**Advanced Concepts in Data Analytics**

**Final Project: New York City Case Study using CRISP-DM**

**Report 1: Data and Visualization**

**Report Submitted by:**

Saymon Acharya

Sangram Singh

Gurvinder Singh

Table of Contents

[Abstract/Executive Summary 3](#_Toc164842890)

[Business Understanding 4](#_Toc164842891)

[What is the purpose of the SQF program? 4](#_Toc164842892)

[How would you define and measure the effectiveness of such a program? 4](#_Toc164842893)

[What data would you need to be able to judge its effectiveness? 4](#_Toc164842894)

[Data Understanding 6](#_Toc164842895)

[Describe the meaning and type of data (e.g., scale, values) for each attribute in the data file. 6](#_Toc164842896)

[Verify data quality. Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems? 7](#_Toc164842897)

[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. 10](#_Toc164842898)

[Visualize the most important attributes appropriately (at least 5 attributes). 11](#_Toc164842899)

[Important: Provide an interpretation for each chart, explaining each attribute and why you chose the visualization you did. 11](#_Toc164842900)

[Explore relationships between attributes. Look at the attributes and then scatter plots, correlation, cross-tabulation, group-wise averages, etc., as appropriate. 16](#_Toc164842901)

[Compare the reasons for an SQF and what type of force was used by the officer. 19](#_Toc164842902)

[Conclusion 20](#_Toc164842903)

# Abstract/Executive Summary

The Stop-Question-Frisk (SQF) initiative, launched by the New York City Police Department (NYPD), aimed to curtail crime through the detention and search of individuals suspected of illegal acts. This approach intensified in the early 2000s, reaching over 500,000 stops in 2012, but soon faced criticism and legal challenges over claims of racial prejudice, especially against Black and Hispanic people. A critical lawsuit in 2013 ruled the method unconstitutional, leading to a sharp reduction in the number of stops.

This study examines SQF incidents from 2012 to detect any ongoing biases in its enforcement, with a particular focus on Black and Hispanic populations. Using spatial data analysis, the research assesses the impact of precinct locations on the frequency and characteristics of stops. Early results show significant variations across geographical areas, hinting at the influence of location on the enforcement of SQF policies.

The study also assesses the consequences of stops, examining the measures taken during and after, to see if discriminatory actions continue despite reforms. Preliminary outcomes indicate racial disparities, pointing to a continued need for policy reassessment and modification.

These insights are vital for policymakers, law enforcement, and the community, emphasizing the need for further adjustments to NYPD practices to uphold justice and legality in the SQF initiative. The study advocates for ongoing research and monitoring to maintain transparency and equity in police procedures.

# Business Understanding

## What is the purpose of the SQF program?

The Stop-Question-and-Frisk (SQF) program implemented by the New York City Police Department (NYPD) was designed to reduce crime by stopping, questioning, and searching civilians suspected of engaging in criminal activities. The rationale behind the policy is based on the idea that such proactive measures could deter criminal behavior, confiscate illegal items, and potentially prevent crimes before they occur.

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

The effectiveness of the SQF program can be defined and measured through several key indicators:

**Reduction in Crime Rates:** The primary measure would be a noticeable decline in local crime rates after the implementation and during the operation of the SQF program, especially in areas with high stop rates.

**Seizure of Contraband:** The number of illegal weapons and contraband items recovered because of stops can serve as a direct measure of the program’s impact on public safety.

**Community Safety Perceptions:** Surveys and studies measuring public perception of safety in neighborhoods where SQF is heavily applied could also indicate effectiveness, showing whether residents feel the program makes their community safer.

**Legal and Procedural Compliance Rates:** Monitoring the frequency of stops that result in lawful arrests compared to those deemed unjustified or leading to complaints could help assess the program's adherence to legal standards and fairness.

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

To comprehensively judge the effectiveness of the SQF program, the following data would be essential:

**Crime Statistics:** Detailed crime reports before and after the implementation of the SQF policy, segmented by types of crime and geographical areas, to observe any correlations between stops and crime rates.

**Stop Data:** Comprehensive data on stops, including the number of stops, reasons for stops, outcomes (e.g., no further action, summons, or arrest), and recovery of weapons or contraband.

**Demographic Information:** Data on the race, age, and gender of individuals stopped to analyze the fairness and targeting efficiency of the policy.

**Geospatial Data:** Locations of stops to explore spatial relationships and impacts, potentially using GIS technology to visualize and analyze data.

**Community Feedback:** Results from community surveys on perceptions of police presence and effectiveness, feelings of safety, and the perceived impact of the SQF policy on the community.

**Legal Compliance and Complaints:** Records of complaints and lawsuits related to the SQF practices to monitor compliance with legal standards and respect for civil liberties.

Collecting and analyzing these types of data would provide a multidimensional view of the SQF program’s impact, enabling policymakers and law enforcement to assess its effectiveness and make informed decisions on its future implementation.

# Data Understanding

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

The table used is given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| frisked | knifcuti | pf\_wall | sex | race |
| Y | N | Y | M | A |
| N | Y | Y | M | U |
| Y | Y | N | F | W |
| Y | N | Y | Z | I |
| N | N | N | F | P |

Table 1: Attributes for Report 1

The attributes we filtered out from the given dataset are:

1. **frisked** – It means “Was suspect frisked?”  
   Scale of attribute:
   1. **Yes:**   
      Total Count - 297244
   2. **No:**   
      Total Count - 235667
2. **Knifcuti** – It means “Was a knife or cutting instrument found on suspect?”  
   Scale of attribute:
   1. **Yes:**   
      Total Count - 4705
   2. **No:**   
      Total Count - 528206
3. **pf\_wall** – It means “Physical force used by officer - suspect against wall”  
   Scale of attribute:
   1. **Yes:**   
      Total Count - 12627
   2. **No:**   
      Total Count - 520284
4. **sex** – It means “Suspect’s Sex”  
   Scale of attribute:
   1. **Male:**   
      Total Count - 487065
   2. **Female:**   
      Total Count - 38062
   3. **Z/Others:**

Total Count **-** 7784

1. **race** – It means “Suspect’s Race”  
   Scale of attribute:
   1. **A: Asian**

Total Count - 17058

* 1. **B: Black or African American**

Total Count - 284229

* 1. **I: American Indian**

Total Count - 2257

* 1. **P: Pacific Islander**

Total Count - 35772

* 1. **Q: Questioned/Unknown**

Total Count - 129368

* 1. **U: Unknown**

Total Count - 3759

* 1. **W: White**

Total Count - 50366

* 1. **Z: Other**

Total Count - 10102

## Verify data quality. Are there missing values? Duplicate data? Outliers? Are those mistakes? How do you deal with these problems?

Verification of data quality is done by the given screenshot of the python code that we ran on the above attributes that we derived from the given dataset. The result of this python code provides us with the actual statistics of missing values, duplicate data, outliers, if any. The code is as follows:

import pandas as pd

# Load the data from a CSV file

# Replace 'path\_to\_your\_csv.csv' with the actual path to your CSV file

df = pd.read\_excel('report1\_attributes.xlsx')

# Check for missing values

missing\_values = df.isnull().sum()

# Check for duplicate rows

duplicate\_rows = df.duplicated().sum()

# Assuming binary columns are only 'Y' or 'N' and categorical columns have specific categories

binary\_columns = ['frisked', 'knifcuti', 'pf\_wall']  # Replace with your actual binary columns

categorical\_columns = {

    'sex': ['M', 'F', 'P'],  # Replace with your actual sex categories

    'race': ['B', 'Q', 'W']  # Replace with your actual race categories

}

# Function to detect outliers in binary columns

def binary\_outliers(df, columns):

    outliers = {}

    for col in columns:

        outliers[col] = df[~df[col].isin(['Y', 'N'])][col].unique().tolist()

    return outliers

# Function to detect outliers in categorical columns

def categorical\_outliers(df, columns):

    outliers = {}

    for col, categories in columns.items():

        outliers[col] = df[~df[col].isin(categories)][col].unique().tolist()

    return outliers

binary\_outliers\_detected = binary\_outliers(df, binary\_columns)

categorical\_outliers\_detected = categorical\_outliers(df, categorical\_columns)

# Create a summary report

report = {

    'missing\_values': missing\_values,

    'duplicate\_rows': duplicate\_rows,

    'binary\_outliers': binary\_outliers\_detected,

    'categorical\_outliers': categorical\_outliers\_detected

}

print(report)

Result of the above python code:

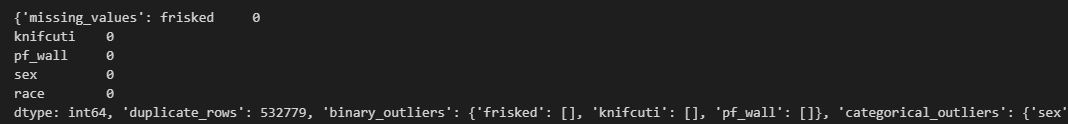


Figure 1: Verification of data quality

The result shows there are no missing values for any of the taken attributes. The duplicate rows are 532779 for all the attributes collectively. It also shows some negligible outliers.

Below is the code and its outcomes for how we dealt with the situation of duplicates and other outliers:

# Remove duplicate rows

df = df.drop\_duplicates()

# Correcting or removing outliers in 'sex' column

# Assuming 'Z' is not a valid category, we can either remove these rows or replace them with a valid value

# Example of removing:

df = df[df['sex'] != 'Z']

# Example of replacing 'Z' with 'Unknown' (or another appropriate value):

df['sex'] = df['sex'].replace('Z', 'Unknown')

# Similar corrections would be applied to the 'race' column based on the context of the data.

# Define a dictionary for correcting known errors or mapping values

race\_corrections = {

    'P': 'Known\_Category\_1',  # Replace 'Known\_Category\_1' with the actual category, if 'P' stands for something known

    'A': 'Known\_Category\_2',  # Do the same as above

    # 'U': 'Unknown', # Uncomment this line if you decide to label 'U' as 'Unknown'

    # 'Z': 'Other',  # Uncomment this line if 'Z' should be categorized as 'Other'

    'I': 'Known\_Category\_3'  # Replace with the actual category, if applicable

}

# Apply corrections

df['race'] = df['race'].replace(race\_corrections)

# For values that are errors and do not have a known correction, remove those rows

# For example, if 'Z' is an error and you want to remove it:

df = df[df['race'] != 'Z']

# Alternatively, label them as 'Other' if you do not want to remove rows

df.loc[df['race'].isin(['U', 'Z']), 'race'] = 'Other'

df.head()

**Outcomes:**

A screenshot of a computer

Description automatically generated

Figure 2: Handling duplicates/outliers.

## 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.

Below are some of the important statistics for the attributes taken for this report. We ran a python code to evaluate all these statistics.

# For binary columns, we'll calculate the count of 'Y' and 'N' and treat 'Y' as 1, 'N' as 0 to calculate mean and variance

binary\_columns = ['frisked', 'knifcuti', 'pf\_wall']

binary\_stats = {}

for col in binary\_columns:

    df\_binary = df[col].apply(lambda x: 1 if x == 'Y' else 0)

    binary\_stats[col] = {

        'counts': df[col].value\_counts(),

        'mean': df\_binary.mean(),

        'variance': df\_binary.var()

    }

# For categorical columns, we'll calculate the mode, counts, and the range (unique values)

categorical\_columns = ['sex', 'race']

categorical\_stats = {}

for col in categorical\_columns:

    categorical\_stats[col] = {

        'mode': df[col].mode().iloc[0],

        'counts': df[col].value\_counts(),

        'range': df[col].unique()

    }

# Compile the statistics into a report

report = {

    'binary\_stats': binary\_stats,

    'categorical\_stats': categorical\_stats

}

# Print the report

print(report)

Outcomes with the important Statistics:

A screenshot of a computer

Description automatically generated

Figure 3: Important Statistics from the attributes

**Some of the interpretations:**

* **Mean of 0.55777** could represent a probability of an event occurring. In context of Frisked attribute, there is an equal probability of person being frisked or not.
* In statistical analysis, variance is a measure of the dispersion or spread of a set of data points. Specifically, it quantifies how much the values in a dataset differ from the mean value. For event pertinent if knife cutting instrument found on suspect is a variance of 0.00875093 represents the average squared deviation of individual data points from the mean.
* Count is a fundamental aspect of statistical analysis, with applications ranging from basic data summarization to advanced statistical inference and decision-making in various fields.

## 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.

We have utilized python (matplot, seaborn libraries) to visualize the attributes and take out some interpretation from each of them as follows:

import matplotlib.pyplot as plt

# Cross-tabulation of 'frisked' and 'sex'

ct\_frisked\_sex = pd.crosstab(df['sex'], df['frisked'])

# Stacked bar chart

ct\_frisked\_sex.plot(kind='bar', stacked=True, figsize=(8, 5))

plt.title('Frisked Counts by Sex')

plt.xlabel('Sex')

plt.ylabel('Count')

plt.legend(title='Frisked')

plt.show()

# Creating pie charts for 'frisked', 'knifcuti', and 'pf\_wall'

fig, axes = plt.subplots(1, 3, figsize=(18, 6))

# Frisked

axes[0].pie(df['frisked'].value\_counts(), labels=df['frisked'].value\_counts().index, autopct='%1.1f%%', startangle=90)

axes[0].set\_title('Proportion of Frisked Individuals')

# Knife Found

axes[1].pie(df['knifcuti'].value\_counts(), labels=df['knifcuti'].value\_counts().index, autopct='%1.1f%%', startangle=90)

axes[1].set\_title('Proportion of Individuals Found with a Knife')

# Physical Force

axes[2].pie(df['pf\_wall'].value\_counts(), labels=df['pf\_wall'].value\_counts().index, autopct='%1.1f%%', startangle=90)

axes[2].set\_title('Proportion Using Physical Force Against a Wall')

plt.show()

fig, axes = plt.subplots(3, 1, figsize=(10, 18), sharex=True)

# Frisked by Race

sns.countplot(ax=axes[0], x='race', hue='frisked', data=df)

axes[0].set\_title('Frisked by Race')

axes[0].set\_xlabel('Race')

axes[0].set\_ylabel('Count')

axes[0].legend(title='Frisked')

# Knife Found by Race

sns.countplot(ax=axes[1], x='race', hue='knifcuti', data=df)

axes[1].set\_title('Knife Found by Race')

axes[1].set\_xlabel('Race')

axes[1].set\_ylabel('Count')

axes[1].legend(title='Knife Found')

# Physical Force by Race

sns.countplot(ax=axes[2], x='race', hue='pf\_wall', data=df)

axes[2].set\_title('Physical Force Used Against Wall by Race')

axes[2].set\_xlabel('Race')

axes[2].set\_ylabel('Count')

axes[2].legend(title='Physical Force')

plt.show()

**Resulted Plots:**

A graph of a number of people

Description automatically generated with medium confidence

Figure 4: Count of suspects frisked vs sex

* This stacked bar chart visualizes the relationship between sex and the practice of frisking. By displaying how many suspects of each sex were frisked versus not frisked, this chart allows us to see gender patterns within frisking practices. Such insights are essential for identifying any gender biases in policing actions. Such visualization helps in determining if one gender is disproportionately targeted in frisk operations, which is a critical aspect for ensuring fair and unbiased law enforcement practices.

A blue circle with a yellow line

Description automatically generated

Figure 5: Pie charts to count different incidents

* These pie charts offer a visual representation of the proportions of individuals who were frisked, found with a knife, or subjected to physical force against a wall. By displaying the percentage for each outcome, these charts help quickly assess the prevalence of these policing actions. This visualization aids in identifying potential patterns or disparities in policing tactics and informs discussions about police conduct and the need for different interventions in various situations.

A screenshot of a graph

Description automatically generated

Figure 6: Policing actions against different races

* By displaying policing actions (frisked, knife found, physical force) across different races, this chart helps analyze potential biases or trends in how different racial groups are treated during police encounters.

## Explore relationships between attributes. Look at the attributes and then scatter plots, correlation, cross-tabulation, group-wise averages, etc., as appropriate.

We ran python code to explore relationships between the attributes and plotted some visualization with explanation why we used them. We created various correlation, cross-tabulations plots to generate meaningful relationship between the attributes as follows:

import matplotlib.pyplot as plt

import seaborn as sns

# Assuming df is your DataFrame and it has columns 'frisked\_num', 'knifcuti\_num', 'pf\_wall\_num'

correlation\_matrix = df[['frisked\_num', 'knifcuti\_num', 'pf\_wall\_num']].corr()

plt.figure(figsize=(8, 6))

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm')

plt.title('Heatmap of Correlation Between Frisked, Knife Found, and Physical Force')

plt.show()

# Cross-tabulation of 'pf\_wall' and 'knifcuti'

ct\_pfwall\_knife = pd.crosstab(df['pf\_wall'], df['knifcuti'])

# Stacked bar chart

ct\_pfwall\_knife.plot(kind='bar', stacked=True, figsize=(8, 5))

plt.title('Physical Force vs. Knife Found')

plt.xlabel('Physical Force Against a Wall')

plt.ylabel('Count')

plt.legend(title='Knife Found')

plt.show()

# Preparing the data: counting occurrences

data = {

    'Frisked': df['frisked'].value\_counts(),

    'Knife Found': df['knifcuti'].value\_counts(),

    'Physical Force (Wall)': df['pf\_wall'].value\_counts()

}

# Converting dictionary to DataFrame for plotting

plot\_data = pd.DataFrame(data)

# Plotting

plot\_data.plot(kind='bar', figsize=(10, 6))

plt.title('Comparison of Frisked, Knife Found, and Physical Force Use')

plt.xlabel('Category')

plt.ylabel('Count')

plt.xticks(rotation=0)  # Keeps the category labels horizontal

plt.show()

**Resulted relationship plots:**

**A red and blue squares

Description automatically generated**

Figure 7: Correlation between Frisked, Knives found and physical force on wall

* This heatmap provides a visual representation of the correlation coefficients between three policing actions: frisking, finding a knife, and using physical force against a wall. By examining the colors and values in the heatmap, we can assess how strongly these variables are related. For instance, a highly positive correlation between frisking and using physical force might suggest that stops involving frisking are more likely to involve physical confrontations. This analysis is crucial for understanding the interdependencies of different policing actions and can guide policy adjustments or training programs.

**A graph of a blue rectangular object with orange text

Description automatically generated with medium confidence**

Figure 8: Suspects forced to wall vs Knife Found on them

* This visualization uses a stacked bar chart to show the counts of individuals subjected to physical force against a wall based on whether a knife was found with them. This chart can reveal if the use of such force correlates strongly with finding potential weapons like knives. Analyzing these patterns helps in assessing the justification and context of force usage in policing. It is a crucial analysis to understand whether physical measures are preemptively or reactively applied, and if such actions are warranted by actual threats.

A graph of different colored bars

Description automatically generated

Figure 9: Comparing different attributes related to suspect.

* This grouped bar chart allows us to directly compare the frequency of these three key policing actions: frisking, finding a knife, and using physical force against a wall. It is helpful to analyze how common each action is and to explore potential correlations visually, such as whether frisking more frequently results in finding knives or using physical force.

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

To compare the reasons for SQF encounters with the use of force by officers, it is essential to examine stop data in the context of NYPD practices. The data indicates a high incidence of SQF stops among Black and Latinx individuals, with most stops resulting in no evidence of wrongdoing. This pattern suggests that stops may not be consistently based on substantiated suspicion.

Despite a reduction in total stops, the increase in frisks and searches points to a more aggressive approach in the year 2013 as per the given data. The data specifies various natures of the force used, but the trend towards more intrusive measures is clear. The effectiveness of SQF is questionable, given the low rate of arrests and summonses compared to the number of stops, especially when the policy aims to deter crime and seize illegal items.

Outcome from Data Analysis:

* **Main reasons for an SQF:** Race
* **Type of force used by the officer:** Use of physical force against the wall.

To justify this, we utilized a correlation between knife found over suspect vs force against the wall data and python code to visualize it:

df['knifcuti\_num'] = df['knifcuti'].map({'Y': 1, 'N': 0})

df['pf\_wall\_num'] = df['pf\_wall'].map({'Y': 1, 'N': 0})

# Cross-tabulation between 'knifcuti' and 'pf\_wall'

ct\_knifcuti\_pfwall = pd.crosstab(df['knifcuti'], df['pf\_wall'])

# Visualization of the cross-tabulation

ct\_knifcuti\_pfwall.plot(kind='bar', stacked=True, figsize=(8, 6))

plt.title('Use of Physical Force (Against Wall) When a Knife/Cutting Instrument is Found')

plt.xlabel('Was a knife/cutting instrument found?')

plt.ylabel('Number of Stops')

plt.show()

**Outcome:**

**A graph of a blue rectangular object with a orange line

Description automatically generated with medium confidence**

Figure 10: Correlation between Knife found and Forced against wall

Hence, the visualization clearly shows that even though knife was not found on the suspect, the officer still used force against the wall most of the time. This was due to the race of the suspect.

# Conclusion

In conclusion, our study of the NYPD's Stop-Question-and-Frisk (SQF) initiative has revealed several significant findings. Primarily, there are marked racial disparities in the implementation of the SQF policy, with Black and Hispanic individuals being stopped more frequently than their counterparts. Moreover, a considerable proportion of these stops did not result in the recovery of contraband or lead to arrests, calling into question the effectiveness and justifications of the stops.

**Key Findings:**

Racial Disparities: Black and Hispanic populations are disproportionately represented in SQF incidents.

Outcome of Stops: A large majority of stopped individuals were found to be innocent, undermining the premise that higher stop rates contribute to crime reduction.

Use of Force: The application of force, including frisks and searches without finding contraband, suggests a more aggressive approach with questionable grounds.

Effectiveness of SQF: The low correlation between stops and the seizure of illegal items suggests that the initiative may not significantly contribute to public safety.

**Recommendations based on the study's findings are as follows:**

Policy Reassessment: There should be a comprehensive review and reform of the SQF policy to address and correct racial biases.

Enhanced Training: Officers should receive enhanced training focused on bias recognition, de-escalation techniques, and more stringent criteria for initiating stops.

Community Engagement: Initiatives to increase community engagement and improve public relations should be undertaken to rebuild trust.

Implementing these recommendations could lead to a more equitable and effective SQF policy that respects the rights of all New Yorkers while maintaining public safety.