**Advanced Concepts in Data Analytics**

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

**Report 2: Association Rule Mining**

**Report Submitted by:**

Saymon Acharya

Sangram Singh

Gurvinder Singh

Table of Contents

[Abstract/Executive Summary 3](#_Toc164835496)

[Data Preparation 3](#_Toc164835497)

[Construct the required transaction data set for frequent itemset and association rule mining. 3](#_Toc164835498)

[Modelling 4](#_Toc164835499)

[Create frequent itemsets and association rules. 5](#_Toc164835500)

[Use tables and visualizations to help explain your results. 6](#_Toc164835501)

[Evaluation 7](#_Toc164835502)

[What findings are the most interesting? Why? 7](#_Toc164835503)

[Conclusion 7](#_Toc164835504)

# Abstract/Executive Summary

This report delves into the NYPD's Stop-Question-Frisk (SQF) policy using data analysis to unearth patterns of enforcement and potential biases.

By focusing on variables like race, sex, and location, the study processed the SQF data to generate frequent itemsets and association rules, revealing racial disparities and uneven geographic enforcement—despite reforms aimed at reducing discrimination. These findings suggest persistent issues in policy implementation, underscoring the need for nuanced reforms and continuous monitoring to ensure just and fair law enforcement practices.

The report highlights the value of data-driven insights in guiding policy reassessment and advocating for the equitable application of policing strategies across New York City.

# **Data Preparation**

## Construct the required transaction data set for frequent itemset and association rule mining.

We get a transaction data set from the main SQF dataset after implementing a python code as below for frequent itemset and association rule mining. The output gives us a csv file with transaction data set.

import pandas as pd

# Load the dataset (replace 'your\_data.csv' with the path to your uploaded file)

df = pd.read\_excel('data\_report2.xlsx')

# Assuming the dataframe `df` is as per the structure seen in the screenshot:

# Convert 'Y'/'N' columns to boolean True/False

boolean\_columns = ['frisked', 'searched', 'pistol', 'pf\_hcuff']

for col in boolean\_columns:

    df[col] = df[col] == 'Y'

# Create a new dataframe where each row represents a transaction

# and each transaction contains a set of items

transactions = []

for index, row in df.iterrows():

    transaction = set()

    for col in boolean\_columns:

        if row[col]:

            transaction.add(col)

    # Add other categorical attributes as items

    transaction.add(f"race\_{row['race']}")

    transaction.add(f"sex\_{row['sex']}")

    transaction.add(f"crimsusp\_{row['crimsusp']}")

    transaction.add(f"city\_{row['city']}")

    transactions.append(transaction)

# Now, `transactions` is a list of sets, where each set is a transaction

# Example: a CSV where each row is a transaction and items are comma-separated

with open('transactions.csv', 'w') as f:

    for transaction in transactions:

        f.write(','.join(transaction) + '\n')

The transction.csv file thus created will be used for the modelling and evaluation parts of the CRISP DM analysis.

A screenshot of a table

Description automatically generated

Figure 11: Transaction.csv file created

# Modelling

## Create frequent itemsets and association rules.

The code below helps create frequent itemsets and association rules for the above resulted transaction dataset:

from mlxtend.preprocessing import TransactionEncoder

from mlxtend.frequent\_patterns import apriori, association\_rules

import matplotlib.pyplot as plt

import seaborn as sns

# Step 1: One-hot encode the transaction data

te = TransactionEncoder()

te\_ary = te.fit(transactions).transform(transactions)

df = pd.DataFrame(te\_ary, columns=te.columns\_)

# Step 2: Use the apriori algorithm to find frequent itemsets

frequent\_itemsets = apriori(df, min\_support=0.5, use\_colnames=True)  # adjust the min\_support as necessary

# Step 3: Generate association rules

rules = association\_rules(frequent\_itemsets, metric="confidence", min\_threshold=0.7)  # adjust the min\_threshold as necessary

# Step 4: Display tables of frequent itemsets and association rules

print(frequent\_itemsets)

print(rules)

This creates a list of frequent item sets and association with different related values:

A computer screen with white text

Description automatically generated

Figure 12: Resulted Frequent Item Sets

## Use tables and visualizations to help explain your results.

Below are the simple visualizations to help explain our results from above:

A graph with blue rectangular bars

Description automatically generated with medium confidence

Figure 13: Frequent Itemset taken from Transaction dataset

A graph with a blue dot

Description automatically generated

Figure 14: Association rules - Support vs Confidence

This plot graphs with 'Support' on the x-axis and 'Confidence' on the y-axis. The point represents an association rule, and the color of the point reflects the 'Lift' value, as shown by the color scale on the right. The rule depicted in this plot has high confidence (above 0.95), moderate support (around 0.53), and a lift slightly above 1 (indicating a positive correlation between the items in the rule). A lift of 1 means the items are independent of each other.

# Evaluation

## What findings are the most interesting? Why?

The most interesting findings in the report seem to be the preliminary outcomes indicating racial disparities in the Stop-Question-Frisk (SQF) incidents. Despite the initiative undergoing scrutiny and legal challenges for its alleged racial bias, especially against Black and Hispanic individuals, the early results of this study suggest that discriminatory practices may still persist. These findings are particularly notable because they imply that the reforms introduced after the 2013 lawsuit may not have been effective in eliminating biases from NYPD's enforcement of SQF policies.

Another intriguing aspect is the variation in SQF incidents across different precinct locations. This hints at the possibility that enforcement is not uniform across the city and that certain locations may be more prone to intensive policing practices. The influence of location on law enforcement behavior is a critical insight as it can guide more geographically targeted reforms and resource allocation.

Both these points are crucial for understanding the dynamics of SQF initiatives and for informing policy decisions. They underline the need for continuous evaluation and modification of police practices to ensure fairness and legality. The report advocates for ongoing research and monitoring, highlighting the importance of transparency and equity in police procedures, which are fundamental to public trust and the effectiveness of law enforcement agencies.

# Conclusion

In summary, the analysis of the Stop-Question-Frisk data from 2012 highlights enduring racial disparities and geographic inconsistencies in NYPD's practices, despite past reforms. The persistence of these issues underscores the urgent need for policy reassessment and a commitment to ongoing, data-informed adjustments to the SQF initiative. This study serves as a critical reminder of the importance of equitable and just law enforcement practices, and it advocates for sustained scrutiny and improvement of policing strategies to ensure they align with the principles of fairness and legality.