Assignment 4: Tableau for Business Decisions

## Import Data Into Tableau

### Process:

1. Import ride data and merge all monthly datasets into one dataset containing all bixi rides for the year
2. Import station data as a single CSV file
3. Inner join the merged ride data with the station data on start\_station\_code (from rides) and code (from station data)

### Challenges:

* I found the first time I merged the ride datasets to be a little challenging. I am familiar with joins methods from SQL so that part was intuitive but merging as a “union” was some terminology I had to get used to
* Getting tableau to remember this union proved challenging. For one reason or another some times when I opened up Tableau again after closing it the union had been disturbed and I had to re-specify the filepaths

## Sort the number of rides per day to figure out the top 3 days of the month.

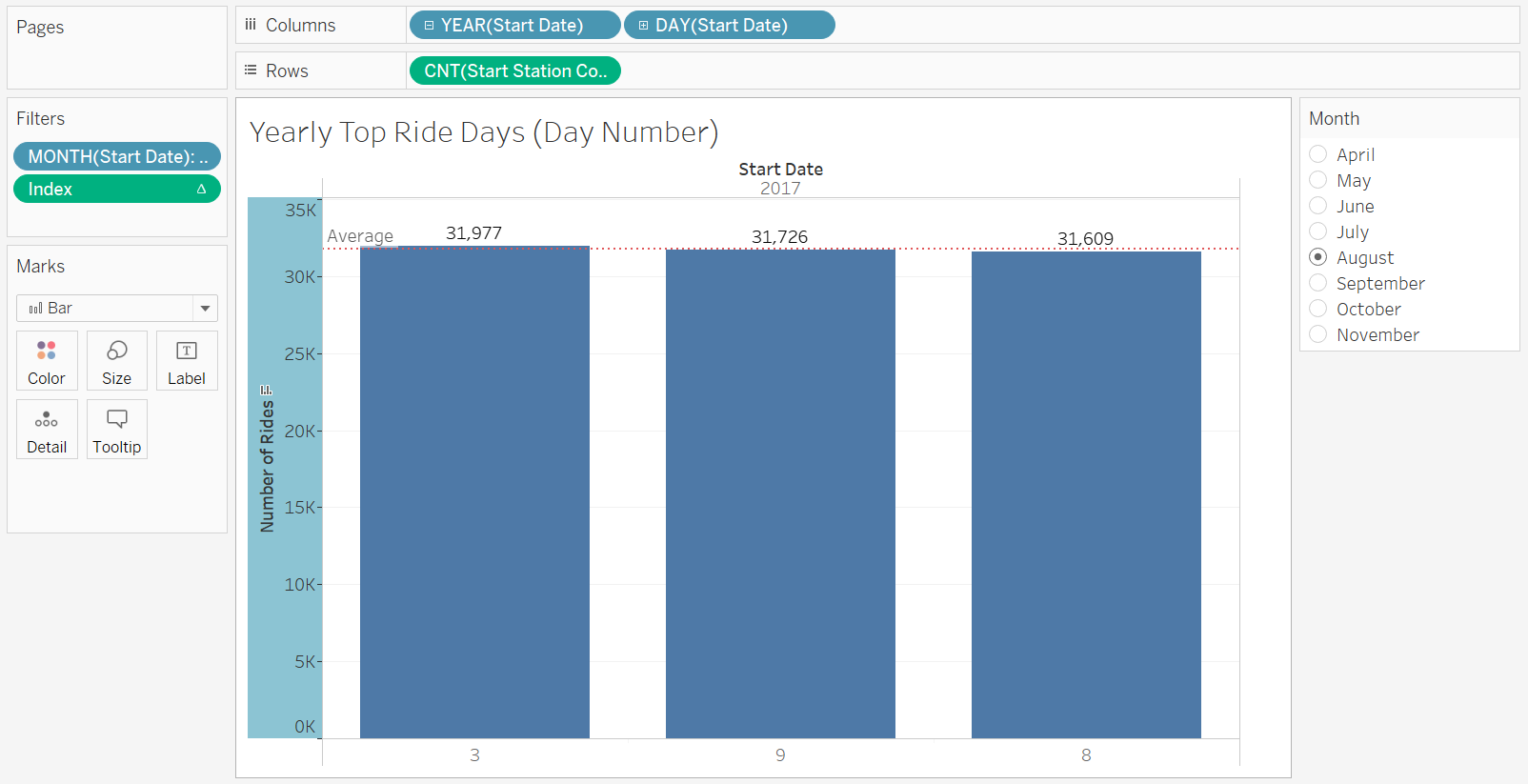
### Process:

1. Open a new tableau sheet and configure the columns by date in the format year, month, day
2. Add the metric count by start station code to the rows field
3. Create a filter on the month of the start date & display the filter on the side of the sheet
4. Configure the filter to make it show by only one month at a time, rename it, and remove the “all” view
5. Sort by largest count to smallest
6. Add an index to the filter to only show the top 3 results in the graph per month
7. Add average line in y axis

### Challenges:

* Intuitively understanding what was happening to the date fields in the columns was difficult. Eventually I understood that if the type of date wasn’t there (for example Quarter) then you couldn’t filter by that type of information. Since I was trying to filter by month and show individual dates in each column I knew I needed to have both the month and day field present in the column
* At first understanding the Index function filter was difficult. Seeing how the index interacts with the filtering order of the graph was a little unintuitive for me – I thought that maybe the index filter could (itself) filter the top 3 day values (perhaps it can but I haven’t figured it out yet).

### Results:



## For the month, which Starting point Station has the most rides being originated?

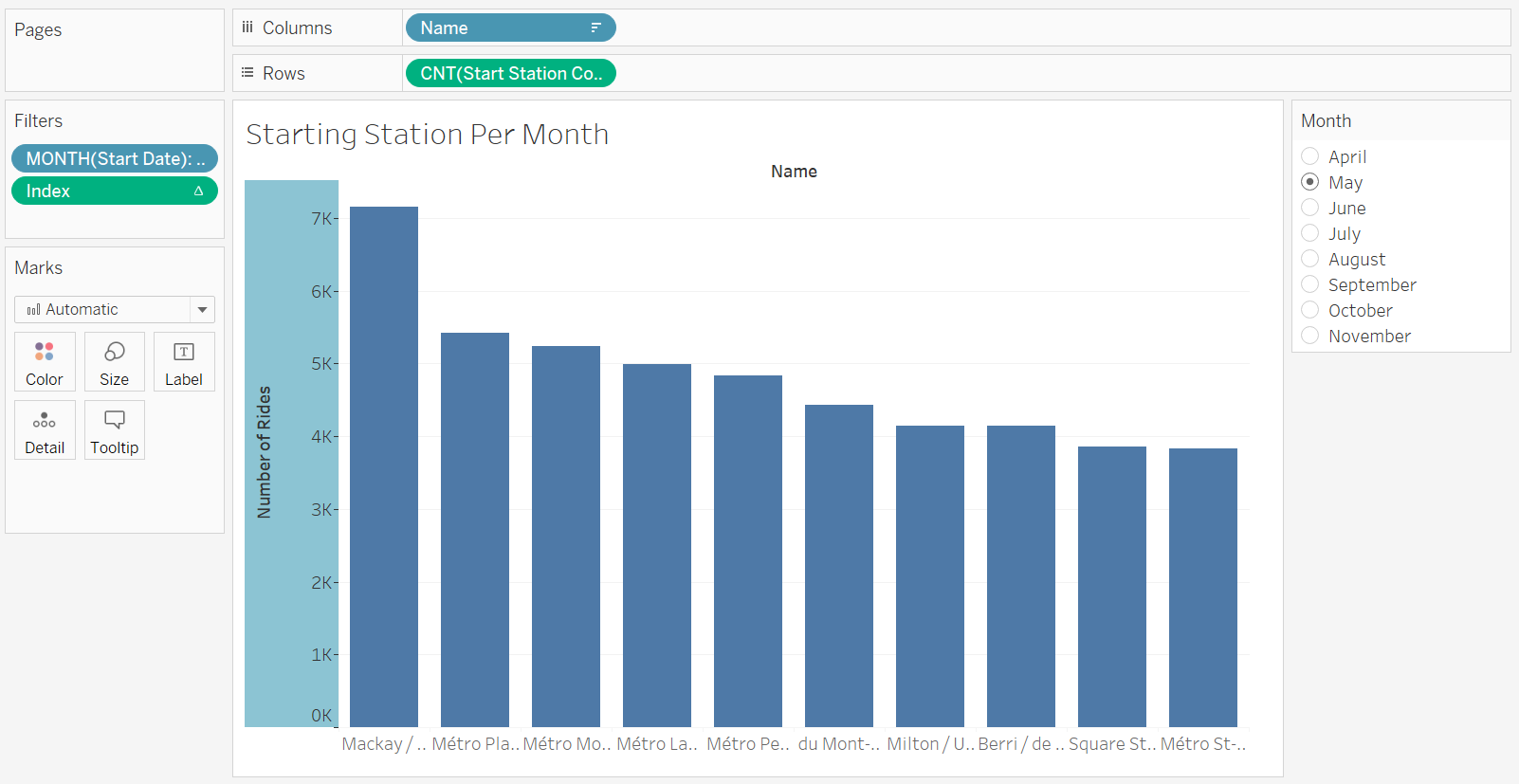
### Process:

1. Duplicate graph showing the top ride days per month
2. Replace the column field with the name of the station from the inner joins
3. Rename all necessary axes & titles
4. View the top station count for each month by flipping through the month filter

### Challenges:

* The only challenge I faced with this task was trying to figure out how I could leverage the first sheet I made to create this new one. I realized that I want (essentially) the same style of graph as the first problem but I wanted to expose station information instead of date information. Switching the columns fields out allowed me to do this

### Results:



## For the month, provide the top 10 Starting point Stations (you can do the same for End Stations). Can you map them?

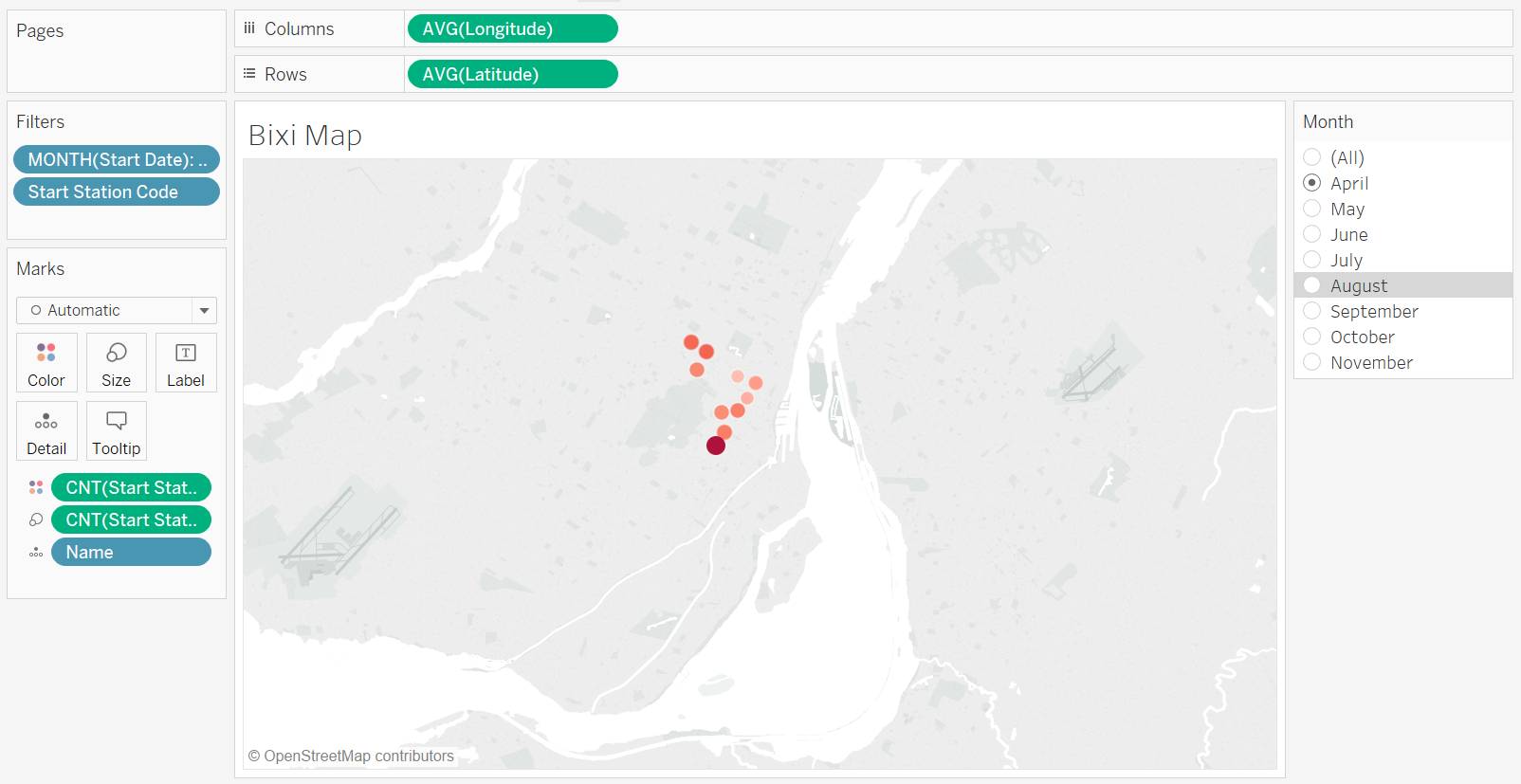
### Process:

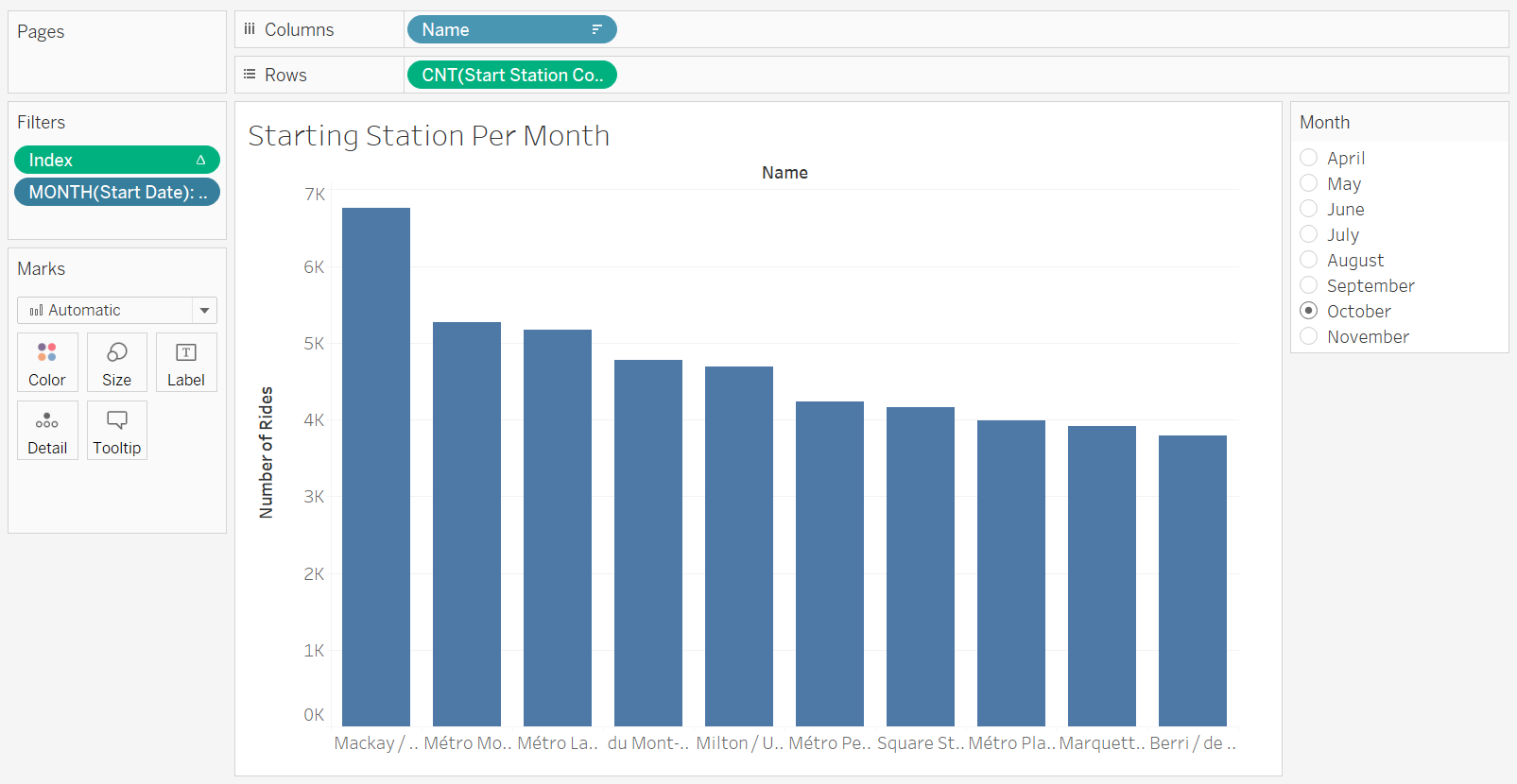
1. Leverage the graph created in question 2 to flip through each month to read off the top 10 stations by ride count
2. Leverage the graph in question 2 & switch out the count from start station code to end station code. Flip through each month to read off the top 10 stations by ride count
3. Mapping the stations:
   1. Put the longitude value in the columns field and the latitude in the rows field. Configure the graph to be a map of montreal
   2. Add the name of the station as a detail
   3. Add the count of the start station to be a color and size mark (redder and larger represents a higher count of rides starting from that station
   4. Add a filter by month of the start date
   5. Add a “Top” filter by start station code and limit it to 10 stations

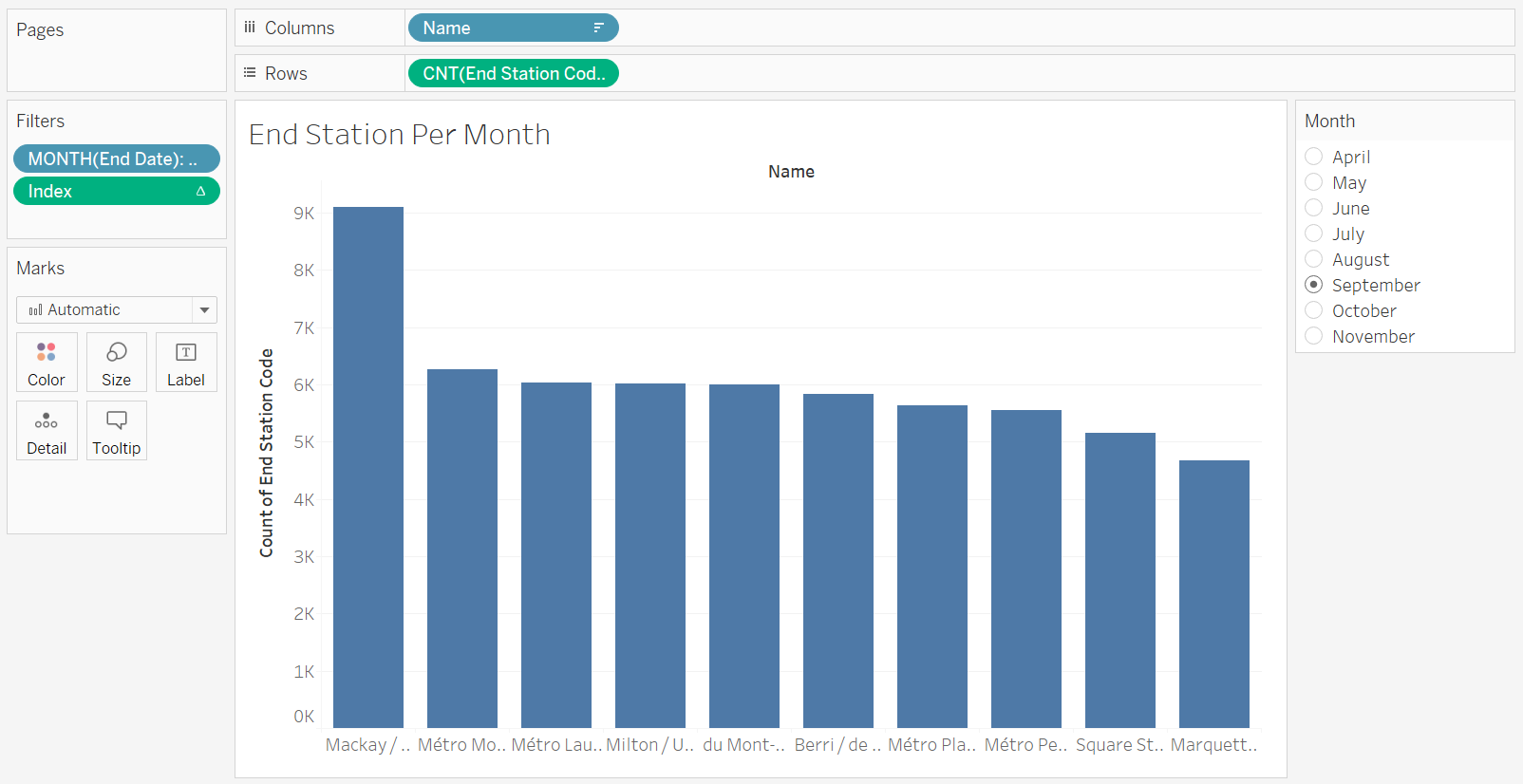
### Challenges:

* The largest challenge in this section was creating the map and specifically filtering it by the top 10 count of stations
* Configuring the filters was tricky – I had to play around with a bunch of features until I figured out how to use the “top” filter by start station code
* After adding the top filter the map still wasn’t behaving correctly – it didn’t look like it was updating with the change of the month on the date filter
* After trying to debug the filter sections I realized my real issue was that my marks section wasn’t configured appropriately & was therefore displaying the wrong information (I wasn’t configuring size and color by the count of the start station code it was just by the start station code attribute itself)

### Results:







## Which day of the week (Monday to Sunday) has the most rides?

### Process:

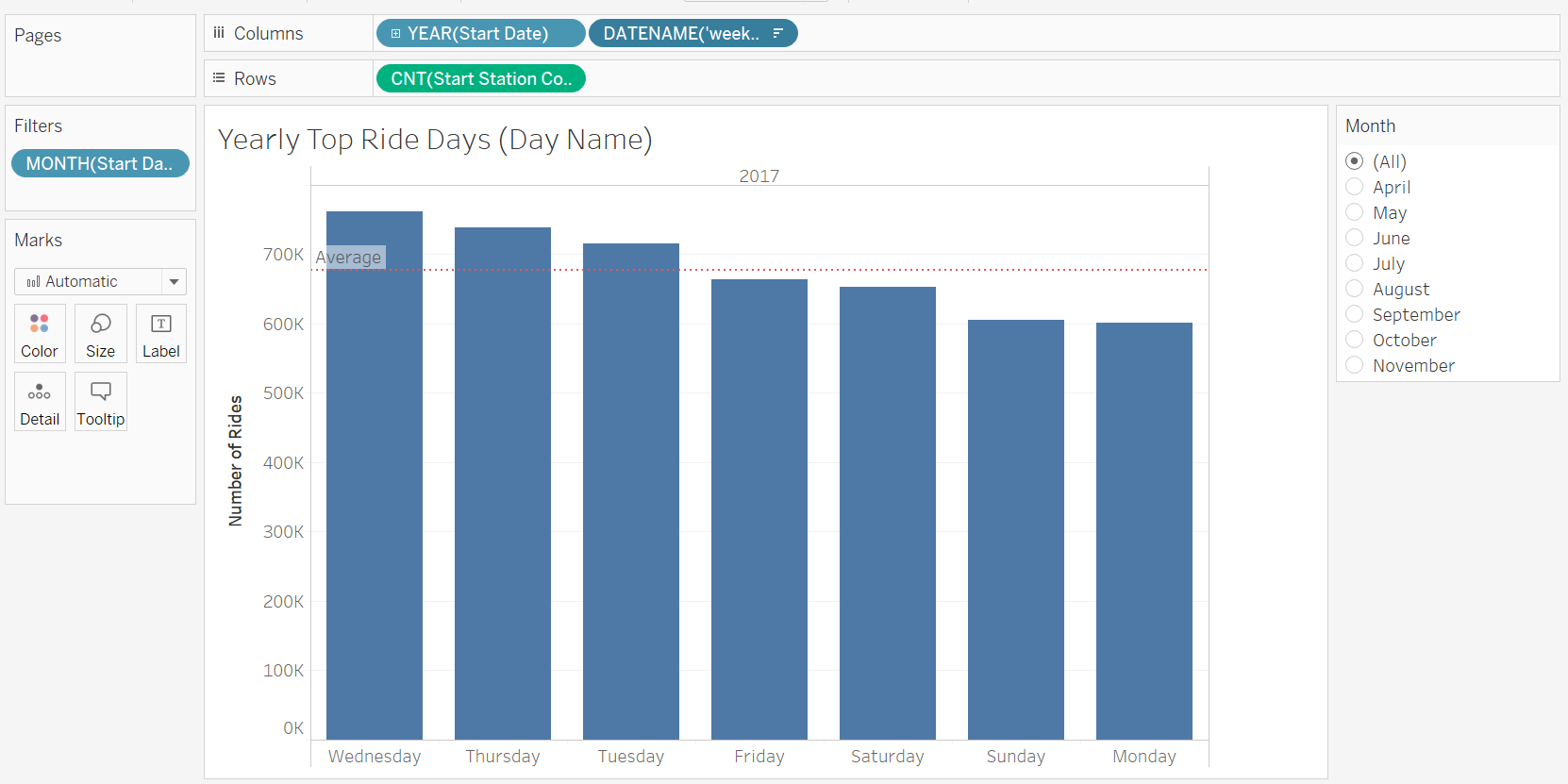
1. Create a new worksheet and add the Year and DateName of the start station code to the column field (for datename the formula being DATENAME('weekday',[Start Date]))
2. Add the count of the start station code to the row field
3. (Optional) add a filter on the month of the start date. For our purposes this was not necessary but I left the “All” field exposed in the filter to view information this way

### Challenges:

* The challenge in this activity was adjusting from displaying the date number of the start date to the date name (Monday through Sunday)
* At first I thought maybe I would have to add a new column name to the dataset but thought there must be a better way to expose this information – and sure enough after some googling I landed on the format DATENAME('weekday',[Start Date]).

### Results:

**Wednesday has the most rides**



## Which day of the week (Monday to Sunday) has the least rides?

The process/challenges for this section can be seen in the previous answer.

### Results:

**Monday has the most rides**

## Characterize the data: how many rides in total, average time per ride, % member rides,…

### Process:

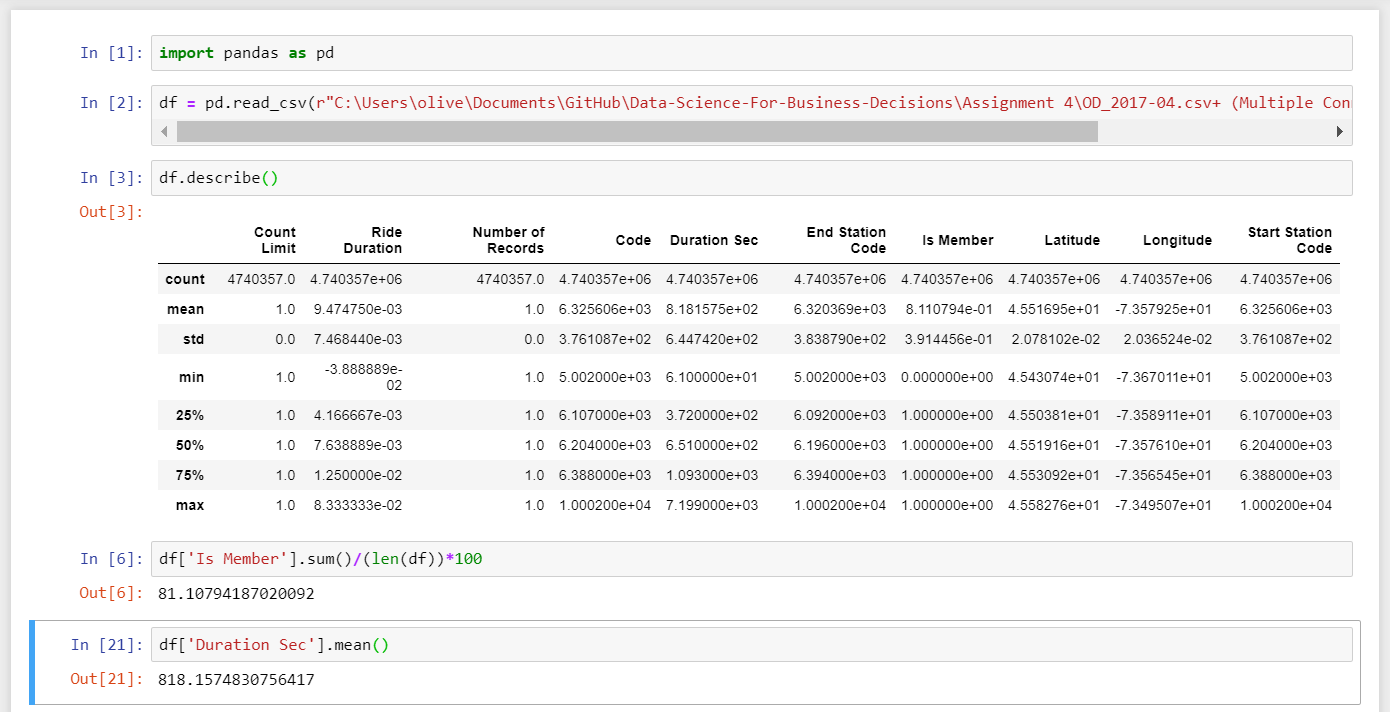
1. Export joined dataset from Tableau to CSV
2. Open a Jupyter Notebook and import the CSV using pandas
3. Use describe function on the imported dataframe
4. Calculate the % Member based on the Is Member column

### Challenges:

* I spent some time trying to figure out the smartest way to display the required information in Tableau then finally came to the conclusion that I did not think it was the best tool for the job. Instead I took the CSV and imported it into a pandas dataframe and used the describe function
* I was looking for a generic “describe the data” dashboard functionality from Tableau but I couldn’t find one that suited what I wanted – so I decided to switch tools

### Results:

Below you can see my described dataframe showing that there were 4,740,357 rides in total, with average ride duration of 818 seconds, and an 81% is-member rating.



## Provide the number of rides per day (day of the month 1-30)

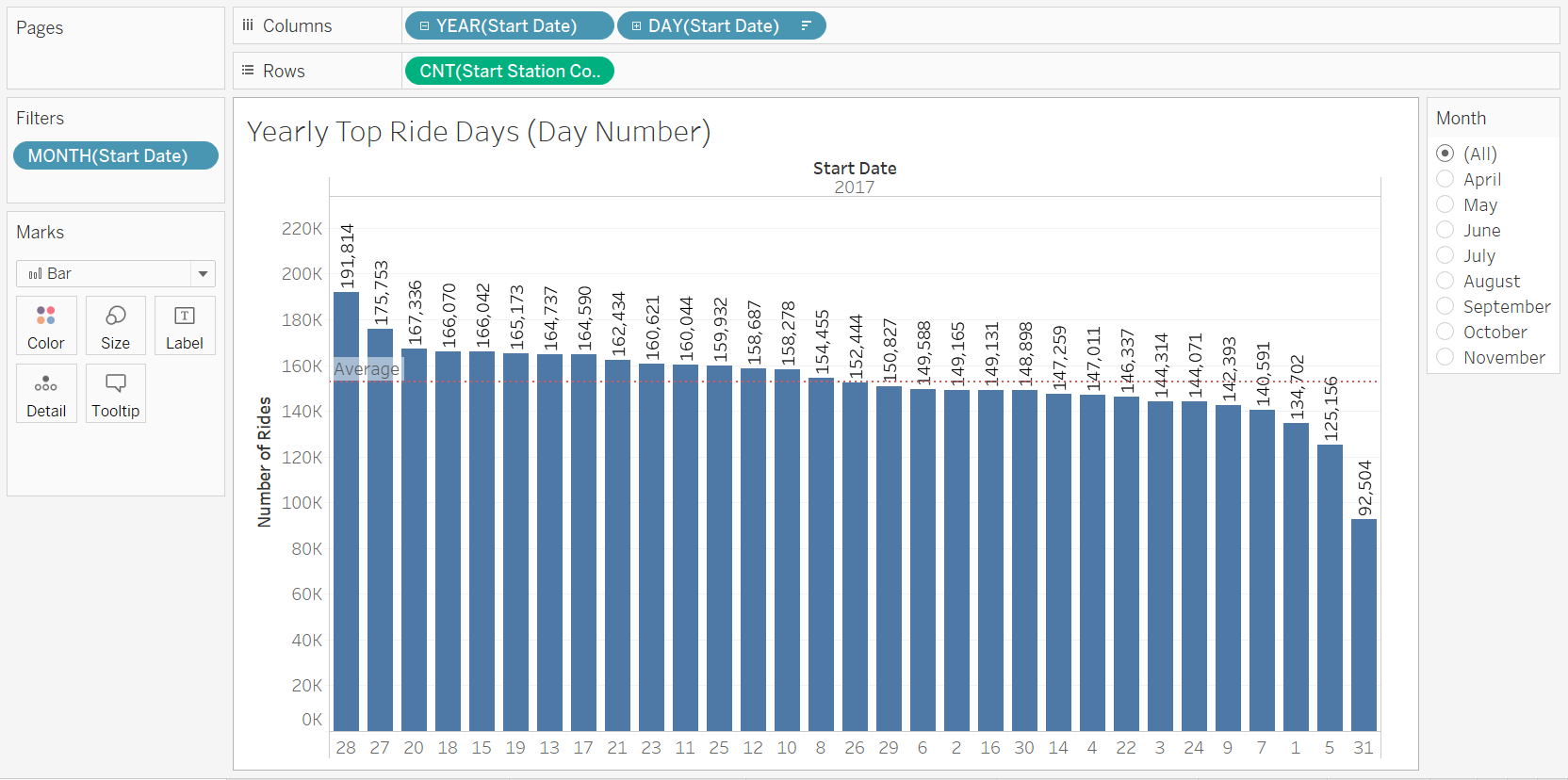
### Process:

1. Reformat results from question 2 to remove the monthly filter and display all days (1-31)

### Challenges

* See response from question 2 for challenges

### Results:



## Please provide the names of the top 10 Starting Point Stations

### Process:

1. Create new worksheet and put the Name in the column field
2. Add the count of the start station code for the row field
3. Filter the results in descending order and place an index filter set to limit 10 values

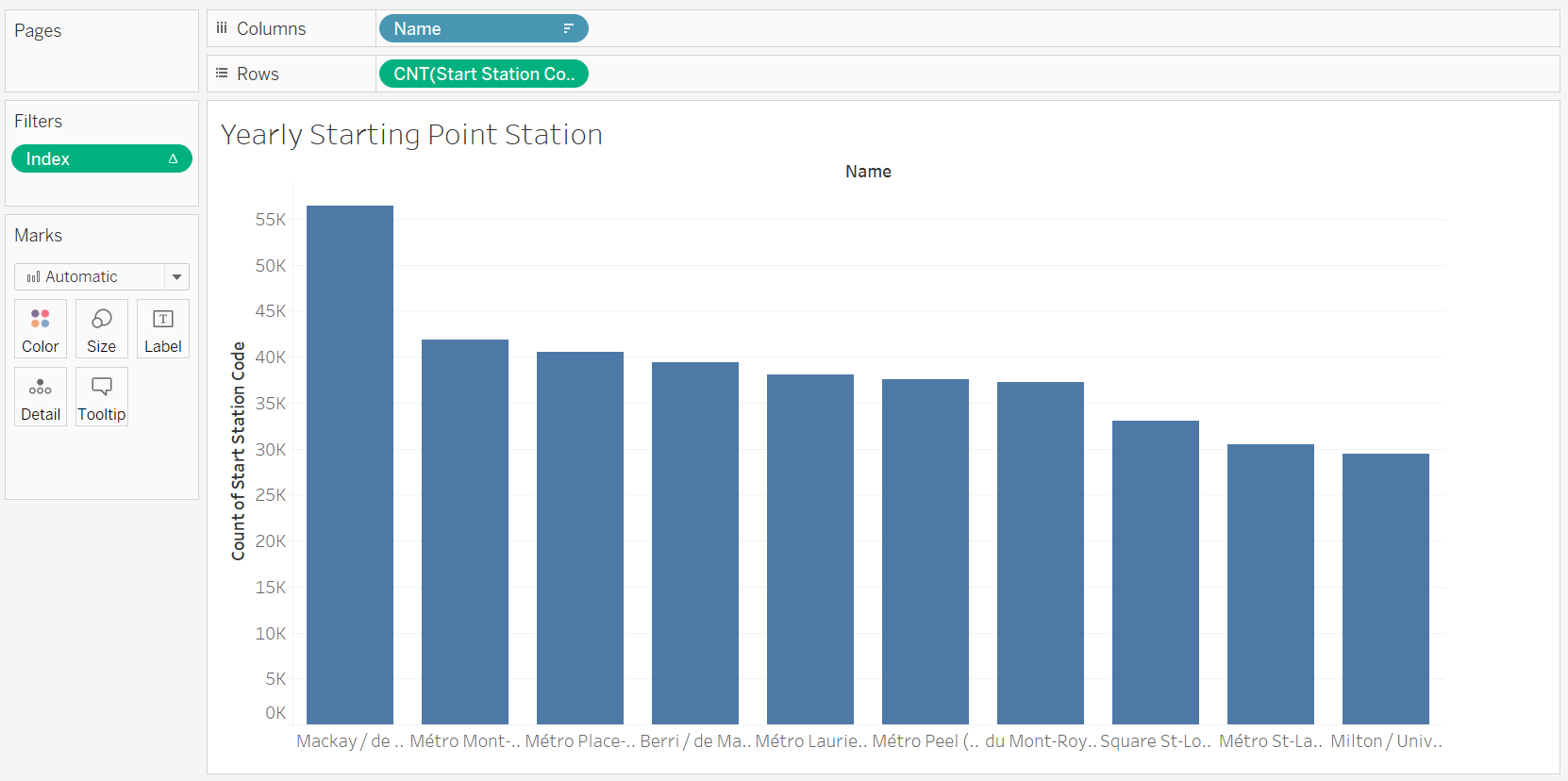
### Challenges:

* Given the exercises done previously in this assignment this was a straight-forward task without any major challenges

### Results:

Top 10 Stations by Ride Count:

|  |  |
| --- | --- |
| **Name** | **Count of Start Station Code** |
| Mackay / de Maisonneuve | 56,380 |
| Métro Mont-Royal (Rivard / du Mont-Royal) | 41,836 |
| Métro Place-des-Arts (de Maisonneuve / de Bleury) | 40,525 |
| Berri / de Maisonneuve | 39,413 |
| Métro Laurier (Rivard / Laurier) | 38,063 |
| Métro Peel (de Maisonneuve / Stanley) | 37,570 |
| du Mont-Royal / Clark | 37,287 |
| Square St-Louis | 33,056 |
| Métro St-Laurent (de Maisonneuve / St-Laurent) | 30,469 |
| Milton / University | 29,473 |



## How many bike rides originate and end at the same station? Are there any patterns?

### Process:

1. Create new sheet and create a new metric called “Start&End” that contains the following formula:

IF [Start Station Code]==[End Station Code] THEN 1 ELSE 0 END

1. Add the month of the start date to the column field
2. For the row field enter the following:

SUM([Start&End])/COUNT([Start Station Code])

1. Format the y axis of the graph to display percentages and change the names of the axis/chart

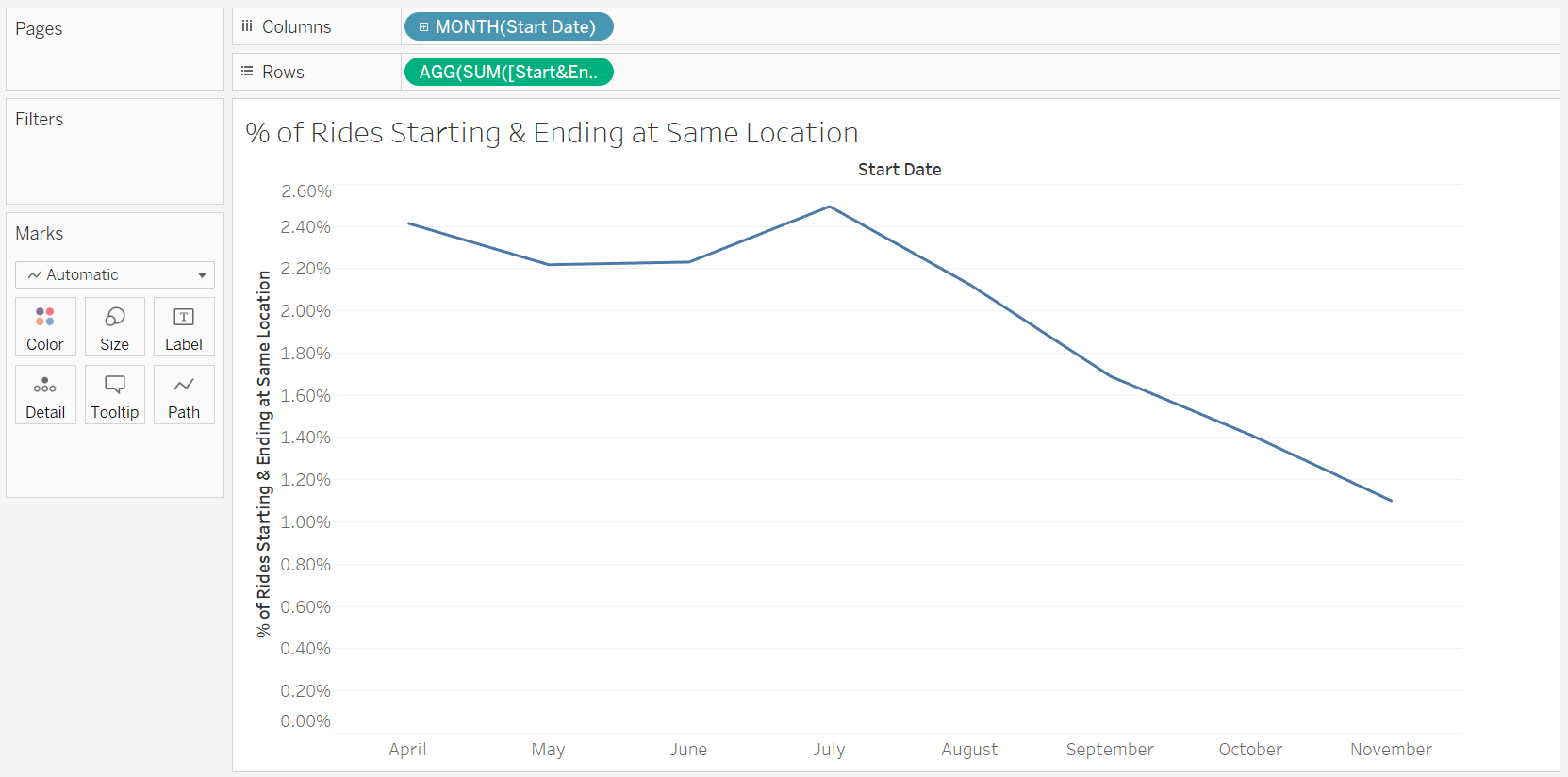
### Challenges:

* I knew I wanted to display the trend of the starting & ending Boolean over the months of the year but I struggled to structure my data in a way that this would show well in chart format
* I started looking at Pie charts and displaying one pie chart per month, but the “True” slice was so small that I couldn’t visually pick out a trend
* Finally I was able to articulate a formula that returned the % of rides starting & ending at the same location and plotted that as a line graph

### Results:

I noted that there was a decreasing trend over the year in terms of rides starting and ending at the same location. There is a local bubble in the summer season where this percentage briefly increases but overall it seems more people are starting & stopping their Bixi at the same stop towards the beginning of the season.

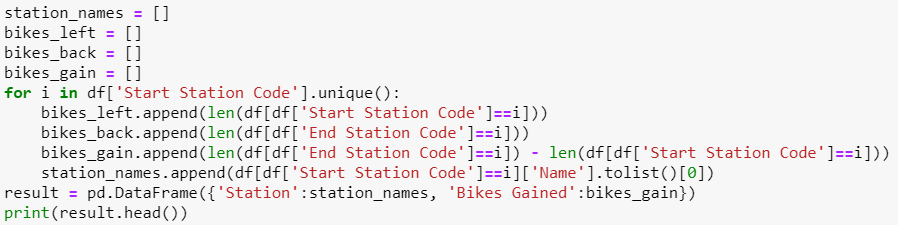
A reasonable hypothesis might be that people are just excited to get back on their Bixi at the start of the season so they’re just taking it out for fun. Once the excitement wears off throughout the season they’re using it for more utility purposes which is why that start & stop percent decreases through to November (commuting takes over as the primary use of Bixi instead of just taking it out for a spin in your neighborhood)



## On a daily basis, which stations are net-gainers (more bikes in than bikes out) and which stations are net-losers (more bikes out than bikes in)

### Process:

1. Export dataset from Tableau
2. Open data in pandas via Jupyter Notebook
3. Calculate bike gain deltas for each station. See code below:



1. Use Pandas built-in bar chart to plot the 5 largest net gainers and losers by station for all of 2017

### Challenges

* I started this process in Tableau and was having a hard time articulating to the software exactly how I wanted it to handle the data
* After trying some aggregations (COUNT([Start Station Code])– COUNT([End Station Code]) I realized I was spending a lot of time on this issue without progressing much when this would be much simpler if I used pandas
* In pandas I finished this process in a couple minutes

### Results:

