DATA-475-A
Advanced Concepts in Data Analytics

Final Project: New York City Case Study Using CRISP-DM

Spring 2020

Niroojen Thambimuthu(000859345)

Table of Contents

Abstract:

This final project is about the New York City (NYC) case study for the program known as the Stop-Question-Frisk(SQF). Throughout the reports, we will be analysing it using the CRISP-DM model to measure the efficiency and effectiveness of the program. On paper and on the news, the program is controversial such that it targets young black males unfairly through the city.

Report 1: Data and Visualization Report

Business Understanding:

The first step of the CRISP-DM process is the business understanding phase, where we will go in depth in the New York City Case Study and explain some core objectives needed to be addressed. For this study, we are supplied with the official NYC government dataset for the year 2012, where this specific year contributed to peak numbers of recorded data for the program.

What is the purpose of the SQF program?

The SQF program, also known as the Stop-Question-and-Frisk program, is a policy exercised by the New York City Police Department, or NYPD. The policy permits NYPD police officers to detain and question suspicious civilians or suspects, and possibly search their bodies, mainly for finding and to seize items such as weapons and contrabands. In the United States, this practise is also commonly referred to as the "Terry Stop", and that program has its roots from the late 1960s.

This policy is aimed to make the streets of NYC safer by reducing the number of weaponized crimes and civilian acquisition of contrabands, by frisking people and removing such weapons and contrabands from the public. The program is subject to many controversies, mainly on the targeting of pedestrians based on their racial backgrounds and the overall effectiveness of the program in reducing crimes in the NYC. Therefore, in this case study, we would focus on the overall effectiveness of the performed SQF encounters, and we would also focus on finding clustered groups that are more targeted by the officers than others, such as racial or age groups.

How would you define and measure the effectiveness of such a program?

The effectiveness of such a program is based on how the SQF operations can accurately identify and seize people of interests in large quantities, that poses actual threat to society and contains illegal items. The main point of contact of this program towards the NYC population is through the engagement of the NYC police officers. Therefore, the crucial measure of the program is to document all enforced instances of the policy issued by the officers, and to study from the obtained data. By examining the statistics from the data, the program could learn useful information such as patterns or hot spots that are closely associated with these issues.

A useful consideration from analyzing the data can be the measure of successful and non successful SQF exertions, and maybe finding specific clustered groups related to race, age or by location to name a few. Although finding clustered groups may be helpful for officers in the operation, such clusters must be

proven to be related prior to being exploited. An example of this can be the targeting of a specific racial group, where the data must prove beforehand that civilians of that group poses an actual threat by having a higher ratio of possessing a weapon or a contraband. Other means of program effectiveness can be measured by other factors such as the number of seized items and the decrease in reported crimes rates, just to name a few.

Another important factor to mention is that the program is largely based on the NYC officers' judgement during patrol, therefore such factor must be included when analysing the program effectiveness. Due to the human nature, the officers' judgement can be biased and may negatively impact the efficiency of the program. For example, examining numerous civilians that turns out to be innocent, can correlate to wasting man hours throughout the day and decreasing the efficiency of the program. The efficiency can also relate to who and how the officers chooses to perform the SQF practices, such as targeting a specific group of individuals that are not proven to cause a threat and wasting man hours through unfair and unpredicted police brutality.

We should also evaluate the cost-effectiveness of the program, from the deployment of vast resources and manpower of the NYPD unit, which contains around 36000 officers who are responsible for monitoring about 8.5 million civilians living in NYC. This program is known for its long-lasting negative impact to several targeted population, consequently it must prove that the program itself is reducing crimes and not from other external factors.

What data would you need to be able to judge its effectiveness?

In general, we would need as many details as possible during each encounter between the NYC officers and the civilians. Such data will vary within the different circumstances, such as the distinct features of each individual suspected civilian, data of the involved single or multiple officers in question, data of the time and location of the SQF exercise, and probably some other miscellaneous data attributes.

For data on the suspected civilians, we could analyse if they were stopped, questioned, and frisked and whether they had illegal items in possession. We would also need data of the civilian's profiling, such as race, gender, age, height, weight, build, and any other relevant features that may help with clustering analysis. Recording the total number of arrest and the total number of seized weapons and contraband can also be beneficial to judge how many of these stops results in finding these items. For the data on the officer's section, we should analyse the reason why they decided to carry out SQF practise on a specific civilian, and what actions they used to perform the SQF exchange. Such measures can be analysed by the time spent performing the practise, the level of search and physical force used, whether they arrested the individual or not, and other assorted data attributes. For the data on time and location, it would be useful

to analyse during what time frame the SQF intervention occurs and where exactly within the NYC is it occurring. There are locations where crime rates and suspicious civilians are higher than usual, therefore having data like an address or coordinates can help map such aspects.

From obtaining and analyzing any of these data, one can draw educated and proven conclusions that whether the program efficiently reduces crimes rates by removing illegal items and finding proof that some groups are more targeted than others.

Data Understanding:

The second step of the CRISP-DM process is the data understanding phase, where we will focus more within the provided dataset to get more familiar with the data type contents, systematically deal with data quality issues, and start to process meaningful and hidden information which revolves around the business understanding.

Describe the meaning and type of data (e.g., scale, values) for each attribute in the data file.

We will be examining the dataset from the year 2012, due to been one of the peak years of the program and due to its high numbers of recorded data encounters. A file spec guideline is also supplied alongside the dataset, which describes the column fields of major attributes, and helps us process the skeleton of the dataset. The .csv dataset contains 532911 row records, which includes 112 column attributes ranging from suspect, police actions and location information for us to analyse. Following the data preparation, we have reduced the dataset to 492251 rows with 79 columns, and this dataset will be used for the following case study, attributes shown on Figure 1.1. The data types of the new dataset contains 73 string attributes, 3 float attributes, 2 integer attributes and a single datetime attribute.

The combined rows can be divided into 3 subsets, civilian, location and police data. Rows such as [sex, race, age, weight, haircolr, eyecolor, build, height, contrabn, pistol, riflshot, asltweap, knifcuti, machgun, othrweap, othpers, typeofid], represent data on the civilians. It contains data on the civilian's biological profile and if any illegal weapons of contraband are found in their bodies. Rows such as [city, coord, pct, inout, trhsloc, ac_incid, ac_time, datetime] represent location and time data. It contains data such as coordinates and exact locations of SQF enquiries such as was it inside a building and what time during the day, etc. Rows such as [perobs, perstop, explnstp, arstmade, sumissue, offunif, officrid, frisked, searched, pf_hands, pf_wall, pf_grnd, pf_drwep, pf_ptwep, pf_baton, pf_hcuff, pf_pepsp, pf_other, radio, ac_rept, ac_inves, rf_vcrim, rf_othsw, ac_proxm, rf_attir, cs_objcs, cs_descr, cs_casng, cs_lkout, rf_vcact, cs_cloth, cs_drgtr, ac_evasv, ac_assoc, cs_furtv, rf_rfcmp, ac_cgdir, rf_verbl, cs_vcrim,

cs_bulge, cs_other, rf_knowl, ac_stsnd, ac_other, sb_hdobj, sb_outln, sb_admis, sb_other, rf_furt, rf_bulg, offverb, offshld, forceuse, detailcm] represents police data. This data contains police decisions and actions during the SQF process, such as reason for SQF, any police force used, period of observation and so forth.

```
Index(['pct', 'inout', 'trhsloc', 'perobs', 'perstop', 'typeofid', 'explnstp',
       'othpers', 'arstmade', 'sumissue', 'offunif', 'officrid', 'frisked',
       'searched', 'contrabn', 'pistol', 'riflshot', 'asltweap', 'knifcuti',
       'machgun', 'othrweap', 'pf_hands', 'pf_wall', 'pf_grnd', 'pf_drwep',
       'pf_ptwep', 'pf_baton', 'pf_hcuff', 'pf_pepsp', 'pf_other', 'radio',
       'ac_rept', 'ac_inves', 'rf_vcrim', 'rf_othsw', 'ac_proxm', 'rf_attir',
       'cs_objcs', 'cs_descr', 'cs_casng', 'cs_lkout', 'rf_vcact', 'cs_cloth',
       'cs_drgtr', 'ac_evasv', 'ac_assoc', 'cs_furtv', 'rf_rfcmp', 'ac_cgdir',
       'rf_verbl', 'cs_vcrim', 'cs_bulge', 'cs_other', 'ac_incid', 'ac_time',
       'rf_knowl', 'ac_stsnd', 'ac_other', 'sb_hdobj', 'sb_outln', 'sb_admis',
       'sb_other', 'rf_furt', 'rf_bulg', 'offverb', 'offshld', 'forceuse',
       'sex', 'race', 'age', 'weight', 'haircolr', 'eyecolor', 'build', 'city',
       'detailcm', 'datetime', 'coord', 'height'],
      dtype='object')
```

Figure 1.1 – List of data attributes following data preparation

Verifying Data Quality.

Most datasets would contain many redundancies, and altering them earlier can simplify the dataset, which will result in a better understanding and leverage of the attributes and values within. For this case study, we used the programming language "Python" to prepare the data and generate suitable statistics and visualizations in the latter part of the analysis. When initially loading or accessing a dataset, it is important to check the data types of attribute values and to convert them if needed. We made sure to convert some attribute column values into numeric as a safety measure for future data preparations. For records with missing values, we would drop or delete such rows to ensure a more equal and complete set of data throughout each attribute field in all records. In the case for duplicate data, a module of the python language allows us to return the dataset with duplicate rows removed, however the records from each row came off as unique and distinct.

We can combine multiple columns to form a single column with reduced complications. Such an example is generated by the date and time columns forming a new column called datetime with both fields saved within. Converting values into more identifying features can assist with readability. One example of this is converting suspect height data from feet and inch to cm, where it uses two columns to form a single

column with a single measurement data. Another example is converting x-coordinate and y-coordinate values into longitude and latitude attributes. The main benefit is that other location related attributes, around 15 columns, are now redundant because the location can be traced within a map from the new longitude/latitude values. Some column fields may have values which are considered uncommon, referred to as outliers. An example of this is the age column, where ages between 0 to 999 are clearly out of the ordinary. This can be resolved by adding a filter to the dataset, where only records with ages between 10 to 70 are analysed in this case. The sex column has records with "UNKNOWN" values, so we removed or dropped such rows and reduced the dataset to 492265 rows. The period of observation was filtered to 300 minutes max or 5 hours max, further reducing the dataset to 492251 rows.

A file spec guideline was also supplied alongside the dataset, which describes the column fields of major attributes, and helps us process the skeleton of the dataset. The file contains detailed specifications for certain specific column attributes. Thus, for the other fields, we can remove columns that are not detailed enough nor useful, for ease of analysis. Following this procedure, the dataset would have 79 column attributes and 499074 records rows for us to proceed with the analysis, while initially having 532911 rows and 122 columns.

Simple Attribute Statistics.

The table is filled with statistic values from four chosen attributes, shown by Figure 1.2: Age, Sex, Arrest Made, and the Period of Observations. We chose these specific attributes because it can help us recognize or discover some clusters which stems from these dataset values. It can also give us an overall idea of the actual effectiveness of the performed SQF interactions.

The count value represents the number of records with such attributes, and all attributes having the same count of 492251 is testament of data quality. The mean value represents the average common result, with the mean age of 28.06 and mean period of observation 2.42 representing a central tendency. The median age of 24 and period of observation 1.0 represents the middle value within the attribute ranges. The mode represents the most frequently occurring value with the age being 19 years old, the male sex, a lower number of arrests made than arrested and a 1-minute observation period before SQF confrontation by the officers. We can assume arrest made are usually linked to possessing weapons and contrabands, which can be a giveaway on program efficiency. The min represents the lowest value within the attribute range while the max is the opposite. The age has a range of 10-70, the sex attribute has MALE or FEMALE as possible values, arrest made has YES or NO as Boolean values, and period of observations has 0-300 minutes.

	Age	Sex	Arrest Made	Period of Observations
Count	492251	492251	492251	492251
Mean	28.06	N/A	N/A	2.42
Median	24	N/A	N/A	1.0
Mode	19	MALE	NO	1.0
SD	11.58	N/A	N/A	3.986
Min	10	FEMALE	NO	0
Max	70	MALE	YES	250
Range	10-70	MALE or FEMALE	YES or NO	0-300

Figure 1.2 – Simple Attribute Statistics Table for Age, Sex, Arrest Made and Period of Observation

From the value statistics of this table, we can conclude that the officers in the SQF program, around a visual observation of a 1-3 minute, usually targets males around the late teens and adulthood ages, and most encounters don't result in an arrest. Therefore, we can comprehend that this program has questionable efficiency of finding weapons and contrabands and arresting them.

Attribute Visualizations.

Following data preparation and attribute statistical analysis, incorporating visualizing data attributes can help us identify trends or patterns from the hundred and thousands of record rows within a visual context, rather than to painfully look at each row. For the civilian subset, we will look at the age distribution (Figure 1.3), the sex count plot (Figure 1.4), and the race plot as well (Figure 1.5). For the location subset, we will only visualize the city count plot (Figure 1.6). While for the police subset, we will look at the arrest made count plot (Figure 1.7) and the period of observation count plot (Figure 1.8). We can assume that an arrest made is correlated to having weapons or contrabands found within the civilian's body.

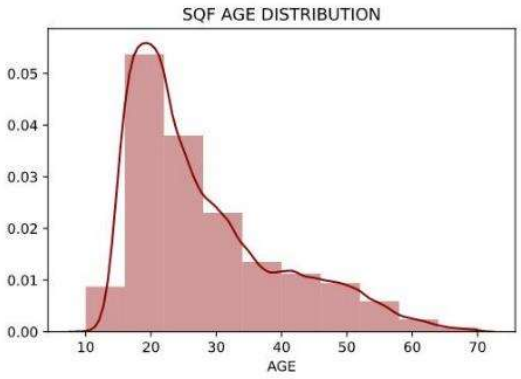

Figure 1.3 – SQF Age Distribution

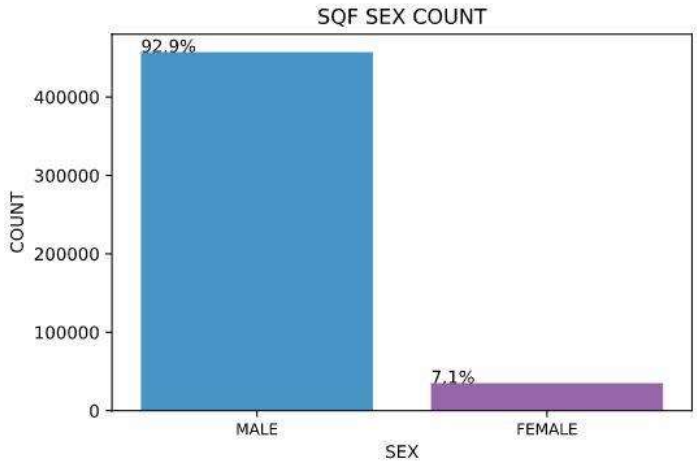

Figure 1.4 – SQF Sex Count

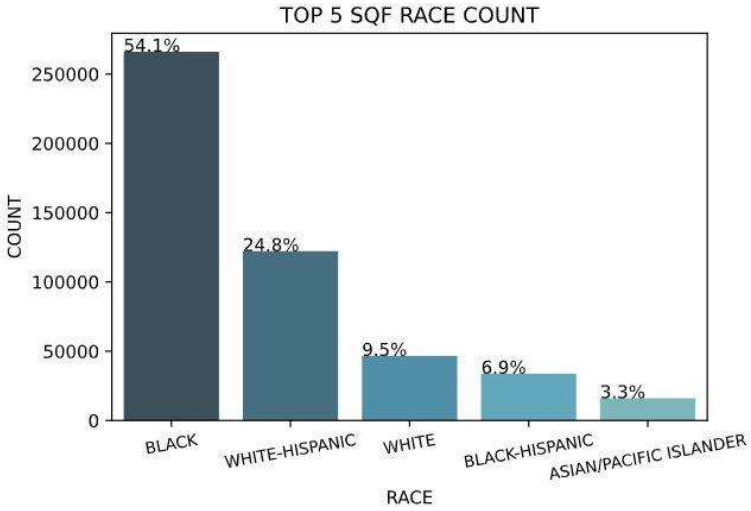

Figure 1.5 – Top 5 SQF Race Count

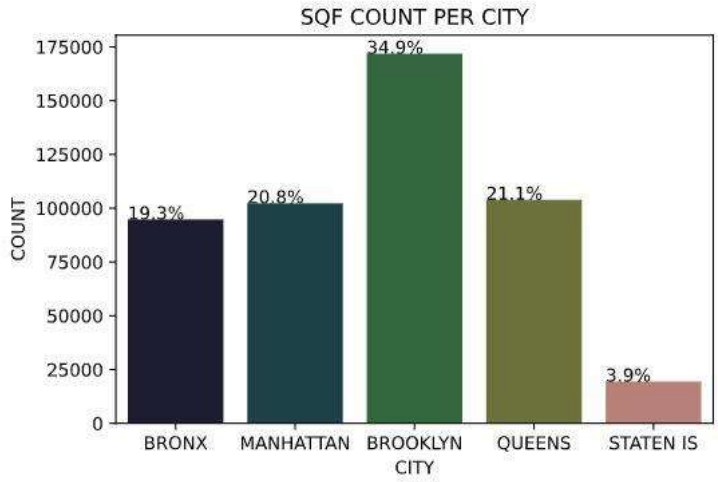

Figure 1.6 – SQF Count Per City

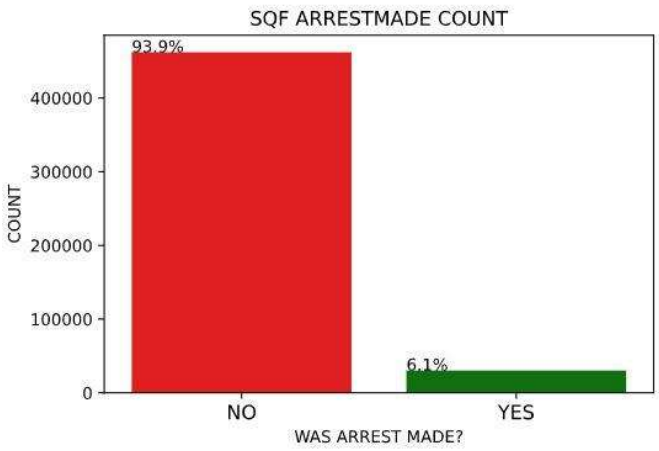

Figure 1.7 – SQF Arrest Made Count

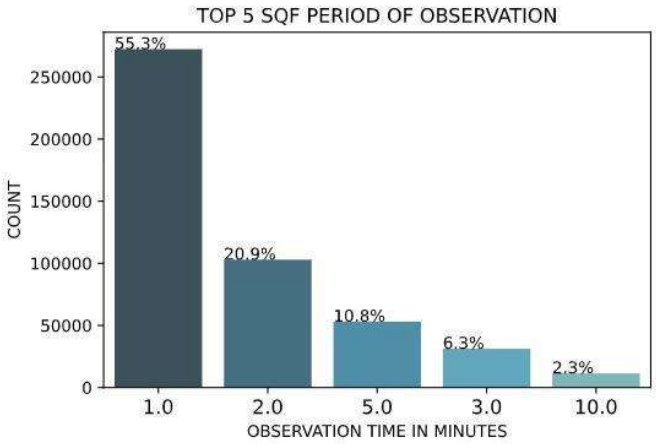

Figure 1.8 – Top 5 SQF Period of Observation for Police Officers

Starting with the civilian subset, the age distribution chart (Figure 1.3) demonstrate that the most targeted age group are the late teens to young adults, 15-25 years old. We did a distribution chart because of the high number of reoccurring age values, it was best to count those numbers in bin groups for every 10 years (0-10,10-20,20-30…years old). The rest of the charts are fully count plots because it has limited reoccurring values which can easily be counted and visually shown in such graphs.

The sex count plot (Figure 1.4) has shown that 92.9% of civilian targets are male, while 7.1% are female, which shows that there is a higher ratio of NYC officers targeting males than females. The race count plot (Figure 1.5) has shown that 54.1% of reported SQF civilian targets are Black, 24.8% are White-Hispanic, 9.5% are White, 6.9% are Black-Hispanic are so on. It's extremely apparent that civilians of Black race are mainly targeted by NYC officers, with over half the records designated for them.

For the location subset, the city count plot (Figure 1.6) demonstrates that Brooklyn City has the highest civilian SQF stops, with 34.9% records in the dataset, followed by Queens at 21.1% and Manhattan which

has 20.8% and so forth. For the police subset, the arrest made count plot (Figure 1.7) shows that 93.9% of officer SQF does not cause an arrest, and with earlier assumption that an arrest made is correlated to having weapons or contrabands, we can presume that most if not all of the 93.9% civilians are innocent. Although 6.1% of civilians were arrested, which questions the effectiveness of this SQF program. The period of observation count plot (Figure 1.8) shows that officers usually take around 1 minute of observation before approaching a civilian for SQF purposes, which also questions the randomness of the officer's judgement.

To conclude this part, black males around the age groups 15-25 are more specifically targeted by the officers in the SQF exercise, and further analysis can provide more verification on this statement.

Relationships between Attributes.

We can further analyse attributes by exploring additionally relationships between multiple attributes, where one can find new patterns or clustered groups which can further help with the case study. We will look at attributes arrest made and sex to analyse all SQF arrest made by sex (Figure 1.9), and we will look at attributes race and city to analyse all SQF race targets by city (Figure 1.10), we finally we will analyse a scatter plot map (Figure 1.11) detailing both criminal possession of controlled substances and criminal sale of controlled substances to look at the locations of these crimes and the amount of contrabands found relative all SQF records.

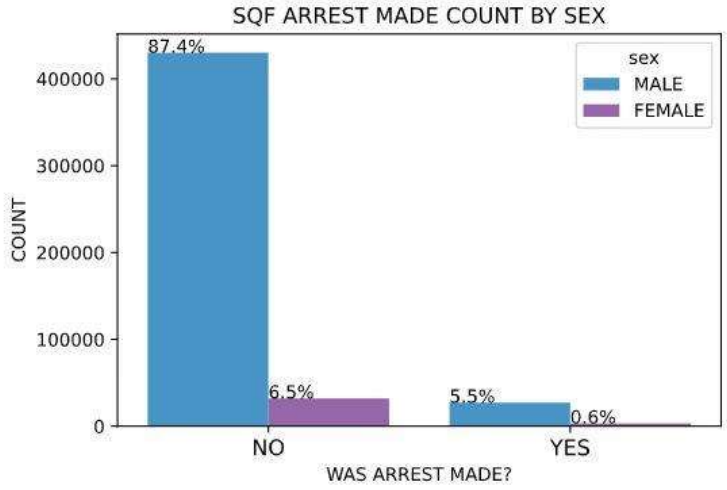

Figure 1.9 – SQF Arrest Made by Sex

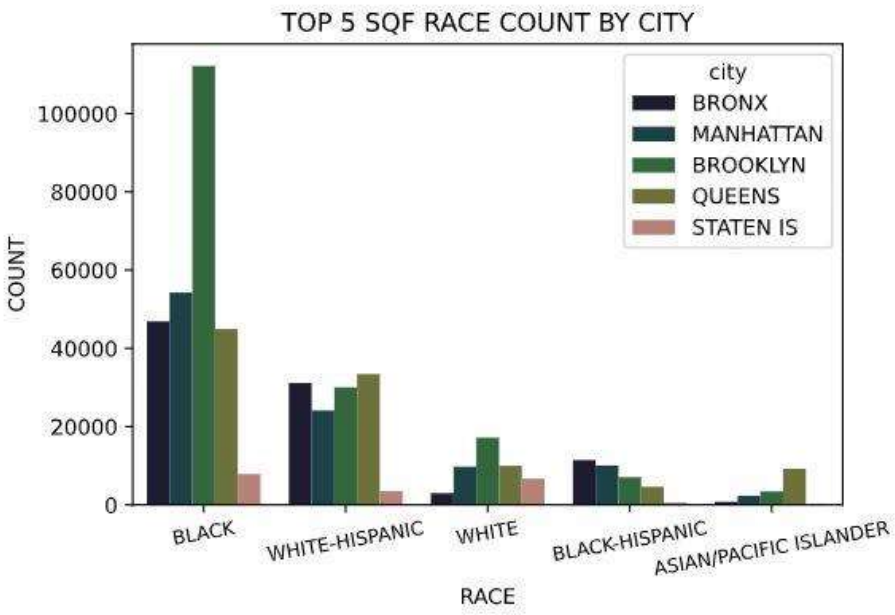

Figure 1.10 – SQF Race Count by City

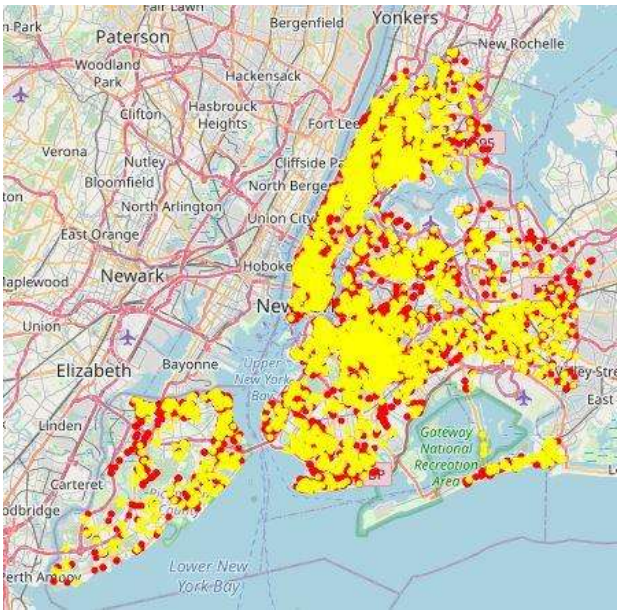

Figure 1.11 – NYC Scatter plot Map for criminal possession of controlled substances (Red) and criminal sale of controlled substances (Yellow)

The Arrest made by sex count plot (Figure 1.9) demonstrate two obvious visual factors, one is the ratio of male civilian SQF targets compared to females, and the other factor is the ratio of arrested males against females. The blue bars of this graph are visually excessive compared to the purple female bars. For non-arrest, males represent 87.4% of the total dataset, which is around 13.4 times larger than female non-arrest

numbers at 6.5%. For the arrested data, males represent 5.5% of the total dataset, which is around 9 times larger than female arrest at 0.6%.

The race by city count plot (Figure 1.10) demonstrate two obvious visual factors, one is the excessive SQF targeting towards the black race, while the other is the city of Brooklyn, which dominates all SQF records in city location. The scatter plot map (Figure 1.11) shows an appropriate count number of criminal possession (Red) and sale (Yellow) of controlled substances. Despite the negativity of the program's racial tendencies, it does surprising well in locating criminal possession and sale of contrabands.

It is obvious that the male sex and the black race are correlated and mostly targeted by NYC officers, and the map showed that the SQF project does produce results with contrabands, but the question still remains of its sufficiency and other efficiency.

Compare reasons for SQF and what type of force was used by the officer.

Firstly, we combined the columns starting with "pf_" which stands for physical force used by officer, followed by columns starting with "cs_" which stands for reason for stop, then finally added columns starting with "rf_" which stands for reason for frisk. Shown in Figure 1.12 is the correlation heat map for SQF reasons against Type of forces. The correlation of 1.0s are between the same columns, thus we should not examine those. Although, as we can see in the heat map, there's plenty of other correlation values ranged from 0.1 to 0.7. We'll look at correlations over 0.5 with greenest colors in the figure, which are the column/row set cs_bulge and rf_bulg with correlation 0.7, and column/row cs_vcrim and rf_vcact with correlation 0.5. The cs_bulge and rf_bulg has a reason for stop and frisk because of a suspicious bulge within the civilian's body. It has a correlation of 0.7, which means that it has a strong positive linear relationship in both columns occurring together. The cs_vcrim and rf_vcact has a reason for stop and frisk because the civilian was engaged in a violent crime. It has a correlation of 0.5, which equates to a moderate positive linear relationship in both columns occurring together. Overall, the correlation heat map can find several combinations of SQF reasons, however modelling techniques may predict better.

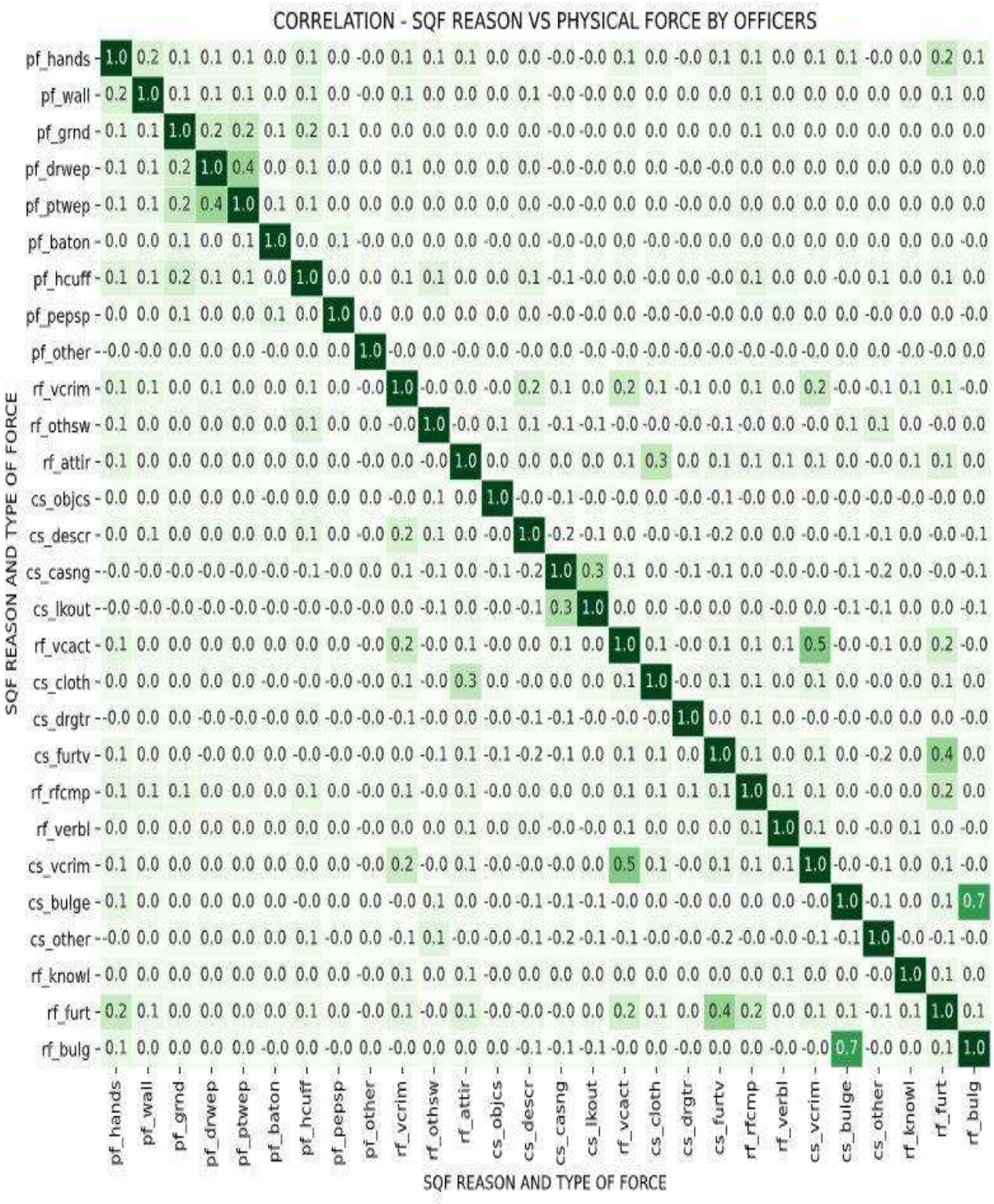

Figure 1.12 – Correlation Heat Map – SQL Reason and Type of Force

Report 2: Association Rule Mining

Data Preparation:

The third step of the CRISP-DM process is the data preparation phase, where we will focus on preparing the provided dataset to suit our analytical needs. Datasets are usually collected from many different sources and for many different intentions, therefore we need to clean, select, construct, and then integrate the raw data into the final dataset for our purposes. It is the most time-consuming part of the CRISP-DM model, and it is executed multiple times to ensure a smooth transition towards the modelling phase. Having bad data equates to bad predictions and analysis in the case study. The data preparation for our dataset was performed in Report 1 and is described in more detail in the Data understanding section.

The association rule mining is a procedure used to identify relations between different items in a dataset. The first step is to convert the dataset to an invoice-based transactional format, which will be needed for the Apriori algorithm, and will be explained in the modelling phase. We performed some One Hot encodings to the race, city, and sex attributes, where these categorical variables are converted into a format that is useful for predictions using machine learning algorithms.

Modelling:

The fourth step of the CRISP-DM process is the modelling phase, where we use a modelling technique to model the final dataset, so that it can output answers, which are closely associated with the business question. It can predict future statistics based on old data or it can find current trends or patterns. It follows a repetitive sequence of selecting the modelling technique, making a test design, building the model, then finally evaluating such model.

The association rule mining is based on the Apriori Algorithm, which works by finding frequently occurring items in a dataset, then finding larger regularly occurring sets of items and so on. The associated rule needs to satisfy both a specified minimum support, which is done to find all frequent itemset, and a specified minimum confidence, which is applied to the frequent itemset to create rules. The Apriori Algorithm has three-unit components: Support, Confidence and Lift. Support represents how frequently the itemset occurs in the dataset, the Confidence represents how probable that an item occurs alongside another when together, and the Lift represents the ratio increase when an item occurs alongside another.

To create a frequent itemsets, we must apply the frequent itemsets mining procedures, where we inputted a minimum support level of 10%. The algorithm will mainly focus on attributes race, city and sex and we should get predictions on where and which civilians are being targeted for SQF practices by NYC officers. The result of the frequent itemset is shown by Figure 2.1. What we learn is that from the dataset on SQF targeting by NYC officers, almost 92.9% are Males, about 54% are Black, around 50% are both Male and Black, and around 21% are Black Males from Brooklyn.

	support	itemsets
7	0.928855	(sex_MALE)
1	0.540610	(race_BLACK)
11	0.503211	(race_BLACK, sex_MALE)
3	0.349111	(city_BROOKLYN)
13	0.324751	(sex_MALE, city_BROOKLYN)
2	0.248251	(race_WHITE-HISPANIC)
12	0.231286	(sex_MALE, race_WHITE-HISPANIC)
9	0.227847	(race_BLACK, city_BROOKLYN)
17	0.211754	(race_BLACK, sex_MALE, city_BROOKLYN)
4	0.211132	(city_QUEENS)
5	0.207754	(city_MANHATTAN)
14	0.200373	(sex_MALE, city_QUEENS)
6	0.192536	(city_BRONX)
15	0.190053	(city_MANHATTAN, sex_MALE)
16	0.178376	(sex_MALE, city_BRONX)
0	0.134175	(pf_hands)
8	0.129241	(pf_hands, sex_MALE)
10	0.110220	(city_MANHATTAN, race_BLACK)
18	0.101371	(city_MANHATTAN, race_BLACK, sex_MALE)

Figure 2.1 – List of Frequent Itemsets with Support values

To further filter out the frequent itemsets, we can find matches of multiple attribute values that are commonly occurring in the dataset. Figure 2.2 shows itemsets of 3 items. We notice that Black Males are values that are reoccurring trends, however we have the cities of Brooklyn and Manhattan which represents 21.1% and 10.1% of the total dataset, respectively.

	support	itemsets
17	0.211754	(city_BROOKLYN, race_BLACK, sex_MALE)
18	0.101371	(city_MANHATTAN, race_BLACK, sex_MALE)

Figure 2.2 – List of Frequent Itemsets with 3 item values

Followed the computation of frequent itemsets, we must then apply the association rules. We inputted a minimum confidence of 93% in the algorithm towards the frequent itemsets to form rules. Figure 2.3 shows the association rules mining output. The output shows 5 association rules with high confidences of over 93%, and lift values over 1.0 which means that the antecedents and consequents of this figure appear together more often, and support which signifies the amount of times these itemsets occur in the dataset.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(pf_hands)	(sex_MALE)	0.134175	0.928855	0.129241	0.963224	1.037001	0.004611	1.934524
1	(race_BLACK)	(sex_MALE)	0.540610	0.928855	0.503211	0.930820	1.002115	0.001062	1.028393
2	(race_WHITE-HISPANIC)	(sex_MALE)	0.248251	0.928855	0.231286	0.931662	1.003022	0.000697	1.041074
3	(city_BROOKLYN)	(sex_MALE)	0.349111	0.928855	0.324751	0.930224	1.001473	0.000478	1.019615
4	(city_QUEENS)	(sex_MALE)	0.211132	0.928855	0.200373	0.949043	1.021733	0.004262	1.396159

Figure 2.3 – Output following association rules mining sorted by lift

To visualize theses results, we made two scatterplots visualizing the relationship between support and confidence (Figure 2.4) and the relationship between support and lift (Figure 2.5).

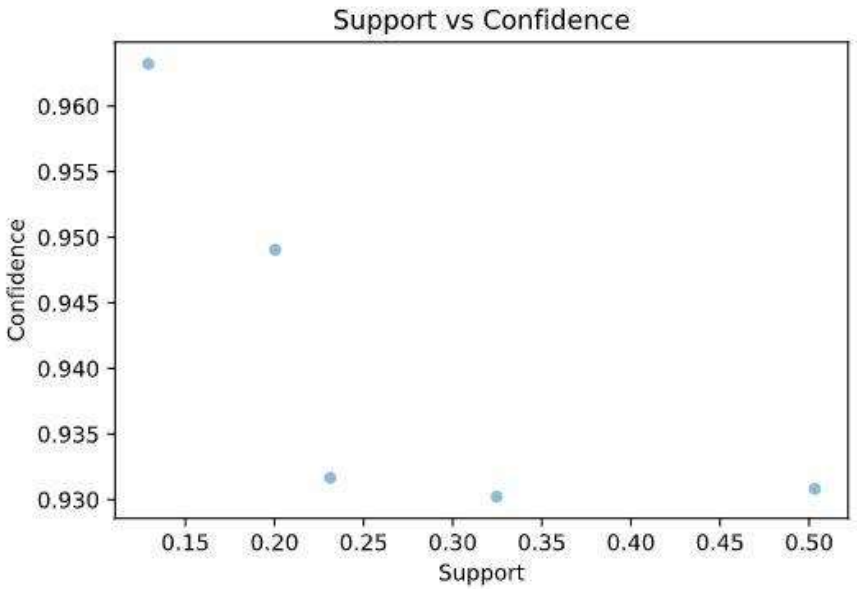

Figure 2.4 – Support vs Confidence

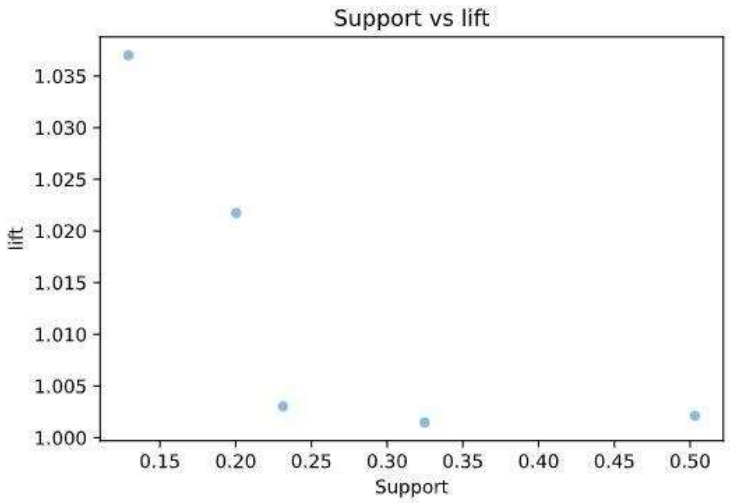

Figure 2.5 – Support vs Lift

From Figure 2.4, we see support greater than 10% and confidence greater than 93%, which means that those association rules have a very strong chance of occurring and this model can be used as a predictor. From Figure 2.4, we still see the support greater that 10%, yet the lift being higher than 1 for all association rules meaning it is likely to occur in a prediction.

Evaluation

The fifth step of the CRISP-DM process is the evaluation phase, where we will judge the results of the model, containing trends and patterns, and deem them accurate and satisfying regarding the business question. The association rule mining for the SQF dataset was fruitful in finding strong rules with a high prediction rate of occurrences. What I found the most interesting is that both Black and White-Hispanic males have the highest rate of SQF targeting by the officers. Unsurprisingly, both Brooklyn and Queens are the hot spot for this SQF program.

Report 3: Cluster Analysis

Data Preparation:

As mentioned in Report 2, the third step of the CRISP-DM process is the data preparation phase, where we will focus on preparing the provided dataset by cleaning, selecting, constructing, and then integrating the raw data into the final dataset for our purposes. The data preparations were done in earlier reports, where we combined multiple column attributes to form a single attribute, dropped missing values and deleted redundant columns, and dealt with outliers by filtering, such that the final dataset has 492251 record rows with 79 column attributes. The dataset can also be considered subdivided into civilian, police and location data. The data types of the final dataset are divided into 73 string attributes, 3 float attributes, 2 integer attributes and a single datetime attribute. In the modelling section, we will be analysing the crime of "CRIMINAL SALE OF CONTROLLED SUBSTANCE", thus we copied that dataset into another data frame and filtered the crime into the "detailcm" attribute which stands for detail of the crime.

Modelling:

As mentioned before. the fourth step of the CRISP-DM process is the modelling phase, where we use a modelling technique to model the final dataset, so that it can output answers, which are closely associated with the business question. The results can predict future statistics or current trends or patterns, based on data. Cluster Analysis is a technique used to group set of data points, such that they have similar characteristics within the dataset. It is a form of an unsupervised learning, where the machine learning algorithm works with unlabeled dataset, and has several clustering algorithms in the form of K-Means, DBSCAN and Hierarchical clustering.

The K-Means clustering will sort all data points into several groups, and the number of such groups is represented by the K variable. The groups are in equal variance, while minimizing the inertia which defines the distance between data points within a group. The DBSCAN, commonly referred to as Density-Based Spatial Clustering of Applications with Noise, is a clustering algorithm which will group the data points that are closely populated together, then produce clusters from it. The Hierarchical clustering will build cluster in a hierarchical format by initially allocating each data point as a cluster, then recursively combining pair of cluster data while considering a small linkage distance.

There are two clustering metrics that we should consider, which are Homogeneity and Silhouette. The homogeneity is something the clustering outcome must satisfy whenever the clusters has data values that are members of a single class. The silhouette is a value which represents how similar a data point is relative to its neighbor data points (cohesion) and compared to other clusters (separation).

For our first cluster analysis, we will cluster the location for the crime of "CRIMINAL SALE OF CONTROLLED SUBSTANCE". We have performed a hierarchical clustering algorithm for this modelling, and we recorded the silhouette scores which will help determine a suitable number of clusters needed. A line plot (Figure 3.1) representing the silhouette score and number of clusters can also aid in visually determining the suitable number of clusters. We notice straightaway that the silhouette scores are multiples of 5, with cluster number 5 having the highest value percentage of 50.68%. The best number of cluster value can also be performed with the Max Python function, which will also return the higher scored cluster k value of 5 (Figure 3.2).

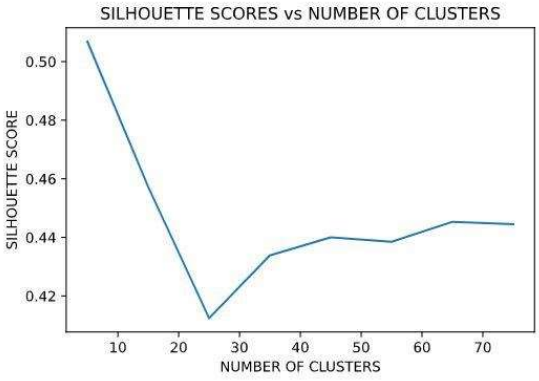

Figure 3.1 – Silhouette and Clusters chart

```
best_k = max(silhouette_scores, key=lambda k: silhouette_scores[k])...

5

silhouette_scores

{5: 0.5068221725184908,
 15: 0.45718263992573055,
 25: 0.4124306579102912,
 35: 0.43381677753426345,
 45: 0.44002109060328204,
 55: 0.4384767103631284,
 65: 0.44533015738039766,
 75: 0.44450142155700245}
```

Figure 3.2 – Best cluster K value and Silhouette scores

By plotting the clustered groups on a map (Figure 3.3), one can notice 5 distinct coloured groups that shows separate locations with the crime occurring in the NYC. We can probably add more filter attributes from the clustering dataset to make more use of the model. As an example, the model is aimed towards a certain crime at this point, but we can add attributes such as age or sex to further optimize.

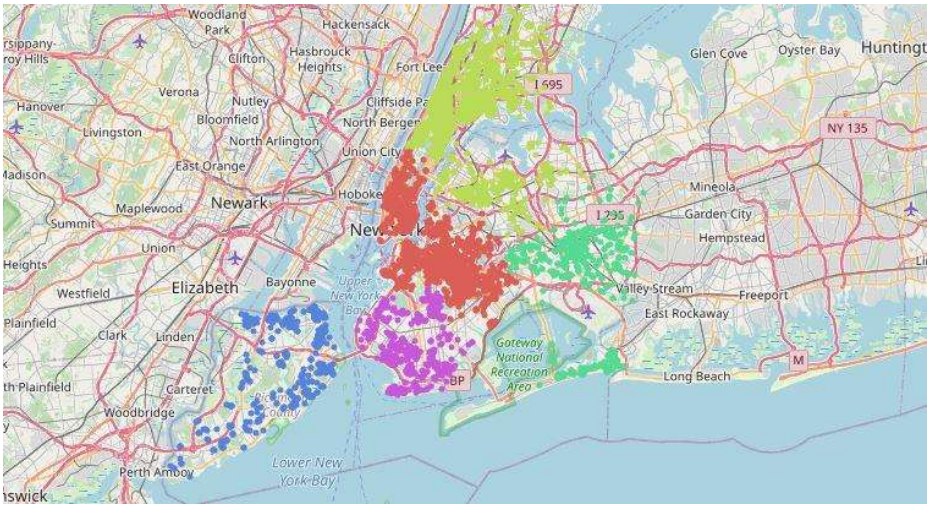

Figure 3.3 – Map with Clustered dataset on CRIMINAL SALE OF CONTROLLED SUBSTANCE

For our second cluster analysis, we will cluster on the reasons why civilians were stopped by NYC officers for SQF purposes. We have performed a K-means clustering algorithm for this modelling, and we recorded the silhouette scores which will help determine a suitable number of clusters needed. A line plot (Figure 3.4) is used to represent the silhouette score and number of clusters, and the first elbow can be noticed at cluster value 25, which may be a suitable number of clusters. From the silhouette scores, we notice that the highest value percentage of 74.06% belongs to cluster number 25. The python Max function also returns k value of 5 when it compares all suitable cluster numbers (Figure 3.5).

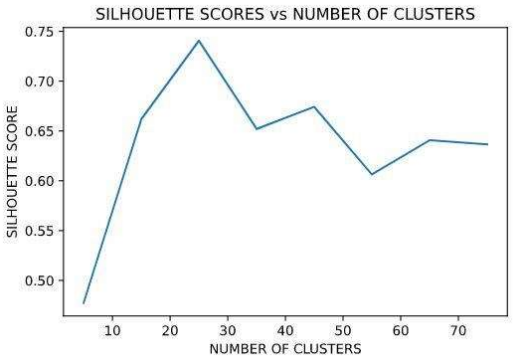

Figure 3.4 – Silhouette and Clusters chart

```
best_k = max(silhouette_scores, key=lambda k: silhouette_scores[k])...

25

silhouette_scores

{5: 0.4775107000213545,
 15: 0.6618728788151234,
 25: 0.7406366128194455,
 35: 0.6520054784309365,
 45: 0.6742608407756617,
 55: 0.606401934459625,
 65: 0.6406894297262201,
 75: 0.6365697454792264}
```

Figure 3.5 – Best cluster K value and Silhouette scores

By plotting the clustered groups by reasons of stop on a map (Figure 3.6), one should notice 25 distinct colored groups in the NYC, but it's somewhat difficult to see. Like the previous modelling, we can add more filter attributes from the dataset to make more use of this model.

Figure 3.6 – Map with clustered dataset on the reasons for stop

Evaluation

As mentioned before. the fifth step of the CRISP-DM process is the evaluation phase, where we will judge the results of the model, containing trends and patterns. The cluster analysis was useful to find cluster location data on the crime of "CRIMINAL SALE OF CONTROLLED SUBSTANCE", along with the reasons for stop by the NYC officers. What I found interesting is the division of 5 location sectors for the crime in NYC, which further analysis will be required to understand the reasoning behind it. However, the mapping for the reasons for crimes was overpopulated with 25 clusters, which represent all reasoning behind the SQF exercise on civilians along with other factors which requires further analysis. These findings can used to gather similar clusters of patterns or trends in a locational setting, and further analysis can help reveal answers such as the density of crimes in a hot spot location.

Report 4: Predictive Modelling

Data Preparation:

As mentioned in previous reports, the third step of the CRISP-DM process is the data preparation phase, where we will focus on preparing the provided dataset by cleaning, selecting, constructing, and then integrating the raw data into the final dataset for our purposes. Most of the data preparations were done in earlier reports, where we combined multiple column attributes to form a single attribute, dropped missing values and deleted redundant columns, and dealt with outliers by filtering, and so forth.

However, for this modelling, we what to predict if a person will be armed. So, we added columns starting with "rf_", which stands for reason for frisk, columns starting with "cs_", which stands for reason for stop, into a dataframe called "x". In the same dataframe, we also added number and category attributes such as age, height, weight, race, city and build. Then we added the Boolean true or false values of civilians having any sort of weapons in their person during SQF investigation, into another dataframe called "y". To measure the accuracy of models, we would split the dataset into training and testing component, where the model is created from the training component, and the testing is execute alongside the training model to deliver classification metrics which dictates the correctness or the classification metrics. The split of the dataset is 75% for the training and 25% for the testing for both x and y dataframe, and the python implementation called train_test_split will generate the 4 components on our behalf.

Modelling:

As mentioned before. the fourth step of the CRISP-DM process is the modelling phase, where we use a modelling technique to model the final dataset, so that it can output answers, which are closely associated with the business question. The results can predict future statistics or current trends or patterns, based on data.

Classification analysis is a supervised learning technique used to predict which class a single of multiple data points belong to. The predictions can be measure using several types of classification metrics, which are accuracy, precision, recall and Fscore. Accuracy is the measure which predicts correct predictions out of the total data points. Precision is the measure of the correct proportion of predictions against the number of data points classified as relevant. Recall is the measure of the correct proportion of predictions against the number of relevant data points, as opposed to data points classified as relevant. The Fscore is a measure that uses both precision and recall to determine the prediction effectiveness.

There are several classification algorithms for this analysis, and we will focus on 3 specific ones which are the decision tree classifier, the logistic regression classier and the Naïve Bayes classifier. The decision tree classifier performs classification analysis in tree-like structure consistent of branches and leaf nodes that represent the conditions of a dataset into small subsets, which are incrementally tested and predicted. The advantage of this model is that it considers all possible outcome in a recursive fashion and can handle numerical and categorical data. The logistic regression classifier performs classification analysis by predicting the probability of conditions occurring in a discrete set of classes. The advantages of this model are that it is very efficient with minimal computing, efficient to train the training dataset and easy to implement. The Naïve-Bayes classifier performs classification analysis by using Bayes' Theorem of probability, which represents a family of algorithm, to predict the set of classes. The advantages of this model are that it can predict real-time situations, can handle continuous and discrete data, is scalable depending on the number of predictions and data points and can work with smaller training datasets.

What we want to predict if a person is armed using the 3 described classification models. The first model we looked at is the Decision tree classifier (Figure 4.2) where we tested on a training set prior to deploying it with the testing dataset. The predictions from the training dataset shows an accuracy of 0.999 which is high, a precision of 0.999 where a score close to 1 is perfect, a recall of 0.965 which is also perfect, and finally a Fscore of 0.982 with an ideal score being 1. Overall, the training model is accurate, however the testing dataset shows and accuracy of 0.944, a precision of 0.114, a recall of 0.142 and a f1 score of 0.127. The testing dataset, apart from the accuracy, how scored lower on everything else and this may suggest that although it can predict if a person is armed or not, it may not classify or prove in a proper way.

```
DecisionTreeClassifier
..Training Result:
....acc: 0.9990032178727369
....precision: 0.9998029944838456
....recall: 0.9651958919741347
....f1: 0.9821946971163151
..Testing Result:
....acc: 0.9446056085094626
....precision: 0.11483364140480591
....recall: 0.1426930806775768
....f1: 0.12725643323518115
```

Figure 4.1 – Decision Tree Classifier Results

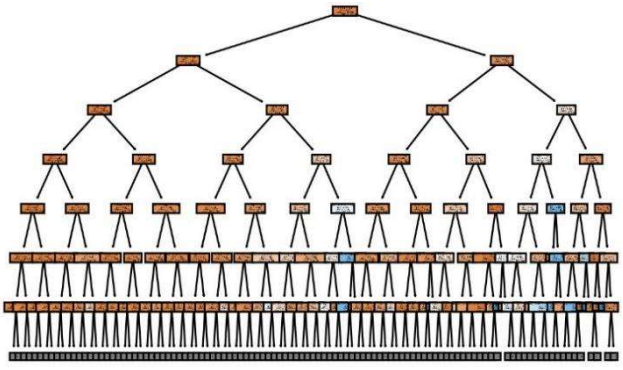

Figure 4.2 – Decision Tree, too clustered to read

The second model is the Logistic Regression classifier for the same task of predicting an armed person, and similarly tested on a training set prior to deploying it with the testing dataset. The predictions (Figure 4.3) from the training dataset shows an accuracy of 0.971 which is high, a precision of 0.435 which is less than halfway the ideal score close to 1, a recall of 0.022 which is very low, and a Fscore of 0.042 which is considerably low. Interestingly, the testing dataset is extremely similar to the training dataset. The testing dataset shows and accuracy of 0.971, a precision of 0.414, a recall of 0.20 and a Fscore of 0.038. We may conclude that this model may have similar predictive capabilities when predicting a person is armed. We also plotted a bar plot for the coefficients in the logistic regression classifier (Figure 4.4) which depicts several attribute values to occur more frequently or others less than ideal.

```
LogisticRegression
..Training Result:
....acc: 0.9713262619586769
....precision: 0.4354243542435424
....recall: 0.022441993153290225
....f1: 0.04268402966178333
..Testing Result:
....acc: 0.9714617716129137
....precision: 0.41420118343195267
....recall: 0.020097616996841802
....f1: 0.038335158817086525
```

Figure 4.3 – Logistic Regression Classifier Results

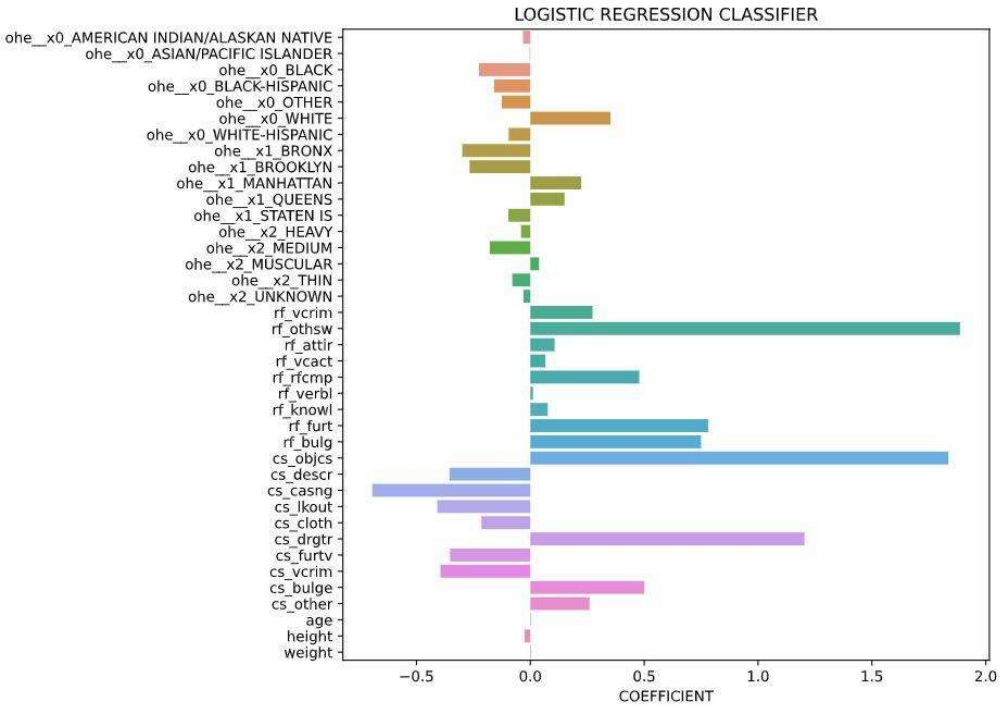

Figure 4.4 – Logistic Regression Classifier Chart

The third model is the Naïve-Bayes classifier for the same task of predicting an armed person, and similarly tested on a training set prior to deploying it with the testing dataset. The predictions (Figure 4.5) from the training dataset shows an accuracy of 0.970 which is high, a precision of 0.356 which is lower, a recall of 0.035 which is very low, and a Fscore of 0.064 which is considerably lower. Similarly, to the previous modelling, where the training and testing results are almost similar. The testing dataset shows and accuracy of 0.970, a precision of 0.363, a recall of 0.037 and a Fscore of 0.0681. We may conclude that this model may have similar predictive capabilities when predicting a person is armed. We also plotted a bar plot for the coefficients in the Naïve-Bayes classifier (Figure 4.6) which depicts a many attribute values with a coefficient less than 0.

```
MultinomialNB
..Training Result:
....acc: 0.9707005644820526
....precision: 0.3565300285986654
....recall: 0.035564853556485536
....f1: 0.06467790747946390
..Testing Result:
....acc: 0.9709010831850353
....precision: 0.3638888888888889
....recall: 0.037611254665518234
....f1: 0.06817590424147801
```

Figure 4.5 – Naïve-Bayes Classifier Results

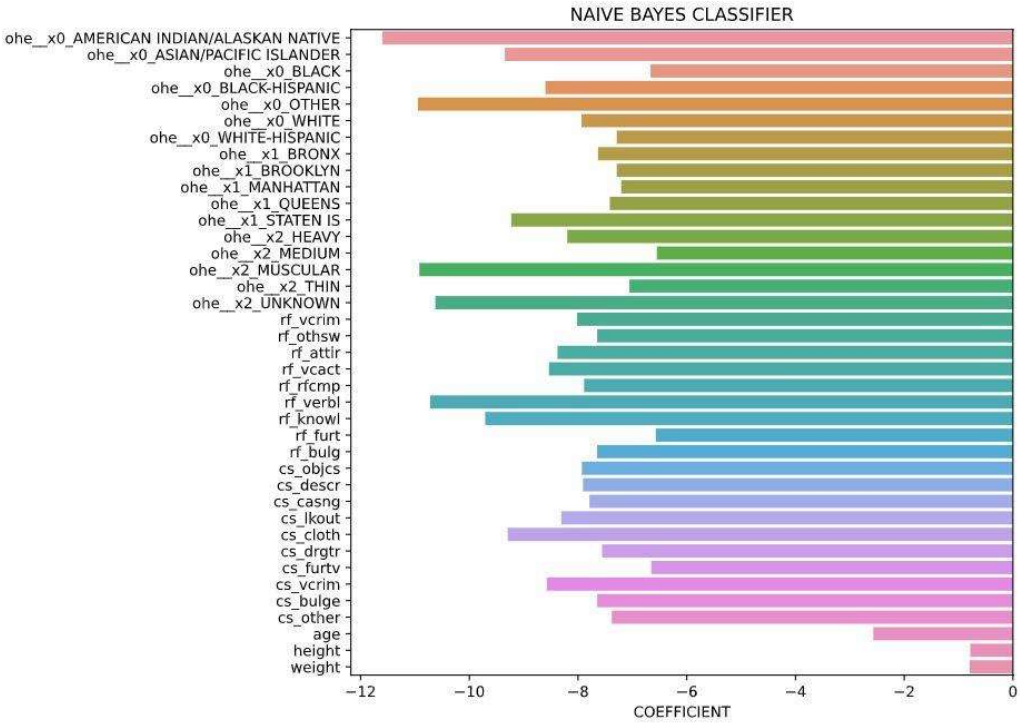

Figure 4.6 – Naïve-Bayes Classifier Chart

Evaluation

As mentioned before. the fifth step of the CRISP-DM process is the evaluation phase, where we will judge the results of the model, containing trends and patterns. The last two of the predictive modelling might be useful for police use because the predictions from both the training datasets are almost the same at the predictions of the testing dataset, which may correlate to similar result data for new testing datasets. Nevertheless, the results from these predictive models are not the greatest, and other specific and filtered data can help reduce some unknown clustered data points from the dataset. For a large scale operation like the SQF program, daily updates may be necessary to track complex network of a city like NYC.

References For All Reports:

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://www.foxnews.com/us/what-is-stop-and-frisk

https://en.wikipedia.org/wiki/Stop-and-frisk_in_New_York_City

https://www1.nyc.gov/site/nypd/about/about-nypd/about-nypd-landing.pagehttps://journalistsresource.org/studies/government/criminal-justice/stop-question-frisk-police-tactics-research-review/

https://silo.tips/download/list-of-variables-variable-description-values