Analysis of crime data in NYC

1st Siddesh Gannu
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
sgannu@stevens.edu

2nd Arnav Arora
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
aarora1@stevens.edu

3rd Siddhesh More
Department of Mathematics
Stevens Institute of Technology
Hoboken, United States
smore4@stevens.edu

***Abstract*—We are living in the world where various different crimes are being committed on daily basis. This questions the safety of people. So how can we know that the area we are living in is safe or not? To solve this we are using the NYC crime data to visualize and identify most safest and most dangerous areas in the city according to the crime rates in that particular area. Identifying the hotspots in the city depending on the locations of crimes occurred in the city is one of the main aims of the project. Creating a useful visual representation of each type of crime according to the location of the crime. Differentiating various crimes according to their violence level. Creating a detailed heat map for better understanding of the crimes being committed all over the city. For this we are using three machine learning algorithms which are, random forest, support vector machines (SVM) and KNN algorithm.**

I. Introduction

Our main objective is to gain useful insight using the NYC's crime dataset about the safety of each borough, crime "hotspots" or locations which the most arrests occur, the types of offenses that occurred the most frequently in each borough and the location of them, etc. The idea is to classify certain areas with a color to classify how safe that specific area is. For example, red would signify that an area is unsafe as there were a lot of arrests made because people were getting robbed in that area. We want to achieve this goal through the use of machine learning algorithms and python. The machine learning algorithms used will enable us to classify a location as safe or not and which streets/locations can be classified as hotspots. This project is meant to provide a variety of insights using algorithms such as KNN, Random forests and SVM's to classify the safety of a borough, street, or locality. Our aim is to possibly present this to the city of NYC so that they may use its results to plan out patrol routes and figure out proper allocation of city resources.

II. Related Work

We found some similar work (Fig. 1) done on the same data which can be seen in the graph. We have been trying to work on something similar to this one, for comparison we have also added our output (Fig. 2) of the similar graph.

Reference: https://a1080211jeff.medium.com/exploring-nyc-analysis-of-crime-data-in-new-york-city-6134642b9833

III. Our Solution

Based on the features such as time of arrest, type offense committed, arrest borough, age group, gender of the suspect,

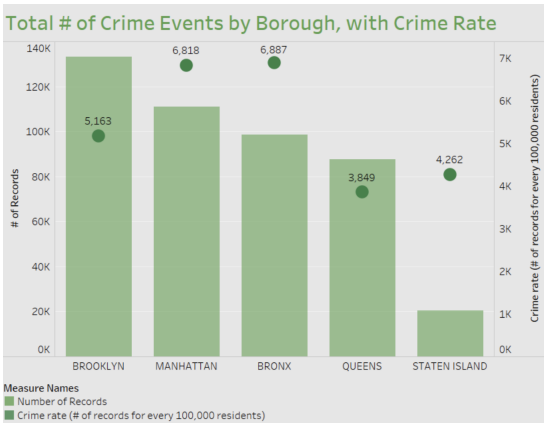

Fig. 1. Similar work found

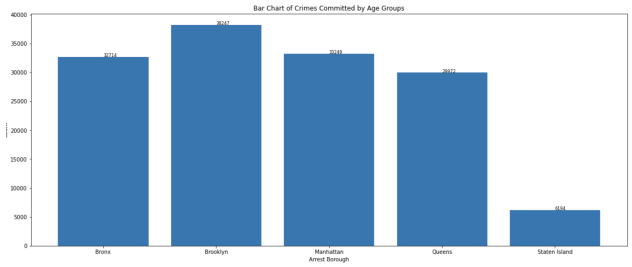

Fig. 2. Our output

etc. our goal is to predict how violent the crime committed would be. We also aim to identify safe and unsafe areas in the city. Safe and unsafe areas in the city can be identified by the heat map. Red color shows the most unsafe area whereas green color shows safe area. Based on how dangerous the crime is we have created a dependent variable which has three levels of violence: Most violent, Medium violent and Least violent.

A. Description of Dataset

The dataset is very large and is also frequently updated with new arrests 4 times every year. This data set has 24 columns and over 140K rows. The data set consists of arrest-key (Randomly generated persistent ID for each arrest), arrest-date(Exact date of arrest for the reported event), pd-cd (Three digit internal classification code (more granular than Key Code)), pd-desc(Description of internal classification corresponding with PD code (more granular than Offense Description)), ky-cd(Three digit internal classi-

	arrest_key	arrest_date	pd_cd	pd_desc	ky_cd	ofns_desc	law_code	law_cat_cd	arrest_boro	arrest_precinct	jurisdiction_code	age_group	
0	220756993	2020-11-19T00:00:00.000	155	RAPE 2	104.0	RAPE	PL 1303001	F	B	41	0	18-24	
1	220422940	2020-11-12T00:00:00.000	157	RAPE 1	104.0	RAPE	PL 1303502	F	Q	112	0	25-44	
2	218804160	2020-10-06T00:00:00.000	157	RAPE 1	104.0	RAPE	PL 1303501	F	M	7	2	25-44	
3	218641095	2020-10-02T00:00:00.000	594	OBSCENITY 1	116.0	SEX CRIMES	PL 2631600	F	M	5	0	25-44	
4	217890704	2020-09-16T00:00:00.000	155	RAPE 2	104.0	RAPE	PL 1303001	F	K	77	0	25-44	
5	217887490	2020-09-15T00:00:00.000	157	RAPE 1	104.0	RAPE	PL 1303502	F	K	77	0	25-44	
6	215164081	2020-07-10T00:00:00.000	NaN	NaN	NaN	NaN	PL 2410201	M	M	17	0	45-64	
7	214738271	2020-06-30T00:00:00.000	157	RAPE 1	104.0	RAPE	PL 1303501	F	K	77	0	18-24	
8	213573222	2020-05-27T00:00:00.000	177	SEXUAL ABUSE	116.0	SEX CRIMES	PL 1306504	F	Q	101	0	45-64	
9	213022815	2020-05-13T00:00:00.000	155	RAPE 2	104.0	RAPE	PL 1303001	F	K	77	0	25-44	

Fig. 3. Dataset Used

fication code (more general category than PD code)), ofns-desc(Description of internal classification corresponding with KY code (more general category than PD description)), law-code(Law code charges corresponding to the NYS Penal Law, VTL and other various local laws), law-catcd(Level of offense: felony, misdemeanor, violation), arrest-borro(Borough of arrest. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens)), arrest-precinct(Precinct where the arrest occurred), jurisdiction-code(Jurisdiction responsible for arrest. Jurisdiction codes 0(Patrol), 1(Transit) and 2(Housing) represent NYPD whilst codes 3 and more represent non NYPD jurisdictions), age-group(Perpetrator's age within a category), perp-sex(Perpetrator's sex description), perp-race(Perpetrator's race description), x-coord-cd(Midblock X-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)), y-coord-cd(Midblock Y-coordinate for New York State Plane Coordinate System, Long Island Zone, NAD 83, units feet (FIPS 3104)), latitude(Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)), longitude(Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)) and geocoded-column(New Georeferenced Column).

The source of this data set is NYC Open Data. As this data set was too large, we dropped some of the unnecessary features. There were some null value entries in the dataset, so we removed them and cleaned the data. All the missing and irrelevant data was cleaned from the main data set. We have used 12 features towards the implementation of algorithms. Dataset includes information about the kind of crime committed like felony, robbery, rape, assault etc. Also, the location of arrest, borough and precinct of arrest and the date of arrest.

Following are the features we have used for data visualization and for training the algorithmic models: Exact date of arrest for the reported event, Description of internal classification corresponding with PD code, Description of internal classification corresponding with KY code, Borough of arrest. Ex.: B(Bronx), Q(Queens), Precinct where the arrest occurred, Perpetrator's age within a category, Perpetrator's sex description, Perpetrator's race description, Mid-block X-coordinate for New York State, Mid-block Y-coordinate for New York State, Latitude coordinate for Global Coordinate System, Longitude coordinate for Global Coordinate System.

B. Machine Learning Algorithms

We are using following three algorithms:

1. Random forest
2. K Nearest Neighbor (KNN)
3. Support Vector Machine (SVM)

A custom-made column in the data called "violence level" was made which categorized all offenses into three categories "Most violent", "Least Violent", and "Medium Violence". This "violence level" column was set as our target variable and we set the columns with the offensive description, race, sex, age, and arrest. When trying to predict arrest borough, we created a random forest using 10 trees in the forest, a max depth of 30 to essentially prune the tree and reduce overfitting. We chose random forest as randomly chooses features to grow out the trees. This also helps with reducing overfitting and leads to a more accurate prediction. We also set Bootstrapping to True in order to ensure we are not providing the whole dataset to each tree in the forest. This reduces variance and avoids overfitting by ensuring a good bias-variance trade off. We did something similar for predicting arrest precinct except we used 200 trees and a max depth of 800 just to verify that the accuracy remains similar to the outputs from the SVM and KNN.

K Nearest Neighbor (KNN) algorithm is a supervised algorithm. Based on the most relevant features we will be using KNN to predict the violence level of the crime committed. This will be working according to the nearest neighbor. When implementing the KNN algorithm we experimented with different parameters to see what would give us favorable outcomes. An increase to the "n-neighbors" parameter from the default 5 to 40 and a change to the "weights" variable from uniform to distance resulted in slightly better accuracy, and to slightly reduce the runtime of the algorithm we increased the "leaf-size" without sacrificing any accuracy. An adjustment in these variables improved our accuracy for arrest borough prediction by approximately 9 percent and the arrest precinct prediction by approximately 2 percent.

Support Vector Algorithm (SVM) algorithm will be implemented using ANOVA Radial Basis Kernel to alter the dimension of the data. By doing this, we hope to create a hyperplane in such a way that it will accurately predict the violence level of crimes based on relevant features. The SVM essentially used the RBF kernel function to generate hyperplanes and split the classes. Since we had a lot of features after doing our one hot encoding, it took the SVM a very long time to train as we had expected. The SVM model was also overfitting just as we had expected too. The accuracy score however was very close to the one we had for random forest. To reduce overfitting, we could have simply used lesser features to train the model. That would have added some error. When we were trying to predict precinct, the SVM's accuracy score was very similar to what we were getting for the random forest and KNN. This is expected as any crime and occur near any precinct. So, there may not be any relation between the features and the precinct the arrest was booked into.

C. Implementation Details

Project includes, applying algorithms such as KNN, Random Forest, and Support Vector Machine (SVM) to obtain insights and make predictions. First comes the pre-processing step, where all the missing and irrelevant data was removed.

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format. Data cleaning is a fundamental step to generate accurate analysis or visualization. Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. When collecting data from several streams and with manual input from users, information can carry mistakes, be incorrectly inputted, or have gaps. There were some features which were un-necessary or irrelevant for our analysis and visualization. To clean this data, we used data reduction method. Which states that 'The highly relevant attributes should be used, rest all can be discarded'. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we used data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. For performing attribute selection, we checked each data attribute according to the significance level for the project and the attributes which were not relevant enough, we discarded those attributes by using 'drop' method. Moreover, some of the data was missing and that was causing inaccurate results. To solve this issue, we again used data reduction and discarded those rows. The reason we did not use data transformation method is because it would be unfair to the data on which we were working on. Missing values were causing incorrect visualization of some features. Considering this problem, all the missing or null values were removed successfully. Some of the independent variables were irrelevant for further implementation of algorithms. Therefore 11 of these features were dropped from final data to be used further. We have made different visualizations for better understanding of the data. List of the crime and count of each crime is represented through a bar graph using matplotlib. Pie chart shows the percentage of criminals according to age groups. Comparison of number of criminals in each group is shown through bar graph. Also Number of crimes committed according to different borough is represented through bar graph. Dependent variable violence-level shows us the level of the crime committed which could be least violent or medium violent or most violent. These are classified according the kind of the crime committed. Using folium library we have created a heat map of New York City which shows the density of crimes committed. On the heat map red color shows more density of crimes similarly yellow color shows medium density of crimes and green color shows very less density of crimes. From this heat map we can clearly see that which area of the city is dangerous or unsafe. Following are the algorithms which we are going to use in the project.

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. When implementing the KNN algorithm we experimented with different parameters to see what would give us favorable outcomes. An increase to the "n-neighbors" parameter from the default 5 to 40 and a change to the "weights" variable from uniform to distance resulted in slightly better accuracy, and to slightly reduce the runtime of the algorithm we increased the "leaf-size" without sacrificing any accuracy. An adjustment in these variables improved our accuracy for arrest borough prediction by approximately 9 percent and the arrest precinct prediction by approximately 2 percent.

When trying to predict arrest borough, we created a random forest using 10 trees in the forest, a max depth of 30 to essentially prune the tree and reduce overfitting. We chose random forest as randomly chooses features to grow out the trees. This also helps with reducing overfitting and leads to a more accurate prediction. We also set Bootstrapping to be true in order to ensure we are not providing the whole dataset to each tree in the forest. This reduces variance and avoids overfitting by ensuring a good bias-variance trade off. We did something similar for predicting arrest precinct except we used 200 trees and a max depth of 800 just to verify that the accuracy remains similar to the outputs from the SVM and KNN.

"Support Vector Machine" (SVM) is a supervise machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. The SVM essentially used the RBF kernel function to generate hyperplanes and split the classes. Since we had a lot of features after doing our one hot encoding, it took the SVM a very long time to train as we had expected. The SVM model was also overfitting just as we had expected too. The accuracy score however was very close to the one we had for random forest. To reduce overfitting, we could have simply used lesser features to train the model. That would have added some error. When we were trying to predict precinct, the SVM's accuracy score was very similar to what we were getting for the random forest and KNN. This is expected as any crime and occur near any precinct. So, there may not be any relation between the features and the precinct the arrest was booked into.

We have also done some exploratory data analysis based on the dataset. Using different features in the data set we have made some visualizations of that data for better understanding of the information. In Fig. 7 we can see the bar chart of each unique crimes committed. On x-axis there is list of offenses and on y-axis the count can be seen. There are various crimes like sex crimes, assault, robbery, fraud, use of dangerous weapons, etc. As it can be observed that the highest number goes over 6000, which is assault 3 and related offenses and the lowest number of crimes committed in New York City are just 1 and those are jostling, homicide-negligent-vehicle, and administrative codes. These crimes were committed in New York City.

In Fig. 5 we can see the pie chart of arrests made by age group. In data there is a feature of age group of the criminal. We have made a simple pie chart of the percentage of criminals in each age group. In this we can see that the most of the crimes were committed by 25 to 44 age group and the least crimes were committed by 65 plus age group. Fro this we can conclude that the most active age group in committing crimes is a young age group. 57.01 percent of all the crimes committed were of 25 to 44 age group. Surprisingly there were much more crimes committed by 18 to 24 age group than expected which is 19.77 percent. There were very less crimes committed by less than 18 years of age group which is 3.07

percent. 45 to 64 age group is also much active like 18 to 24 age group. Percentage of 45 to 64 age group in committing crimes is 18.78 percent. We have also plotted a bar graph of the same which can be seen in Fig. 4 Here we can see the actual count of crimes committed by each age group. On x-axis we can see the various age groups and on y-axis the count of the crimes committed. As we have seen in pie chart 25 to 44 age is group contributes the most in committing crimes. The number of crimes committed by this age group is 22309 which is a lot as compared to 65 plus age group. Criminals in 65 plus age group is 538.

In Fig. 11 we can seen the pie chart of arrest made by race. In this it can be observed that the most criminals are from black race and the least is American Indian/Alaskan native. Secondly, white hispanic is the race which committed considerable amount of crimes. White race have committed 10.60 percent of crimes. There is also an unknown race which contributes 0.37 percent of crimes. Black hispanic crime has committed 8.65 percent of crimes. Asian / Pacific islander race have contributed 5.07 percent towards committing crimes. In Fig. 12 we can see the bar chart of arrests made by race. Here we can see the actual count of the crimes committed by each race. As we have seen in pie chart the highest race which have committed the most crimes is black race. Then count of crimes committed by black race is 19636 and the least is 99 which was by American Indian / Alaskan native. The unknown race has committed 145 crimes as well.

The Fig. 9 shows a bar chart depicting the disparity between the number of males arrested vs the number of females arrested. Because the disparity was so great, we had to add a logy parameter when plotting to ensure that the female bar is visible. As we can see here, there may be a certain police bias towards males than females or it is just that females tend to commit much fewer crimes than males. This data can be used to try and maybe understand the psychology between the 2 genders and see why females don't commit crime as much and what drives males to commit more crimes. To us, this graph shows us that women tend to be calmer and more prudent than men in a conflict as we saw that most arrests in NYC are 3rd degree assaults followed by dangerous drugs and felony assaults. A 3rd degree assault is basically a reckless assault carried out by an individual with a deadly weapon. It's not wrong to assume that something must have motivated such a confrontation with another person. A disagreement perhaps may have triggered a fight to break out where the police are called. In 2017, a study was done in US high schools where it shows 30 percent male students got into physical fights with others as compared to 17.2 percent of female students. This just shows that even statistically from a young age, males tend to fight more often than females.

In this Fig. 8, we have a bar chart which depicts the number of arrests made in each borough. New York City is comprised of 5 boroughs which is Bronx, Brooklyn, Manhattan, Queens and Staten Island. Brooklyn has the largest population followed by, Queens, Manhattan, Bronx, and lastly Staten Island. Since Brooklyn has the largest population, we can make the inference that there are more arrests in Brooklyn because there is a larger population. However, what becomes interesting is that Queens which has the second biggest population in all of NYC has lesser than Manhattan. As we can see, Manhattan has the second largest amount of arrests in all 5 boroughs. This could be because of population density. Even though Queens has a larger population, the population density in Manhattan is much greater which may cause more arrests as more people bump into each other. Manhattan is also where a lot of tourists visit which could also explain the number of arrests there. Staten Island being the least populated also shows the least number of arrests out of all the boroughs. From this bar chart, we can infer than important features which affect number of arrests per borough is the population size, population density, and size of the borough.

In Fig. 10 this chart, our aim was to visually show the number of arrests made per precinct. Now the first thing that might catch your eye is precinct 22 having only 8 arrests. When we first saw the disparity in the number of arrests, we researched more about the precinct. What we found is that precinct 22 is actually called Central Park precinct. Their role is primarily to keep central park safe which is why they must have had very few arrests. Central park isn't exactly a very crime ridden place. It's simply a part where most people go to get a small escape from the hustle and bustle of the city. However, we still thought 8 arrests is far too low a number for such a large area. This was until we found that the data was last updated on May 4th. Since then there have been more arrests which has bumped that number up to 12 just for this year. When I took a look at the historical number of arrests by year, we can see that crime in the central park area has been drastically reducing indicating that NYC is becoming a much safer place. That being said, judging from historical numbers like last year and even this year could be skewed as a lot of people have been under lockdown and quarantine. In 2020, the number of total arrests in central park was just 56 as compared to the 127 in 2001, and the 187 in 1998. In 2021, since the year is not over yet, 12 arrests seems like a number that will likely grow and so we can accept this. As for the other precincts with large arrests, we found a majority of them to be located in Brooklyn and Manhattan which are the 2 boroughs with the most number of arrests. This falls in line with our earlier bar graph showing the number of arrests per borough.

In this Fig. 13, we created clusters of the arrests which are then grouped together based on locality. This map is interactive and can show us some very interesting insight. We were able to locate some centers of organized crimes, some areas which showed us that there was a high amount of drug trafficking, assaults etc. These are basically areas that are simply not cop friendly. The people living there may have an issue with authority and some there are frequent times when the police are called by concerned citizens about robberies, assaults, etc. We were able to locate 1 such area in Staten Island where we found a cluster of 32 arrests. All very similar to each other: robberies, assaults and assaults on public officials. The city needs to understand how to properly allocate resources and assist the people in these areas. Perhaps they lack opportunities in their locality and end up in bad spots. Maybe there is a lacking of the education system or whatever it may be. There are many of these types of clusters. We had found another cluster right under a bridge where there were plenty of apartments. There was a string of robberies and assaults in this apartment complex. This clearly shows that the people in that apartment are not happy. Those constant fights and robberies are wasting valuable time of police officers and city resources. Therefore,

the city needs to understand what is going on there to reduce the amount of calls from those areas. These hotspots are areas with frequent calls to the NYPD. To reduce this strain, the city government should step up and try to solve the root of the issues. Otherwise, these frequent calls to hotspots can even put the life of an officer in danger as well. Other common patterns we found were people speeding on the FDR so there were clusters of vehicle violations such as seatbelts, speeding, and occasionally driving under the influence.

Here, in Fig. 6 we created a sort of a “heat map” to visually depict the number of arrests in certain parts. The areas in red show areas where there were large amounts of arrests where green areas are places with few arrests. Then there are yellow areas which are areas with a medium number of arrests. Now it is not necessarily the case that red areas are “unsafe” and green areas are “safe”. As I had mentioned earlier, population density plays a big role and since Manhattan has a high population density, the number of arrests there will be large but also the distances between arrests will be smaller since the borough of Manhattan is actually quite small in comparison with the other boroughs. It only has a land area of 59.13 km square. So because Manhattan is so small, we will see smaller distances between arrests resulting in a “red” area. This is similar to the problem statisticians faced at the center for naval analysis in ww2 when they tried to find out how they can improve the survivability of the bombers. They later found out they were adding armor to the places that were getting shot and not to the places that weren’t because the places that weren’t getting shot, those planes weren’t coming back to base and being recorded in the data. This is likely the same scenario. Some places are getting targeted by others more often and so there are more arrests in those areas. Whereas, other areas may be unsafe but the NYPD does not have frequent patrols through those places which means there is fewer arrests there but the area is actually unsafe. Other boroughs are just so big that there might be some criminal activity that goes on unnoticed by the police. However, as we can see on the map, there are also some arrest hotspots in larger boroughs like Brooklyn and Queens.

IV. Comparison

We have used all three algorithms to predict 2 aspects namely for arrest by borough and for arrest by precinct. In both of them we have observed very interesting results. For arrest by borough there was good performances by all three algorithms. In which Random forest algorithm shows the highest results of accuracy. On the other hand, support vector machine algorithm resulted in overfitting and took a lot of time to get trained as compared to other two algorithms. Also, for arrest by borough, k - nearest neighbor performed well with 84 percent of accuracy. The highest accuracy is 94 percent and it is of random forest algorithm. As support vector machine resulted in overfitting it is shows accuracy of 100 percent. The probable reason for overfitting would be, we have trained the support vector machine algorithm with a lot of features and that is why it took too much time to train as well. We have also applied the algorithms for testing the accuracy for arrest by precinct as well. In this, the accuracy of all the algorithms are pretty low as compared to accuracies for arrest by borough. In case of arrest by precinct all of the algorithms shows low accuracies. For arrest by precinct, support vector machine algorithm shows the best performance in case of

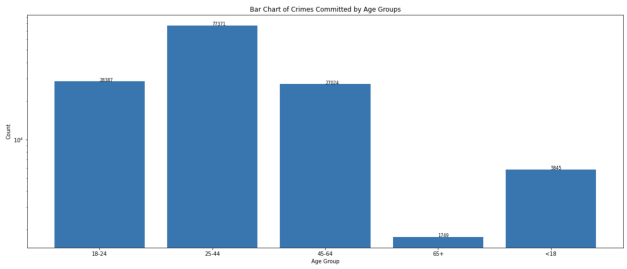

Fig. 4. Bar Chart of crimes committed by Age Groups

accuracy and the lowest accuracy was of k - nearest neighbor algorithm. Random forest algorithm performed as average of all the three algorithm. Highest accuracy is of support vector machine algorithm with 21 precent where the lowest is k - nearest neighbor with 19 percent accuracy. Random forest resulted in pretty decent accuracy score with 20 percent.

V. Future Directions

In future our goal would be to identify patterns of the types of crime committed in certain locations and predict the possibility of the crime. We would find the most accurate method for prediction. We will need to clean the data more to find clear relationships between the data and features. Provided some more time, we could predict the possible crime at possible locations. Our aim is to possibly present this to the city of NYC so that they may use its results to plan out patrol routes and figure out proper allocation of city resources.

VI. Conclusion

According to accuracy of the algorithms the the highest one is Random Forest which is 94 percent for arrest by borough. KNN also performs well on the dataset with the accuracy of 86 percent. SVM is showing 100 percent accuracy because of overfitting. The possible reason for low accuracy in arrest by precinct is an independent variable and has very little correlation with the input features. Hence, for this dataset Random Forest algorithm performs the best for arrest by borough. SVM performed best for arrest by precinct.

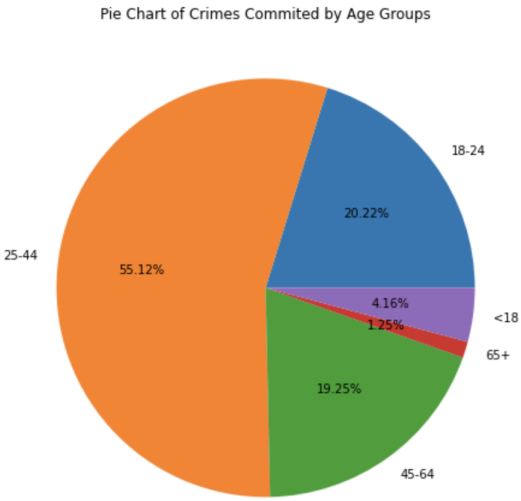

Fig. 5. Pie Chart of crimes committed by Age Groups

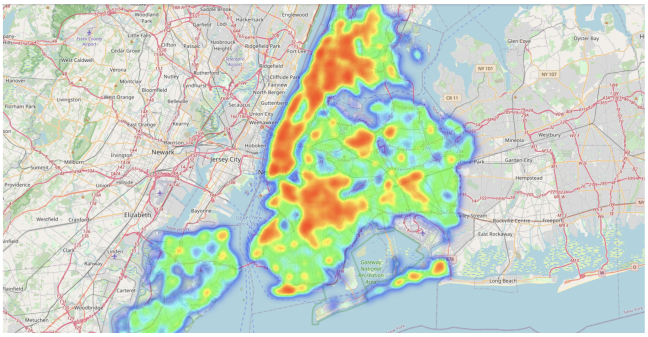

Fig. 6. Heat Map

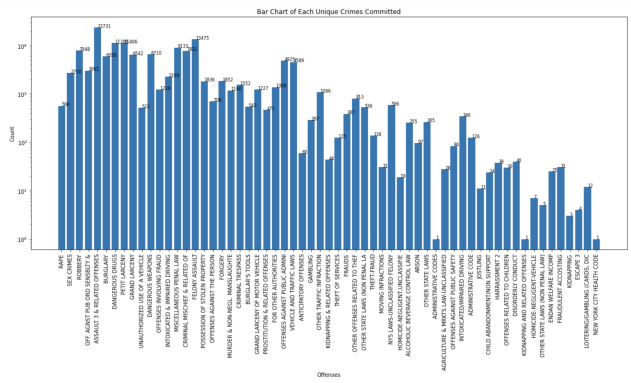

Fig. 7. Bar chart according to each crime

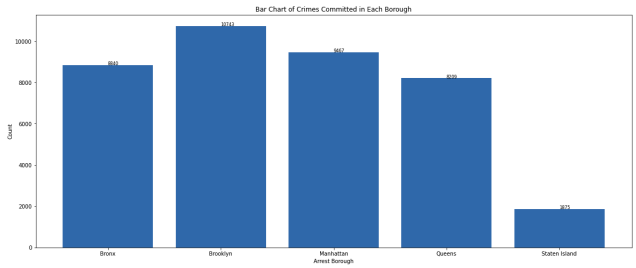

Fig. 8. Bar Chart of arrest by Borough

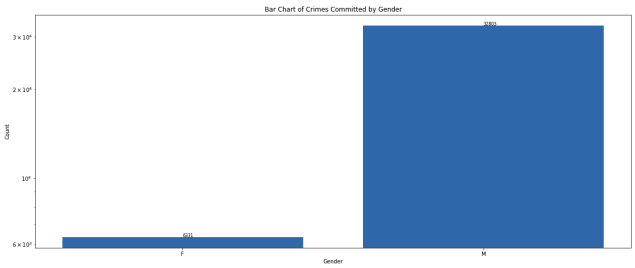

Fig. 9. Bar Chart of arrest by Gender

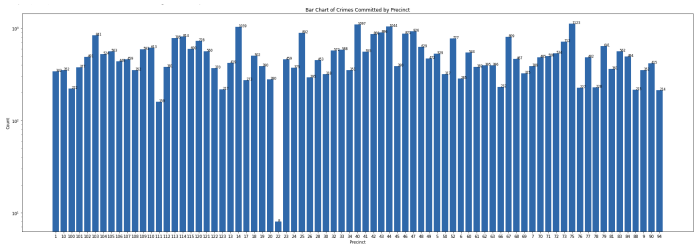

Fig. 10. Bar Chart of arrest by Precinct

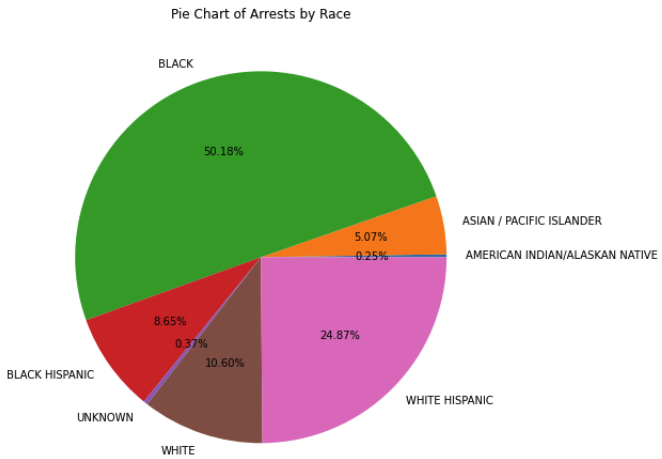

Fig. 11. Pie Chart of arrest by Race

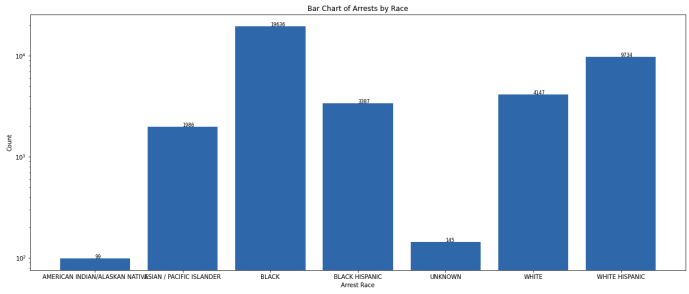

Fig. 12. Bar Chart of arrest by Race

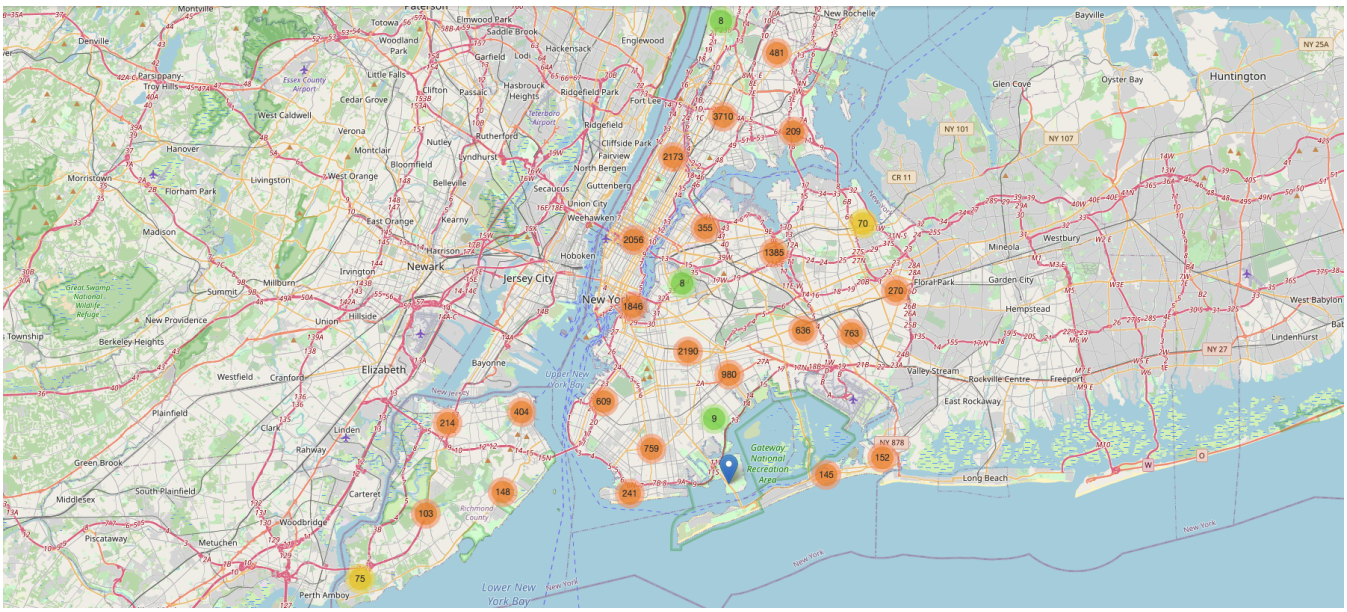

Fig. 13. Heat Map

	arrest_date	pd_desc	ofns_desc	arrest_boro	arrest_precinct	age_group	perp_sex	perp_race	x_coord_cd	y_coord_cd	latitude	longitude	violence_level
0	2021-03-26T00:00:00.000	SEXUAL ABUSE	SEX CRIMES	M	5	25-44	M	WHITE	984946	200203	40.716195914000025	-73.99749074599998	Medium Violent
1	2021-03-25T00:00:00.000	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	K	84	18-24	M	BLACK	987792	190460	40.689452980000056	-73.98722940999993	Most Violent
2	2021-03-18T00:00:00.000	ASSAULT 3	ASSAULT 3 & RELATED OFFENSES	B	48	18-24	M	BLACK	1011811	246833	40.844139945000045	-73.90038861799998	Most Violent
3	2021-03-17T00:00:00.000	STRANGULATION 1ST	FELONY ASSAULT	K	79	25-44	M	WHITE	1000272	187022	40.68000266400003	-73.942236182	Most Violent
4	2021-03-14T00:00:00.000	STRANGULATION 1ST	FELONY ASSAULT	B	49	25-44	M	BLACK	1021639	250223	40.85340834100003	-73.86484873899997	Most Violent
5	2021-03-05T00:00:00.000	ROBBERY,OPEN AREA UNCLASSIFIED	ROBBERY	B	47	18-24	M	BLACK	1026486	262591	40.88733281800006	-73.84725001299995	Least Violent
6	2021-03-05T00:00:00.000	BURGLARY,UNCLASSIFIED,UNKNOWN	BURGLARY	K	90	25-44	M	BLACK	998016	196598	40.70629036500002	-73.95035033099998	Medium Violent
7	2021-03-05T00:00:00.000	ROBBERY,OPEN AREA UNCLASSIFIED	ROBBERY	M	18	25-44	M	BLACK	988248	215413	40.75794273400004	-73.98557030899997	Least Violent
8	2021-03-04T00:00:00.000	TRESPASS 2, CRIMINAL	CRIMINAL TRESPASS	M	26	45-64	M	BLACK	997129	237012	40.817217733000064	-73.95347213499997	Least Violent
9	2021-03-03T00:00:00.000	RAPE 1	RAPE	B	52	25-44	M	BLACK	1017542	255919	40.86905853200005	-73.87963014799993	Most Violent
10	2021-03-03T00:00:00.000	WEAPONS POSSESSION 1 & 2	DANGEROUS WEAPONS	M	25	<18	M	BLACK	1002065	231446	40.80193203300007	-73.93565410599996	Most Violent
11	2021-02-23T00:00:00.000	STRANGULATION 1ST	FELONY ASSAULT	K	70	18-24	M	BLACK	995679	162130	40.611686615000046	-73.95883781899995	Most Violent
12	2021-02-23T00:00:00.000	STRANGULATION 1ST	FELONY ASSAULT	K	75	25-44	M	BLACK	1017119	183909	40.67141166300007	-73.88151172399995	Most Violent
13	2021-02-19T00:00:00.000	ROBBERY,OPEN AREA UNCLASSIFIED	ROBBERY	B	41	25-44	M	BLACK	1013086	236614	40.81608766100004	-73.89582435399994	Least Violent
14	2021-02-09T00:00:00.000	ASSAULT 2,1,UNCLASSIFIED	FELONY ASSAULT	K	81	45-64	M	BLACK	1005312	190540	40.68964821100008	-73.92405412199997	Most Violent
15	2021-02-06T00:00:00.000	SEXUAL ABUSE	SEX CRIMES	K	69	45-64	M	BLACK	1010641	175525	40.64842111900003	-73.90489710899999	Medium Violent

Fig. 14. Data set with dependent variable