Project: Stop, Question, and Frisk: New York City **Case Study**

In this project, I Performed an analysis of the NYPD SQF policing strategy using datasets from 2012. The project utilized the CRISP-DM framework to generate findings & insights for the report.

Tools/Libraries: Python, Pandas, NumPy, Seaborn, Matplotlib.
Analysis: Association Rule Mining, Cluster Analysis Report & Predictive Modelling in python
Data Source: https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
Datasets year: 2012

Project completed by: Ernest Eze

Table of Contents

Executive Summary

The Stop-Question-and-Frisk (SQF) program of New York City's Police Department is a policing policy/tactic of temporarily stopping, detaining, questioning pedestrians and potentially searching them for weapons or contraband. This happens if the police officer(s) have a "reasonable suspicion" that the pedestrian in question "committed, is committing, or is about to commit a felony or a Penal Law misdemeanor.

In New York City, stop and frisk gained momentum under Rudy Giuliani, who served as mayor of New York City from 1994 to 2001. The policy was carried out by a unit within the NYPD called the Street Crimes Unit (SCU). In October 2000, a federal investigation of the SCU "determined that its officers engaged in racial profiling as they conducted their aggressive campaign of street searches across the city." The investigation of the unit was prompted by the February 1999 killing of Amadou Diallo, a 23-year-old Guinean immigrant who was shot at 41 times by SCU officers in front of his apartment building.

In 2011, at the height of the program, over 685,000 people were stopped, with nearly 88% found to be innocent. The NYPD's application of stop and frisk was found unconstitutional in 2013 due to a policy of indirect racial profiling.

This Project analyzed only the 2012 SQF datasets and utilized the CRISP-DM framework to examine and generate findings and insights for further consideration.

Findings revealed from the project revealed how Blacks and Hispanics who resided in specific neighborhoods in New York City and fit into a unique demographic were targeted, profiled and sometimes arrested.

Report 1: Data and Visualization
(Business Understanding)

#	Question	Answer
	What is the purpose of the SQF program?	The Purpose of the SQF is to: • Reduce violence in targeted high-crime neighborhoods. • Detect concealed weapons on a person • Ultimately to prevent crimes and save Lives
	How would you define and measure the effectiveness of such a program?	• The Program will be best measured by the results it achieves based on the purpose of why it was set up in the first place.
	What data would you need be able to judge its effectiveness?	• 2012 Data from NYPD on the number of stops, question and frisk it made and if those events resulted in crime prevention in the long term.

Report 1: Data and Visualization
(Business Understanding)

#	Question	Answer
	What is the purpose of the SQF program?	The Purpose of the SQF is to: • Reduce violence in targeted high-crime neighborhoods. • Detect concealed weapons on a person • Ultimately to prevent crimes and save Lives
	How would you define and measure the effectiveness of such a program?	• The Program will be best measured by the results it achieves based on the purpose of why it was set up in the first place.
	What data would you need be able to judge its effectiveness?	• 2012 Data from NYPD on the number of stops, question and frisk it made and if those events resulted in crime prevention in the long term.

Report 1: Data and Visualization
(Data Understanding)

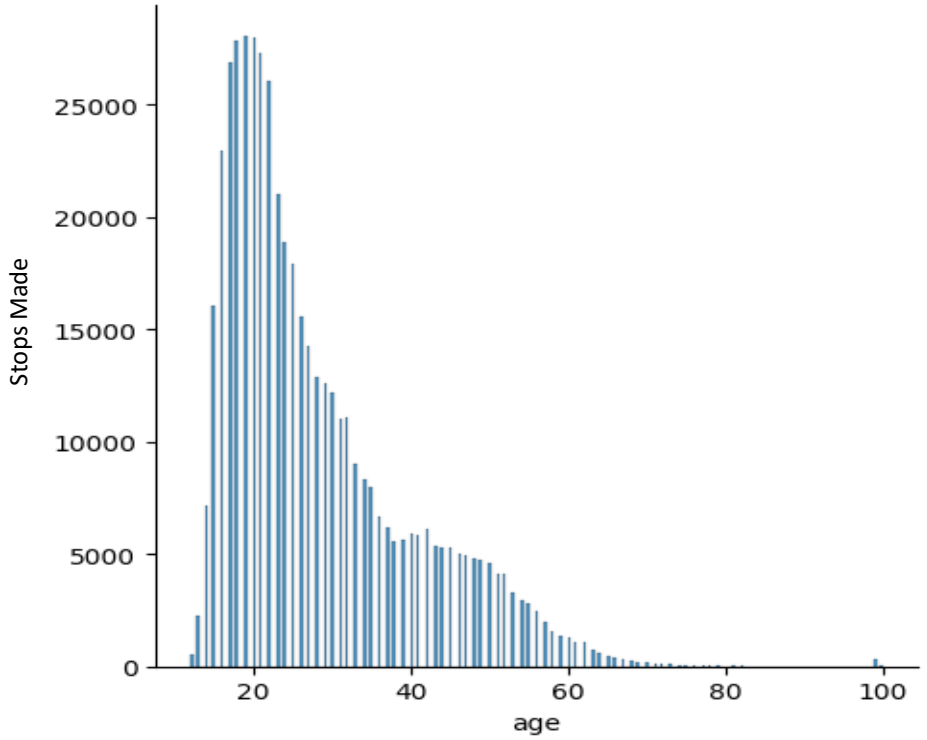
#	Question	Answer
	Describe the meaning and type of data (e.g., scale, values) for each attribute in the data file	• Data was saved in csv to get a better picture as it was a large datasets. • Data type used were numbers(Integers and floats) and where data was not an Int or float, it was coerced to NaN. (Not a Number). Which is a floating-point value.
	Verify data quality. Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems?	• Data had some missing, duplicate and outlier values. • These values were dropped and a new DataFrame updated after they were dropped. • DateTime object was also created for all the rows better analysis.
	Give simple, appropriate statistics (e.g., range, mode, mean, median, variance, counts) for the most important attributes in these files, and then describe what they mean or whether you found something interesting?	• Average age and height of pedestrians SQF'd the most was 28 years and they measured 174.3cm respectively. • This again clearly shows that the SQF program was mostly targeted at young people in early adulthood as the age and height description fits profile of people in this age category.

Report 1: Data and Visualization
(Data Understanding)

#	Question	Answer
	Visualize the most important attributes appropriately. (at least 5 attributes) **Important:** Provide an interpretation for each chart, explaining each attribute and why you chose the visualization you did	• Attributes Visualized were “Race”, “Age”, “Month”, “Day of the Week” and “Hour”. • Race: Report Shows that people of color and white Hispanic were SQF’d the most. This can be attributed to population demographics of the location and or perhaps racial profiling. • Age: Findings revealed that pedestrians between the ages of 18 – 32 years were SQF’d the most as average age was 28. • Month: The months of Jan, Feb & March witnessed the highest SQF. This can perhaps be attributed to the weather as those are colder periods. • Day of the Week: Fridays had the most SQF. This can be attributed to being the beginning of a weekend and young people are usually out in the night by this hour. • Hour: The time of the day most people were stopped was between the hours of 7pm -1am. This can be related to the fact that young people are usually out at night by this time for happy hour and good time.
	Explore relationships between attributes. Look at the attributes and then scatter plots, correlation, cross-tabulation, group-wise averages, etc., as appropriate	• .There is a relationship between the attribute from the report. The primary attribute being the Race and City and Race and age. • Brooklyn City had the most SQF for black pedestrians. Although there is a large population of blacks in the city, this does not still account for the high number of SQF in this city. • Most persons SQF’d were young blacks & white Hispanics. This can be seen in relation to the time of the day they got SQF’d.
	Compare the reasons for an SQF and what type of force was used by the officer	• Top 3 reasons for an SQF were suspicions of Criminal Possession of Weapon (CPW) ,Robbery & Burglary. Which corresponds to: • Top 3 incidences of where specific type of physical force was used by the police on pedestrians: “Hands”, “Handcuffs” and “Pushed against the wall”.

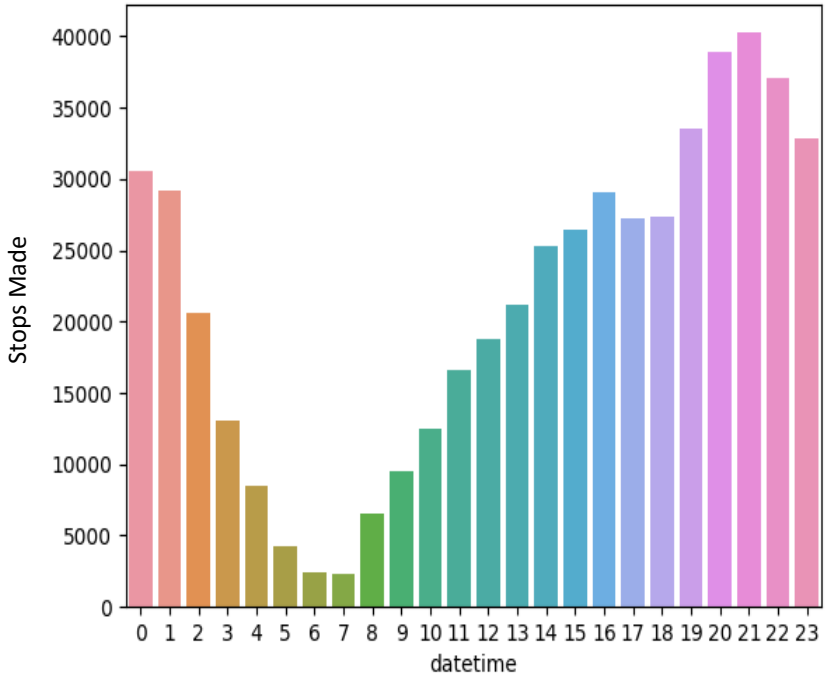
Report 1: Data and Visualization
(Data Visualization)

Figure 1.1: Age Distribution of Pedestrians Stopped



The average age of pedestrians stopped in 2012 was 28 Years

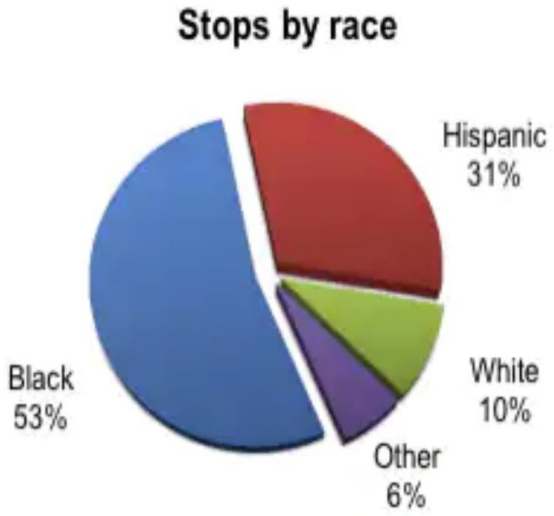
Figure 1.2: Frequency of pedestrian Stops per hour



Most Pedestrians were stopped between the hours of 7pm -1am

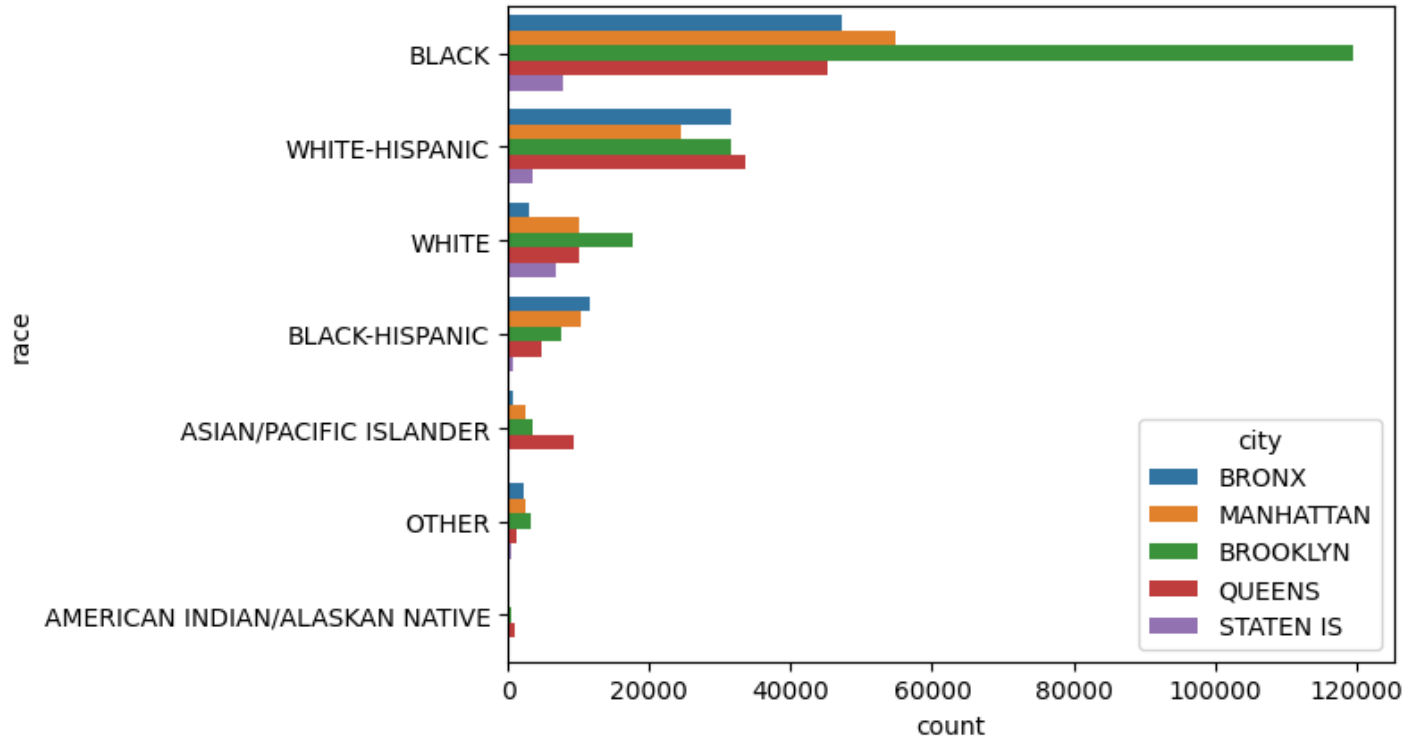
Report 1: Data and Visualization
(Data Visualization)

Figure 1.3: Stops by Race



Blacks and Hispanics faced the most Police SQF

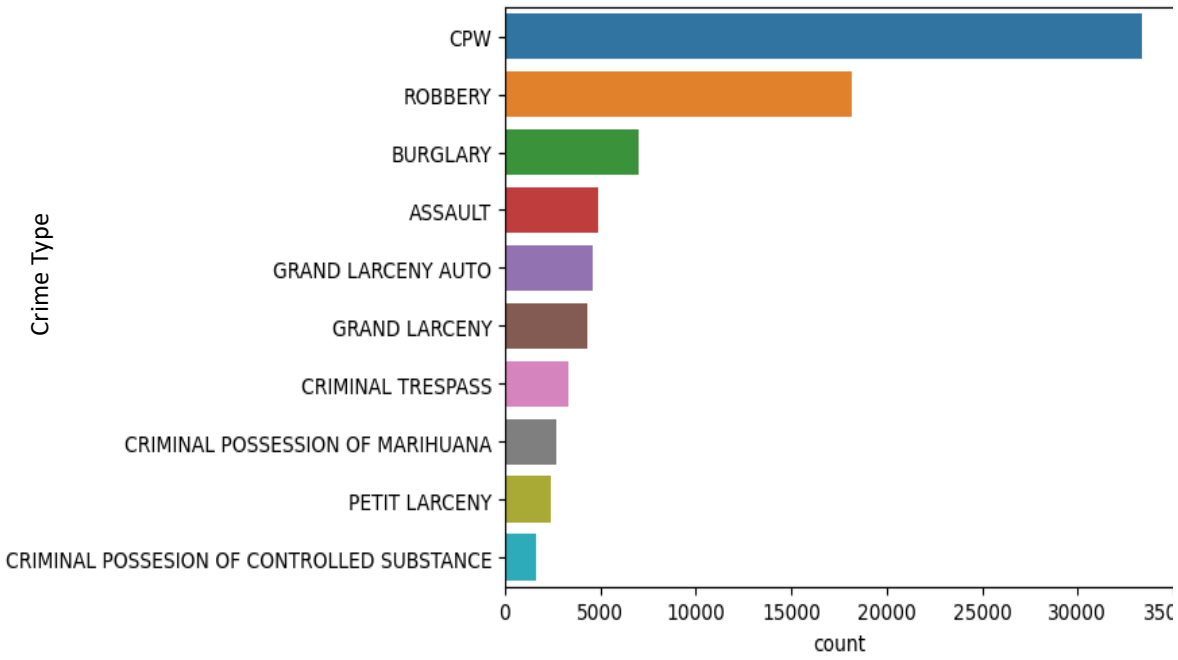
Figure 1.4: Stopped by Race and City



Most Incidences occurred in Brooklyn and Blacks faced the most SQF

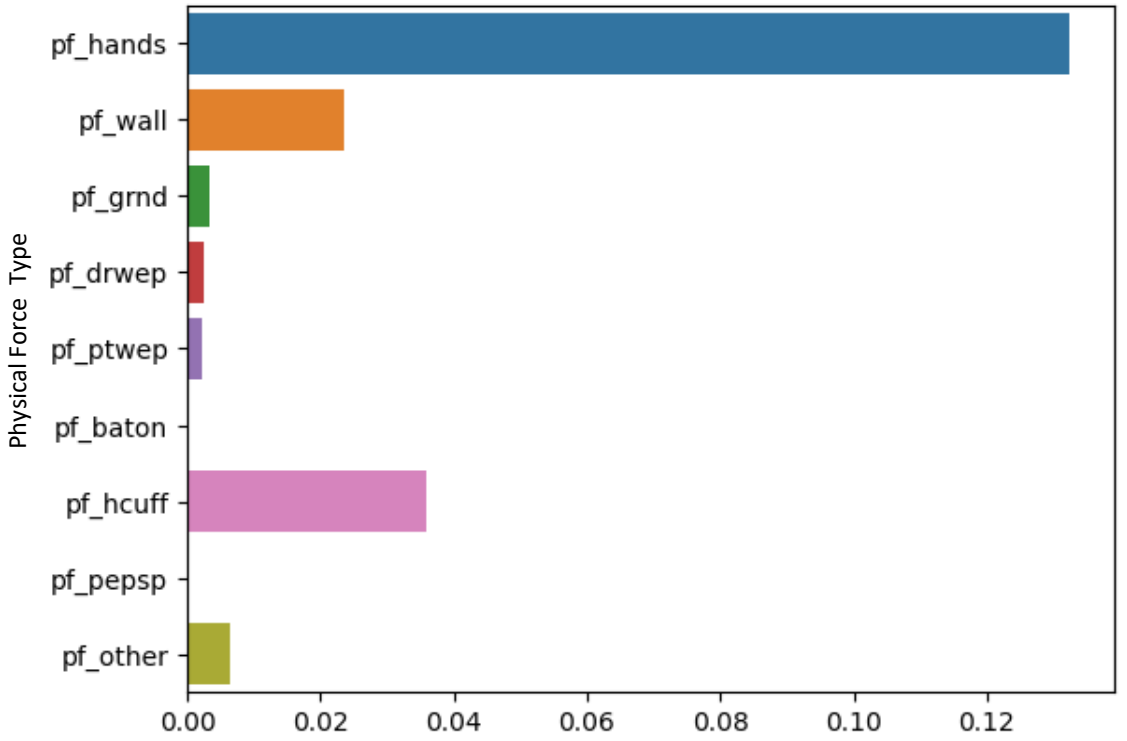
Report 1: Data and Visualization
(Data Visualization)

Figure 1.5: Crime Type and SQF



Suspicion of Criminal Possession of Weapon (CPW) was the leading factor for an SQF

Figure 1.6: Type of Physical Force



Physical Force by use of Hand was the primary method used by the police

Report 2: Association Rule Mining
(Data Preparation)

#	Question	Answer
	Construct the required transaction data set for frequent itemset and association rule mining.	Transaction dataset for frequent itemset constructed on the basis of physical forces, race, city and suspicions if pedestrian is armed

Report 2: Association Rule Mining
(Modelling)

#	Question	Answer
	Create frequent itemset and association rules.	Frequent itemset and association constructed on the basis of physical forces, race, city and suspicions if pedestrian is armed
	• Use tables and visualizations to help explain your results.	The scatter plot 'support and confidence' plot shows a very low correlation between being stopped, arms or contraband being found on pedestrians and race within the city and the use of Physical force by officers.
	What findings are the most interesting? Why?	• The result shows that in the SQF program, for most pedestrians stopped (Blacks & Hispanics), only about 10% were found to have possession of firearm or contraband.

Report 2: Association Rule Mining
(Modelling)

Figure 2.1: Table: Armed vs Race

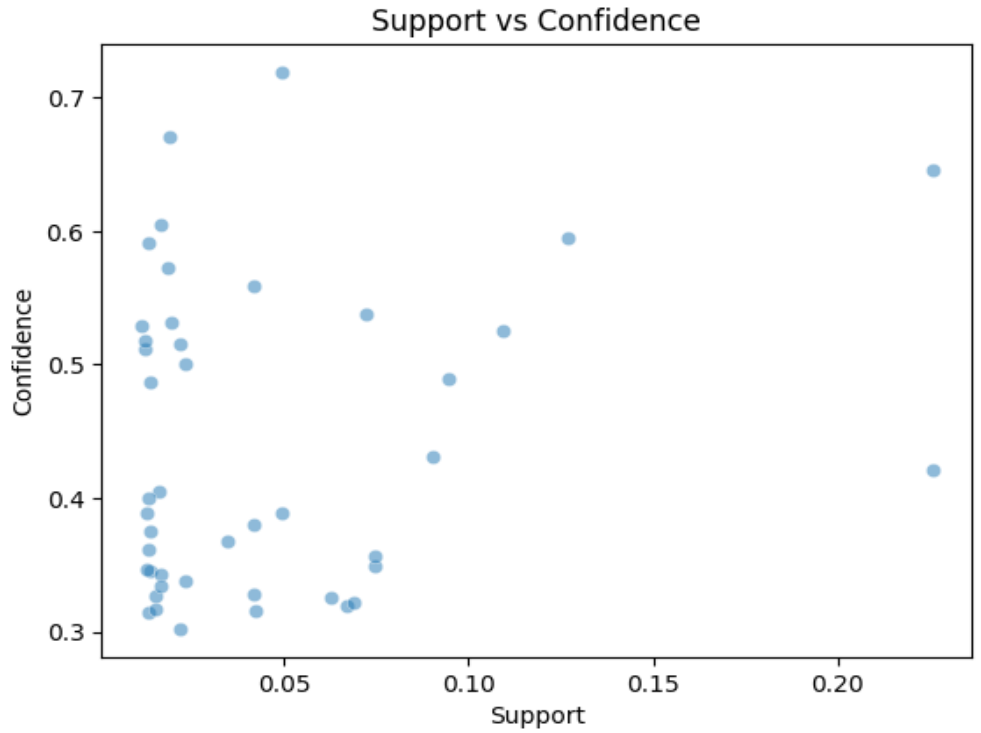
Frisked or Searched Suspects Found Having Contraband or Weapons

Outcome	Reference Group (%)	Comparison Racial Groups (adjusted) (%)	
	White	Black	Hispanic
Any contraband	6.4	5.7[a]	5.4[a]
Weapon	1.2	0.9	1.1
	Black	Hispanic	White
Any contraband	3.3	3.2	3.8
Weapon	0.7	0.7	0.8

SOURCE: Computed from NYPD (2006).

Table showing pedestrians by race SQF'd on suspicion of possessing contraband or weapons

Figure 2.2: Scatterplot: Race vs armed



Low correlation between race and suspicion of firearm possession or contraband

Report 3: Cluster Analysis
(Data Preparation)

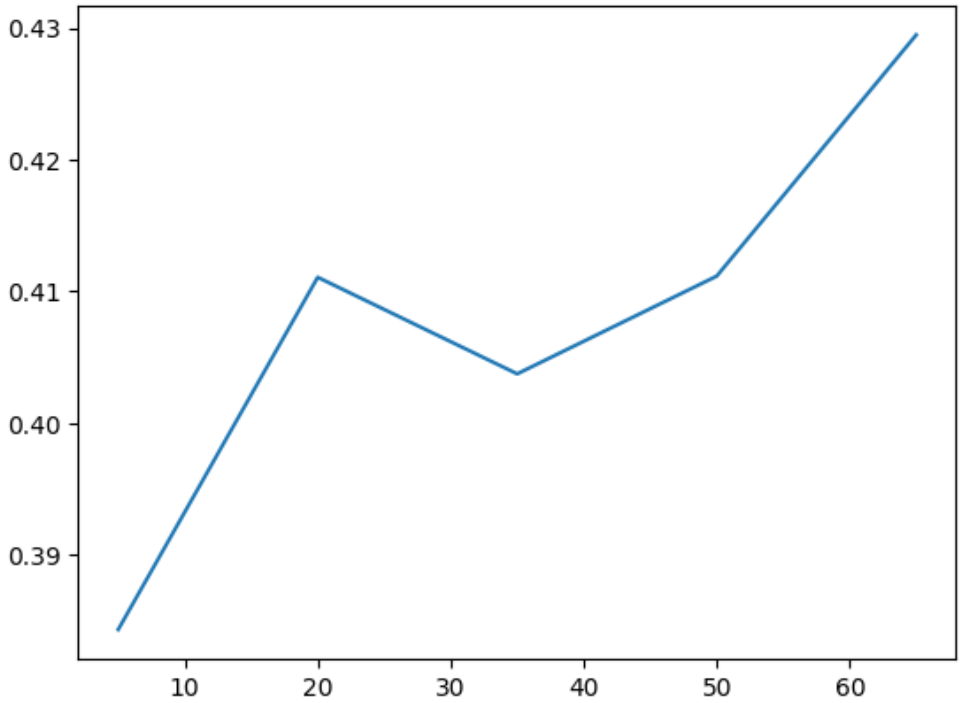
#	Question	Answer
	Define and prepare your class variables. Note: You may have to combine different columns.	Criminal Possession of Stolen Property(CPSP) was selected as the variable.
	Remove variables that are not needed/useful for the analysis.	Criminal Possession of Stolen Property(CPSP) was used for the cluster analysis
	Describe the final dataset that is used for classification and include the scale/range for the new combined variables.	Hierarchical clustering was used for the analysis

Report 3: Cluster Analysis.
(Modelling)

#	Question	Answer
	Perform cluster analysis. Cluster the location for a crime of your choice. Note: Found clusters might be different depending on the time of day. Cluster stopped people by reasons for stop. What else can you use cluster analysis for in the data set?	Criminal Possession of Stolen Property (CPSP-) Hierarchical clustering was used for the analysis. The Silhouette of CPSP crime revealed a silhouette score of 0.89. This indicates that the crime of CPSP was relatively clustered within the specified location between the city and the number of police precinct in the city.
	How did you determine a suitable number of clusters for each method?	The range of values tested was between 5 Cities and 15 police precincts per each city for the hierarchical clustering and the optimum number of clusters was shown to be 65 (As shown in the line chart above)
.	Use internal validation measures to describe and compare the clusters (some visual methods would be good).	The biggest cluster (9) had 490 data points. (clusters).

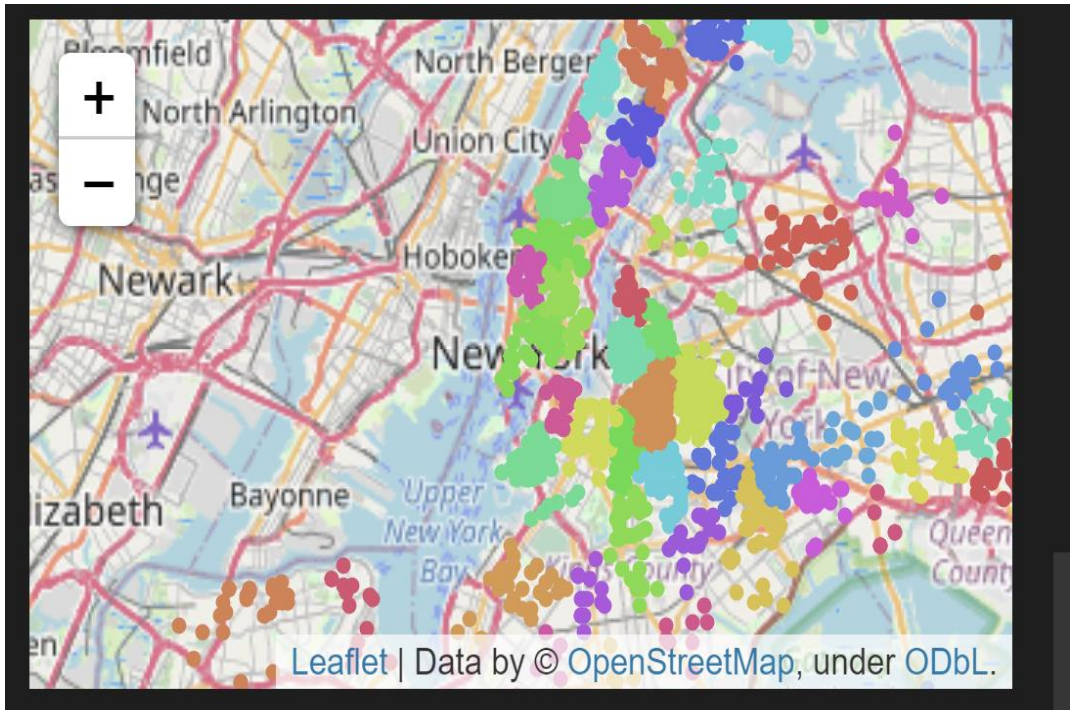
Report 3: Cluster Analysis
(Modelling)

Figure 3.1: CPSP in City & Precinct Clustering



Optimum number of clusters was shown to be 65 in a hierarchical clustering model

Figure 3.2: CPSP Clustering



CPSP suspicion SQF relatively clustered

Report 4: Predictive Modelling
(Data Preparation)

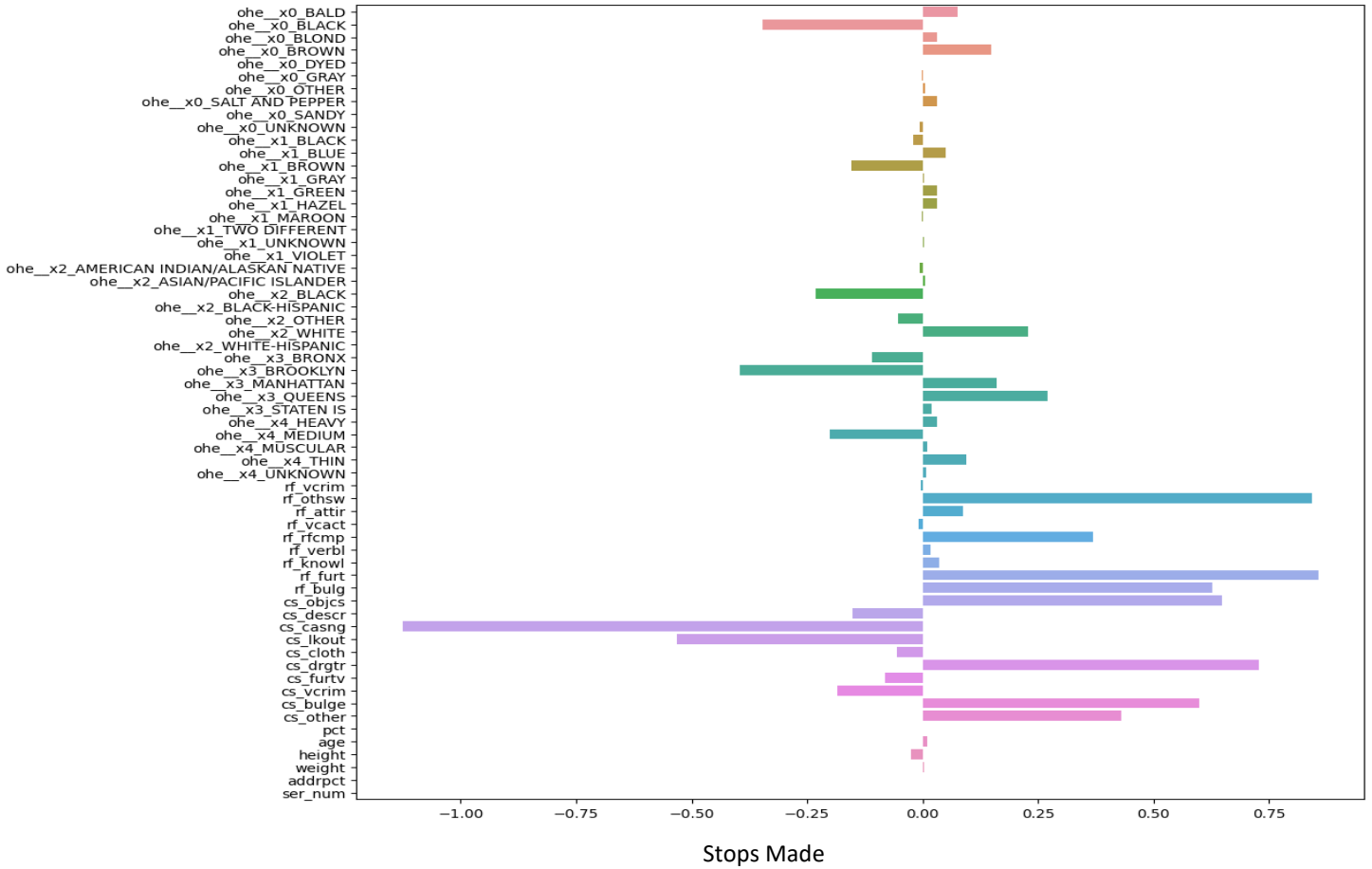
#	Question	Answer
	Define and prepare your class variables. **Note:** You may have to combine different columns.	Variables prepared were related to armed and weapons and they were: Contraband (“Contrabn”), Pistol (“pistol)” Rifle (“Riflshot”), Assault Weapon (“Asltweap”), Knife Or Cutting (“Knifcuti”) , Machine Gun (“Machgun”) Another Type of weapon (“Othrweap“). Other Columns were combined as the below columns were added: Facts column = Precinct ("pct“), Age ("age“), Height ("height“), and Weight ("weight“) Category Columns = Hair Color ("haircolr“), Eye Color ("eyecolor“), Race ("race“), City ("city“), and Build ("build“)
	Remove variables that are not needed/useful for the analysis	The weapons column was removed as we have already identified the variables, we are doing our analysis on.
	Describe the final dataset that is used for classification and include the scale/range for the new combined variables.	Datasets was split into training and testing datasets and then encoded

Report 4: Predictive Modelling
(Modelling)

#	Question	Answer
	Create at least three different classification models (different techniques) for each of the classification tasks. Discuss the advantages of each model for this classification task.	Decision Tree, Logistic regression and Naïve Bayes were the three classification models used for the analysis. Decision trees was extremely useful for this data analytics because it broke down the complex data into more manageable parts by splitting into test and training sets. Logistic regression was important in this task as it was used to calculate and predict the probability of a pedestrian carrying a weapon. An advantage of the Nayes Bayes is that it simple and easy to implement and doesn't require as much training data. Although it's prediction for the armed datasets selected was low.
	What are the most important variables found by each model	Results from the decision tree shows that the top reason for stop and frisk by the police is suspicion of a pedestrian carrying other type of weapon other than a pistol, machine gun, knife, assault weapon, rifle, or contraband. This is closely followed by any action the police thinks is indicative of a drug transaction, suspicious bulge beneath a clothing or when casing a victim or location.
	Assess how well each model performs (use training/test data, cross validation, etc., as appropriate).	Decision tree performed very well on the training datasets as it returned a 99% accuracy. Although the testing data was quite low @ 14%. Logistic regression returned a balanced but low result on both the training and testing datasets @ 42% and 44% respectively.

Report 4: Predictive Modelling
(Modelling)

Figure 4.1: SQF REASON



Stops Made

The bar plot shows that some of the top reasons for stops were:
casing a victim or location, drugs transaction and carrying suspicious object.
while top reasons for frisk are: - Suspicion of Other type of weapon, furtive look, and a suspicion of bulge.

Report 3: Evaluation.
(Modelling)

#	Question	Answer
	How useful is your model for the police? How would you measure the model's value if it were used?	My models are useful to the police as it shows some of the challenges of the SQF program. For example, the number of young black and Hispanic pedestrians racially profiled. Also, As shown by the model results, some of the reasons for SQF were not good enough as it violates the right of persons.
	How would you implement your model to improve policing? What other data should be collected?	The use of "Stop, Question and Frisk" (SQF) as a crime reduction or deterrence strategy in itself is good but it becomes problematic when implemented wrongly or poorly to profile certain demographics of the population. My model will help guide and show areas of improvement. Especially with regards to reason for a stop, question and frisk. Additional data should thus be collected on pedestrians feedback and review of their experience when they were SQF'd.
	How often would your model need to be updated?	My model will need to be updated based on feedback and surveys from members of the public on an ongoing basis.

Conclusion and References

Conclusion:

The stop, question and frisk practice of the New York City Police Department ("NYPD") has over the years been the subject of significant public debate and litigation because the program became the subject of a racial-profiling controversy. At the height of the program in 2011, over 99% of recorded non-lethal interpersonal violence in the city was reportedly committed by the New York Police Department itself.

The 2012 data of the SQF program which formed the basis of this report seeks to contribute to the ongoing dialogue within law enforcement and among all stakeholders about some of the outcomes of the program as well as providing some suggestions on next steps.

The findings reveal, among other things, that approximately 84% of Stops were young black and Hispanic pedestrians' and specific clusters in New Yorksuch as Brooklyn appeared to be locations were SQF were carried out the most. The report also revealed that a leading reason for SQF is when the police Cases a victim or location.

These findings, along with a host of other relevant factors and events, merit consideration in the broader and ongoing dialogue about policing in general.

References:

- https://www1.nyc.gov/assets/nypd/downloads/pdf/public_information/TR534_FINALCompiled.pdf
- https://bridge.georgetown.edu/research/factsheet-nypd-stop-and-frisk-policy/
- https://www.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page
- https://www.washingtonpost.com/news/wonk/wp/2013/08/13/heres-what-you-need-to-know-about-stop-and-frisk-and-why-the-courts-shut-it-down/
- https://ag.ny.gov/pdfs/OAG_REPORT_ON_SQF_PRACTICES_NOV_2013.pdf