CAN WE FORCAST CRIME WITH PARKING VIOLATIONS?

GITANJALI NARAINRAM | JIARUI SHAO | SKYE MORGAN

DATA CHOICE

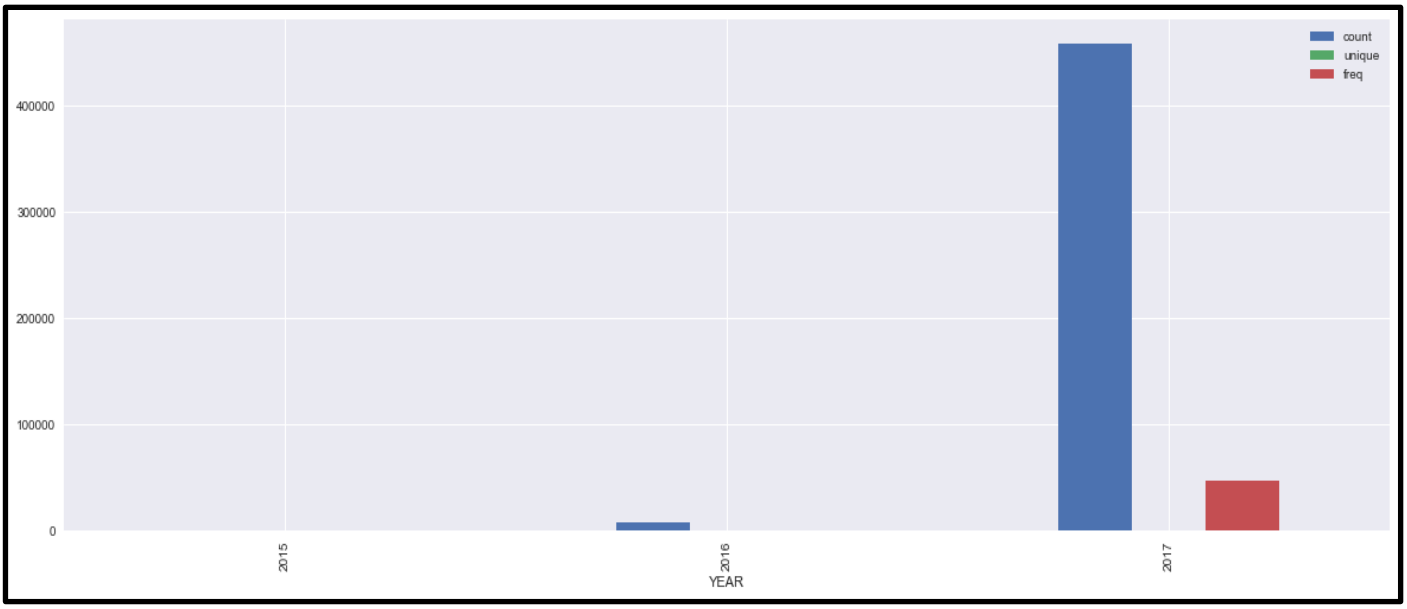
Not having any parameters can be a blessing or a nightmare. Initially, my group felt the former, however, after exploring the various data sets available, we could not narrow down a specific topic. One topic of common interest was NYC crime complaints. There were multiples variables we could work with and analyze.

After browsing through NYC Open Data, we found the reported parking violations for 2017 fiscal year. We were curious if there was a correlation between parking violations and the amount of crime in each borough. Our hypothesis was there would be a higher number of crime in places where there were higher number of parking violations. There are various studies showing the correlation between demographics in certain neighborhoods and how they compare to the crime rates.

Of course we can never isolate variables enough to attribute the cause of one event as a direct result of another. However, there can be a positive correlation between certain variables which account for a percentage of the outcome. To some degree, everything is interconnected and no one event especially one as big as a crime, can be exclusive.

DATA PREPARATION

We decided to narrow down the data from 2015 to 2017 which consisted of a decent amount of data. However, after seeing the initial distribution by year below, there was a discrepancy in the counts for 2015 and 2016 compared to 2017. As it can be seen, there is substantially more crimes reported in 2017 compared to the other years. This may be due to a



higher awareness and just more technology available in recent years. Currently, more and more data is being collected from all types of devices. There could have been similar number of crimes committed in previous years, but they may not all be reported. Hence, we decided to eliminate the previous years, because we could not directly isolate the reason for less data and it would cause problems when comparing the various years.

Focusing on the 2017 crime data, there were only 8365 rows. There were multiple rows which were blank and had NA values. This caused problems when developing an initial scatter plot matrix. This matrix was a good way to see the raw distributions and explore the data set. Then we would narrow down and focus on interesting observations. Because of this, we also had to delete various columns to declutter to data frame. It was not necessary to import all of the columns into the data frame. This decreases processing and loading time for the graphs and preparation.

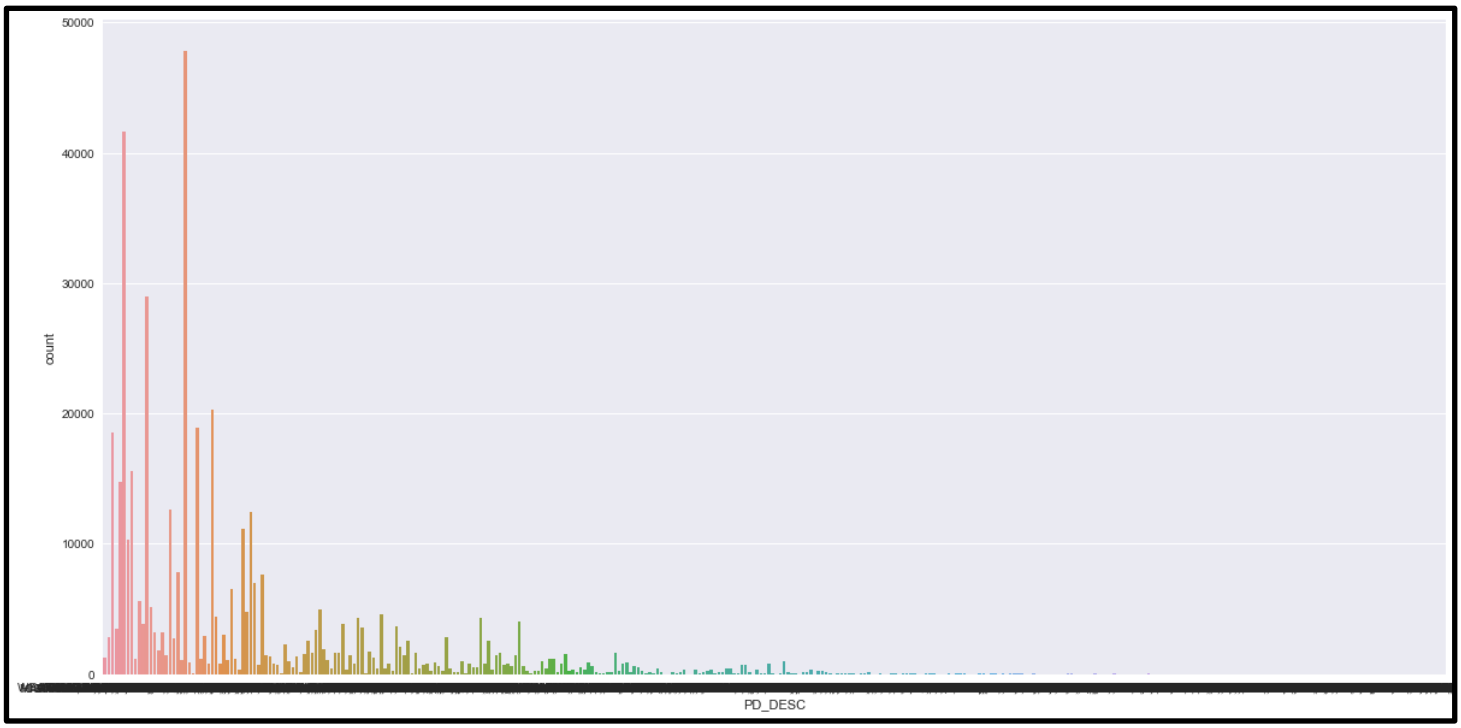
We started with over ten million records in the parking violations file which only consisted of data for 2016 and 2017. After filtering out 2016, we were left with about 45,000 records. This is a major difference, when showing visualizations and processing the file. It effectively gave us an adequate sample set to compare to the crime rates.

DATA ANALYSIS

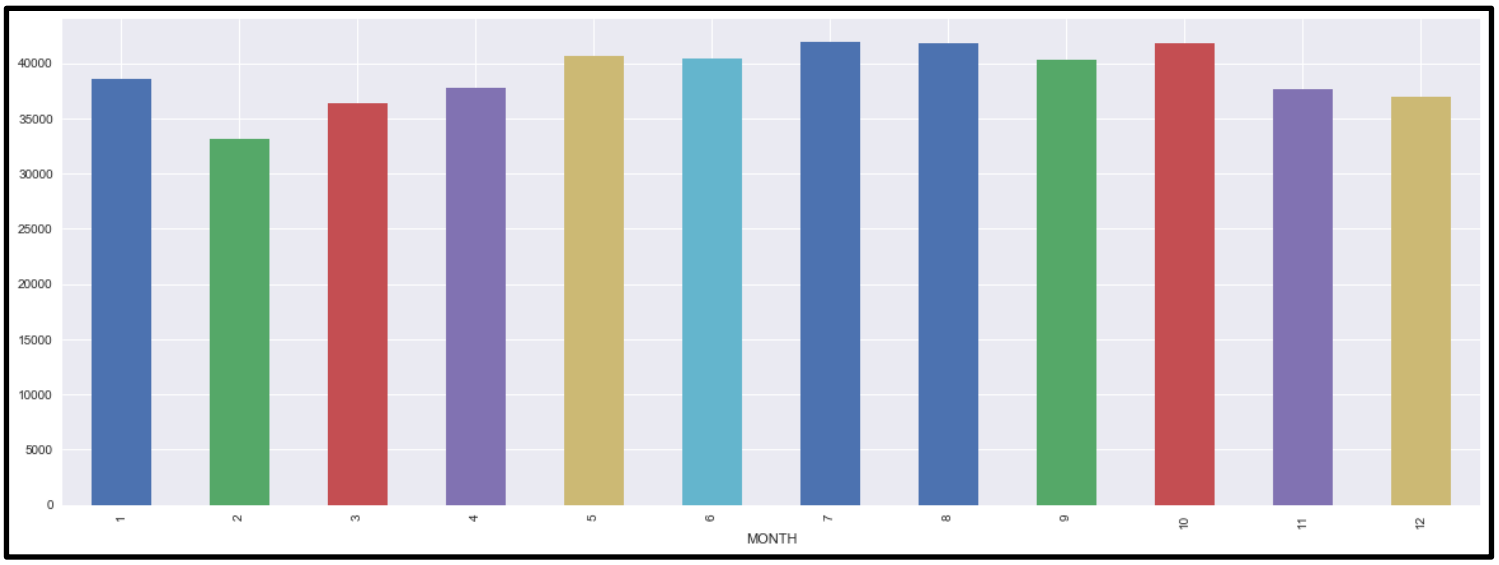
The graph below shown in the beginning is the distribution of crimes in New York City between 2015 and 2017. From this graph we can see that there's many types of crime committed,

and the frequency of each crime varies a lot. Because of sufficient data volume, the data variability, and the data is uncategorized, so it's valuable for us to analyze these data based on what we learned in class.

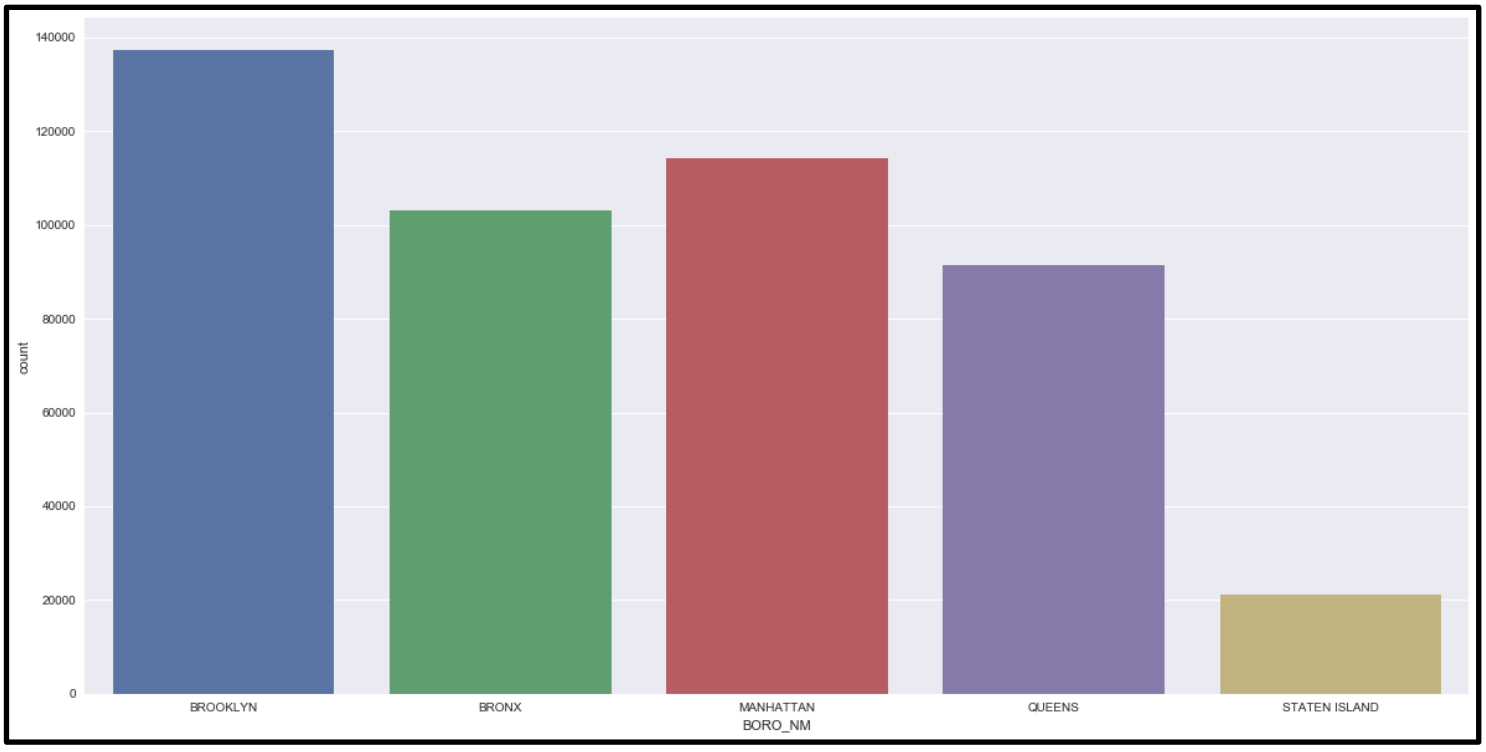
One of the first graphs describes the crime by month. We can see that February is the



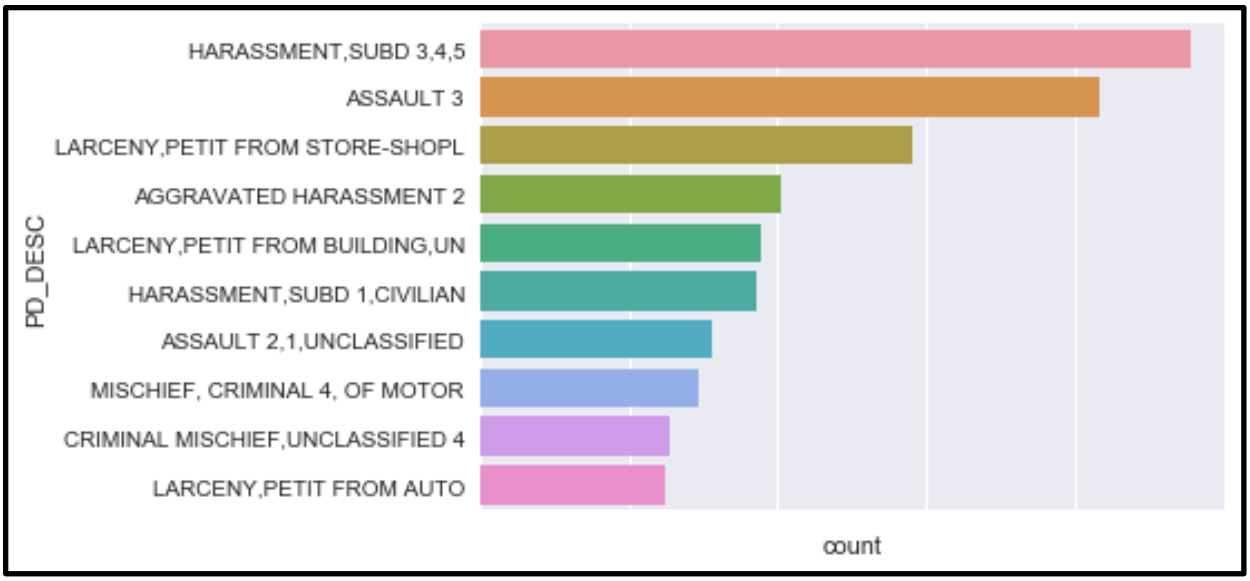
month in which crime happened the least, the number of crimes is under 40,000 from November to April. And the crime number is above 40,000 form May to October. Based on that, we can draw a conclusion that the crime occurred more in summer than winter in all.



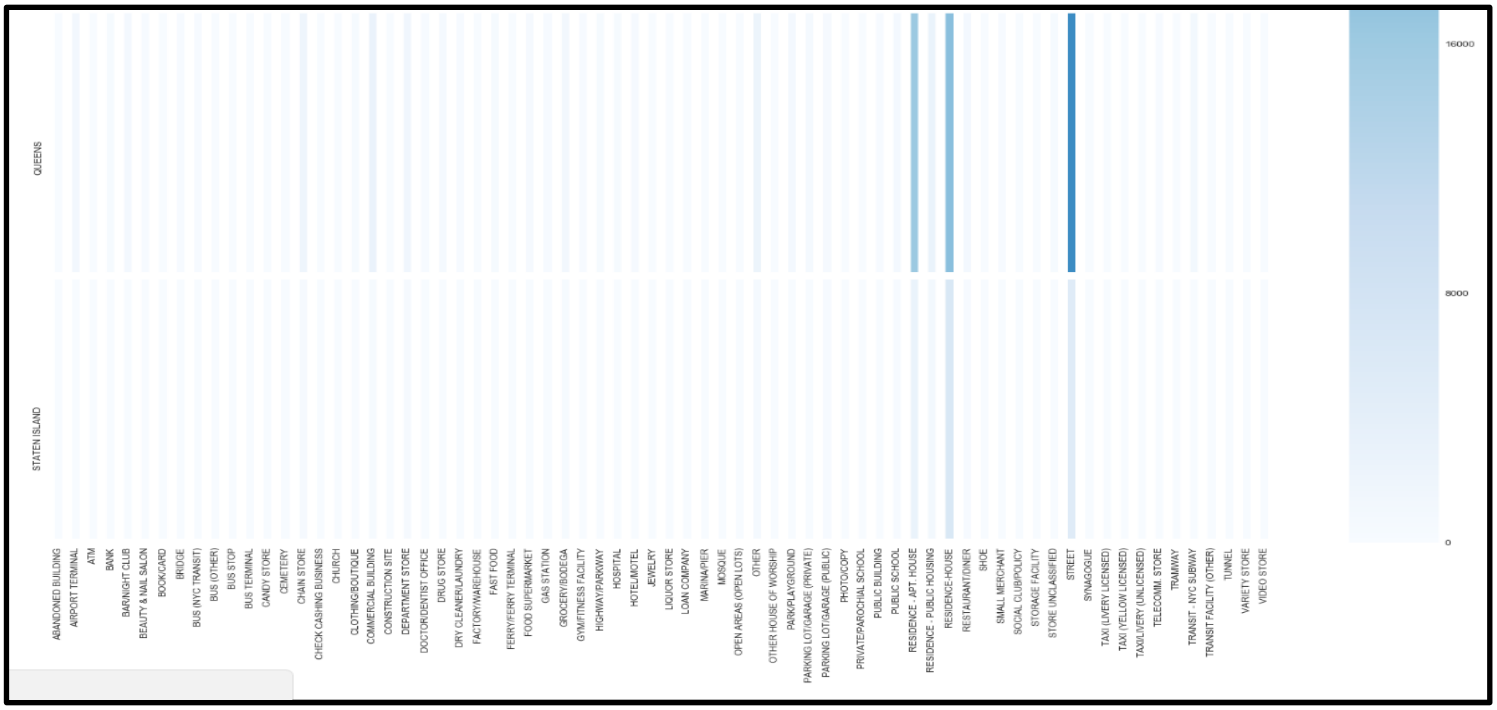
The next graph we were interested was the count of crimes per borough. We wanted to see if certain boroughs had a significant number of crimes compared to others. If so, we were interested in the borough of the most prevalent types of crimes. Based on the analysis of crime distribution by borough, we came to the conclusion that Brooklyn is the dangerous borough in New York and Staten Island is the safest place since there are in all 137,283 crimes occurred in Brooklyn and only 21,262 crimes happened in Staten Island.



As for primary type of the crimes, we found that Harassment and Assault is the most frequent crime occurred in New York in that their counts are higher than 40,000 and far beyond others.



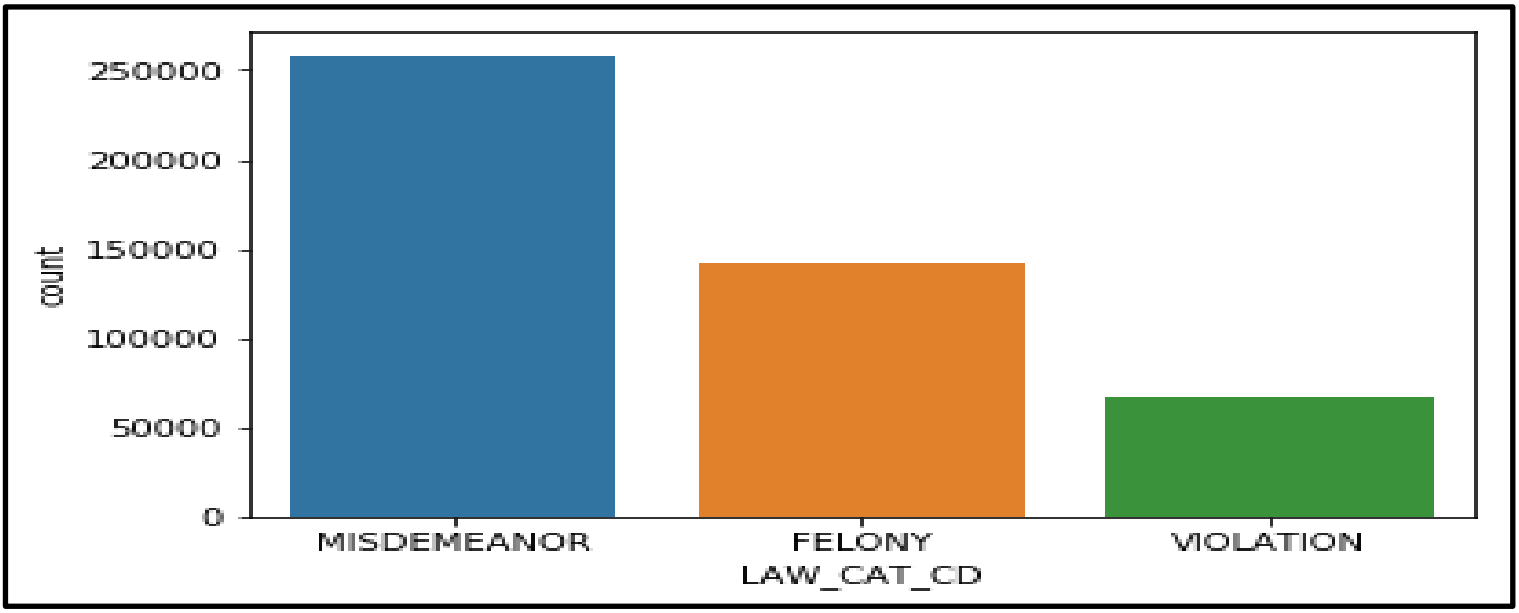
The following graph is the counts of the premises type separated by each borough. This was done by creating a crosstab data frame with the Borough Name column and the Premises Type column. The result of the crosstab was a dictionary and we put it into a dataframe to reduce difficulty later. As it can be seen below, there 70 types of premises types. As a chart this data would be useless and take too long to analyze. However, from the heat map below, we can easily see there are significantly more crimes on the streets and people's houses, especially public housing. This is a significant trend across all the boroughs, however, Brooklyn has the darkest patches for the street and public housing crimes.



PREM_TYP_DESC	ABANDONED BUILDING	AIRPORT TERMINAL	ATM	BANK	BAR/NIGHT CLUB	BEAUTY & NAIL SALON	BOOK/CARD	BRIDGE	BUS (NYC TRANSIT)	BUS (OTHER)	...	TAXI (LIVERY LICENSED)	TAXI (YELLOW LICENSED)	TAXI (UNL
BORO_NM														
BRONX	30	4	116	415	358	262	7	58	353	41	...	145	57	
BROOKLYN	84	3	108	444	977	460	14	126	275	61	...	170	113	
MANHATTAN	100	1	148	858	2597	391	168	137	213	69	...	157	294	
QUEENS	50	1208	55	412	935	239	10	221	193	43	...	74	53	
STATEN ISLAND	36	0	12	49	89	58	4	27	78	22	...	13	3	

5 rows × 70 columns

In order to understand more about the nature of crime that NY is exposed to, we filtered the data down by Law Category. By doing this, we were able to recognize more clearly, the severity of the crimes which occured in the city. This graph manages to easily highlight how inflated Misdemeanor crimes are to that of Felony and Violation. In fact, Misdemeanors crimes



have been recorded at over 3 times that of Violations, and nearly over 2 times Felonies.

Additionally, we decided look not only at the Law categories crime count by themselves but to chart them against the individual boroughs. Through doing this, we could easily identify which borough held the mantle for least crime count per law category.

CONCLUSION

Going into this analysis, there seemed to be a lot of ground to cover with these two data sets. While there was much to analyze with the crime data set, there were not as many variables that could be aggregated similarly for the parking violations data set.

Here is a count of parking violations grouped by the time of day when they were issued. There seems to be more violations reported in the morning compared to the later parts of the day. Initially we wanted to compare this distribution to the reported time of the crimes. As a group, we thought there would be a positive correlation to this. There are more parking violations during these hours, because they are the peak hours of the day, there are more cars and people, so the chances of a parking violation is higher. As a group, we were planning on developing a train data set from the crime data and use that to develop a clustering algorithm to forecast the parking violations in the same neighborhoods.

However, there was also multiple problems with the data set after further analysis. The dictionary did not adequately explain all of the columns in the data. Hence it was harder to really understand what some fields were saying. For instance, "Q" and "QN" both stand for Queens. The data set is also fairly new on the site and they are still collecting data for this. Hence we decided to focus more on the crime data set instead.