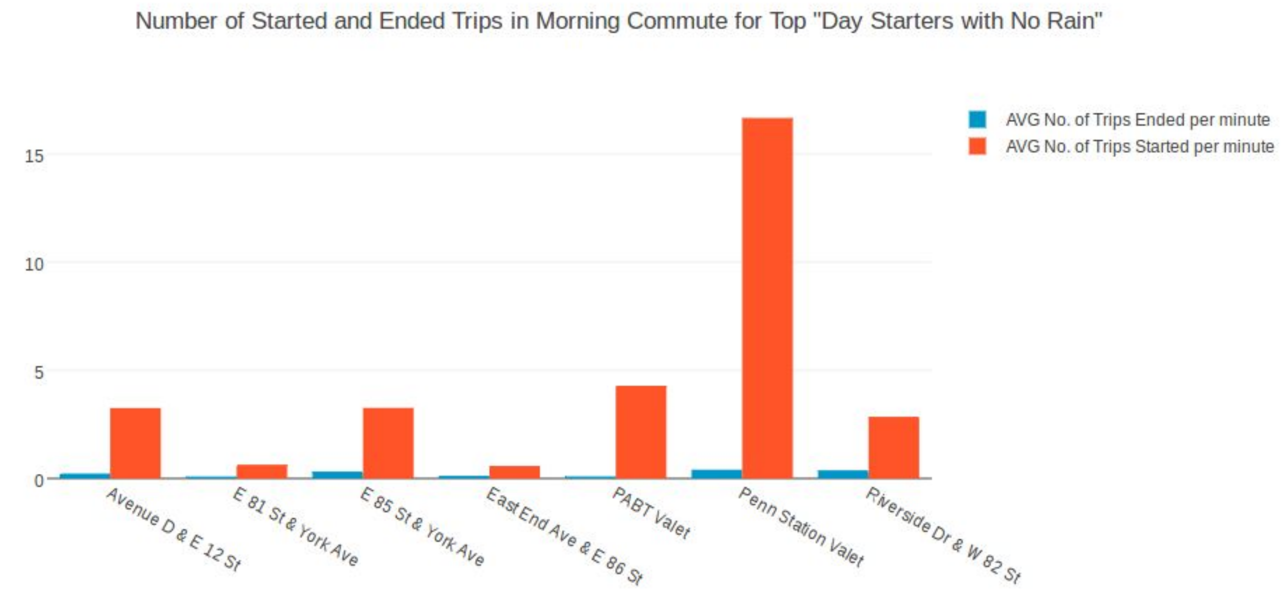
Now, we zoom the analysis on a few stations with extreme "morning trips starter behaviour". If we filter our dataset by the stations that have >80% of trips started in the morning, and >80% of trips ended in the afternoon, we get the following.

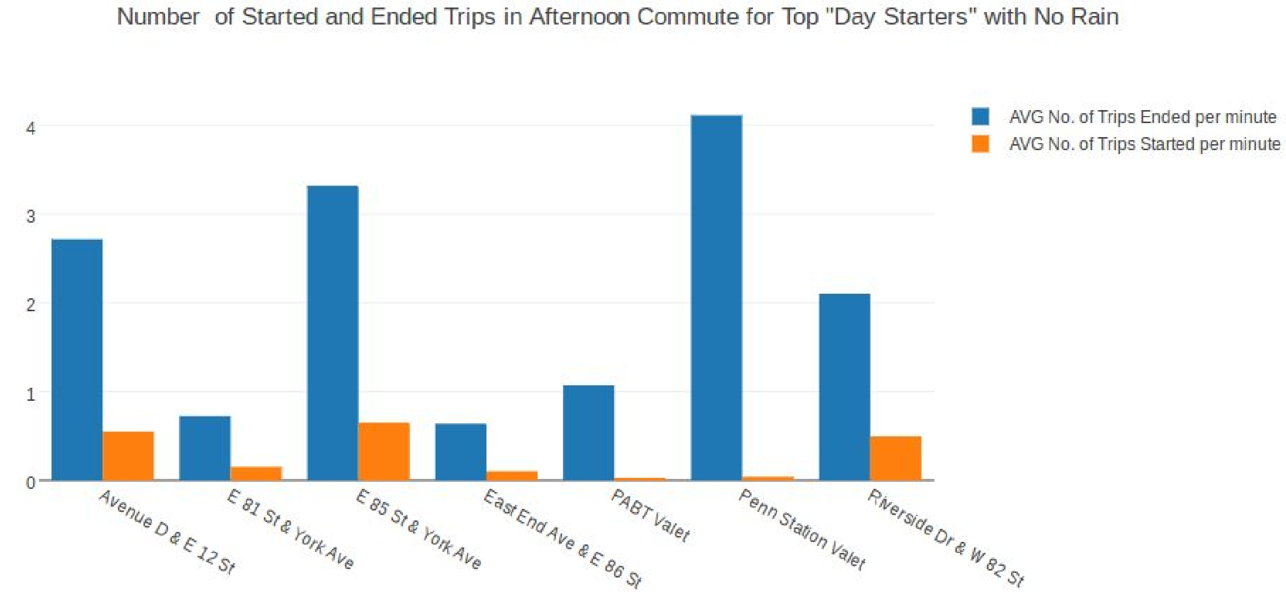



% of Started and Ended Trips in Afternoon Commute for Top "Day Starters"

Charts 18,19

However, this does not tell us which is the amount of trips that each station starts/ends and how does weather impact them. It is important that we analyze the average amount of trips started/ended by raining condition by hour. ***Clarification: The following charts say "per minute" but it is "per hour".***


Number of Started and Ended Trips in Morning Commute for Top "Day Starters with No Rain"



Charts 20, 21

Now we have an idea of the average number of trips started and ended in these stations when weather is good. The next chart shows how this number changes when rain is high.

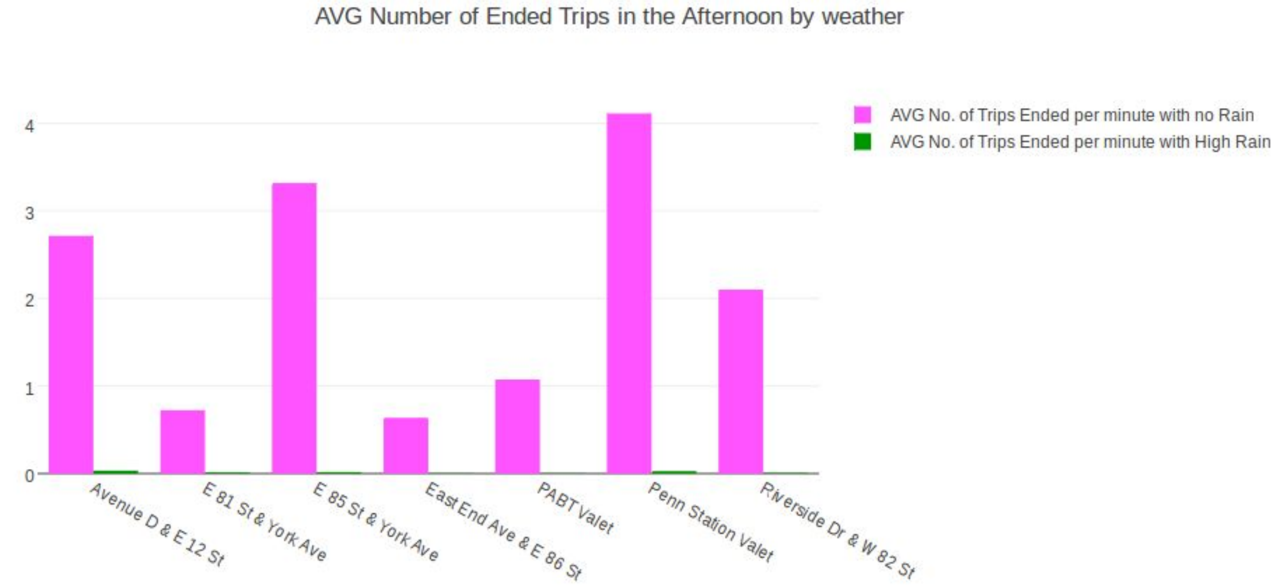


Chart 22

What this chart tells us, is that when it is raining strongly during the afternoon commuting times, the number of ended trips in these stations diminishes dramatically. In other words, all the bicycles that are supposed to return to their place and stay there overnight (so that people can use them to start their trips the next morning) are not returning to their places. This is of crucial importance to Citi Bike.

If we take the example of one particular station, the Penn Station Valet and analyze the time series of started/ended trips, we get the following:

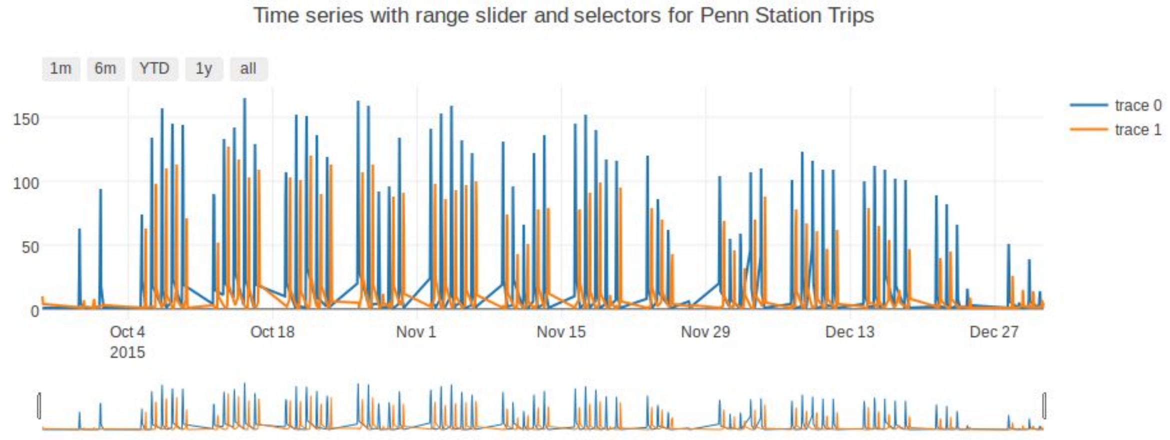


Chart 23

If we zoom to see the intraday behavior:

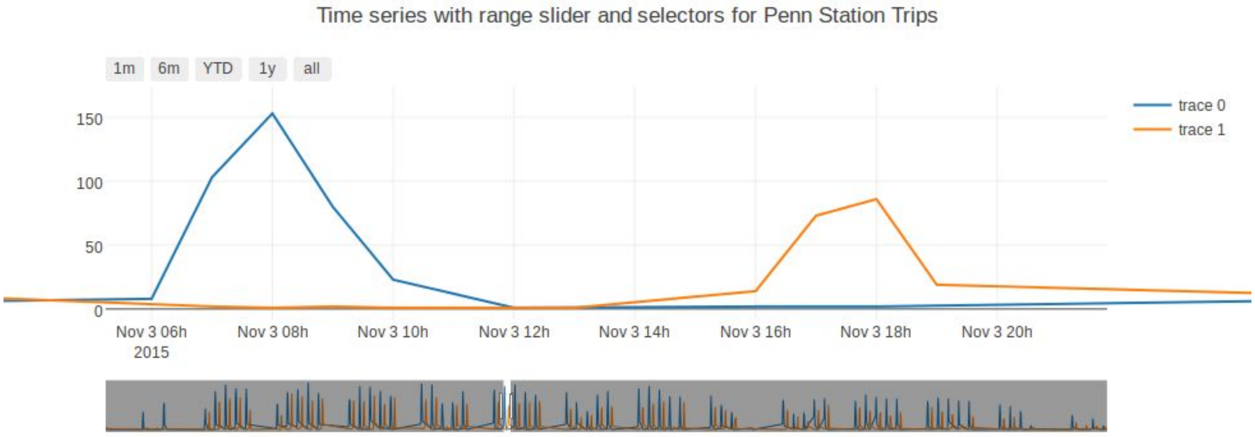


Chart 24

The blue line shows the number of started trips (with a peak in the morning commuting hours) and the orange line shows the number of ended trips (with a peak in the afternoon commuting hours).

So the important point here is how rain affects this behavior. We took an example: On October 28th it was mostly rainy the whole day. In particular, there was high rain at the afternoon commuting times. It we see the number of ended trips at that time, it is considerable lower than on the other days:

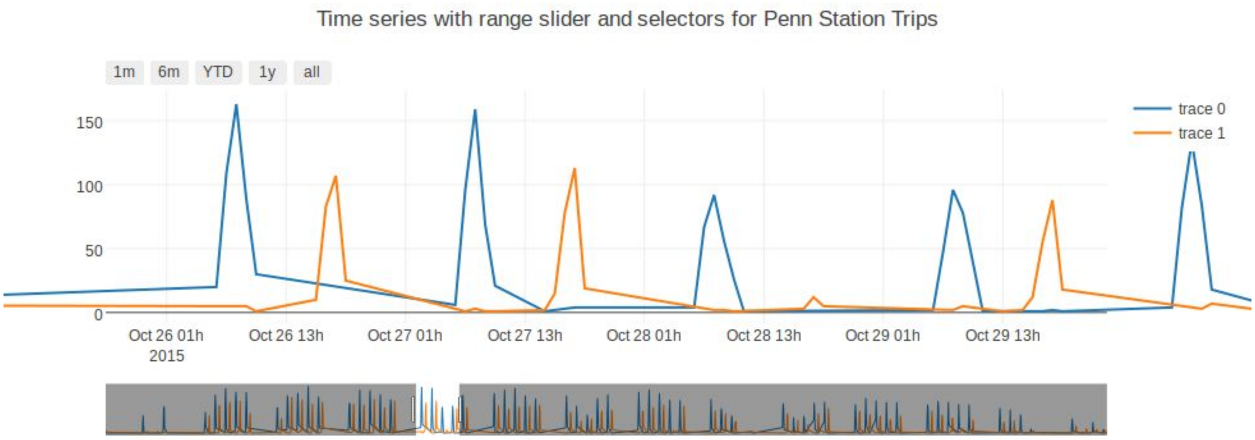


Chart 25

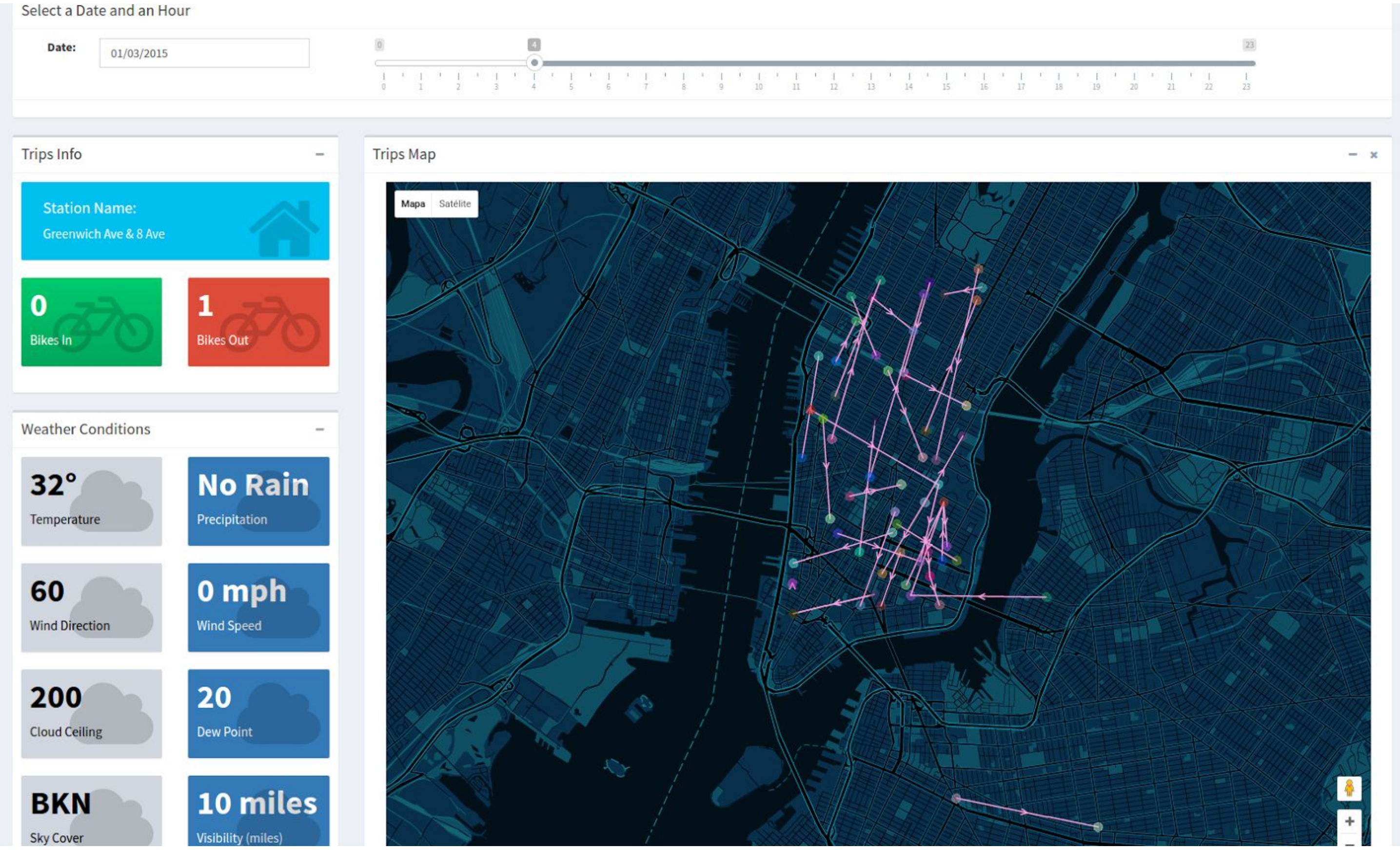
Furthermore, the number of started trips in the morning of October 29th, is also smaller than the average of the other days (probably similar to the one of October 28th, when it was rainy as well). On October 29th in the morning it was not raining.

This might be related to the fact that not all the bikes that were expected to return to the station on the October 28th afternoon actually returned. It is possible that the next day, not everyone that wanted to take a bike at that station had one available, so the number of started trips was relatively smaller.

This finding motivated us to build a Visualization in D3

Visualizing start and end stations in D3

We created an interactive visualization in D3 for the year 2015. User must select a date, and hour, and the system will display all the trips that started and ended between each pair of stations at that time. By moving the mouse over the stations, it is possible to see the exact amount of trips. The system also shows the weather conditions at that particular time.



Select a Date and an Hour

Date: 01/03/2015

4

Trips Info

Station Name: Greenwich Ave & 8 Ave

0 Bikes In

1 Bikes Out

Weather Conditions

32° Temperature

No Rain Precipitation

60 Wind Direction

0 mph Wind Speed

200 Cloud Ceiling

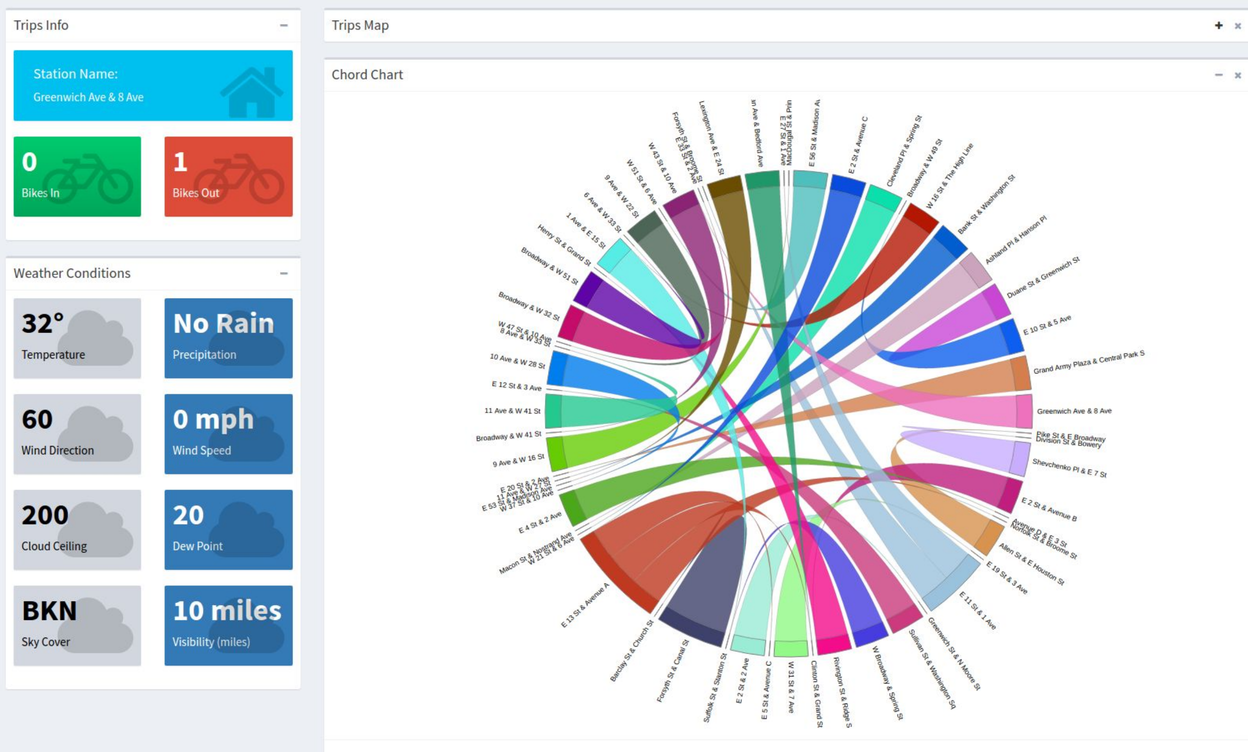
20 Dew Point

BKN Sky Cover

10 miles Visibility (miles)

Trips Map

This Second part of the visualization allows to see the flow of trips between pairs of stations to have a better idea of the quantity of trips at that particular time



Trips Info

Station Name:
Greenwich Ave & 8 Ave

0 Bikes In

1 Bikes Out

Weather Conditions

32° Temperature

No Rain Precipitation

60 Wind Direction

0 mph Wind Speed

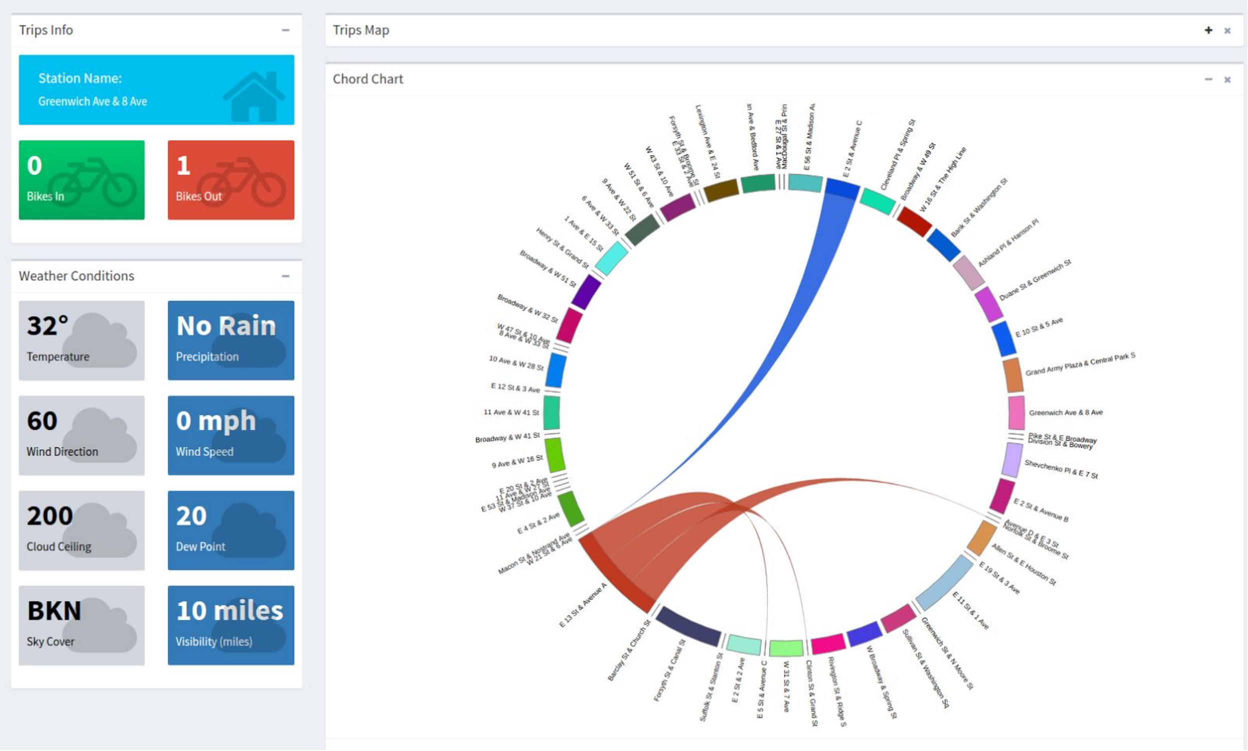
200 Cloud Ceiling

20 Dew Point

BKN Sky Cover

10 miles Visibility (miles)

User can move the mouse over each station, and see the amount of trips started and ended, both with different colours.



Trips Info

Station Name:
Greenwich Ave & 8 Ave

0 Bikes In

1 Bikes Out

Weather Conditions

32° Temperature

No Rain Precipitation

60 Wind Direction

0 mph Wind Speed

200 Cloud Ceiling

20 Dew Point

BKN Sky Cover

10 miles Visibility (miles)

5) Summary, Conclusions and Next steps

After careful analysis of how weather conditions affected Citi Bike trips during 2015 we arrived to the following conclusions:

Rain

- The average trips started per minute when raining are at least less than half than the ones occurring when no raining.
- Women are more sensitive to rain than men. The proportion of women using the service diminishes considerably when raining.
- Short-Term customers are more sensitive than Subscribers to rain.
- The average trip time diminishes when raining, which is partially (but not completely) explained by a decrease in average travel distance.
- Rain seems to affect in similar manner users from all ages. There is no significant decrease in the proportion of any particular age when raining.

Wind and Temperature

- Genders and age are insensitive to temperature and wind changes.
- There are almost 20 times more trips occurring on warm weather compared to cold weather.
- The variance of the trip duration in a warm day is around two times the variance during a cold day (36 min vs 64 min).

Rain affecting stations behaviour

- There are stations that are clearly used for commuting purposes, showing extreme behaviours during afternoon and morning rush hours.
- Because of this reason some stations work as “day docks” and other as “night docks”.
- Rain clearly affects this relationship in percentage terms.
- Furthermore, rain affects the quantity of trips ended at different stations, which changes the normal flow of bikes (and though the quantity of bikes available at those docks).
- Ultimately, this can have a lasting effect on the following day or the following afternoon commutes, since users might not find the expected amount of bikes at their frequented docks.

Implications to Citi Bike and Next steps.

The fact that users do not find their expected amount of bikes at their frequented docks can affect both customer satisfaction and Citi Bike's profits. Citi Bike could address this issue by moving bikes from one dock to another when weather conditions are not appropriate.

Because of that, we suggest as a next step to build a Machine Learning model that can predict the amount of trips started and ended at a particular station at a particular time, given weather conditions. In that way, we could roughly estimate any shortage of bikes at a particular station based on the weather forecast, and visualize this shortage in the D3 visualization that we built.

This tool could help Citi Bike move their bikes when required, and improve customer satisfaction.

6) Individual Contributions

The three members of the Project contributed equally.

7) References

- Sandar Tin Tin, Alistair Woodward (2012). Temporal, seasonal and weather effects on cycle volume: an ecological study

- Todd W. Schneider (2013). A Tale of Twenty-Two Million Citi Bike Rides: Analyzing the NYC Bike Share System

- I Quant NY. Mapping Citi Bike's Riders, Not Just Rides

LOGS OF THE PROJECT

1) Project Proposal

Goal of the Project

After many discussions, we decided to work with City Bike and NYC weather datasets. The goal of the project is to find interesting patterns between the weather in NYC and changes in the City bike trips. Some open question that we are willing to answer, include:

- How does variance of the City Bike usage changes across stations/neighbourhoods depending on weather conditions. For example, does usage of bikes in downtown neighbourhoods decrease more/less than the usage in residential neighbourhoods when it is raining?
- Is a certain age/gender group variance in the number of trips more sensitive to weather conditions than the others?
- Can we quantify which is the impact of bad weather on trip durations?
- Does the proportion of the different types of customers (subscriber/customer) changes with different weather conditions?

Extra Datasets

We are also looking forward getting some data about bike trips in general (not only City Bikes) in order to compare how City Bikes usage changes with certain weather conditions vs. changes in private bikes usage. One possible dataset that we are trying to get is a sample dataset from Strava. We are also open to finding more weather dataset that could provide more granular information.

Visualization

After we analyse the dataset and patterns, we are looking forward to putting together a nice visualization. For that purpose, we will build a map with either D3, Basemap (Python) or Mapbox - to be defined. The idea is that users could select a certain time/day and we will show: a) which weather conditions at that point b) and how many trips were initialize and ended in each station c) On the side, any relevant statistic associated with the questions expressed above.

To be discussed with professor: Is it necessary to include reproducibility of the visualization as well or only of the statistics/answers to the questions? In that case, how could we do it?

Tasks to perform and Timeline

04-11-2016 to 04-18-2016: Use Hadoop or Spark to preprocess and aggregate City Bikes data in order to have the same granularity as the weather data. Preprocessing includes fixing data issues such as invalid and Null values.

04-18-2016 to 04-25-2016: Perform Exploratory analysis in order to find interesting patterns between City Bikes and weather data. The goal is to answer the questions described above and any other question that might arise in the process.

04-25-2016 to 05-02-2016: Select the right visualization tool to show that relationships and make it as user friendly as possible.

05-02-2016 to 05-09-2016: Create visualization based on processed and aggregated data and beta test it with some classmates.

05-09-2016 to 05-13-2016: Finalize Project Report, Ensure Documentation and Reproducibility.

Status Report - 04/17

- No issues encountered so far with the data.
- Starting to clean the Weather dataset
- Building the Map Reducers to process the City Bikes Data and aggregate it.

Status Report - 05/01

- Conducting exploratory analysis with weather and city bikes usage
- We cleaned and interpolated the weather data in order to create one record per each minute occured in 2015. Used python for this purpose
- Added the GMT timestamp column to the citybike data using spark, and joined both datasets.
- Conducted exploratory analysis based on aggregated data. Some early findings:
    - Rain does not change the distribution of user by age.

- Rain has negative correlation with average travel time. This is partially explained by the fact that users make smaller distances when raining.
- In general, more trips are made with higher temperatures (except very extreme temperatures)
- Rain diminishes the proportion of women users significantly.
- Rain diminishes the proportion of non-suscriber users significantly.