BikeShare and Parking Tickets in New York

MIDS W200.3 (Kleeman)
Kanitha Mann, Nach Mohan & Daniel Olmstead
April 18, 2018

Objective

The purpose of this project is to offer an analysis of available public data on CitiBike (NYC) bikeshare usage and New York parking violations. These two datasets will be analyzed separately for information particular to each of them, and then in tandem for a specific neighborhood in Brooklyn that installed CitiBike docks in August 2016 to see whether and how the installation of docks had any effect on parking violations.

Background

New York is a famously difficult place to find parking. The city writes about 10 million parking tickets a year (~6 million to passenger cars), despite only 45% of households having a vehicle (that's an average of 4 tickets per household with vehicle per year). A third of these tickets occur in Manhattan. Total annual income to the city from parking tickets alone is over $400m, which sounds like a lot but is small compared to the subway's annual farebox revenue of $6.2B or the MTA toll revenue of $2B.

The parking nightmare is a function of the traffic. New York is the third-worst city for congestion in the world (tied with Moscow at 91 hours per year stuck in peak-hour traffic for the average driver, beat out only by Los Angeles). Nobody beats New Yorkers for economic cost, though, as the value of fuel, time, freight and business fees add up to a $16.9B cost to the city.

In an effort to address congestion, in 2013 New York partnered with private company Motivate and sponsor Citigroup to introduce CitiBike, one of the first and largest city wide public bikeshare programs in the world. Starting with 332 stations and 6,000 bikes, the program was far more popular than expected (70k members in the first month alone), and CitiBike scrambled to keep up with demand. Annual expansions have increased the footprint to 706 stations carrying 12,000 bikes, and currently CitiBike is working on implementing a dockless bikeshare system (like Spin in San Francisco) that would have no set geographic boundaries.

In August 2016, CitiBike expanded into Community District 306 in Brooklyn, a diverse set of neighborhoods including posh Park Slope and industrial Gowanus. Notably it also includes Red Hook, one of the rare bywaters in New York with no subway access.

Hypothesis

Our hypothesis is that the introduction of CitiBike into a neighborhood should reduce parking violations, as it provides an alternative to driving, particularly in transit-starved areas like Red Hook.

Focus

We have divided the project into three sections, with attendant questions:

1. **What are the patterns for New York City Parking Tickets?**
    a. Where and when do they occur most frequently?
    b. What can we say about who gets them?
    c. What are the most common violations, and in what ratio?
2. **What are the patterns for CitiBike?**
    a. Where do people go with them, and when?
    b. How has the service grown since 2015, in terms of footprint and usage?
    c. For the target zone in Brooklyn, what are people using their bikes for?
3. **Does CitiBike affect parking tickets?**
    a. Does the rate of tickets change in terms of either quantity or revenue?
    b. Do the type of violations change?
    c. Does the geographic location of tickets change?

Data Sources

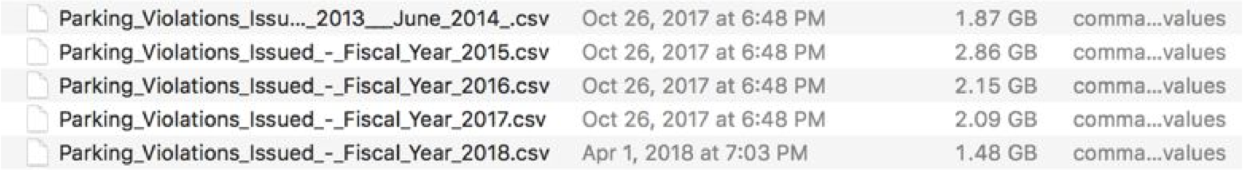
Our data comes from several sources:

- CitiBike System Data
- NYC Parking Tickets (FY2013-FY2017, FY2018)
- NYC Political and Administrative Districts Metadata
- NYC Parking Violation Codes and Fines

Data Exploration and Cleanup

Parking Tickets

New York City (NYC) Department of Finance collects data on every parking ticket issued in NYC and makes this publicly available to aid in ticket resolution and to guide policymakers. Each year's dataset has 43 columns and between 10-11M records, taking up over 2GB of space per file. Data for the years 2013 to 2017 are available in the following 5 files roughly organized by fiscal year (July 1-June 30). Overall 50 million records had to be analyzed. Getting this down to a manageable size was our first priority.



Parking_Violations_Issu..._2013___June_2014_.csv	Oct 26, 2017 at 6:48 PM	1.87 GB	comma...values
Parking_Violations_Issued_-_Fiscal_Year_2015.csv	Oct 26, 2017 at 6:48 PM	2.86 GB	comma...values
Parking_Violations_Issued_-_Fiscal_Year_2016.csv	Oct 26, 2017 at 6:48 PM	2.15 GB	comma...values
Parking_Violations_Issued_-_Fiscal_Year_2017.csv	Oct 26, 2017 at 6:48 PM	2.09 GB	comma...values
Parking_Violations_Issued_-_Fiscal_Year_2018.csv	Apr 1, 2018 at 7:03 PM	1.48 GB	comma...values

For our analysis, we only used the following columns:

- Summons Number (a unique index)
- Violation Code (integer)
- Violation Time (string)
- Violation County (string)
- Issue Date (datetime)
- Vehicle Make (string)
- Vehicle Body Type (string)
- Vehicle Color (string)
- House Number (string)
- Street Name (string)
- Intersecting Street (string--unused)

Data Preprocessing and Optimization:

Sanity check was performed and null values and erroneous data were dropped from the analysis datasets. Descriptive statistics was performed for above columns to identify the data distributions, frequency counts of unique values, outliers, missing data, data type mismatch and proxy variables. The following data issues were noted and handled:

County and Borough information: While we expect to see only 5 unique values representing the 5 counties of NYC (NY, K, Q, R, BX) we found that data had more unique values. Majority of data was captured with appropriate coding and hence inconsistently coded data were ignored. The county information was used as a proxy to represent the Boroughs ('Q': 'Queens', 'NY': ‘Manhattan’,'BX': 'Bronx', ‘K’: 'Brooklyn', 'R': 'Staten Island').

Issue Date: Data was in the format of MM/DD/YYYY (datetime) and had to be sliced into Month and Year for our time series analysis. We encountered severe time lag in pd.datetime conversion, so we ended up slicing the required part and converting to numeric that helped to improve the running time. Depending on the analysis required, pertinent data from the 5 files were extracted and constructed year wise (2014 to 2017) into separate dataframes and stored in separate CSV files. After appropriate grouping based on the analysis required, aggregate data were further extracted and used for plotting.

Violation Time: Data was captured as string in a 12 hour format and with A or P appended to the end of the string to represent AM or PM. In order to make the processing easier, time data was converted into a 24 hour format and utilized only the hour information to analyse the data.

Vehicle Color: Sanity checks revealed that there were more than 2900 unique colors recorded, as there were numerous variations of the same color (BLACK, BK, BLK etc). In order to make a meaningful analysis, we filtered only colors that had more than 20,000 violations and consolidated different variations of same color.

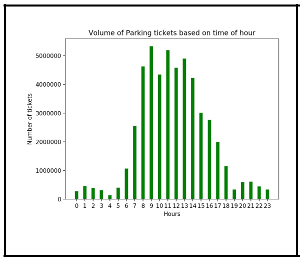
Vehicle Make: Similar to vehicle color, sanity checks on the vehicle make revealed more than 7000 unique values. Mirroring the same approach, we filtered only vehicle makes that had more than 20,000 violations that could be compared easily.

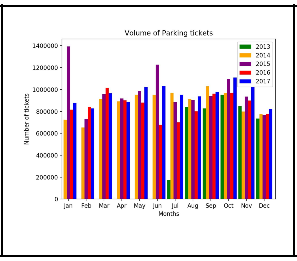
Violation code: In order to make a meaningful analysis, unique violation codes that had more than 5000 violations were filtered and utilized in the analysis.

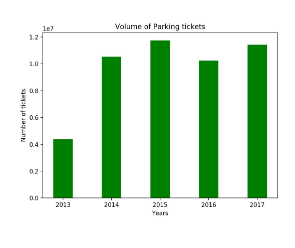
Challenges: Geo-coordinates of the location of the parking tickets were not readily available. Many Geocoding APIs including the Google API had restrictions for number of coordinates that can be retrieved (2500 per day). After some research, we found that the New York State geocoding server could be used for our project without any restrictions. However retrieving this info for all locations in the entire dataset would have been time consuming and not necessary. So, we filtered out for the locations that had a minimum of 100 violations or more, that resulted in approximately 5000 unique addresses per year and used this to find out the geo-coordinates.

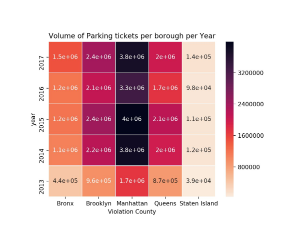
Using street name and house number available in the dataset, latitude and longitude details were retrieved for the county of interest and added this information to the dataset to represent the parking tickets in heat maps.

Initial Analysis: Analysis was done to understand the trend and parking tickets volume by time of day (grouped hourly), by month (Jan to Dec), by Year (2013 to 2017) and by the 5 boroughs in NYC.









CitiBike

CitiBike trip histories are available for free in CSV format. Each trip history file consists of one month's worth of data, usually around 1 million rows. CitiBike had already cleaned the data of rides taken by staff members, rides conducted at test stations, and rides that were 60 seconds or less. The data contains 15 columns of the following (note that columns from "Bike Id" down were not used in the analysis):

- Trip Duration in seconds (string)
- Start Time and Date (string, converted to datetime)
- Stop Time and Date (string, converted to datetime)
- Start Station Name (string)
- Start Station ID (integer)
- Start Station Latitude (float)
- Start Station Longitude (float)
- End Station Name (string)
- End Station ID (integer)
- End Station Latitude (float)
- End Station Longitude (float)
- Bike ID (integer)
- User Type (string stating "Customer" or "Subscriber")
- Gender (integer stating 0 for unknown 1 for male, 2 for female)
- User Year of Birth (integer)

Conducting the analysis required data from the years 2015 (prior to Brooklyn stations), 2016 (the beginning of Brooklyn stations), and 2017 (further expansion into Brooklyn, Harlem, and Queens). Therefore, one dataframe for each year was constructed.

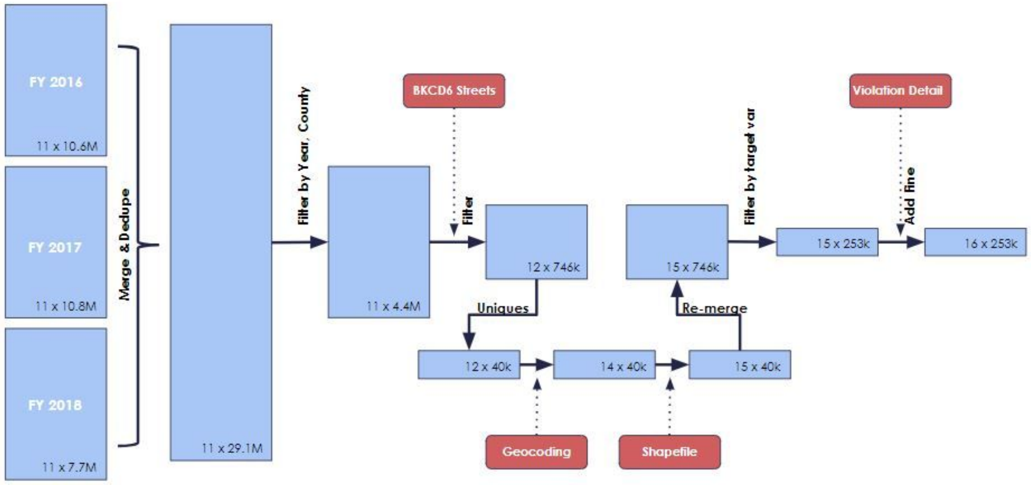
For the pertinent years, each month's CSV file was converted into a dataframe and stored in a list. Afterwards, Start Time/Date and End Time/Date columns were converted to datetime format so that it would be easier to derive time-based analysis. The dataframe in each year's list was then concatenated together, and saved as a pickle file for future use.

Some other preparation included creating a dataframe of unique coordinates per year, not only to represent how many stations were operational that year, but to then derive how frequently these stations were a start and/or end point. Such weights are important to create heatmaps of these locations.

There were some additional cleaning processes that occurred once analysis began. Some column names were inconsistently named across files and had to be rectified. In addition, some of the latitude and longitude coordinates were not sensible and had to be removed. An observation of 0,0 is 340 miles south of Ghana, and another coordinate was in the middle of Lower New York Bay.

CitiBike/Parking Data Combination

Getting the subset of data to run an analysis on a particular neighborhood (Brooklyn Community District 6--coded as CD306) required several steps.

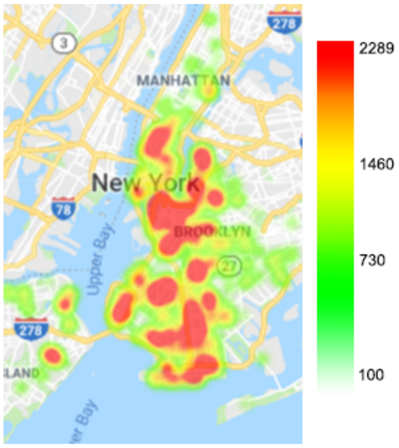
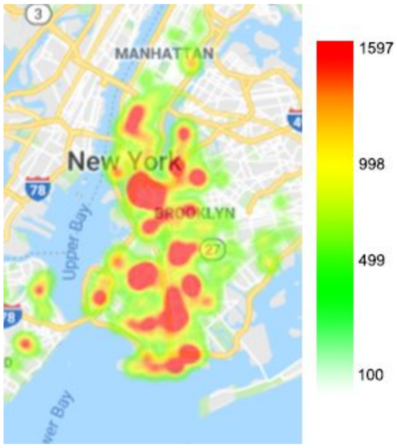
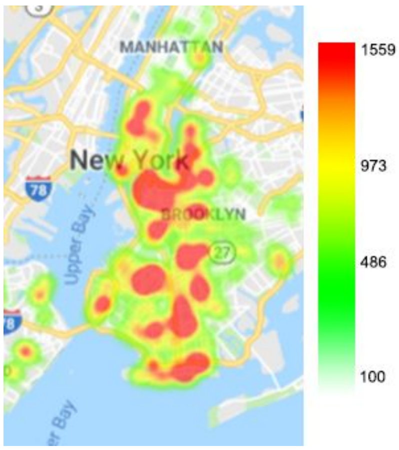


1. Parking Data for FY2016, FY 2017, and FY2018 was imported into three separate dataframes, dates parsed by a special function and NaN Issue Dates imputed to a date that would later be filtered out. At over 30M records, this took a while.
2. The three dataframes were concatenated and deduped, then filtered in stages by dates one year on either side of CitiBike dock installation (21.5M records), by Kings County (ie, Brooklyn--4.4M records), and then by a hand-constructed dataframe of street names in CD306. This gave us a set of 750k records that at least shared a county and street with the target district, but since a number of those streets are very long, and not necessarily unique in Brooklyn, we couldn't say definitively they belonged.
3. A new Address column combining House Number and Street Name with city and state was added and then used to dedupe a new dataframe, making a new subset of ~40k unique addresses. This was fragmented into 8 CSV files of 5000 records each.
4. The files were run in batches through a geocoding services API provided by the state of New York which returned a latitude and longitude as a JSON response to an http query. These coordinates were added to each CSV, and then all the fragments were loaded and rejoined to recreate the dataframe of unique addresses, this time with 'lat' and 'lon' columns. The hit rate was quite good, only failing to return a coordinate pair on a little over a thousand addresses.
5. Using Geopandas and Shapely, these records' coordinates were compared against the Shapefile of CD306 and flagged either '1' or '0' in a new 'target' column to set apart the ones that were definitively in the target zone. (This same function was used on the CitiBike data to pull out the target district docks).
6. A left merge of the unique addresses back into the original dataframe of parking tickets on 'Address', followed by filtering out all the '0' targets, yielded a final dataset of 253k parking tickets in CD306 from September 1, 2015 to August 31, 2016.
7. As an additional step, Pandas' read_html function was used to scrape the tables of parking violation codes and fines from the web. After some cleanup, this was merged into the final dataset to create a "fine" column that contained the putative value of each ticket.

New York Parking Ticket Analysis

Broader heatmap of parking tickets in NYC in Brooklyn from 2015-2017.

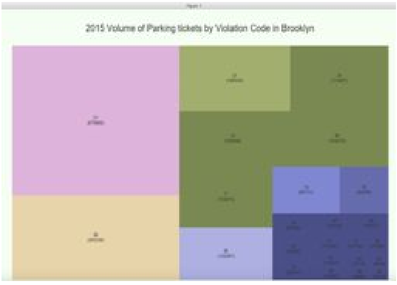
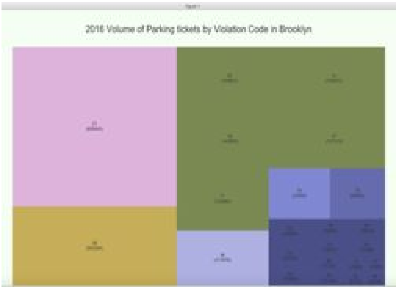
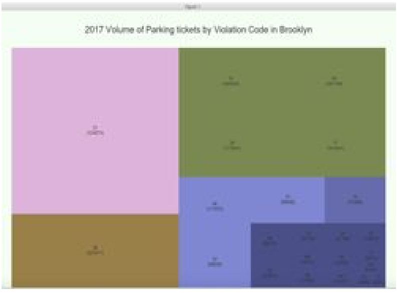
- Weight for the heatmap is based on the no. of violations per location

2015	2016	2017
 MANHATTAN New York BROOKLYN Upper Bay LAND 278 78 3 27 2289 1460 730 100	 MANHATTAN New York BROOKLYN Upper Bay 78 3 27 1597 998 499 100	 MANHATTAN New York BROOKLYN Upper Bay 78 27 1559 973 486 100

Inference: Intensity of violation tickets has reduced to a certain extent in upper parts of Brooklyn in 2016 when compared to 2015, but, it has increased again in 2017

Violation code - how does it compare from 2015 to 2017 in Brooklyn?

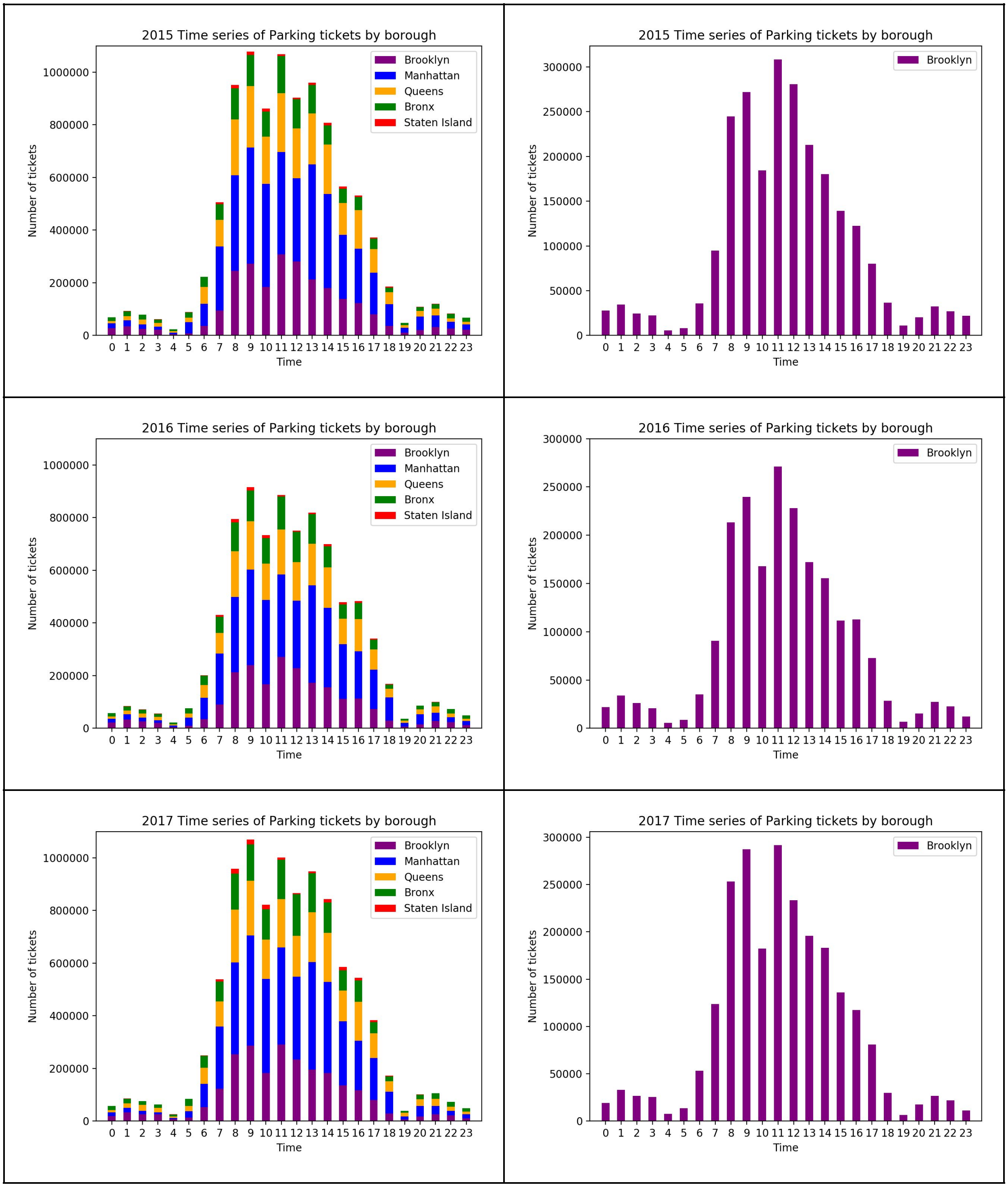
- Treemap representing the parking ticket volumes by violation codes. Violation codes that have >5000 entries are only included in analysis.

Inference: The top 5 violation codes are 21, 38, 20, 14 and 40. The code 37 (Parking in excess of the allowed time) has reduced to a greater extent from 2015 (3rd position) to 2017 (6th position)

Parking violations by time of day - how does it compare to city at large?

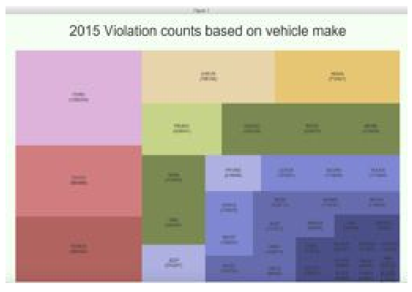
- Stacked bar chart representing total parking tickets over 24 hour period grouped by borough (left side) and bar chart representing total parking tickets over 24 hour in Brooklyn (right side).

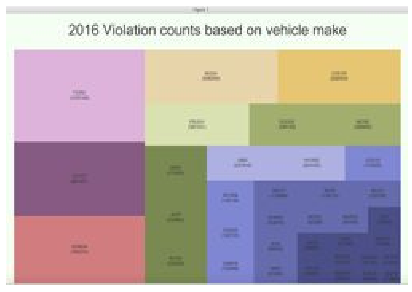


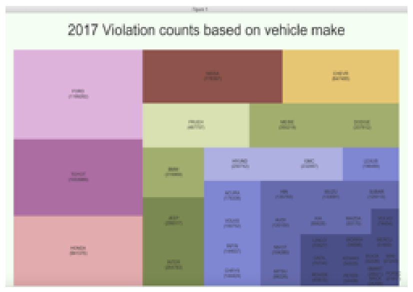
Inference: Parking ticket volume follows normal distribution (as expected) as it starts to increase during peak hours both in morning and evening and low at off-peak hours. Interestingly, ticket volume diminishes at 10 am, which could likely be attributed to street sweeping schedule. Around 8am-12pm in Brooklyn, volume of tickets has decreased in 2016 and if we overlay CitiBike information on top, we can identify any correlations, if they exist.

Vehicle Make with most parking violations

- Treemap plot representing parking violations by make of an automobile in NYC in general from 2015-2017. Automobile make with >20,000 violations in each make are displayed.





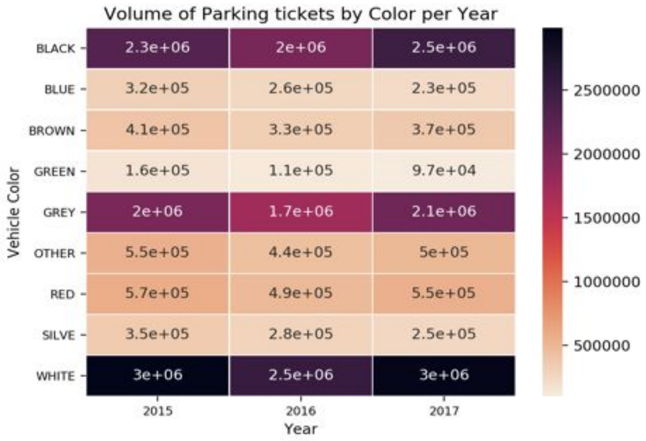


Inference:

- The economy range cars (Ford, Toyota, Honda, Chevrolet, Nissan) contributed the majority of parking ticket violations in all 3 years as expected.
- The volume of tickets by these makes has reduced in successive years which may indicate more uptake of CitiBike by these users.

Vehicle Color with most parking violations

- Heatmap plot representing parking violations by vehicle color from 2015-2017. Automobile make with >20,000 violations in each color are displayed



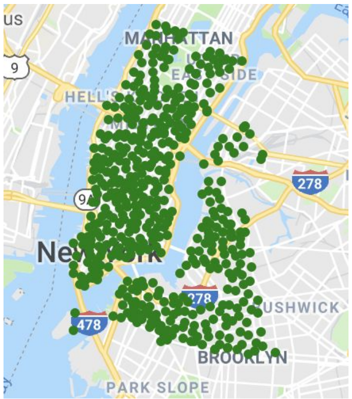
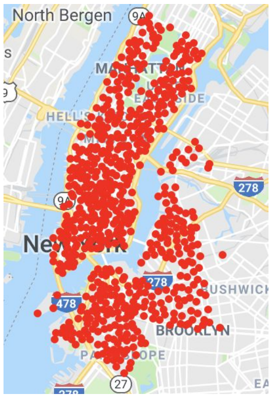
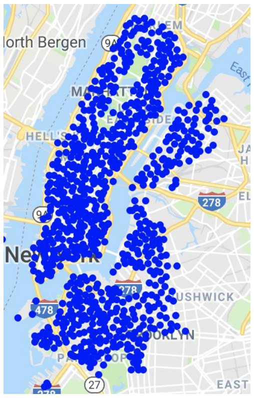
Inference:

- The white, black and grey automobiles have contributed to majority of parking ticket violations in all 3 years which is an expected outcome. Overall, the ticket volume for all colors has decreased in 2016, but increased in 2017.
- The number of violations received by red automobiles is proportionately higher when compared to other colors.

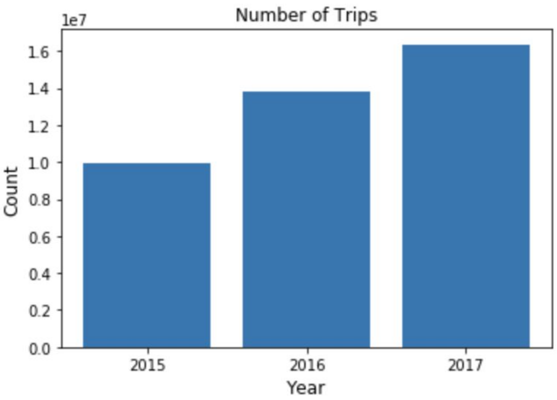
CitiBike Analysis

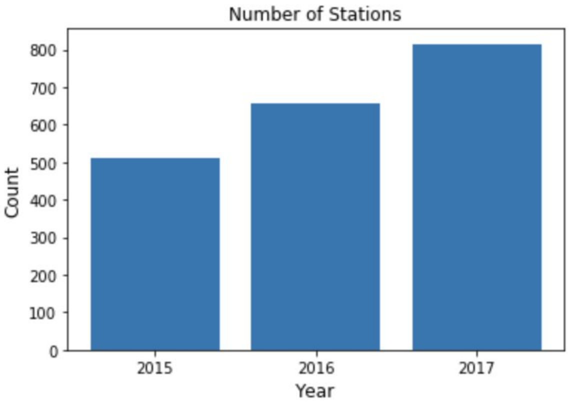
The pivotal time of interest in our hypothesis is August 2016, which is when CitiBike installed ~70 stations in Brooklyn. We therefore wanted to look at a benchmark before this change (2015), during this change (2016), and after this change (2017).

The graphs below displays the expansion in stations over the three years of interest. 2016 was the year where Brooklyn received a major expansion in CitiBike docks, while 2017's expansions were spread between the different boroughs.

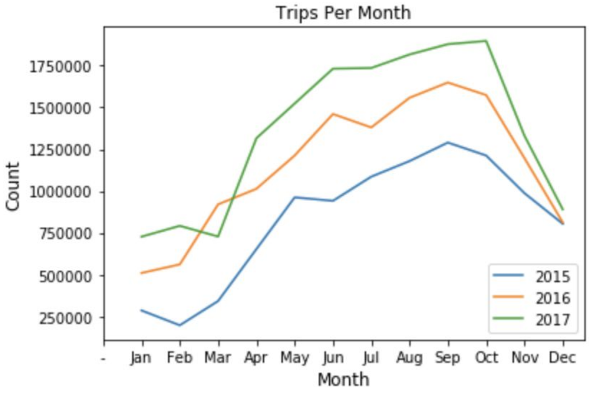
2015	2016	2017
 MANHATTAN EAST SIDE HELL'S New York 278 478 BUSHWICK BROOKLYN PARK SLOPE	 North Bergen MANHATTAN EAST SIDE HELL'S New York 278 478 BUSHWICK BROOKLYN PARK SLOPE 27	 North Bergen MANHATTAN HELL'S New York 278 478 BUSHWICK BROOKLYN 27 EAST

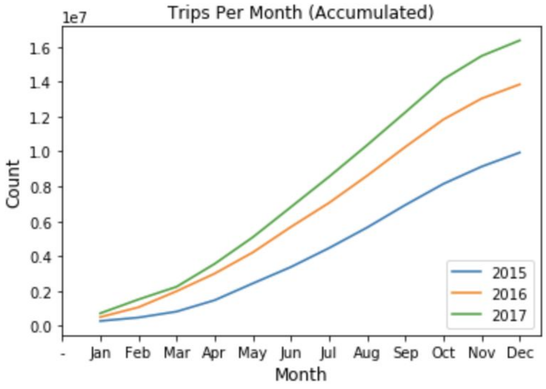
Both the number of trips taken and the number of stations installed around New York City has grown in the overall period of 2015 to 2017. The greatest increase in trips occurred between 2015 and 2016, while the greatest increase in stations occurred between 2016 and 2017. The differences between these increases is minimal.



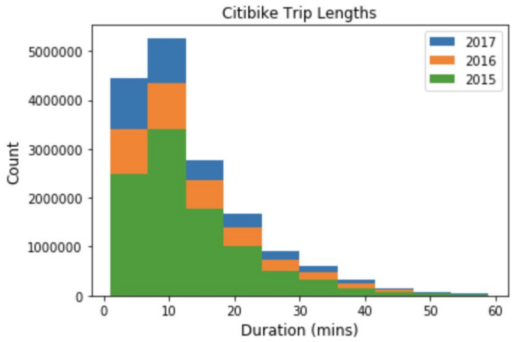


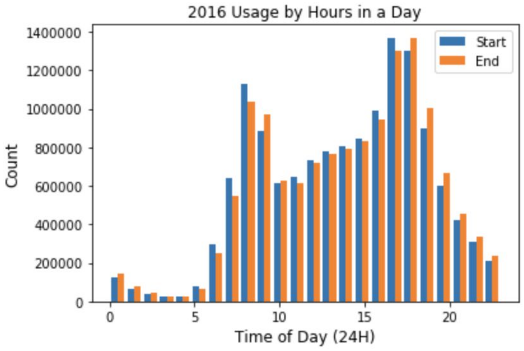
Year over year, CitiBike trips experience a growth in trips, beginning in April and dropping off in October. This behavior correlates to the change in average temperature in New York City, where the weather is more pleasant to use a CitiBike from April to November. If this same graph displayed trips accumulated from January to December for each year, we would see that 2016 grew more trips than 2015, but the 2016-2017 change wasn't as significant.



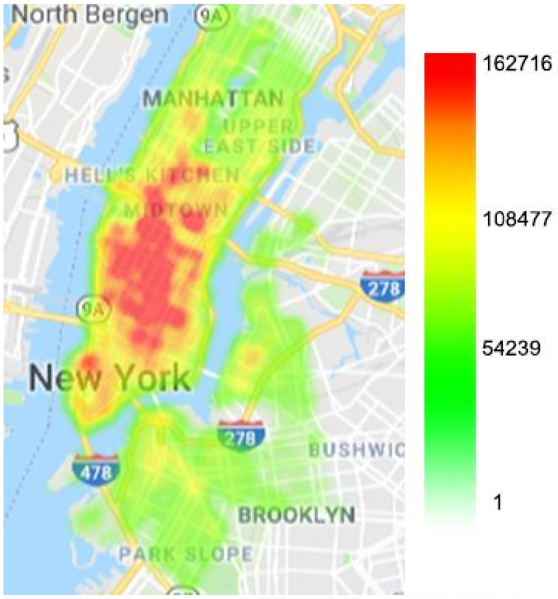
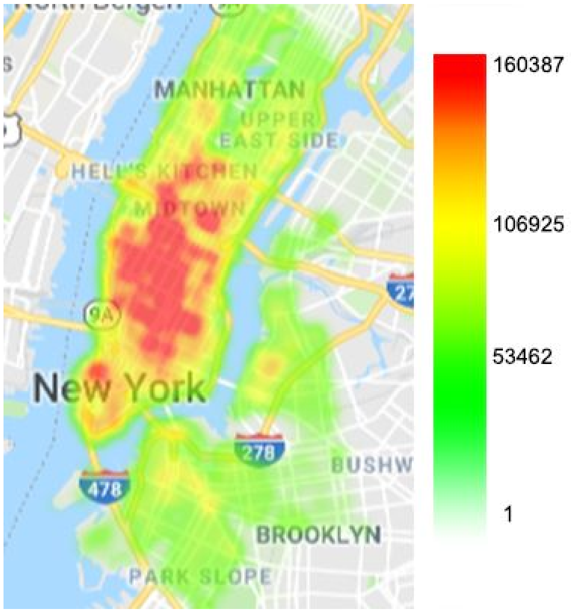


The most commonly occurring trip duration for a CitiBike ride was around 10 minutes. Furthermore, the most commonly occuring time of days when CitiBike rides were taken is the evening commuting hours (1700-1800 hours) and the morning commuting hours (0800-0900 hours). This behavior has stayed consistent from 2015 to 2017.



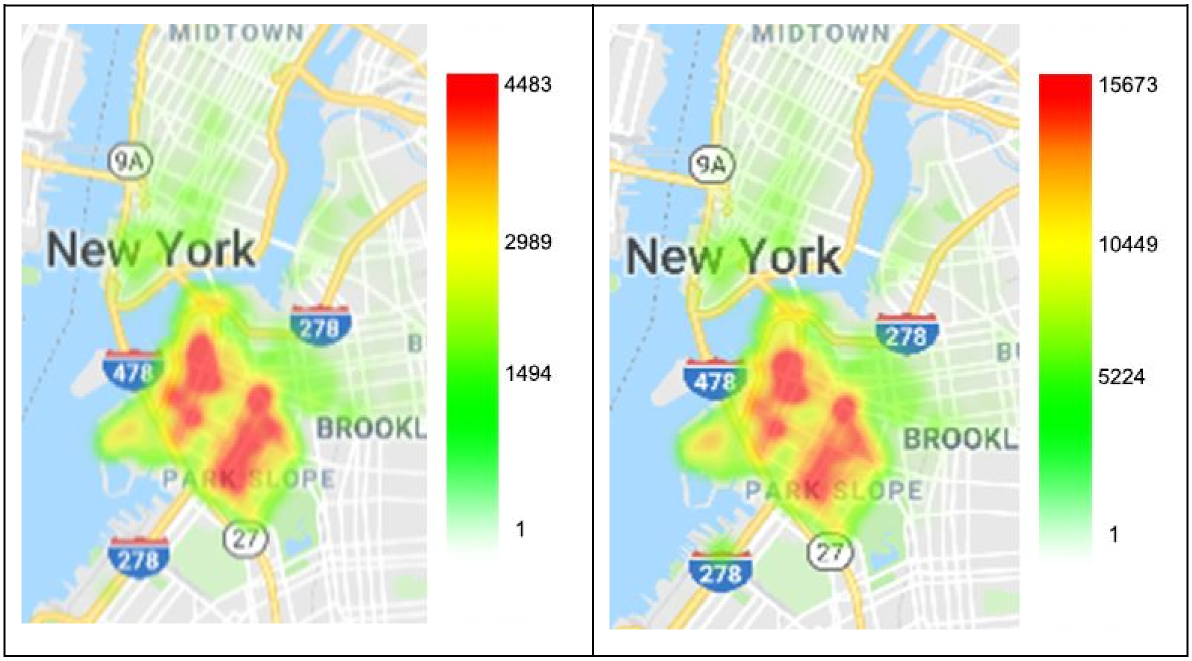


A common, short trip duration indicates that riders won't get very far from their starting location. Indeed, this is shown when creating heat maps of starting and ending locations for all years.

2017 - Start	2017 - End
 North Bergen MANHATTAN UPPER EAST SIDE HELL'S KITCHEN MIDTOWN New York 278 478 BUSHWIC BROOKLYN PARK SLOPE 162716 108477 54239 1	 MANHATTAN UPPER EAST SIDE HELL'S KITCHEN MIDTOWN New York 278 478 BUSHW BROOKLYN PARK SLOPE 160387 106925 53462 1

Finally, we were interested to see where riders travel to, if they began their trip at one of the stations installed in Brooklyn in 2016. As expected, due to a common trip duration of 10 minutes, these riders didn't travel very far out of Brooklyn.

Aug 2016 - Dec 2016	All of 2017

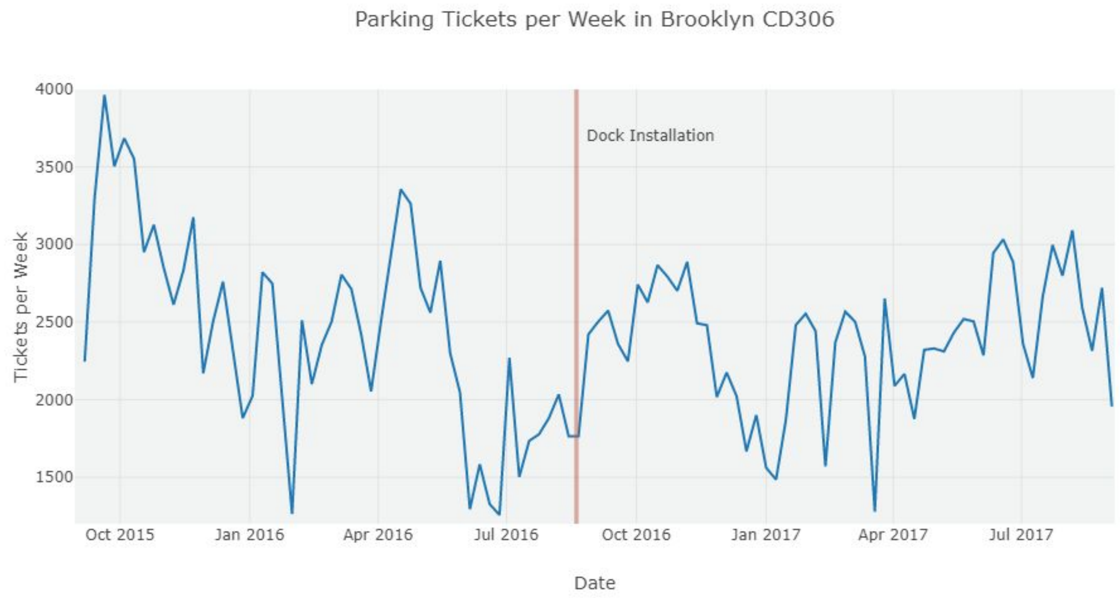


When focusing on just CitiBike data, we can conclude the following:

- Number of trips grew more in the time prior to the Brooklyn dock installation than after it.
- The average trip is a short, 10 minute commute during the rush hours, so riders tend to stay in their borough.

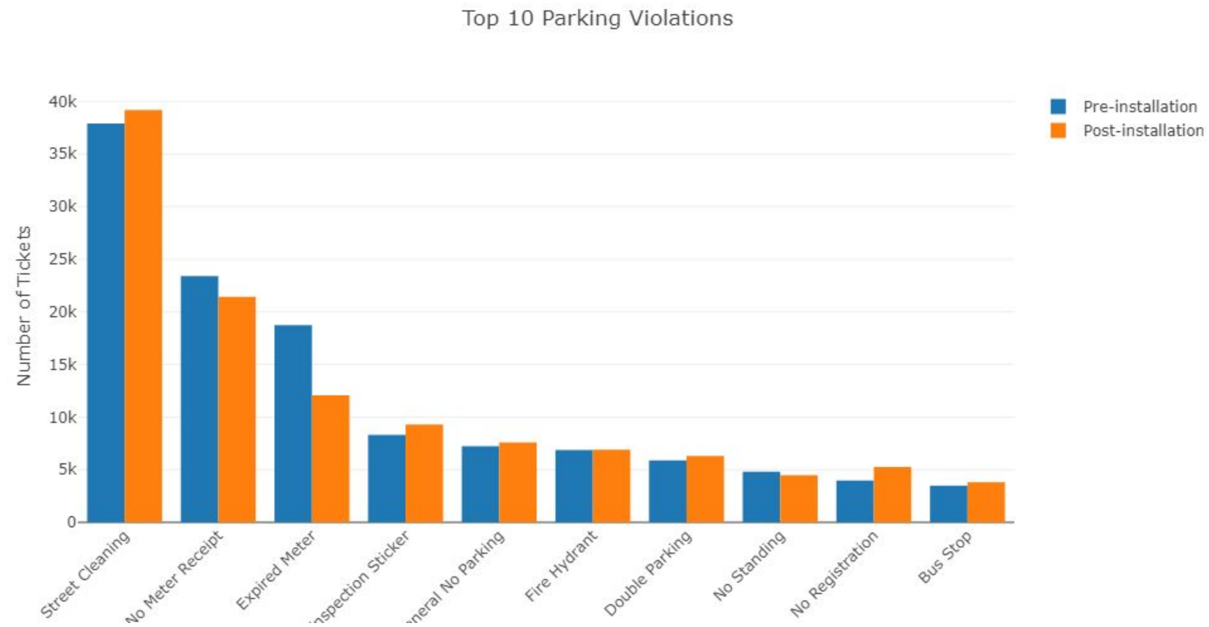
CitiBike/Parking Ticket Analysis

Did the installation of CitiBike docks in Brooklyn Community District 306 have an effect on parking violations? In the year preceding installation, New York issued 128,209 citations in the district, whereas in the year following, it issued 124,705. The 3,504 difference represents a drop of only 3%, so if there was an effect, it was not dramatic. A look at the weekly volume of parking tickets sheds a little more light on the story

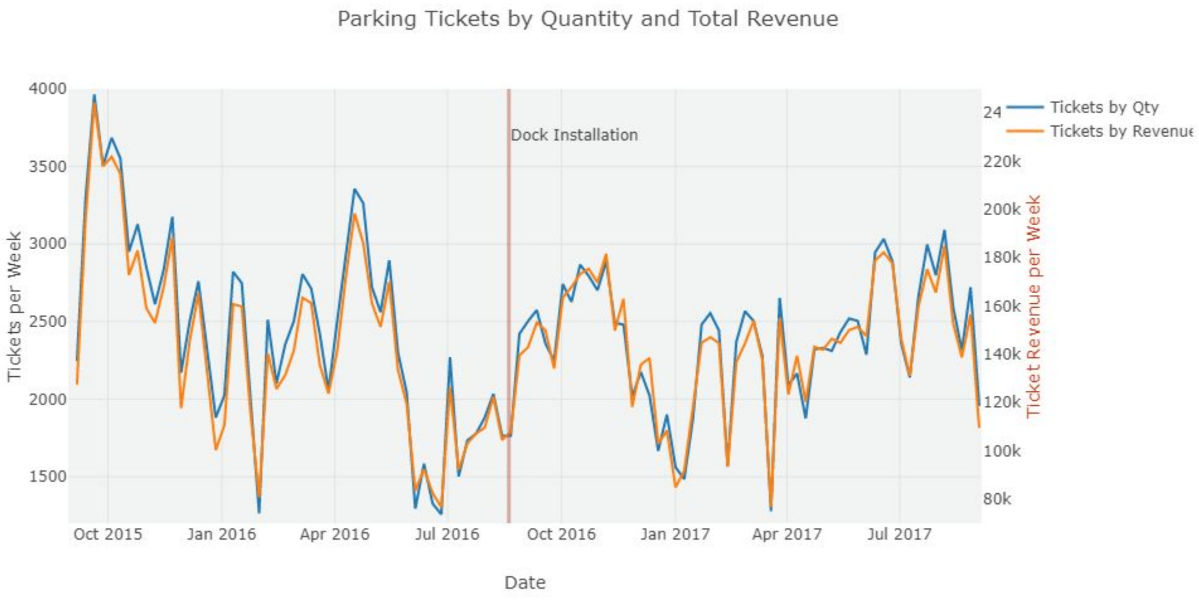


Apart from the relatively quiet period in the summer of 2016, it looks like the pace of tickets in 2016 was generally higher than 2017. You can also see the effect of a harsher winter in 2017, when snowstorms caused the pace of ticketing to drop for one week in 2016 but multiple times in 2017 (generally the police are more lenient about residential street parking violations when the cars are buried under snow). It is curious that ticketing dropped in summer of 2016 while that doesn't appear to have happened in 2015 or 2017. Having lived in that neighborhood that summer, I can’t think of a reason why parking ticket issuance would have been depressed in those months.

What about the type of parking violations, before vs. after? Let's look at the violation codes for each period.

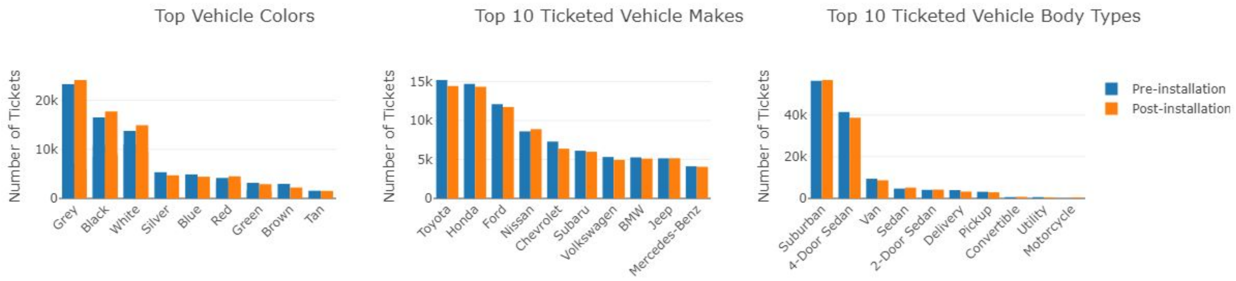

Top 10 Parking Violations

A few more street cleaning and sticker violations, but **remarkably** fewer meter violations - likely a result of fewer metered parking spots where the docks displaced street parking. This begs the question: did installing these docks change the city's income from parking violations in this area? A summary analysis of fines indicates that 2016 revenue in this district was $7,516,395, while 2017 revenue was $7,551,640. So revenue went UP by $35,245, or ~0.5%, despite slightly fewer tickets. Not a significant change, and overlaying weekly revenue on top of weekly volume confirms this: on a revenue basis, it appears the CitiBike installation may have lowered meter violations, but parking enforcement made up for it elsewhere.


Parking Tickets by Quantity and Total Revenue

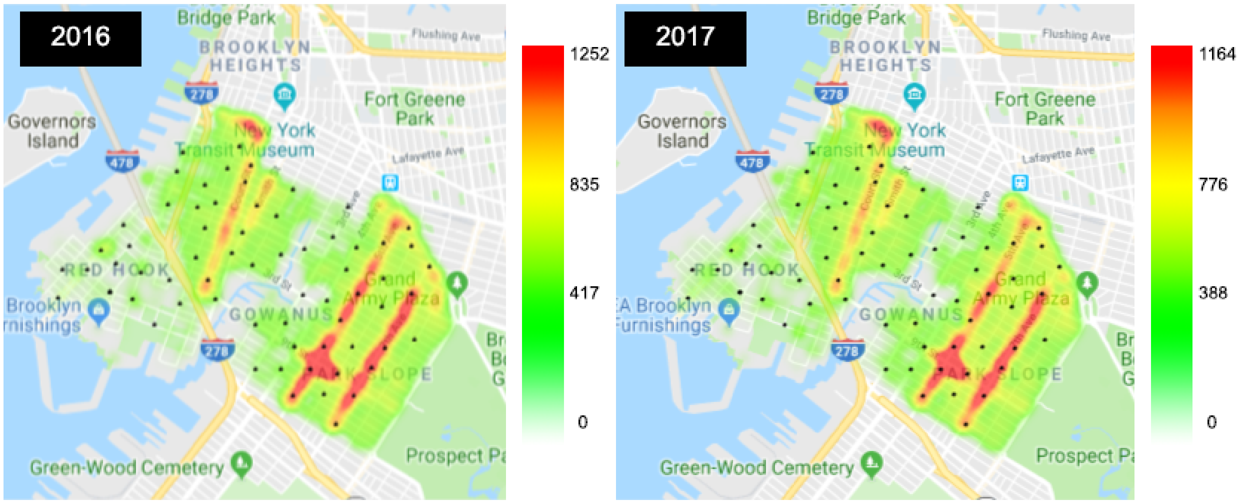
The graphs look nearly identical, indicating that the mix of tickets (as defined by fine amount) remained largely consistent over the two years, or at least didn't shift between high tickets and low tickets. Given the bulk of tickets goes to minor offenses like street cleaning, that's not surprising.

Just out of curiosity, what about incidental characteristics like vehicle color, make and model? We wouldn't expect to see a meaningful change in these, but since we have the data, might as well look:



As expected, no dramatic differences in these variables from one year to the next.

What about geographic distribution of the parking violations? For this, we need a heatmap comparing where the most parking violations occurred before and after the bike dock installation.



(black dots are the location of CitiBike docks)

Conclusion

If CitiBike installation had an effect on parking tickets in New York's Community District 306, it was too small an effect for us to measure. While the number of meter violations appears to have dropped noticeably, that did not have an effect on parking ticket income. Incidental characteristics of vehicle color, make and body type are not significantly changed, and geographic distribution looks very similar--if anything, parking violations in Red Hook *increased* in 2017. There is insufficient evidence to accept our hypothesis that the introduction of CitiBike eases parking.