1. Process the original parking violations data

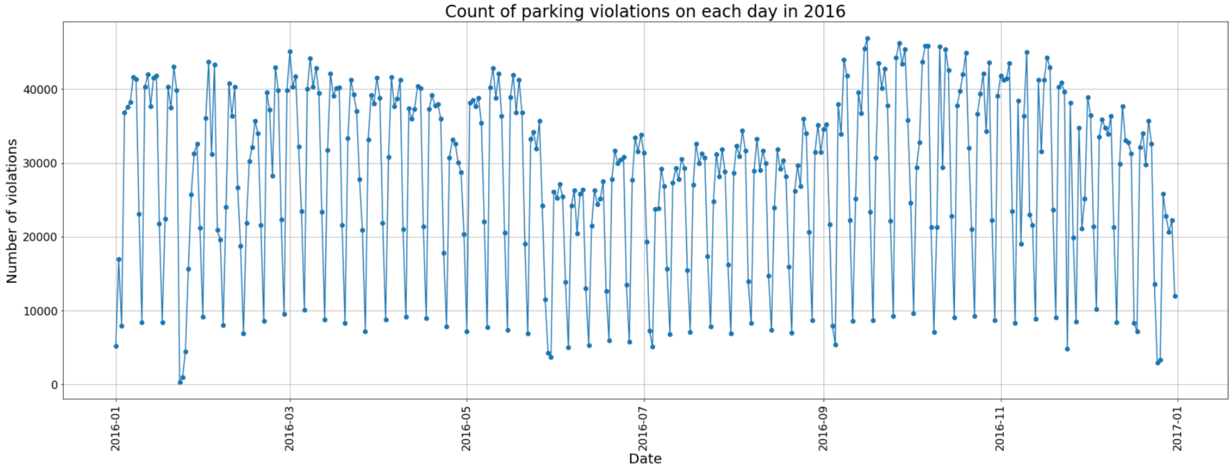
I only found the detailed NYC hourly weather data in 2016, so I planned to only analyze NYC parking violations in 2016. However, when I checked the date range of "NYC_Parking_Violations_2016.csv", I found the data was ranged from June, 2015 to June 2016. Therefore, I also analyzed "NYC_Parking_Violations_2017.csv" to achieve data from June, 2016 to December, 2016. To make sure there were not any missing dates, I also analyzed "NYC_Parking_Violations_2015.csv".

There were time consuming issues when I read data from these three large datasets, so I decided to write intermediary files as datasets for future use where only attributes I wanted to analyze were kept.

I realized that studying registration state might be interesting, so I added this attribute to the intermediary file. Also, to indicate each violation, summons number would be useful so it was set as the key of a dictionary.

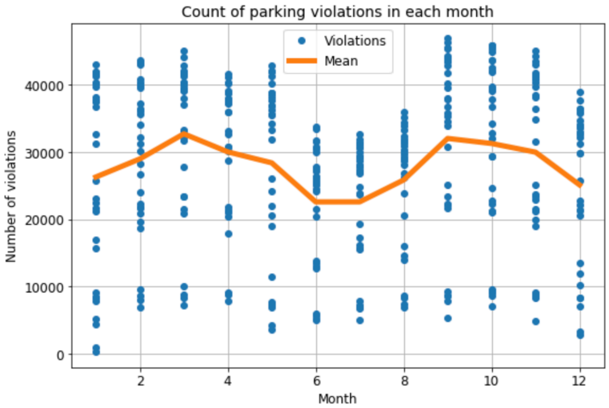
2. Answer if parking violations happen more on specific days

I didn't want to have data with missing values, so I used list-wise deletion. After list-wise deletion, the total number of parking violations in 2016 decreased from 10241012 to 10240895.



After calculating the count of violations occurring on each day and demonstrating the result in the above figure, I found the curve between June and August was obviously lower than other months, so the largest 50 numbers of parking violations were printed to show if there was any date in June, July, or August. Just as expected, the result is none.

I also found the distance between two adjacent lower vertices are similar. Therefore, I printed the weekday of days with minimum count and found most of them were on Sunday, which means parking violations occurred less on Sunday or parking violations were penalized less. Besides Sundays, some special holidays were also shown in the result. I looked up each day and found there were many holidays in the result.



To prove that violations occurring in June, July, and August were less than other months, I drew the above figure to show violations in each month and the mean of each month.

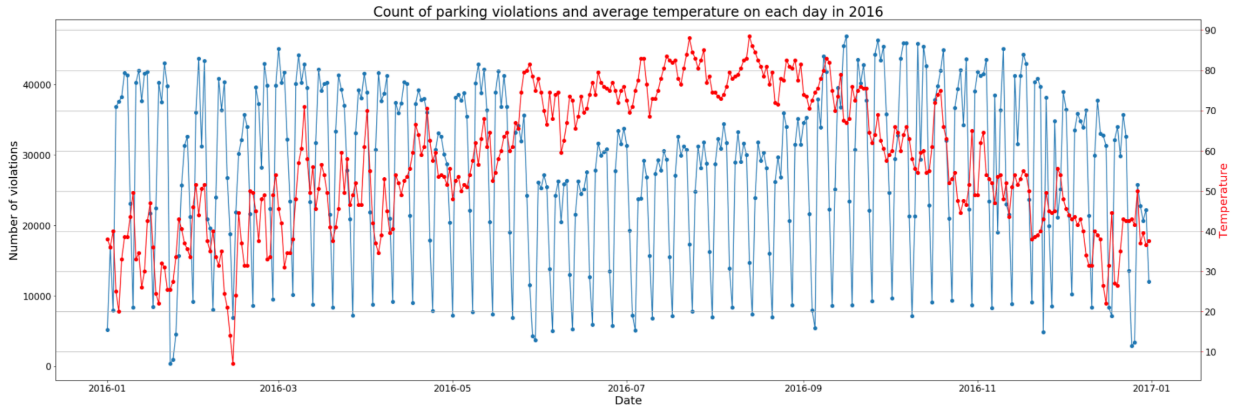
3. Answer if it tends to happen in the morning, afternoon, or evening when there is a parking violation.

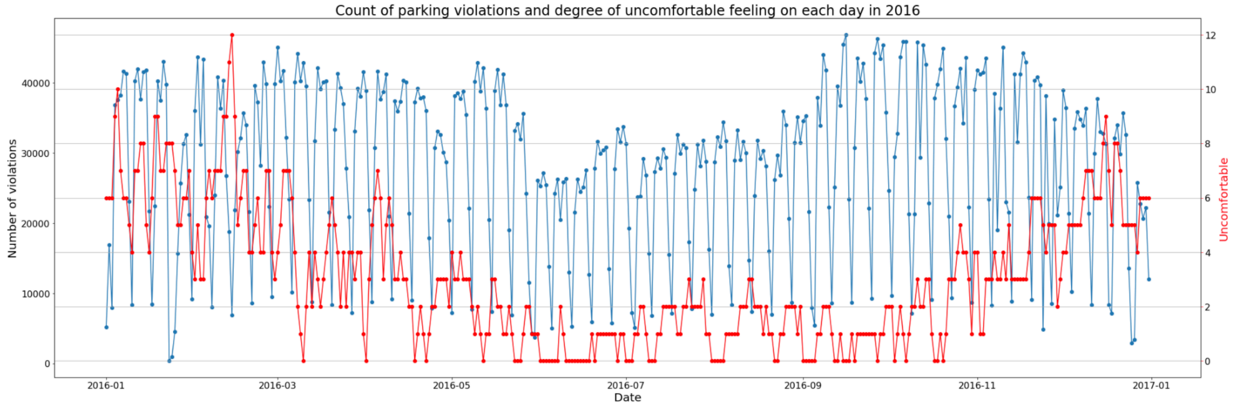
To answer this question, I combined issue date and violation time as the time information for each violation. Violation time is stored as a string like 1037A, so I changed it to be 10:37 AM to store as a datetime object. Also, strings like 0025A should be changed to 12:25 AM because 12-hour clock starts from 01 to 12. Some weird minute formats were also found because errors occurred when time was parsed to datetime object.

I counted the number of violations in each 24 hour and calculated the mean of each hour. The result showed that during 8 AM to 2 PM, there were relatively higher numbers of parking violation.

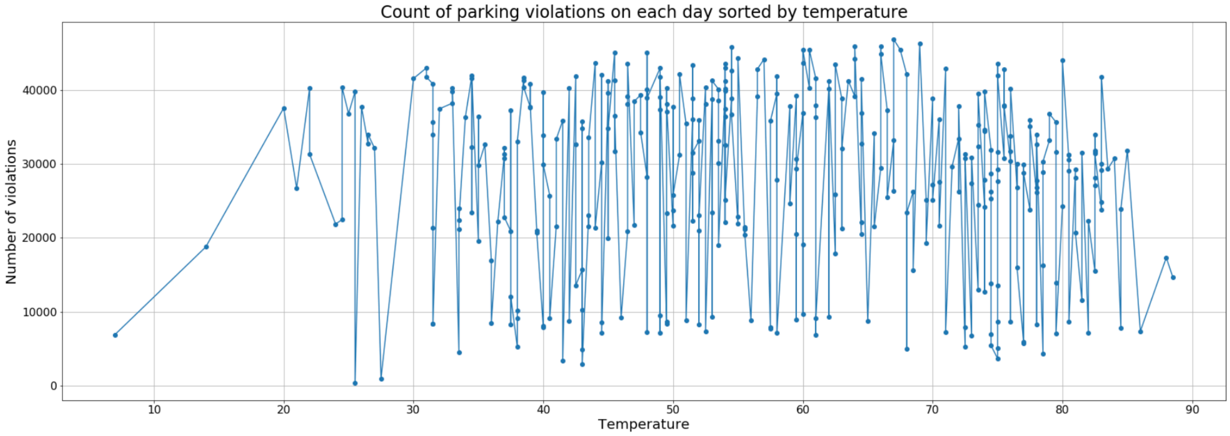
4. Answer if daily average temperature influences the number of parking violations.

I found the dataset storing maximum, minimum, and average temperatures of each day in 2016, so I decided to analyze the relation of parking violations and temperature.





At first, I plotted the number of violations and temperature of each day on the same figure, which is the upper image above. Then I found it was too ambiguous because when temperature is too high or too low, people feel worse than a medium temperature. Therefore, I set 70 as the base temperature and divided the difference between the temperature on that day and the base by 5 to determine the degree of uncomfortable feeling. The figure is the lower image above.



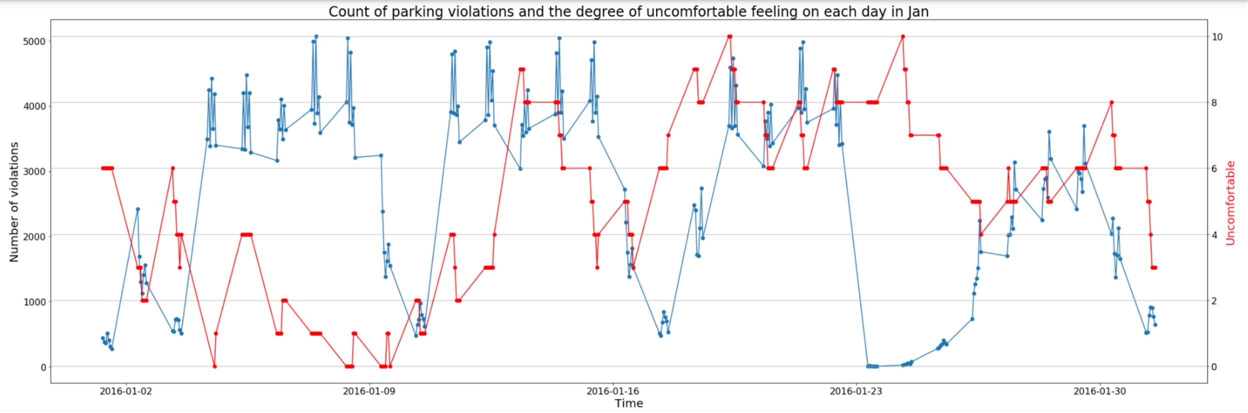
The figure was too complicate, so I decided to only focus on the temperature and the number of violations. I sorted issue day by the average temperature on that day and made the above plot showing the temperature and the number of violations. This plot shows that the relation of parking violations and temperature might be really weak.

5. Answer if hourly outside temperature influences the number of parking violations.

There are 24 hours in a day, so the amount of data might be too large and hard to process. Therefore, the dictionaries of violations and hourly temperature were classified into different months.

Because previous result showed that during 8 AM to 2 PM, there were relatively higher numbers of parking violation, I only considered the violations and temperatures during this period. This made the figure more simple.

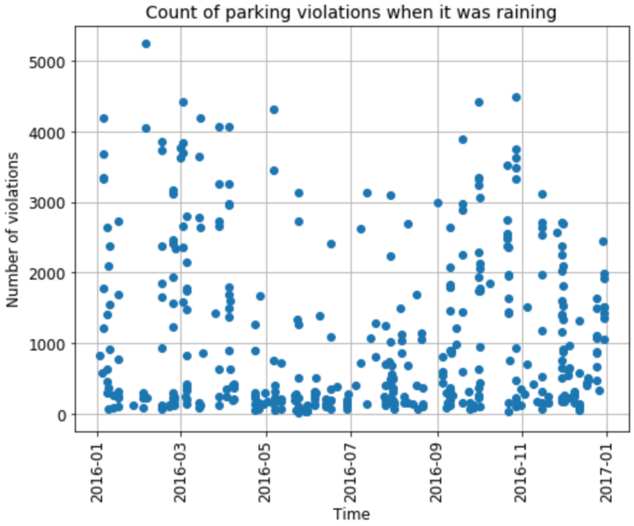
The previous issue of ambiguous y axis was there again, so I changed the temperature to be the degree of uncomfortable feeling and make the below plot.



The above figure shows the relation of violations and the degree in January. It does not reveal any obvious conclusion. When I looked at the higher vertices of the uncomfortable degree, the number of violations was not relatively higher or lower. Oppositely, the number was not relatively higher or lower when I looked at the lower vertices of the degree.

After looking at these figures, I concluded that there was not obvious trend when the temperature was too high or too low.

6. Answer if the number of parking violations is higher or lower when it is raining.



I checked the boolean value which represents if it was raining or not in that hour and made the plot showing the number of violations when it was not raining for every hour. Then, I found the relation between them was still weak.

7. Find the registration states with the largest numbers of parking violations.

I did this just for fun, but still found an interesting fact. When the numbers of parking violations were calculated by the registration state of vehicles, Florida is the fifth. However, the distance between FL the farthest from NYC in these top ten states.