Final Project: New York City Case Study using CRSIP-DM

Report 1: Data and Visualization

Juan J. Holguin

The Southern Alberta Institute of Technology

DATA 475: Advanced Concepts in Data Analytics

Abstract

This paper analyzes the 2012 Stop, Question, and Frisk (SQF) dataset from New York City to evaluate the program's benefits, risks, and overall efficiency. By leveraging statistical methods and data visualization techniques, the study examines detailed records of stops, frisks, and outcomes to provide insights into SQF's impact on crime prevention and public safety. The analysis includes crime rate trends, arrest and seizure data, and demographic patterns to assess the program's effectiveness and potential biases. Key metrics such as stop-to-arrest ratios, geographic distribution of stops, and temporal patterns are explored to understand the operational dynamics of SQF. The findings aim to offer a data-driven perspective on the program's efficacy, highlighting areas for improvement and informing evidence-based policy decisions. This paper seeks to provide a comprehensive and objective evaluation of SQF, emphasizing the importance of data analytics in shaping effective and equitable policing strategies.

Keywords: Data analytics, crime prevention, statistical methods, public safety.

Report 1: Data and Visualization

Introduction

In this paper, we utilize the CRISP-DM (Cross-Industry Standard Process for Data Mining) model to analyze the 2012 Stop, Question, and Frisk (SQF) dataset from New York City. The CRISP-DM model provides a structured approach to data mining, encompassing six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Our analysis begins with the 5W 1H framework to establish a comprehensive understanding of the dataset. By addressing these questions, we aim to uncover insights into the benefits, risks, and efficiency of the SQF program, providing a data-driven perspective that informs evidence-based policy decisions.

Business Understanding

1st W - WHO

#	Question	Answer
Q1.1	Who is involved?	Police officers, community, individuals stopped
Q1.2	Who is affected?	Individuals stopped
Q1.3	Who will benefit?	Community
Q1.4	Who will be harmed?	Innocent people

2nd W - WHAT

#	Question	Answer
Q2.1	What is your topic?	Program's impact on crime rates, particularly violent crime, and its compliance with constitutional standards

Q2.2	What does your topic involve?	Type of crimes, demographics, crime prevention, public safety, potential biases
Q2.3	What is it similar to / different from?	Countries: United Kingdom, Spain, Bulgaria, Hungary, USA: Philadelphia and Los Angeles. (Weisburd, 2023)
Q1.4	What might be affected/changed by your topic?	Laws and regulation changes, police procedures, crime prevention strategies

3rd W - WHEN

#	Question	Answer
Q3.1	When does / did /will it / should take place?	This analysis will be based on 2012 dataset. Trends and forecasts will be compared to historical data
Q3.2	Does when this takes place affect the topic?	This analysis is based on historical data. The model might be extrapolated and adapted to current data.

4th W - WHERE

#	Question	Answer
Q4.1	Where does this take place?	This analysis will be based on New York City dataset.
Q4.2	Does it matter where it takes place?	The model will be specific for New York City data. However, it might be extrapolated to other cities.

5th W - WHY

#	Question	Answer
Q5.1	Why is this topic important/matter?	Understand the effectiveness and shortcomings, address concerns about racial profiling and civil liberties, refine crime prevention strategies.

Q5.2	Why do certain things happen?	Major risk is the false positives and the legal, ethical and public implications

1st H - HOW

#	Question	Answer
Q6.1	How does this topic work/ function/ what does it do?	Proactive policing strategy based on Police officers' observation, authority to stop, gather information and frisk. Data can be collected through police reports, legal databases, and community outreach programs.
Q6.2	How did it come to be?	The program peaked in 2011 with nearly 700,000 stops. However, it faced significant controversy and legal challenges due to concerns about racial profiling and civil. (Gelman, 2012)
Q6.3	How are those involved affected?	Community with a reduction in crime rates. Police officers balancing proactive policing with respecting civil liberties. Individuals being stopped for no reason.

Data Understanding

Data dictionary

#	Variable	Label
1	year	YEAR OF STOP
2	pct	PRECINCT OF STOP
3	ser_num	UF250 SERIAL NUMBER
4	datestop	DATE OF STOP (MM-DD-YYYY)
5	timestop	TIME OF STOP (HH:MM)
6	recstat	RECORD STATUS
7	inout	WAS STOP INSIDE OR OUTSIDE ?
8	trhsloc	WAS LOCATION HOUSING OR TRANSIT AUTHORITY ?
9	perobs	PERIOD OF OBSERVATION (MMM)
10	crimsusp	CRIME SUSPECTED
11	perstop	PERIOD OF STOP (MMM)
12	typeofid	STOPPED PERSON'S IDENTIFICATION TYPE
13	explnstp	DID OFFICER EXPLAIN REASON FOR STOP ?
14	othpers	WERE OTHER PERSONS STOPPED, QUESTIONED OR FRISKED ?
15	arstmade	WAS AN ARREST MADE ?

16	arstoffn	OFFENSE SUSPECT ARRESTED FOR
17	sumissue	WAS A SUMMONS ISSUED ?
18	sumoffen	OFFENSE SUSPECT WAS SUMMONSED FOR
19	compyear	COMPLAINT YEAR (IF COMPLAINT REPORT PREPARED)
20	comppct	COMPLAINT PRECINCT (IF COMPLAINT REPORT PREPARED)
21	offunif	WAS OFFICER IN UNIFORM ?
22	officrid	ID CARD PROVIDED BY OFFICER (IF NOT IN UNIFORM)
23	frisked	WAS SUSPECT FRISKED ?
24	searched	WAS SUSPECT SEARCHED ?
25	contrabn	WAS CONTRABAND FOUND ON SUSPECT ?
26	adtlrept	WERE ADDITIONAL REPORTS PREPARED ?
27	pistol	WAS A PISTOL FOUND ON SUSPECT ?
28	riflshot	WAS A RIFLE FOUND ON SUSPECT ?
29	asltweap	WAS AN ASSAULT WEAPON FOUND ON SUSPECT ?
30	knifcuti	WAS A KNIFE OR CUTTING INSTRUMENT FOUND ON SUSPECT ?
31	machgun	WAS A MACHINE GUN FOUND ON SUSPECT ?
32	othrweap	WAS ANOTHER TYPE OF WEAPON FOUND ON SUSPECT
33	pf_hands	PHYSICAL FORCE USED BY OFFICER - HANDS
34	pf_wall	PHYSICAL FORCE USED BY OFFICER - SUSPECT AGAINST WALL
35	pf_grnd	PHYSICAL FORCE USED BY OFFICER - SUSPECT ON GROUND
36	pf_drwep	PHYSICAL FORCE USED BY OFFICER - WEAPON DRAWN
37	pf_ptwep	PHYSICAL FORCE USED BY OFFICER - WEAPON POINTED
38	pf_baton	PHYSICAL FORCE USED BY OFFICER - BATON
39	pf_hcuff	PHYSICAL FORCE USED BY OFFICER - HANDCUFFS
40	pf_pepsp	PHYSICAL FORCE USED BY OFFICER - PEPPER SPRAY
41	pf_other	PHYSICAL FORCE USED BY OFFICER - OTHER
42	radio	RADIO RUN
43	ac_rept	ADDITIONAL CIRCUMSTANCES - REPORT BY VICTIM/WITNESS/OFFICER
44	ac_inves	ADDITIONAL CIRCUMSTANCES - ONGOING INVESTIGATION
45	rf_vcrim	REASON FOR FRISK - VIOLENT CRIME SUSPECTED
46	rf_othsw	REASON FOR FRISK - OTHER SUSPICION OF WEAPONS
47	ac_proxm	ADDITIONAL CIRCUMSTANCES - PROXIMITY TO SCENE OF OFFENSE
48	rf_attir	REASON FOR FRISK - INAPPROPRIATE ATTIRE FOR SEASON
49	cs_objcs	REASON FOR STOP - CARRYING SUSPICIOUS OBJECT
50	cs_descr	REASON FOR STOP - FITS A RELEVANT DESCRIPTION
51	cs_casng	REASON FOR STOP - CASING A VICTIM OR LOCATION
52	cs_lkout	REASON FOR STOP - SUSPECT ACTING AS A LOOKOUT
53	rf_vcact	REASON FOR FRISK- ACTIONS OF ENGAGING IN A VIOLENT CRIME
54	cs_cloth	REASON FOR STOP - WEARING CLOTHES COMMONLY USED IN A CRIME
55	cs_drgtr	REASON FOR STOP - ACTIONS INDICATIVE OF A DRUG TRANSACTION
56	ac_evasv	ADDITIONAL CIRCUMSTANCES - EVASIVE RESPONSE TO QUESTIONING
57	ac_assoc	ADDITIONAL CIRCUMSTANCES - ASSOCIATING WITH KNOWN CRIMINALS
58	cs_furtv	REASON FOR STOP - FURTIVE MOVEMENTS

59	rf_rfcmp	REASON FOR FRISK - REFUSE TO COMPLY W OFFICER'S DIRECTIONS
60	ac_cgdir	ADDITIONAL CIRCUMSTANCES - CHANGE DIRECTION AT SIGHT OF OFFICER
61	rf_verbl	REASON FOR FRISK - VERBAL THREATS BY SUSPECT
62	cs_vcrim	REASON FOR STOP - ACTIONS OF ENGAGING IN A VIOLENT CRIME
63	cs_bulge	REASON FOR STOP - SUSPICIOUS BULGE
64	cs_other	REASON FOR STOP - OTHER
65	ac_incid	ADDITIONAL CIRCUMSTANCES - AREA HAS HIGH CRIME INCIDENCE
66	ac_time	ADDITIONAL CIRCUMSTANCES - TIME OF DAY FITS CRIME INCIDENCE
67	rf_knowl	REASON FOR FRISK - KNOWLEDGE OF SUSPECT'S PRIOR CRIM BEHAV
68	ac_stsnd	ADDITIONAL CIRCUMSTANCES - SIGHTS OR SOUNDS OF CRIMINAL ACTIVITY
69	ac_other	ADDITIONAL CIRCUMSTANCES - OTHER
70	sb_hdobj	BASIS OF SEARCH - HARD OBJECT
71	sb_outln	BASIS OF SEARCH - OUTLINE OF WEAPON
72	sb_admis	BASIS OF SEARCH - ADMISSION BY SUSPECT
73	sb_other	BASIS OF SEARCH - OTHER
74	repcmd	REPORTING OFFICER'S COMMAND (1 TO 999)
75	revcmd	REVIEWING OFFICER'S COMMAND (1 TO 999)
76	rf_furt	REASON FOR FRISK - FURTIVE MOVEMENTS
77	rf_bulg	REASON FOR FRISK - SUSPICIOUS BULGE
78	offverb	VERBAL STATEMENT PROVIDED BY OFFICER (IF NOT IN UNIFORM)
79	offshld	SHIELD PROVIDED BY OFFICER (IF NOT IN UNIFORM)
80	forceuse	REASON FORCE USED
81	sex	SUSPECT'S SEX
82	race	SUSPECT'S RACE
83	dob	SUSPECT'S DATE OF BIRTH (CCYY-MM-DD)
84	age	SUSPECT'S AGE
85	ht_feet	SUSPECT'S HEIGHT (FEET)
86	ht_inch	SUSPECT'S HEIGHT (INCHES)
87	weight	SUSPECT'S WEIGHT
88	haircolr	SUSPECT'S HAIRCOLOR
89	eyecolor	SUSPECT'S EYE COLOR
90	build	SUSPECT'S BUILD
91	othfeatr	SUSPECT'S OTHER FEATURES (SCARS, TATOOS ETC.)
92	addrtyp	LOCATION OF STOP ADDRESS TYPE
93	rescode	LOCATION OF STOP RESIDENT CODE
94	premtype	LOCATION OF STOP PREMISE TYPE
95	premname	LOCATION OF STOP PREMISE NAME

Data quality

For data quality the following steps were followed:

```
# %% read data
import pandas as pd

crimes = pd.read_csv(
    u"C:\\Users\\juanj\\Documents\\Python\\Final\\sqf-2012-csv.zip"
)

num_rows = crimes.shape[0]
num_rows
```

R: 532911

```
# Check for missing values
missing_values = crimes.isnull().sum()

# Separate numerical and non-numerical columns
numerical_cols = crimes.select_dtypes(include=['number']).columns
non_numerical_cols = crimes.select_dtypes(exclude=['number']).columns

# Handle missing values
# Fill numerical columns with mean
crimes[numerical_cols] = crimes[numerical_cols].fillna(crimes[numerical_cols].mean())

# Fill non-numerical columns with mode
for col in non_numerical_cols:
    if not crimes[col].mode().empty:
        mode_value = crimes[col].mode()[0]
        crimes[col] = crimes[col].fillna(mode_value)

# Check for duplicate data
duplicates = crimes.duplicated().sum()

# Remove duplicates
crimes = crimes.drop_duplicates()

num_rows = crimes.shape[0]
num_rows
```

R: 532911

No duplicates found and data quality revised.

Empty columns drop:

```
# Drop empty columns
crimes.drop('compyear', axis=1, inplace=True)
crimes.drop('comppct', axis=1, inplace=True)
```

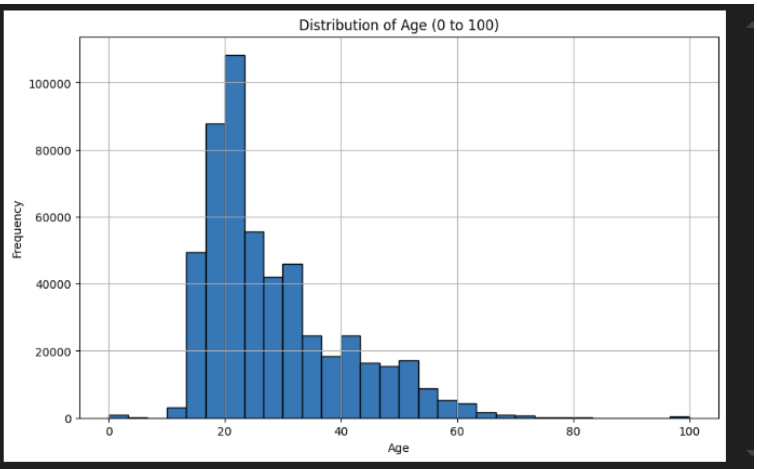
Column with errors drop:

```
# Calculate the difference between '2012-12-31' and the 'dob' column
reference_date = pd.to_datetime('2012-12-31')
crimes['calc_age'] = (reference_date - crimes['dob']).dt.days // 365
# print(crimes[['dob', 'calc_age']].head())

# Count the number of rows where 'age' is over 100
count_over_100 = crimes[crimes['calc_age'] > 100].shape[0]

#50% of dob wrong. drop column
crimes.drop('dob', axis=1, inplace=True)
crimes.drop('calc_age', axis=1, inplace=True)
```

```
✓ # dob period ···

Number of rows where age is over 100: 214252
```

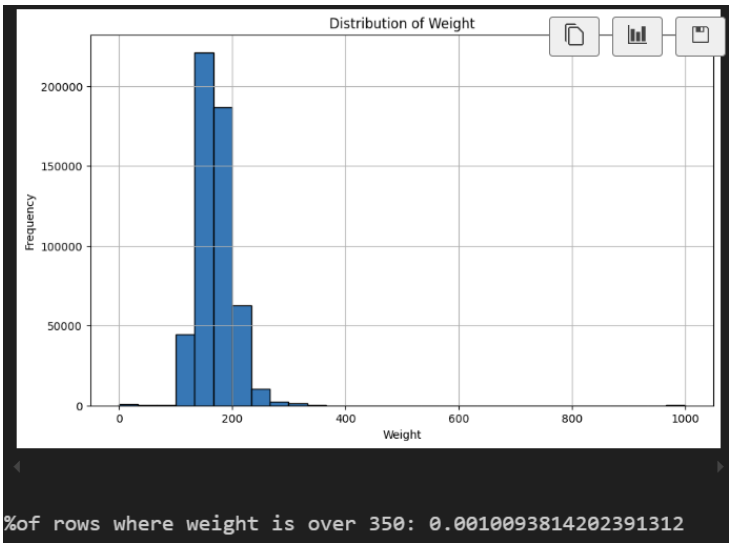
Outliers in age & weight:

```
count_outside_100 = crimes[(crimes['age'] < 0) | (crimes['age'] > 100)].shape[0]
print(f"% of rows where age is outside the range of 0 to 100: {count_outside_100/num_rows}")

# Drop outliers
crimes = crimes[(crimes['age'] >= 0) & (crimes['age'] <= 100)]
num_rows = crimes.shape[0]
num_rows
```




```
% of rows where age is outside the range of 0 to 100: 0.00169
```

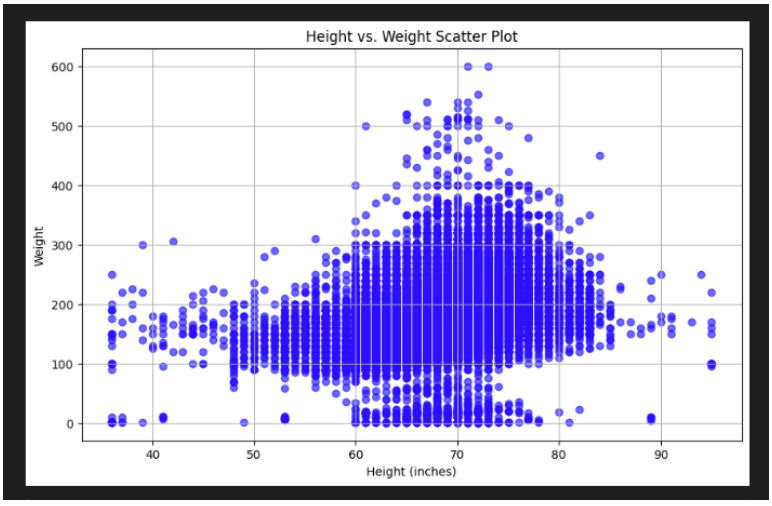
```
count_over_350 = crimes[crimes['weight'] > 350].shape[0]
print(f"%of rows where weight is over 350: {count_over_350/num_rows}")

# Drop outliers
crimes = crimes[(crimes['weight'] <= 350)]
num_rows = crimes.shape[0]
num_rows
```



Rows count after numeric outliers removed: 531472

Data visualization & insights

Weight – height correlation



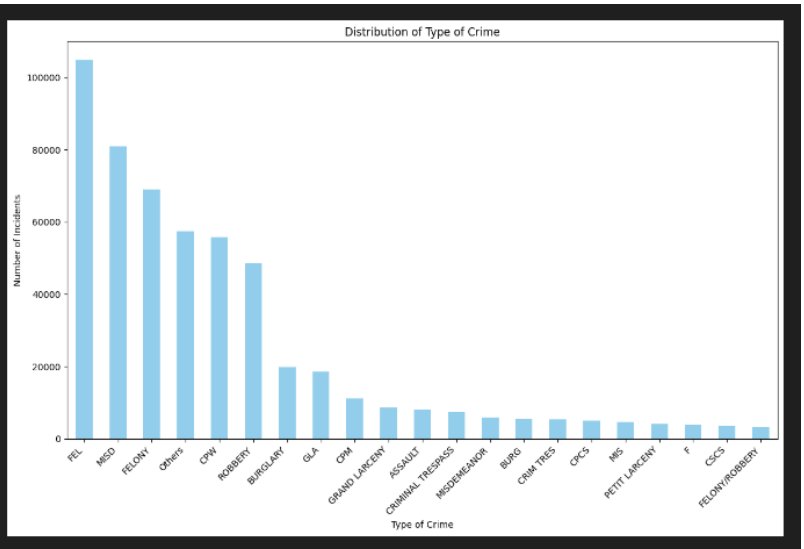
```
correlation = crimes['weight'].corr(crimes['ht_inches'])

print(f"The correlation between weight and height is: {correlation:.2f}")
```

R: 0.45

Nonnumerical features:

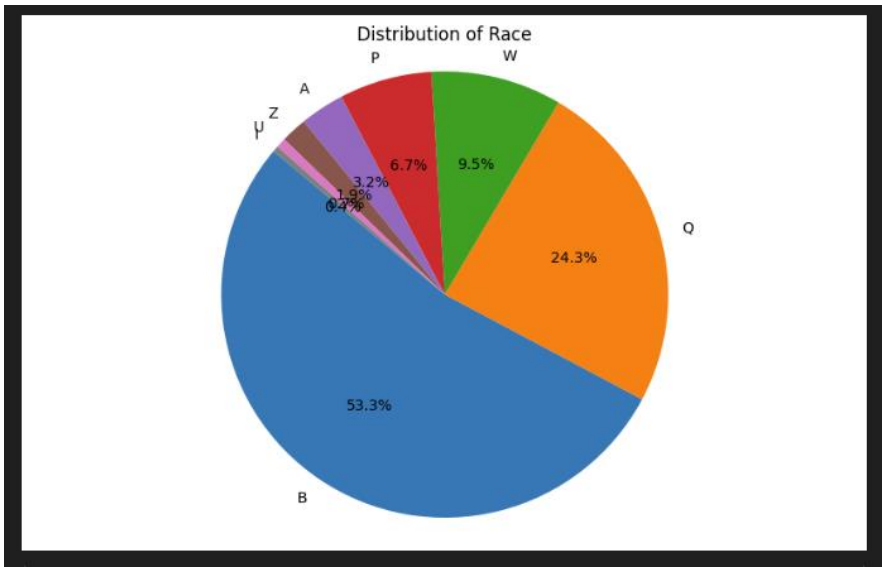
Change the type of crimes to ‘Others’ for crimes with a small count, given there are 6,806 different crime options. Types reduced to 20 + others.

```
# %% non numerical review
import numpy as np

# Set the threshold for grouping smaller categories
threshold = 2500  # to fit to 20 types

# Compute the value counts for the 'crimsusp' column
value_counts = crimes['crimsusp'].value_counts()

# Create a mask to identify categories that meet the threshold
mask = crimes['crimsusp'].isin(value_counts[value_counts >= threshold].index)

# Use numpy.where to efficiently create the 'crimsusp_grouped' column
crimes['crimsusp_grouped'] = np.where(mask, crimes['crimsusp'], 'Others')

# Display the value counts of the new grouped column
grouped_counts = crimes['crimsusp_grouped'].value_counts()
print(grouped_counts)
```



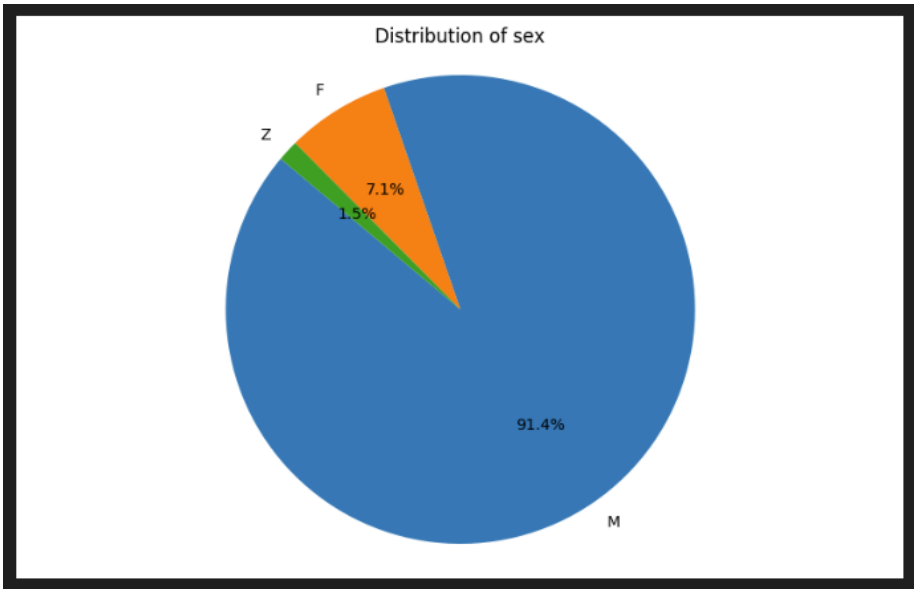

```
Number of unique types in 'crimsusp_grouped': 21
```

```
race_count = crimes['race'].value_counts()
```



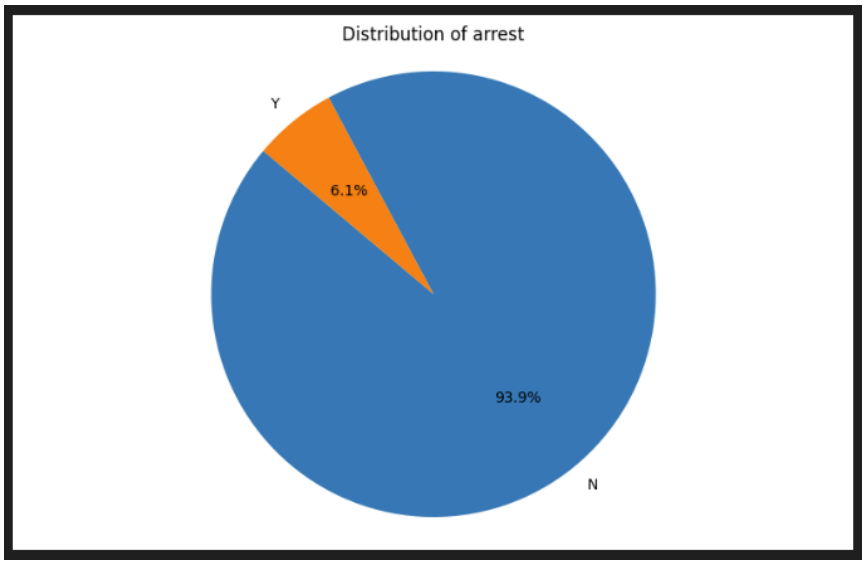
LEGEND		Qty
A	Asian/Pacific Islander	17025
B	Black	283498
I	American Indian/Alaskan native	2255
P	Black-Hispanic	35702
Q	White-Hispanic	129127
W	White	50248
X	Unknown	3756
Z	Other	10042

```
sex_count = crimes['sex'].value_counts()
```



	LEGEND	Qty
Y	Yes	32271
N	No	499382

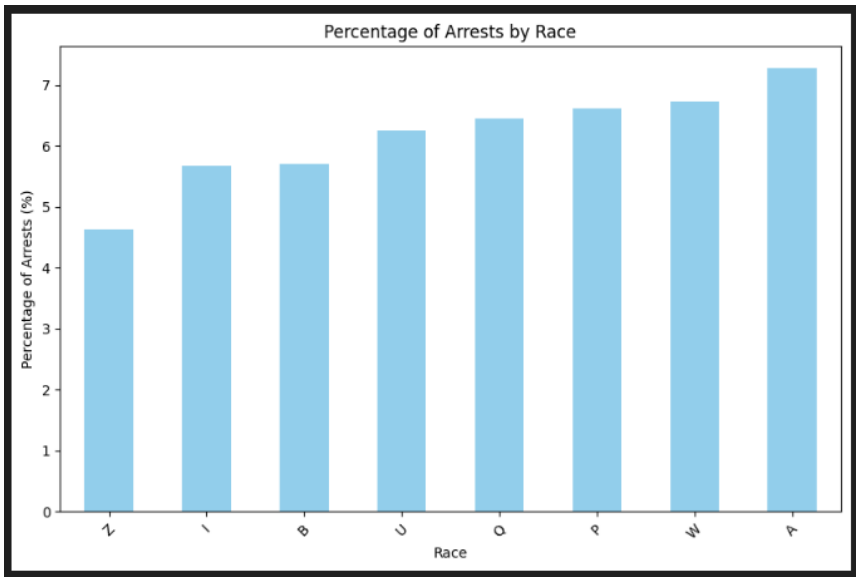
```
arrest_count = crimes['arstmade'].value_counts()
```



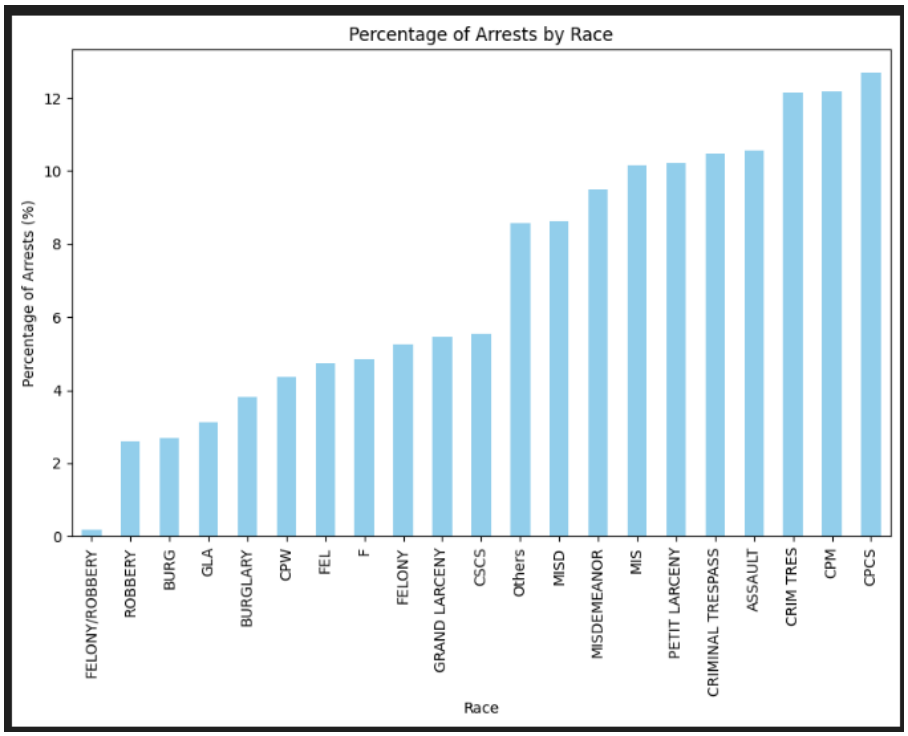
The next step is to browse for relationships between attributes that might help optimize a prediction model and clarify the business questions.

Relation between race and # of arrests:

```
# Calculate the total number of arrests per race
arrests_per_race = crimes[crimes['arstmade'] == 'Y']['race'].value_counts()

# Calculate the percentage of arrests per race
percentage_arrests_per_race = (arrests_per_race / race_count) * 100
percentage_arrests_per_race
```



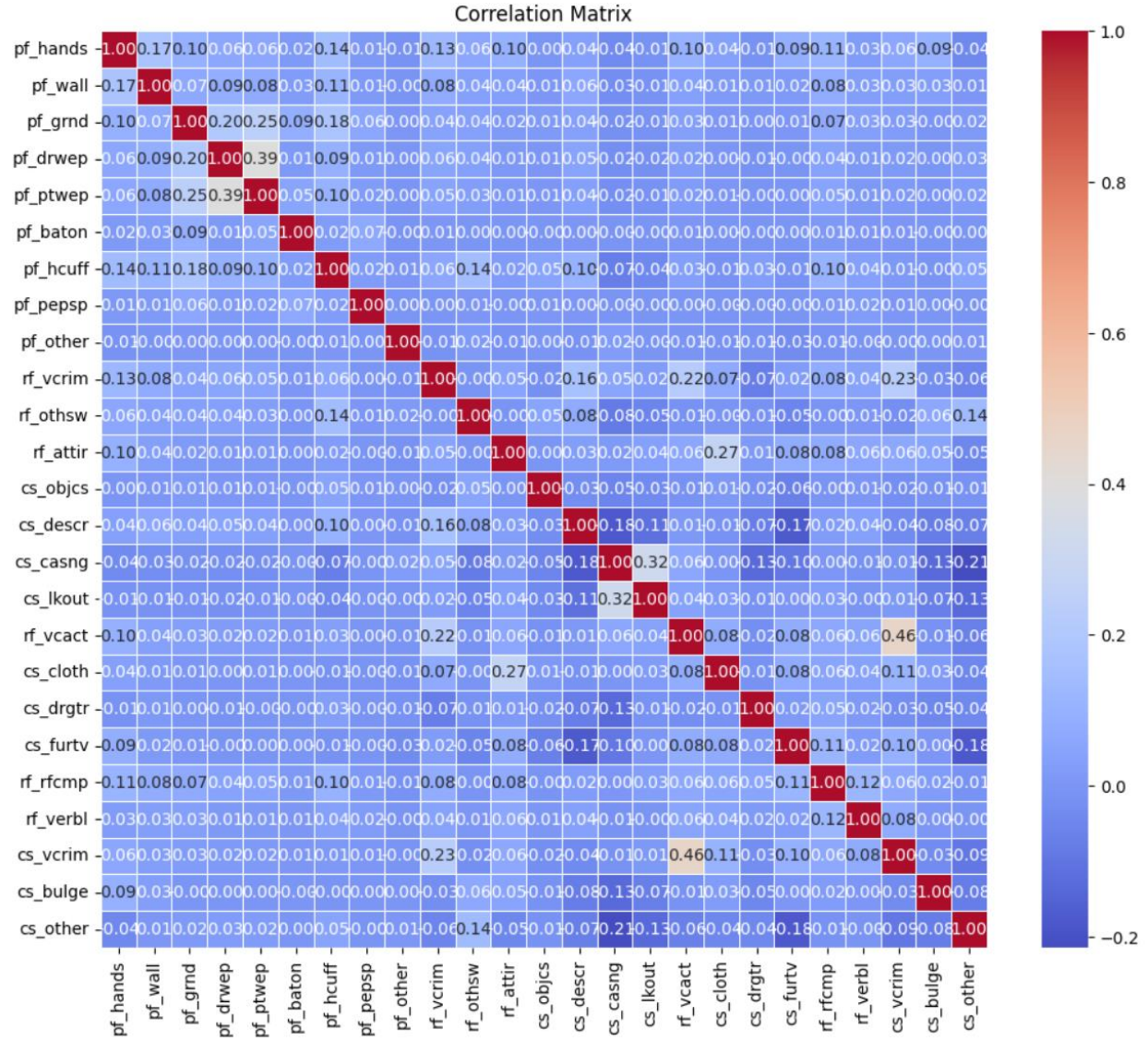

```
arrests_per_crime = crimes[crimes['arstmade'] == 'Y']['crimsusp_grouped'].value_counts()

# Calculate the percentage of arrests per race
percentage_arrests_per_crime = (arrests_per_crime / grouped_counts) * 100
percentage_arrests_per_crime
```



Based on this chart, a further analysis can be performed for correlations between attributes and the higher rate of actual arrests.

Finally, to analyze the reasons for an SQF and what type of force was used by the officer, a correlation matrix isolating these attributes can give important information.

```
# Convert categorical values to numeric
numeric_subset = subset.apply(lambda x: x.replace({'Y': 1, 'N': 0, 'N/A': np.nan}) if x.dtype == 'object' else x)

# Impute missing values with column mean
numeric_subset = numeric_subset.apply(lambda x: x.fillna(x.mean()) if x.dtype == 'float64' else x)

# Calculate the correlation matrix
correlation_matrix = numeric_subset.corr()
```



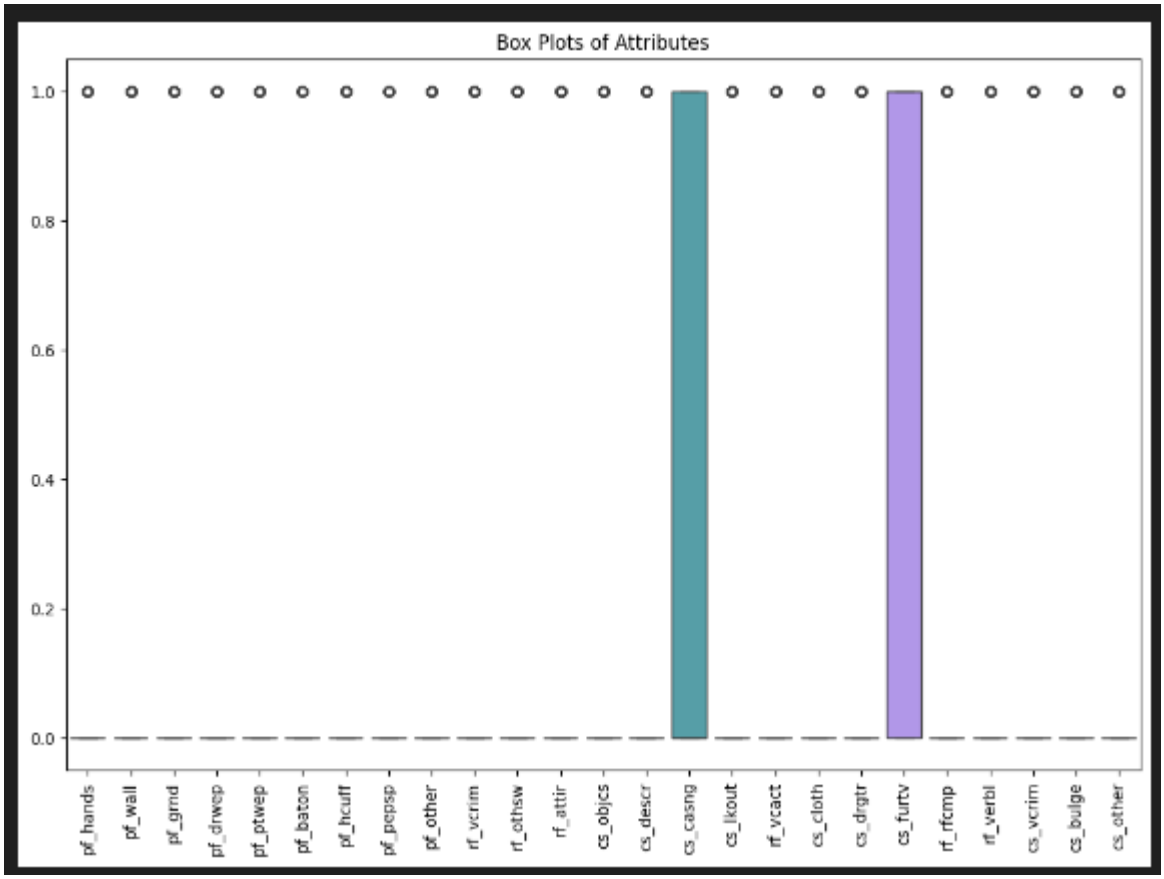
Higher correlations founded:

cs_vcrim – rf_vcact	0.46	both related to violent crime
pf_ptwep – pf_ptdrwep	0.39	both related to officer taking out gun

highest relation between officer action and reason for SQF

rf_othsw – pf_hcuff	0.14	suspicion of weapon with handcuffs

```
# Box plots for each attribute
plt.figure(figsize=(12, 8))
sns.boxplot(data=numeric_subset)
plt.title('Box Plots of Attributes')
plt.xticks(rotation=90)
plt.show()
```



Casing a victim and a furtive movement are the only two attributes with a median of 1 (True).

References

Gelman A, Fagan J, Kiss A (2012). “Stop-and-frisk policy in the context of claims of racial bias.” Journal of the American Statistical Association. 102(479): 813 - 823.

New York (N.Y.). Police Department. New York Police Department (NYPD) Stop, Question, and Frisk Database, 2006. Inter-university Consortium for Political and Social Research [distributor], 2008-06-09. https://doi.org/10.3886/ICPSR21660.v1

Weisburd D, Petersen K, Fay S. Does Scientific Evidence Support the Widespread Use of SQFs as a Proactive Policing Strategy?, *Policing: A Journal of Policy and Practice*, Volume 17, 2023, paac098, https://doi.org/10.1093/police/paac098URLURLURL.