New York Policing Strategy Case Study

A CRISP-DM study on the Stop-Question-Frisk policy employed by the NYPD

Timothy Lau

000836623

DATA 475

**Introduction**

The stop-question-frisk (SQF) program is a practice employed by police departments in the United States and most famously in New York; it typically entails the detaining, questioning, and occasional searching of potential suspects for possession of contraband. Prior to the 1980’s, an officer only had the authority to stop and question someone under the suspicion of possible crime, and the authority to frisk someone under the suspicion of weapons possession. Since then, their authority and subsequent use of this policy has grown under “broken windows”, a criminological theory that proposes “low-level crime and disorder creates an environment that encourages more serious crime”. Originally used for possession of weapons and other illegal materiel, the purpose of SQF has expanded to more preventative measures.

A simple metric for the effectiveness of this policy would be the percentage of convictions of suspects taken in from SQF to the total SQF stops made by the police. This would measure how much crime it is able to prevent. Other methods include regression models and clustering, which would present how each feature affects the probability of someone being pulled over. This would perhaps portray certain biases that the police in the US have been accused of having.

**Data**

The data pertains to stops by the police in New York in 2012, described below.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Meaning** | **Type** |
| **Year** | Year of stop | int, nominal |
| **pct** | precinct of stop | int, nominal |
| **ser\_num** | Serial number | int, nominal |
| **datestop** | Date of stop (mm-dd-yyyy) | int, nominal |
| **timestop** | time of stop (hh-mm) | int, continuous |
| **recstat, inout, trhsloc** | description of location for stop | str, boolean |
| **perobs, perstop** | period suspect was under observation (mmm)/suspect was stopped (mmm) | int, continuous |
| **crimsusp** | was suspect suspected of a crime | str, nominal |
| **typeofid, explnstp, othpers, arstmade, arstoffn** | description of stop; type of id of suspect, explanation of stop, if others were involved, was an arrest made | str, boolean |
| **sumissue, sumoffen** | description of summons | str, boolean |
| **compyear, comppct** | complaint of officer | str, boolean |
| **offunif, officrid** | description of state of officer | str, boolean |
| **frisked, searched, contrbn, adtlrept, pistol, riflshot, asltweap, knifcuti, machgun, othrweap** | description of what the officer found | str, boolean |
| **pf\_""** | description of force utilized by officer | str, boolean |
| **radio** | radio run | str, boolean |
| **ac\_""** | additional circumstances | str, boolean |
| **rf\_""** | reasons for frisk | str, boolean |
| **cs\_""** | reason for stop | str, boolean |
| **sb\_""** | basis for search | str, boolean |
| **forceuse** | reason force used | str |
| **sex, race, dob, age, ht\_feet/inch, weight, haircolr, eyecolor, build, othfeatr** | description of suspect | str, int depending on variable |
| **addrtyp, rescode, premtype, premname, addrnum, stname, stinter, crossst, aptnum, city, state, zip, addrpct, sector, beat, post, xcoord, ycoord** | description of stop location | str |
| **""CM** | crime details, decription | str |

Table 1. Description (meaning, type) of variables found within dataset

There are all kinds of different attributes affecting data quality. Typically, missing data is deliberate. For example, if there was not an arrest made, then arstoffn (offense suspect was arrested for) would be left blank. Missing values that are string values can be fill with the fillna() function. Missing values with int values can typically be filled with the mean of the data. Another method for filling missing value is by taking a regression of the data and simulating the missing values. Outliers in the data are mistakes; for example there are ages of suspects that range outside of being reasonably detained by the police, such as below 5, while other outliers seem to be typos, such as the ones ranging above 100. Another example is a record with 532911 from perobs (period suspect was under observation), which amounts to nearly 9000 hours. These records will be taken out of the data considered for the model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **perobs** | **perstop** | **sex** | **age** | **ht\_feet** | **weight** | **arstmade** |
| **count** | 523807 | 523567 | 523807 | 523807 | 523807 | 523807 | 523807 |
| **mean** | 2.446258 | 5.46332 | 0.927553 | 28.7217 | 5.181975 | 168.8896 | 0.061005 |
| **std** | 5.053645 | 4.276332 | 0.259226 | 23.14235 | 0.39605 | 28.79468 | 0.23934 |
| **min** | 0 | 0 | 0 | 0 | 3 | 50 | 0 |
| **25%** | 1 | 3 | 1 | 19 | 5 | 150 | 0 |
| **50%** | 1 | 5 | 1 | 24 | 5 | 165 | 0 |
| **75%** | 2 | 5 | 1 | 34 | 5 | 180 | 0 |
| **max** | 955 | 99 | 1 | 999 | 7 | 350 | 1 |

Table 2. Simple statistics for attributes

These statistics paint a pretty interesting picture. On average, of the 520,000 stopped by officers, only 6% are arrested. However, suspects are observed for about 2 to 3 minutes before being detained; about 68% of the detained are observed between 0 and 7 minutes. Such a short amount of observation could be explained by the idea of “broken windows”, but could also explain why the number of arrests made were so low. Other significant statistics include that most of the suspects subjected to SQF are typically male and in their 20-30s.

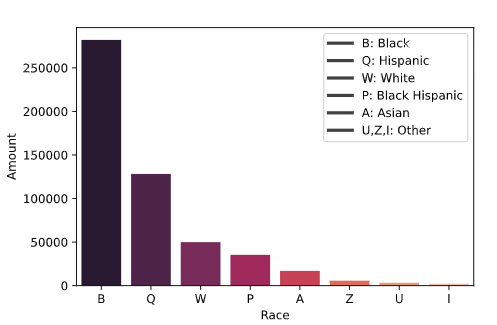


Figure 1. Amount of people subjected to SQF segregated by race.

In Figure 1, it is apparent that black people are disproportionately stopped by the police. They are more than twice as likely to be stopped as the next group of people, and have been stopped more than anyone else combined in 2012.

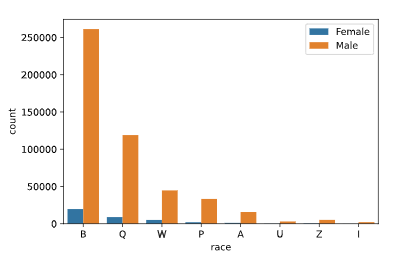


Figure 2. Amount of people subjected to SQF segregated by race and gender.

People subjected to SQF are not only skewed by race, but by sex as well. Those that were stopped in 2012 were overwhelmingly male. It is interesting to note that this is true across the board for race.

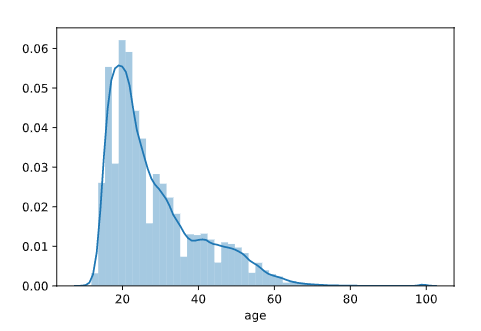


Figure 3. Age distribution of people stopped by SQF.

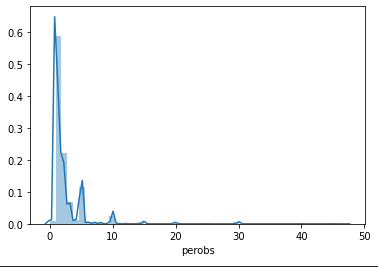


Figure 4. Distribution of period of observation a person is subjected to before being stopped by SQF.

Age distribution is typically skewed towards young adults, whereas the elderly are nearly never stopped. People stopped are typically observed for a couple of minutes, with the majority not being observed at all. There are spikes at over 5 minutes are much rarer.

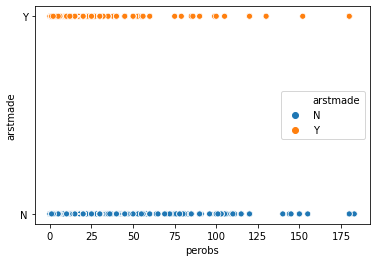


Figure 5. Scatterplot of period of observations separated by those not arrested and those arrested.

It is interesting that though there is such a huge disparity between amount of people being subjected to SQF and those arrested, there seems to be no correlation between the period that the suspect is under observation and those actually arrested.

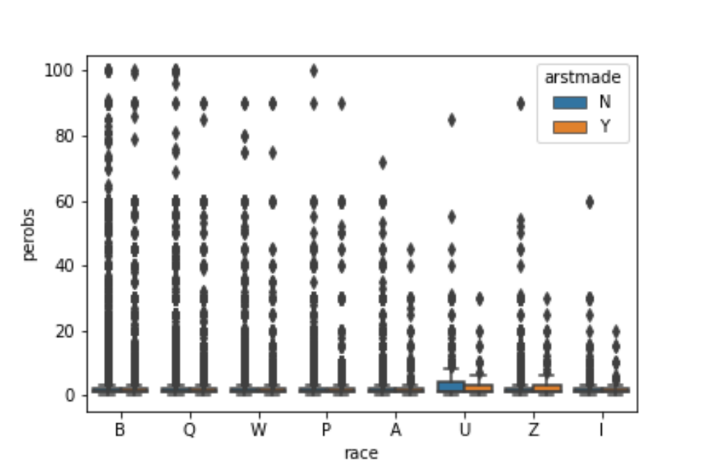
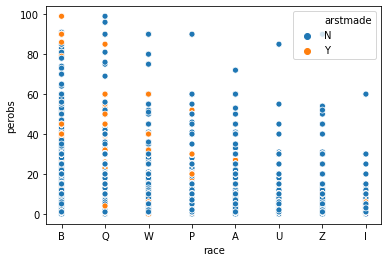


Figure 6, 7. Scatterplot/boxplot of period of observation segregated by race, highlighted by amount of arrests.

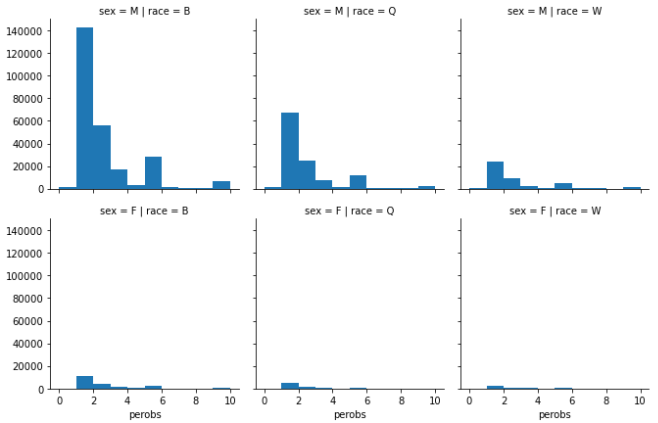


Figure 8. Distribution of period of observation of the most significant races stopped by police, separated by sex.

The scatter for blacks are much more densely populated, suggesting a much higher rate of being SQF’d than others, as shown by Figure 1, 6, and 7. The distribution of period of observation seems to be atypical to race as shown by Figure 7. Figure 8 suggests that, while males are stopped much more often, for those that are stopped, period of observation is also atypical to sex.

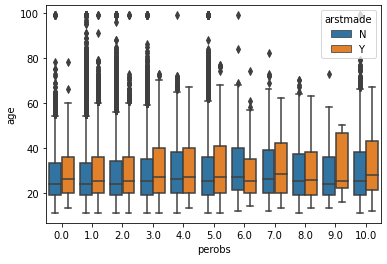


Figure 9. Distribution of period of observation by age, distinguished by arrests.

It seems that older people are observed for longer before being stopped by SQF.

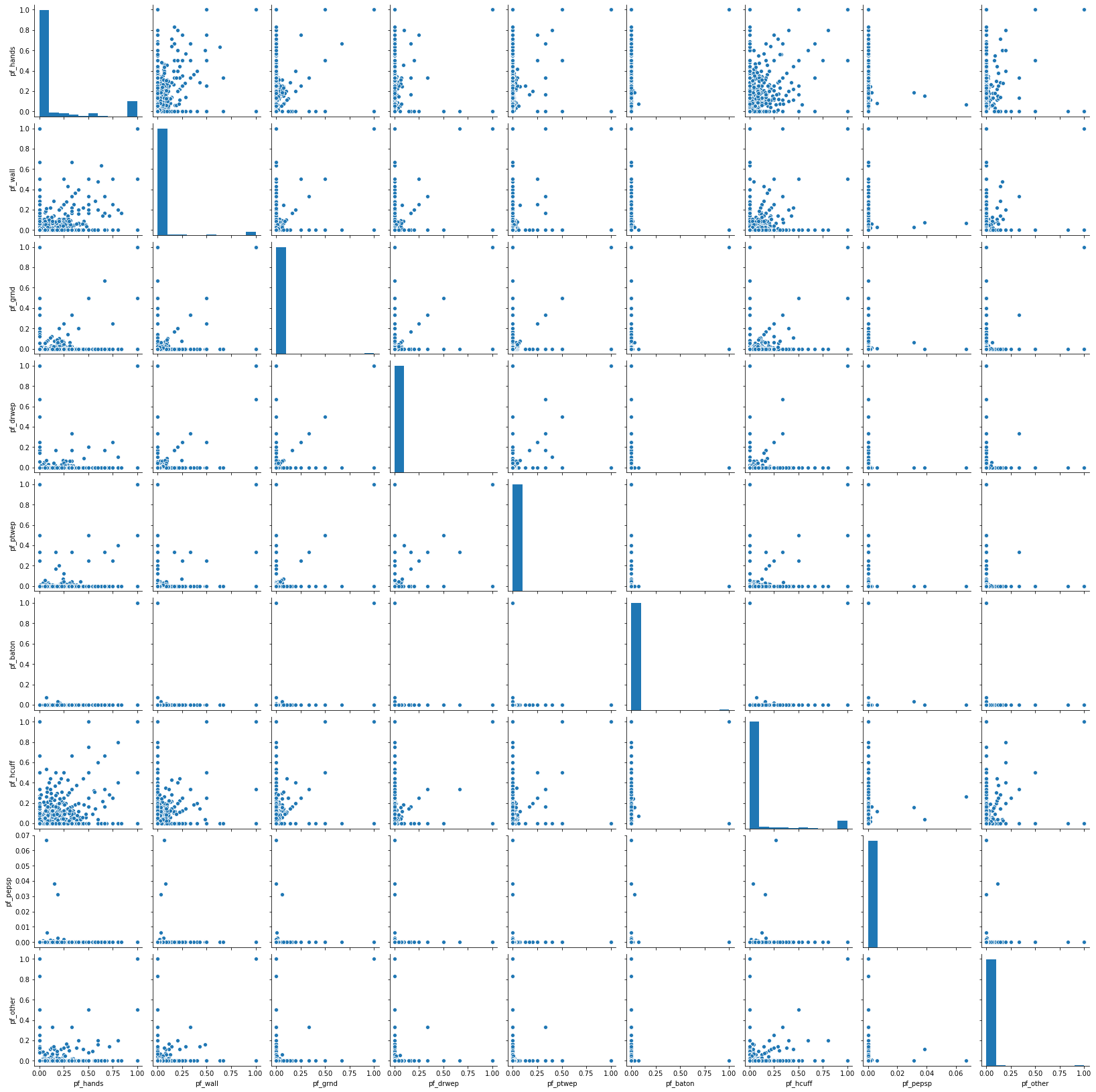


Figure 10. Common usages of force in conjunction to others.

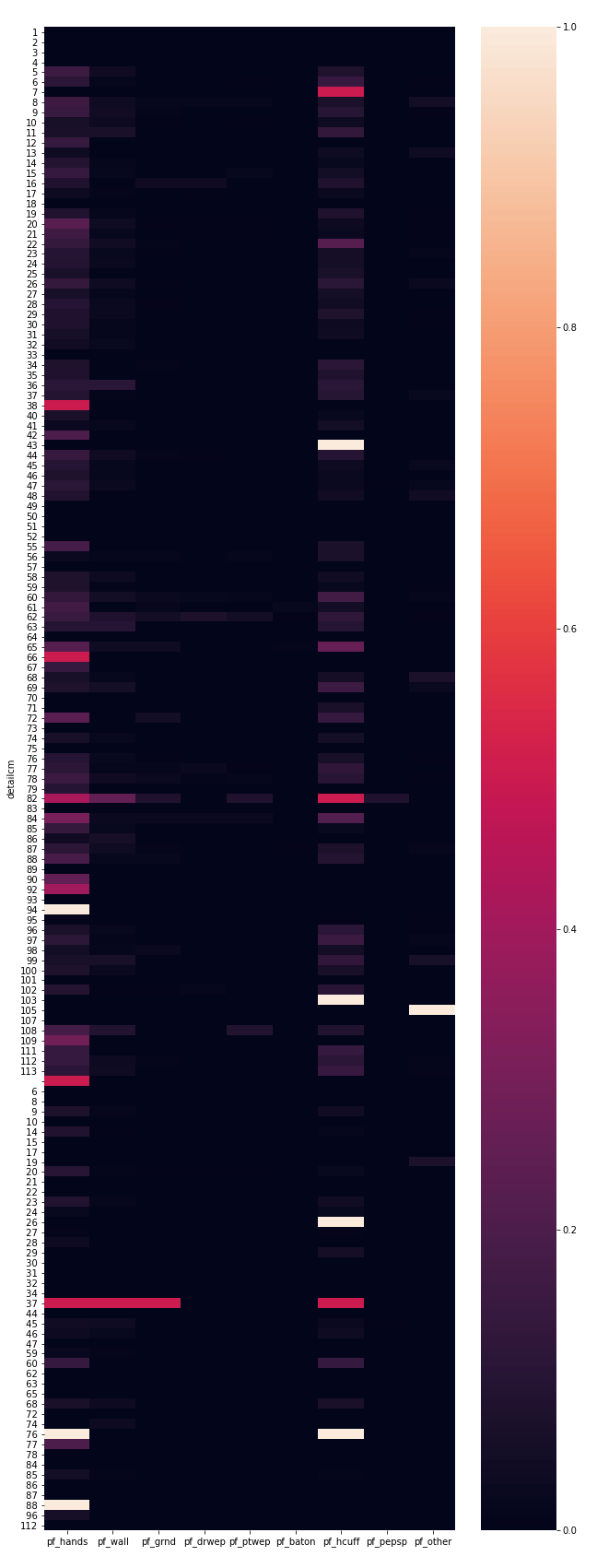


Figure 10. Heatmap of reasons for SQF and what type of force used by officer. [Legend of the heatmap.](https://pastebin.com/yrWxi9w1)

Most SQFs include a use of force with hands and handcuffs. Forcing the suspect against the wall is much next common method after these two. Forgery (37) has a somewhat common use of force involving hands, handcuffs, walls, and pushing the suspect onto the ground. Use of force involving weapons and batons are very rare; it is most common with 62, 82, and 100, which is murder, resist of arrest, and unauthorized use of vehicles respectively. Riots (84) also involve usage of force involving weapons.

**Association Rule Mining**

|  |  |  |
| --- | --- | --- |
|  | **support** | **itemsets** |
| **0** | 0.60 | ({'ac\_incid'}) |
| **1** | 0.44 | ({'ac\_time'}) |
| **2** | 0.36 | ({'cs\_casng'}) |
| **3** | 0.52 | ({'cs\_furtv'}) |
| **4** | 0.39 | ({'rf\_furt'}) |
| **5** | 0.56 | ({'frisked'}) |
| **6** | 0.53 | ({'Black'}) |
| **7** | 0.91 | ({'Male'}) |
| **8** | 0.36 | ({'ac\_time', 'ac\_incid'}) |
| **9** | 0.33 | ({'cs\_furtv', 'ac\_incid'}) |
| **10** | 0.34 | ({'ac\_incid', 'frisked'}) |
| **11** | 0.33 | ({'ac\_incid', 'Black'}) |
| **12** | **0.55** | ({'ac\_incid', 'Male'}) |
| **13** | 0.25 | ({'ac\_time', 'cs\_furtv'}) |
| **14** | 0.25 | ({'ac\_time', 'frisked'}) |
| **15** | **0.40** | ({'ac\_time', 'Male'}) |
| **16** | 0.33 | ({'cs\_casng', 'Male'}) |
| **17** | 0.30 | ({'rf\_furt', 'cs\_furtv'}) |
| **18** | 0.34 | ({'cs\_furtv', 'frisked'}) |
| **19** | 0.29 | ({'cs\_furtv', 'Black'}) |
| **20** | **0.48** | ({'cs\_furtv', 'Male'}) |
| **21** | 0.39 | ({'rf\_furt', 'frisked'}) |
| **22** | 0.37 | ({'rf\_furt', 'Male'}) |
| **23** | 0.31 | ({'frisked', 'Black'}) |
| **24** | **0.53** | ({'Male', 'frisked'}) |
| **25** | **0.49** | ({'Male', 'Black'}) |
| **26** | 0.33 | ({'ac\_time', 'ac\_incid', 'Male'}) |
| **27** | 0.31 | ({'cs\_furtv', 'ac\_incid', 'Male'}) |
| **28** | 0.32 | ({'ac\_incid', 'Male', 'frisked'}) |
| **29** | 0.31 | ({'ac\_incid', 'Male', 'Black'}) |
| **30** | 0.30 | ({'rf\_furt', 'cs\_furtv', 'frisked'}) |
| **31** | 0.29 | ({'rf\_furt', 'cs\_furtv', 'Male'}) |
| **32** | 0.32 | ({'cs\_furtv', 'Male', 'frisked'}) |
| **33** | 0.27 | ({'cs\_furtv', 'Male', 'Black'}) |
| **34** | 0.37 | ({'rf\_furt', 'Male', 'frisked'}) |
| **35** | 0.29 | ({'Male', 'frisked', 'Black'}) |
| **36** | 0.29 | ({'rf\_furt', 'cs\_furtv', 'Male', 'frisked'}) |

Table 3. Frequent itemsets with minimum support = 0.25

With frequent itemsets, several interesting patterns of SQF become apparent. The sets that have over 40% occurrence are highlighted in Table 3. Because 91% of the stops in the data are attributed to males, it is not surprising that most of these itemsets include this item, though it is still interesting to note. Nearly half of SQF stops are black males, and 53% of those stopped are male and have been frisked. Of those black males stopped, 30% of them are frisked. There is also a notable amount of cases that are due to incidence to time and place; because someone was at the wrong place at the wrong time, they were subjected to SQF by the police. Another common occurrence is that stops and frisks are caused by suspicious movement. Nearly 30% of those stopped and frisked are males that have “furtive movements”. What is perhaps comforting is that the pf category, or usage of force, is not a common occurrence – though it may show up more often if the minimum support was lowered below 25%.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **antecedents** | **consequents** | **support** | **confidence** | **lift** |
| **26** | ({'rf\_furt'}) | ({'frisked'}) | 0.393123805 | 1 | 1.792840226 |
| **84** | ({'rf\_furt', 'Male'}) | ({'frisked'}) | 0.374706095 | 1 | 1.792840226 |
| **60** | ({'rf\_furt', 'cs\_furtv'}) | ({'frisked'}) | 0.30031844 | 1 | 1.792840226 |
| **96** | ({'rf\_furt', 'Male', 'cs\_furtv'}) | ({'frisked'}) | 0.287273109 | 1 | 1.792840226 |
| **92** | ({'frisked', 'Black'}) | ({'Male'}) | 0.294069741 | 0.959774867 | 1.05011566 |
| **100** | ({'rf\_furt', 'cs\_furtv'}) | ({'Male', 'frisked'}) | 0.287273109 | 0.956561674 | 1.804570306 |
| **66** | ({'rf\_furt', 'cs\_furtv'}) | ({'Male'}) | 0.287273109 | 0.956561674 | 1.046600019 |
| **97** | ({'rf\_furt', 'cs\_furtv', 'frisked'}) | ({'Male'}) | 0.287273109 | 0.956561674 | 1.046600019 |
| **73** | ({'cs\_furtv', 'frisked'}) | ({'Male'}) | 0.323301639 | 0.955521294 | 1.045461711 |
| **49** | ({'ac\_incid', 'frisked'}) | ({'Male'}) | 0.322207648 | 0.955111304 | 1.04501313 |
| **87** | ({'rf\_furt'}) | ({'Male', 'frisked'}) | 0.374706095 | 0.953150358 | 1.798134799 |
| **85** | ({'rf\_furt', 'frisked'}) | ({'Male'}) | 0.374706095 | 0.953150358 | 1.042867606 |
| **28** | ({'rf\_furt'}) | ({'Male'}) | 0.374706095 | 0.953150358 | 1.042867606 |
| **33** | ({'frisked'}) | ({'Male'}) | 0.530077255 | 0.950343825 | 1.039796902 |
| **79** | ({'cs\_furtv', 'Black'}) | ({'Male'}) | 0.270382859 | 0.938525872 | 1.026866559 |
| **42** | ({'cs\_furtv', 'ac\_incid'}) | ({'Male'}) | 0.30678481 | 0.931370203 | 1.019037349 |
| **55** | ({'ac\_incid', 'Black'}) | ({'Male'}) | 0.307709918 | 0.931239707 | 1.018894569 |
| **16** | ({'cs\_casng'}) | ({'Male'}) | 0.331565684 | 0.930679041 | 1.01828113 |
| **24** | ({'cs\_furtv'}) | ({'Male'}) | 0.47780211 | 0.927109541 | 1.014375644 |
| **35** | ({'Black'}) | ({'Male'}) | 0.493391955 | 0.925078018 | 1.012152899 |
| **36** | ({'ac\_time', 'ac\_incid'}) | ({'Male'}) | 0.32928763 | 0.923122649 | 1.010013477 |
| **8** | ({'ac\_incid'}) | ({'Male'}) | 0.552852165 | 0.921087844 | 1.007787142 |
| **14** | ({'ac\_time'}) | ({'Male'}) | 0.403299988 | 0.920216478 | 1.006833757 |
| **99** | ({'cs\_furtv', 'Male', 'frisked'}) | ({'rf\_furt'}) | 0.287273109 | 0.888560633 | 2.260256494 |
| **62** | ({'cs\_furtv', 'frisked'}) | ({'rf\_furt'}) | 0.30031844 | 0.887594212 | 2.257798182 |

Table 4. Association Rules; top 25 sorted by highest to lowest confidence.

There are four major associations found by the algorithm that have a confidence of 100%. All of them involve a consequent of being frisked. It seems that furtive movements, in conjuncture (perhaps because the data is so skewed in one direction) with being male, is an antecedent to being frisked. These have an extremely high lift, which vouches for the association between the two sets. Black males that have “furtive movements” are 95% likely to be frisked. Most of this table have a consequent of {male} which is, again, most probably because of the one sided nature of the data, though it might suggest that males may exhibit more furtive movements.

“Furtive movements” seem to be the main cause for SQF by the police. This perhaps could be caused by the “broken windows” policy where stopping any and all minor crimes has the butterfly effect of stopping major crimes. However, the many SQFs due to just suspicious movement could be the antecedent to why the number of arrest cases are so low.

**Clustering**

The final dataset used for clustering includes a portion of the variables of the original dataset, as described in this [link](https://pastebin.com/6eXhbSG8). The coordinates are converted into latitude/longitude values.

K-means and hierarchical clustering was used to cluster the location of murders. H-clustering is a more appropriate method to cluster latitude longitude values because the data is not linear and therefore minimizing variance is not the optimal way to model this data.

The suitable number of clusters for k-means was obtained by using the “elbow” method. The elbow method involves iterating the algorithm in a loop, increasing cluster numbers and then plotting inertia, or the “within cluster sum of squares” (basically the criterion for regression), against the number of clusters.

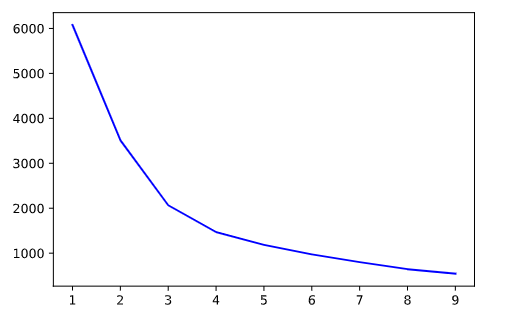


Figure 11. Inertia of k-means model used.

The inertia curve is smooth and there is not a clear “elbow” to use. The best values seem to be within the range of 2-4. Clustering with 2, 3, 4 clusters results in a inertia of 3517, 2067, 1472, respectively, suggesting 4 to be the best number of clusters in this context.

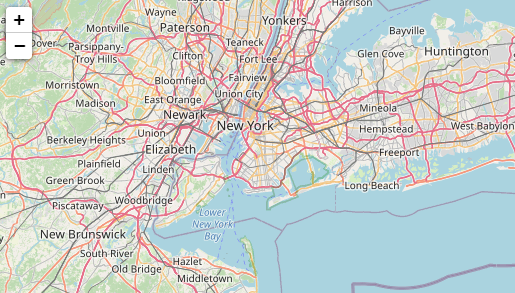
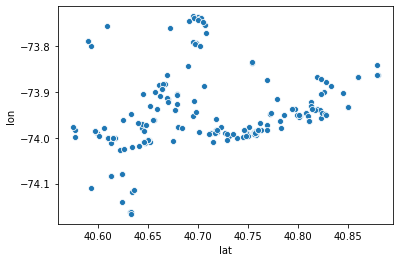


Figure 12, 13. Clustered murders in New York City.

When clustering the reason a suspect was stopped, DBScan was used. DBScan is an algorithm that expands clusters from core samples of high density, which is good for data that contains clusters of similar density. DBScan, by nature, does not need a number of clusters to be specified. The dataframe used in the previous clustering was adjusted due to arrests made due to the small sample size that “crime = murder” allows.

The silhouette score – a clustering criterion that measures the cohesion or how similar a point is to its own cluster compared to other clusters – is 0.98, suggesting that the clusters created are 98% relative to each other.

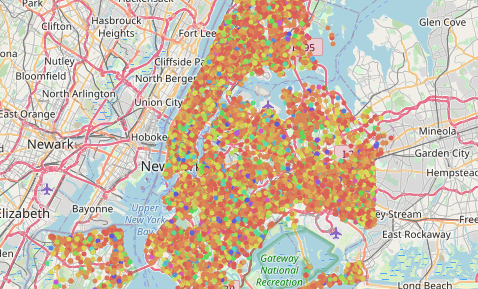
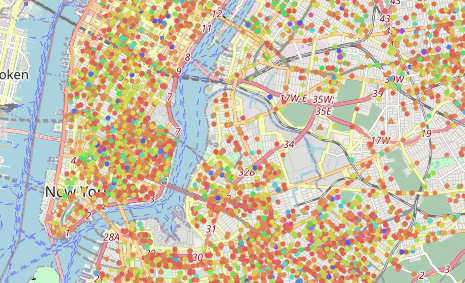
 

Figure 14, 15. Fitting reasons suspect was stopped with data on arrests made.

Though difficult to ascertain in Figure 14, most of the density is in the downtown and surrounding metropolitan area.

When clustering with use of force data based on arrests made using DBScan, the silhouette score was even higher: 0.9955.

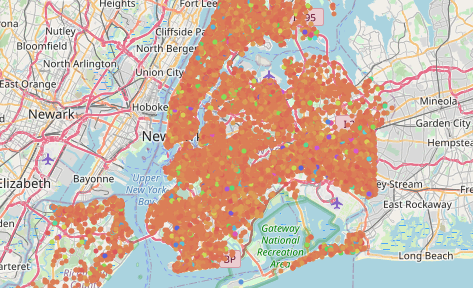


Figure 16. Use of force based on arrests made.

This could be explained due to the fact that use of hands and handcuffs count as “force” in the data used, and being arrested basically requires both usages of force. By reiterating the model with columns pf\_hands and pf\_hcuffs deleted, the silhouette score increases to 0.998.

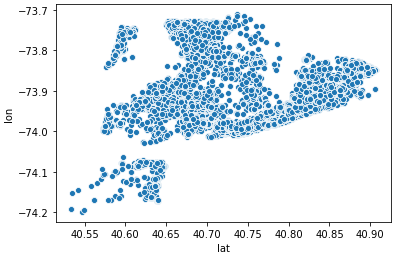


Figure 17. Scatterplot of black suspects being arrested.

When plotting black people being arrested, which counts for around half of the data, the results are very similar to Figure 16.

Results portray that there are extremely high similarities, or very strong cohesion, between people arrested, and the force used to arrest them with. This suggests that predictive modeling can be used to better predict who will be arrested, allowing the NYPD to increase their SQF efficiency.

**Predictive Modelling**

The main idea behind utilizing data is to further the efficiency of the NYPD in stopping crime. So far, it seems that even though there was a substantial increase in amount of people being stopped and frisked by the police, the rate of arrests have stayed low, suggesting that such a method is ineffective at their objective. The introduction of data could help identify if an arrest will be made. The final classification dataset will include similar features as the clustering one:

* Period of observation
* If suspect was carrying any of the following:
  + Contraband
  + Pistol; rifle; assault weapon; knife; machine gun; other potential weapons
* Age
* Sex
* Race
* Height (feet)
* Weight
* Crime description

*Model I: K-nearest neighbors*

Similar to clustering, k-nearest neighbors is a model that identifies entries with similar features; a point is classified based on how its neighbors are classified. The challenge of using this model is finding the number of neighbors to include in the model, or “K”. A higher value of k increases decision boundaries by allowing the algorithm to include more features to a cluster, thereby lowering variance. However, this may overfit the model and increase bias. K values that are too small may introduce too many classifications, creating unnecessary noise – basically leaving a lot of variance unexplained. KNN results can also be hard to interpret.

With neighbors = 5, an accuracy score of 93.7% was achieved at predicting the testing set. The f1 score, or the weighted average between precision and recall, was calculated in 3 different ways:

* Micro: “calculate metrics globally by counting total true positives, false negatives, and false positives”
* Macro: calculate metrics for each variable and find their unweighted mean
* Weighted: calculate metrics for each variable and find their average weighted by support

Respectively, this model obtained f1 scores of 0.937, 0.508, and 0.911, where 1 is the absolute best for f1 and 0 is the worst.

Feature importance is not a feature of KNN classification.

*Model II: Regression (Logistic)*

Regression is a model based on probability. In this case, logistic regression is being used because the regressor is a ‘dummy variable’. Features chosen will skew the regression closer to 0 or 1. Regression models are simple and one of the most interpretable techniques in machine learning. However, other models will outperform regression easily due to its simplicity. It should also be noted that the model will fall apart once too many variables are collinear to each other, and that regression is limited to linear problems.

Due to prior experience with regression models, the race column was instead encoded into their respective dummy variables.

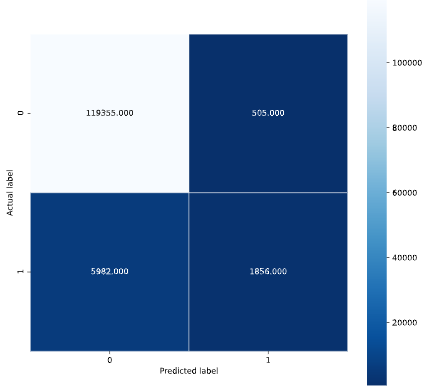


Figure 14. Confusion matrix of the lg-regression model.

The confusion matrix portrayed in Figure 14 gives the result of the model; the algorithm was able to predict ~120 thousand predictions correctly and about 8000 incorrectly.

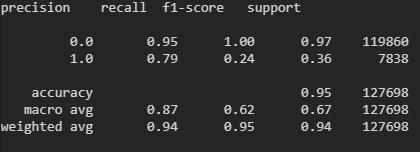


Figure 15. Classification report of the lg-regression model.[[1]](#footnote-1)

The regression model performed similarly to the KNN model. An accuracy score of 94.9% was achieved when predicting the testing set. The weighted f1 score was 0.936.

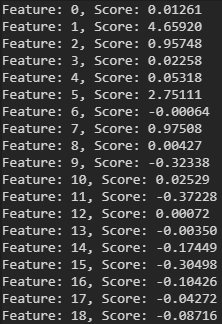
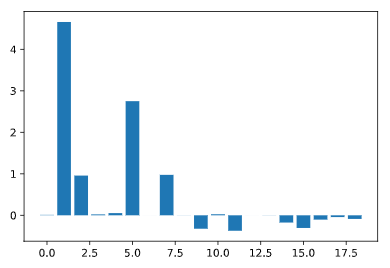
 

Figure 16, 17. Feature importance of the lg-regression model.

The features that most impacted the regression model were contraband, possession of weapons, and age.

*Model III: Decision Trees*

Decision trees are models that classify data based on branches of features, similar to a tree branching out. They are flexible, non-parametric models that allow for a large number of features to be included for modeling. The main advantage of using decision trees is their low wall of entry: they are very easy to interpret due to their ease of creating visuals. In addition to being fundamentally intuitive, DTs are lenient towards smaller and less clean datasets. However, like KNNs, trees are more susceptible to overfitting data – outliers will have a large effect on these models.

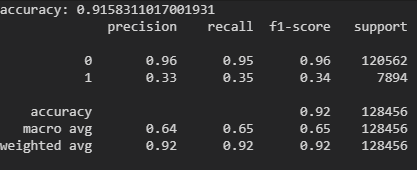
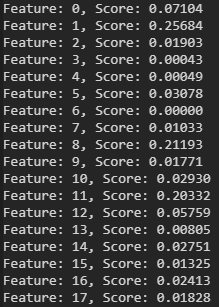
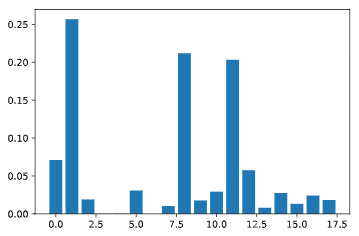


Figure 18. Classification report of the decision tree model.

The decision tree model scored an accuracy score and f1 score of 92.9% and 0.924 respectively. Though this is the weakest performance between all three models, it should be noted that decision trees by themselves are weak – “forests” must be grown for increased predictive power.

Another popular classification algorithm is support vector machines. However, this method is best used for unsupervised data, which is inappropriate for this setting.

Figures 19, 20. Feature importance of decision tree model.

By the decision tree, the top 4 most important features in determining whether or not a suspect stopped by SQF was arrested are contraband, age, and height. Interestingly, period of observation is classed a lot higher than it was in regression. It should be noted that age and weight are could be a consequence of who are more likely to commit arrestable crimes and thus are an erroneous statistical error.

**Conclusion**

The models that are able to portray feature importance both confirm that race and sex have no bearing on whether or not a suspect is arrested, even though those that are stopped are heavily skewed towards males, blacks, and Hispanics. Because the data *is* so skewed, it could be inferenced from these models that the police are looking in the wrong places. Possession of contraband is a very heavy predictor, but since the legalization of marijuana, it may no longer be as such. Because of such changes in the law, and because the data is nearly a decade old, these models are essentially outdated.

A possible method to better police efficient is to include data on what kind of equipment the police are utilizing, which is another means to measuring how effective they are being on cracking down crime. For example, because contraband is such a high predictor for arrest, it would be beneficial for officers to bring around police dogs. Another interesting perspective to take would be to supplement police data with future crime rates and judge the theory behind “broken windows”. If it is deemed to be ineffective in lowering crime rates, perhaps SQF may not be such an efficient way of policing.

**APPENDIX**

**References**

* <https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City>
* <https://en.wikipedia.org/wiki/Broken_windows_theory>
* <https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>
* <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>

1. Classification report in python is read as a “str” and not a “df” – unable to convert to csv. [↑](#footnote-ref-1)