Mini Project 2

Vijay Teja Gottipati, 110622613

Introduction

The goal of this project is to do exploratory analysis on the 'Stop and Frisk' data provided by NYPD as part of New York Open Data and utilize it to make the city a better place. The dataset consists of csv files for data corresponding to each year, starting from 2003 to 2013. Each row contains information related to the event of an officer that either stopped, frisked or searched a person. There are over 100 columns and they can be broadly categorized as columns containing details about:
Location of the stop such as street, city, intersection, zip code, city, etc.
Details about the person: race, sex, age, build, height, weight, build, eye and hair color
Reason for:
Stop – fits a relevant description, suspicious bulge, actions of engaging in violent crime etc.
Frisk – verbal threats, suspicious bulge, inappropriate attire, prior criminal behavior etc.
Search – admission by suspect, hard object, outline of weapon, other
Objects found on person: contraband, gun, assault rifle etc.
Whether an arrest was made etc.

Problem Statement

The primary motivation of this project is to find out:

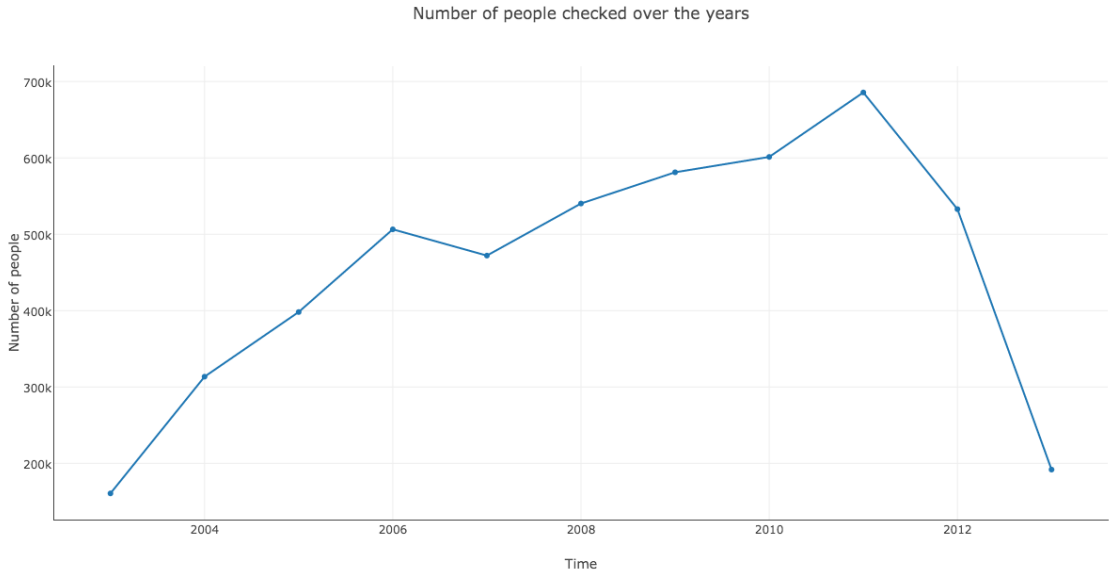
1. Are there any racial biases in the selection of people to stop/frisk/search?
2. Are there any distinguishable characteristics of people that make them more likely to be criminal?

Method, Results and Discussion

Combining all the datasets for all the 11 years together, we get a total of ~4.99 million records. Out of these records there were invalid values: some of which could be corrected and some of which have caused the rows to be discarded. Examples of errors that could be corrected were typos in the persons features, such a invalid eyecolor 'Z' being replaced with 'ZZ' etc. Some of the records that were discarded were where age>150 years or weight >1500 lbs etc. These values could be replaced with their mode/mean instead of discarding the rows. But I chose to discard them since there was abundant amount of data still left. From the 4.99 million records, post cleaning ~70k records were removed resulting in a dataset of ~4.92 million records.

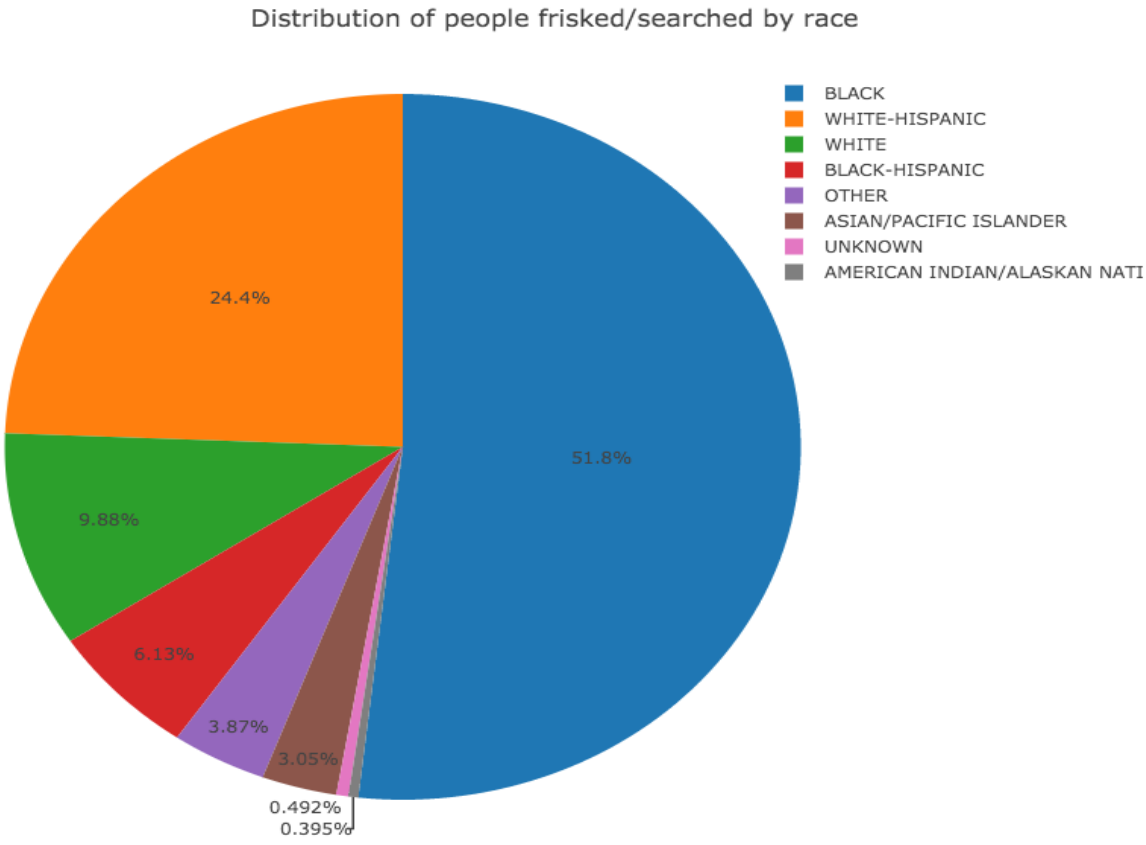
After cleaning the data, we can first plot to see how the number of people stopped changed over time. From Fig 1, we can see that the number increases from 2003 and peaks at 2011. The peek could maybe be attributed to the 9/11 attacks and the subsequent drop to the controversy raised by increasing stop and frisk activity.

Figure 1



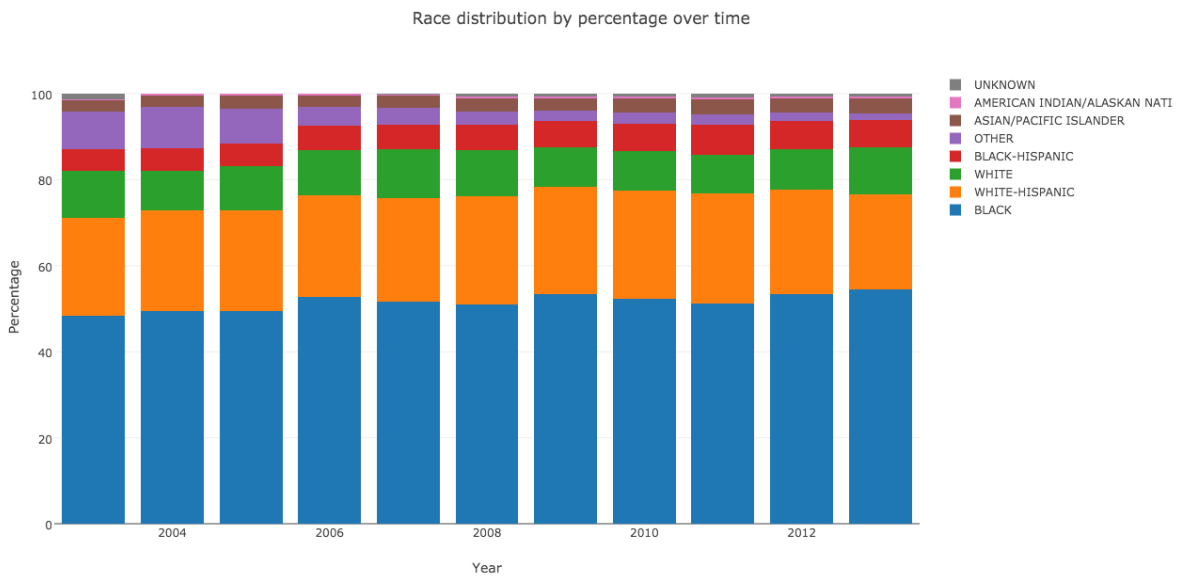
Let us now move towards the 1st questions and examine the distribution of people by race. From Fig 2, we can see that the overwhelming majority of the people that were stopped were black. The next highest races are white/black Hispanics. This is surprising since White is the majority race of NY with more than 70% population and yet there are underrepresented in this data with just 10%. There appears to be strong racial bias in deciding whether to stop a person or not.

Figure 2



Let us check if there is any change in the race distribution over time. From Fig 3 we can make a few comments:
The American Indian percentage is steadily reducing over time.
The number of Asian people checked is increasing over time. Similarly for Black-Hispanic.
The distribution of the other races have remained similar over time.

Figure 3



One aspect we could explore further is whether the race distribution of people who were searched is vastly different from people who were arrested – something that can substantiate racial bias. We can plot the race distribution side by side in a grouped bar chart as in Fig 4. The distribution appears to be more or less the same across stopped, and stopped and arrested people. Black has marginally less percentage in arrested set while the remaining races have marginally more percentage.

Figure 4

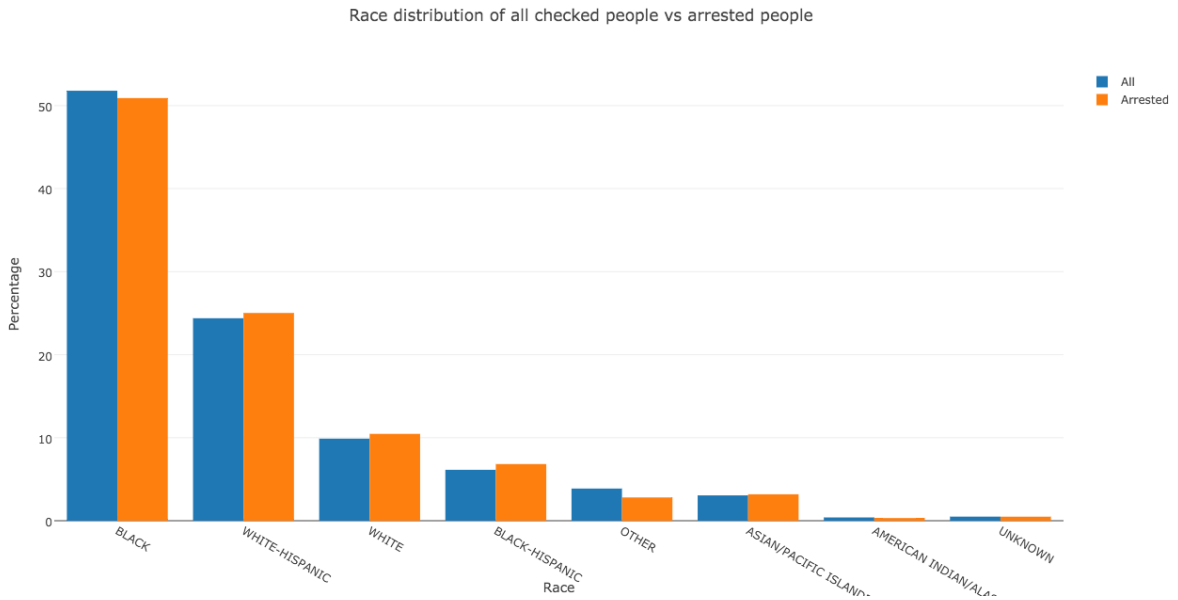
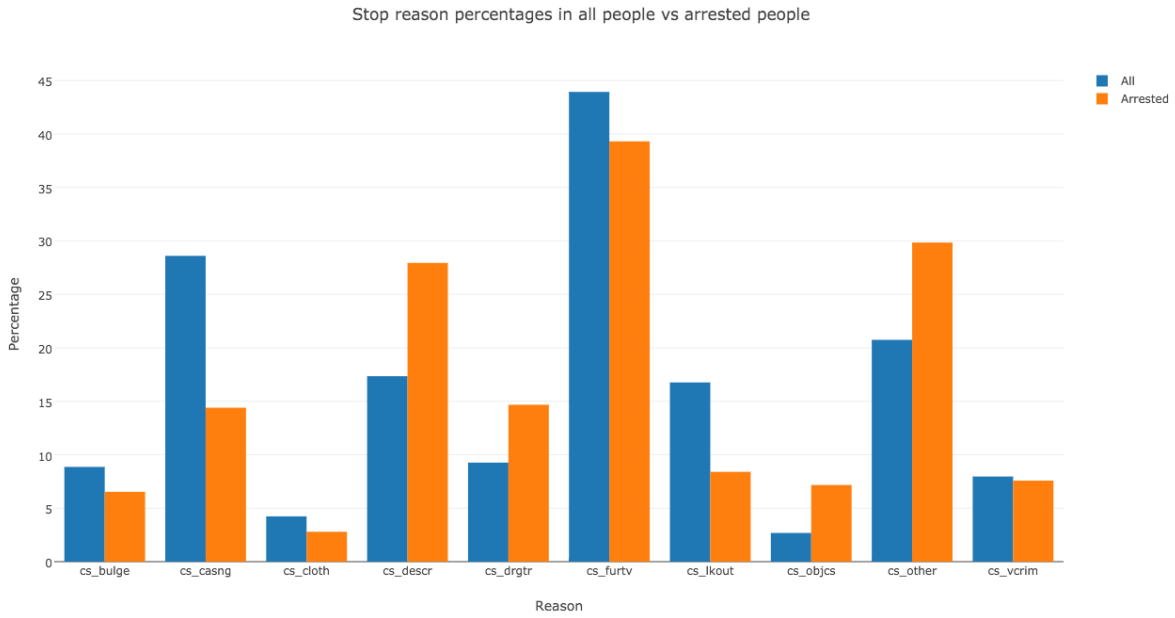
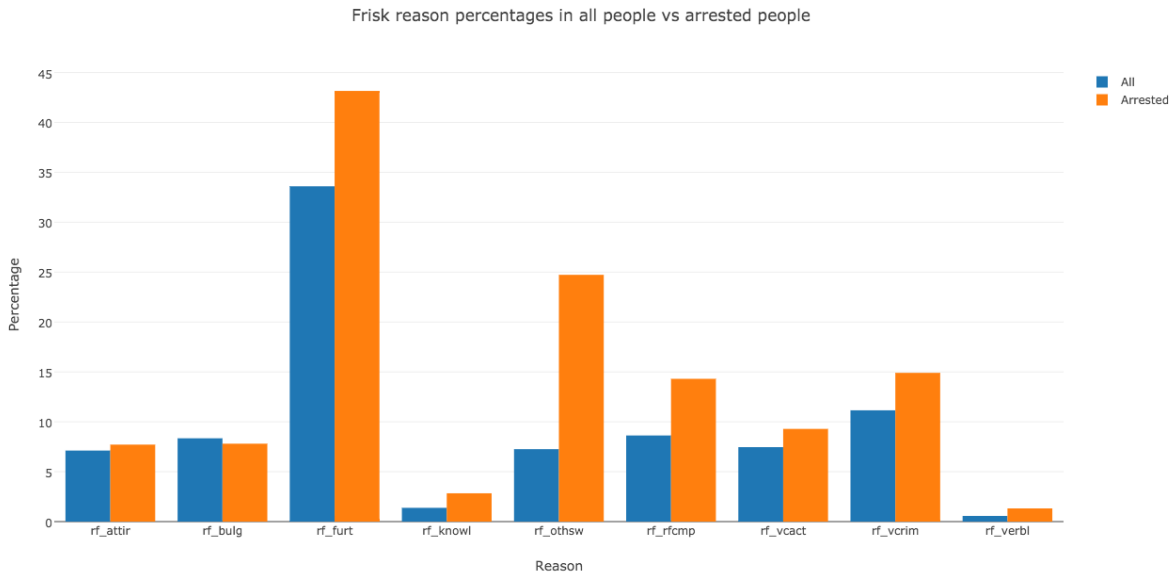


Figure 2



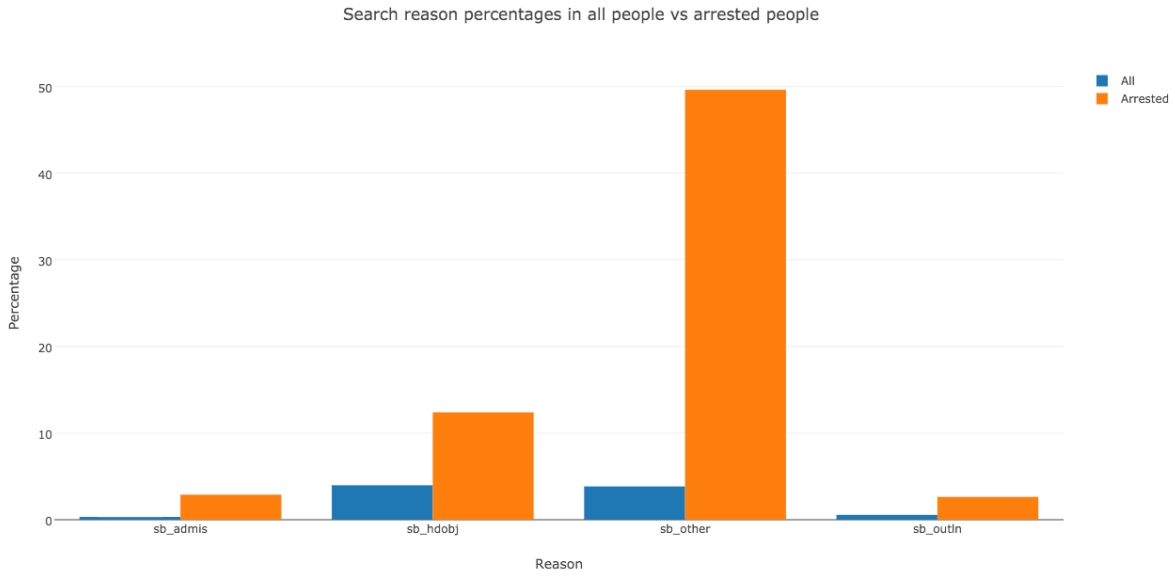
Let us now move on to second question. Out of the 4.7mil people stopped in the data set, only ~288K people were arrested i.e. just 5.86% of the people were arrested. This stat demonstrates the ineffectiveness of the criteria being used to decide if a person is a potential criminal. Let us see what reason flags are more popular in general and which flags are more popular in the arrested set. Fig 5, 6 and 7 are the plots comparing Stop, Frisk and Search reasons. From Fig6, we can see cs_descr(fits relevant description), cs_drgtr(actions indicative of a drug transaction), cs_objcs(carrying suspicious object) and cs_other are better indicators of potential criminals. We can also see cs_casng(casing a victim/location) and cs_lkout(suspect acting as lookout) have low accuracy in predicting criminal behavior.

Figure 6



From Fig 6 and 7, we can see that almost all reasons have more percentage in arrested than general set. This is because both frisk and search are deeper level of inspection than stopping, in that order. ~2.7mil people are frisked out of the 4.9mil people and ~246k frisks (9%) out of these lead to an arrest. Similarly ~400k people are searched out of the 4.9mil people and ~187k (46%) people out of these lead to an arrest.

Figure 7



From the data if we could extract a model that sufficiently identifies what characteristics describe a criminal, it would be very beneficial to both the police and the public. The police need not waste their time and resources trying to stop and question innocent people and similarly, the citizens need not be unnecessarily questioned and go through the inconvenience of invading frisks and search. This model needs to minimize invasive frisking/searching. Hence we cannot rely on features that result from frisking/searching such as whether they have a gun/contraband etc. Hence, the features could be: sex, race, height, weight, age, hair-color, eye-color, build, whether they are other people with them, possible stop reasons and whether the location is inside or outside.

Since most of the features are categorical, and what we want to build is a binary classifier (whether the person is a criminal – going to be arrested), Naïve-Bayes could be a good suitable model.

For applying the model, some of the continuous features like age, weight were discretized. Each categorical variable with more than 2 possible values were further split into n variables where n is the distinct number of values that variable can take. Eg. Race is split into 8 variables – one for each race.

Once all the pre-processing is done to make the input suitable for the model, the data was split into training and test data of the ratio 67:33. The model was trained on the training data and was tested on the test data with the following evaluation results:

Class	Precision	Recall	F1-score	Support
0	0.95	0.88	0.91	1520664
1	0.11	0.24	0.15	94838
Avg/Total	0.90	0.84	0.87	1615502

The confusion matrix is:

	Predicted 0	Predicted 1
Actual 0	1339538	181126
Actual 1	72016	22822

Class 0 indicates 'Not arrested' and 1 indicates 'Arrested'. Looking only at the F1-Score it appears that the model is pretty good. But looking at the predictions for class 1, both the precision and recall are very bad. In fact, with this particular data, even a model that always predicts '0' would have a pretty good accuracy because only 5% of the entire data has '1' label. Building a model that predicts only 0, we get the following evaluation and confusion table:

Class	Precision	Recall	F1-score	Support
0	0.94	1.00	0.97	1520664
1	0.00	0.00	0.00	94838
Avg/Total	0.89	0.94	0.91	1615502

	Predicted 0	Predicted 1
Actual 0	1520664	0
Actual 1	94838	0

We can see that the F1 score of the always-no predictor is higher than the Naive-Bayes model.

Conclusion

In conclusion we have seen that there is a huge difference between the race distribution of people in NY and the race distribution for people being stopped. White people are hardly stopped while the Black and Hispanic are always questioned. The criteria for stopping is inefficient because only 5% of the people stopped are criminals. To help in deciding a better criteria, we tried to build a Naive-Bayes classifier. Even though on the surface it looks like a good model with a good F1-score, it is not suitable for practical use since the Precision and Recall for the arrest class is too low to be effective. This is because the amount of data for arrested class is so small (in terms of %) that any pattern found for arrested set would have a higher support from the non-arrested set due to the sheer size of the non-arrested set.