Design and Implementation of Analytics System

Topic: **Data Analytics on general attributes and predictive analytics to Stop and Frisk attributes of the New York Police Department ‘s Stop, Question, and Frisk practices.**

THE MINH TRAN

TABLE OF CONTENTS

General Guidelines

1. Introduction to Design and Implementation of Analytics Project

Design and Implementation of Analytics System is a Big Data Analytics project that utilizes Big and Real Life datasets of The New York Police Department (NYPD)'s Stop, Question and Frisk database between 2007 and 2016. This project accomplished to analyze the datasets on many attributes through the analyzing tools of analytics processes by many analysis software and technical progresses. In general, this project focuses on the analytics of attributes, relationships between attributes by years and by frisked target, as well as using chosen attributes to make final predictive decision depend on targeted attribute of dataset of 10 years, and of each year by the use of modeling methods and machine learning algorithms to see analytics, demonstrations, evaluations, comparisons, visualization, and predictions of moving forward interested purposes. Among analyzing features simulate modeling software techniques, we could approach to data analytics to materialize easier tasks that run well and be faster with combining, cleaning, storing and having good management on Big dataset with around 4,000,000 rows and 112 attributes. Thus, database analysts can see, understand, and make decision from the insight of datasets, figure out optimal solution, and predict problems can occur in future.

What is Stop, Question, Frisk?

As we know, every day in cities in the U.S, police officers can stop, question, and frisk people by their patrol duties. Police officers can stop anyone and then ask them some questions if police finalize that someone has been suspected in criminal problems in using weapons, having crime activities in any place and any time. Upon the Big and Real life database of the NYPD in New York City, we would consider and analyze attributes on this dataset such as race, stop, frisk, searched, criminal status, sex, city attributes, etc. to understand the juncture of crime and to make modeling decision, predict result of near future, and give out experience activities to police offices.

The Data Analytics model figure designed by the author (**Fig 1**) indicates rectangle progresses in the most necessary relationships of tasks in this project. The diagram possesses the important steps initial data process involve collection from raw data to structured data, cleaned data, and then continuing works are Exploratory Data Analysis (EDA), Modeling Analytics, Data driven making, Report-Visualization, and making decisions process. In summary, all these data analytics tasks are evident for any dataset which need to analyze on any campaign of the users depending on their own expectations.

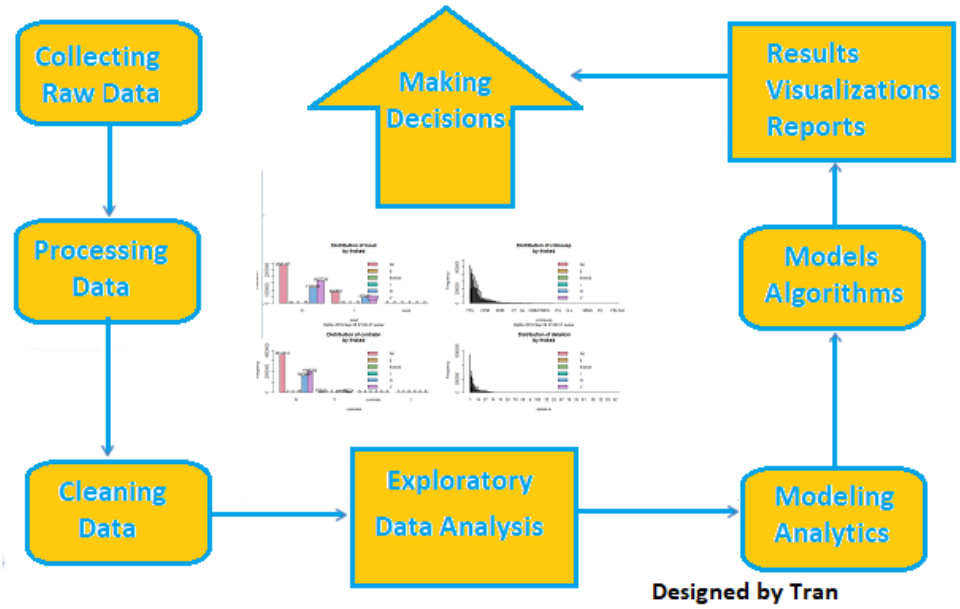


Fig. 1 Data Analytics process.

According to the expectation and intention of Design and Implementation of Analytics System, we need following the diagram above that includes eight phases to move forward to each function of each phases. Each phase has different duties and has deeply relationships due to they are leading hooks in a binding series of throughout data analytics treatment. It cannot waive any phase to go along to make decision from final results or figure out optimal solutions. Upon the success of a phase then arriving next phase so all of them are important and dependent together but the most important is modeling analytics data because overall it is the phase to make final results or optimal decision, hence we need to choose the best modeling method for our own targets since we authorized the parameters of whole analytics system.

2. Collecting, Combining, Cleaning, and Storing Dataset

2.1 Collecting

My capstone project topic is: **Data Analytics on general attributes and predictive analytics to Stop, Question and Frisk attributes of the New York Police Department (NYPD) dataset based on Stop, Question, and Frisk practices.**

Download the dataset:

Dataset: We would use the Big Data dataset of The NYPD Stop Question Frisk Database from 2007 to 2016 involving 112 columns and 3,686,101 rows of the text file size of 1.7GB which have taken at the website: https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

Purpose of this project:

Generally, this project focus on the analytics of attribute in comparing between attributes, relationship of them to take out the results. Moreover, this project accomplishes predictive analytics on the targeted attribute of **the dataset to take decision to Stop, Question and Frisk**.

Base on the dataset, we have initial processes to handle them in the special steps cleaning, simpler analyzing by the progress below as:

The Big Data of 10 csv files of 10 years between 2007 and 2016 with the following their volume:

YEAR	NUMBER OF ROW	VOLUME
2007	**472,000**	**149MB**
2008	**540,000**	**169MB**
2009	**600,000**	**186MB**
2010	**601,286**	**189MB**
2011	**686,000**	**226MB**
2012	**533,000**	**168MB**
2013	**192,000**	**59MB**
2014	**45,800**	**14MB**
2015	**22,600**	**7MB**
2016	**12,400**	**4MB**
TOTAL	**3,686,101**	**1.171GB**

2.2 Combining: Having the Big Dataset in one *.csv file that is good preparing of the raw dataset. It is hard to combine 10 large files in a csv file by Microsoft Excel due to the maximum storing size in Excel is only to have 2^20 = 1,048,576 rows and there are not many ways to do that. Fortunately, in another way, we can use Command Prompt tool in MS-DOS by Microsoft along with the typing the command of "cmd" tool through this simple commands on Start tab of Windows 10 to combine many large csv files in one file (**Fig 2**) as follows:

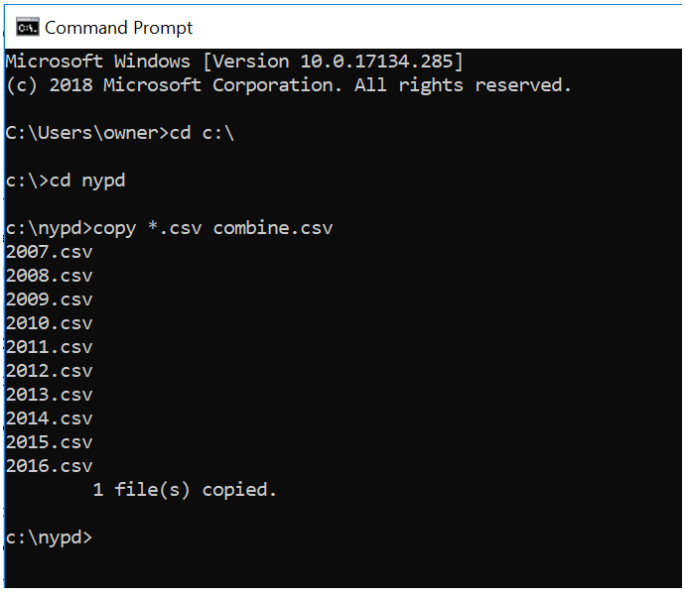


Fig 2: Combining dataset process in Command Prompt

2.3 Storing and Understanding: The dataset comprises 112 attributes with the detail of each that were described in the following table (**Fig 3**):

NYPD Stop Question Frisk Database 2010

File specifications

Variable	Position	Label	Measurement Level	Column Width	Alignment	Print Format	Write Format
year	1	YEAR OF STOP (CCYY)	Nominal	5	Left	A5	A5
pct	2	PRECINCT OF STOP (FROM 1 TO 123)	Nominal	4	Left	A4	A4
ser_num	3	UF250 SERIAL NUMBER	Nominal	7	Left	A7	A7
datestop	4	DATE OF STOP (MM-DD-YYYY)	Nominal	8	Right	F8	F8
timestop	5	TIME OF STOP (HH:MM)	Scale	8	Left	A5	A5
recstat	6	RECORD STATUS	Nominal	7	Left	A1	A1
inout	7	WAS STOP INSIDE OR OUTSIDE ?	Nominal	5	Left	A1	A1
trhsloc	8	WAS LOCATION HOUSING OR TRANSIT AUTHORITY ?	Nominal	7	Left	A1	A1
perobs	9	PERIOD OF OBSERVATION (MMM)	Nominal	7	Right	F7.2	F7.2
crimsusp	10	CRIME SUSPECTED	Scale	24	Left	A30	A30
perstop	11	PERIOD OF STOP (MMM)	Nominal	7	Right	F4	F4
typeofid	12	STOPPED PERSON'S IDENTIFICATION TYPE	Scale	8	Left	A1	A1
explnstp	13	DID OFFICER EXPLAIN REASON FOR STOP ?	Nominal	8	Left	A1	A1
othpers	14	WERE OTHER PERSONS STOPPED, QUESTIONED OR FRISKED ?	Nominal	7	Left	A1	A1
arstmade	15	WAS AN ARREST MADE ?	Nominal	8	Left	A1	A1
arstoffn	16	OFFENSE SUSPECT ARRESTED FOR	Nominal	24	Left	A30	A30
sumissue	17	WAS A SUMMONS ISSUED ?	Nominal	8	Left	A1	A1
sumoffen	18	OFFENSE SUSPECT WAS SUMMONSED FOR	Nominal	24	Left	A30	A30
compyear	19	COMPLAINT YEAR (IF COMPLAINT REPORT PREPARED)	Nominal	8	Right	F5	F5
comppct	20	COMPLAINT PRECINCT (IF COMPLAINT REPORT PREPARED)	Scale	7	Right	F4	F4
offunif	21	WAS OFFICER IN UNIFORM ?	Nominal	7	Left	A1	A1
officrid	22	ID CARD PROVIDED BY OFFICER (IF NOT IN UNIFORM)	Nominal	8	Left	A1	A1
frisked	23	WAS SUSPECT FRISKED ?	Nominal	7	Left	A1	A1
searched	24	WAS SUSPECT SEARCHED ?	Nominal	8	Left	A1	A1
contrabn	25	WAS CONTRABAND FOUND ON SUSPECT ?	Nominal	8	Left	A1	A1
adtlrept	26	WERE ADDITIONAL REPORTS PREPARED ?	Nominal	8	Left	A1	A1
pistol	27	WAS A PISTOL FOUND ON SUSPECT ?	Nominal	6	Left	A1	A1
riflshot	28	WAS A RIFLE FOUND ON SUSPECT ?	Nominal	8	Left	A1	A1
asltweap	29	WAS AN ASSAULT WEAPON FOUND ON SUSPECT ?	Nominal	8	Left	A1	A1
knifcuti	30	WAS A KNIFE OR CUTTING INSTRUMENT FOUND ON SUSPECT ?	Nominal	8	Left	A1	A1
machgun	31	WAS A MACHINE GUN FOUND ON SUSPECT ?	Nominal	7	Left	A1	A1
othrweap	32	WAS ANOTHER TYPE OF WEAPON FOUND ON SUSPECT	Nominal	8	Left	A1	A1
pf_hands	33	PHYSICAL FORCE USED BY OFFICER - HANDS	Nominal	8	Left	A1	A1
pf_wall	34	PHYSICAL FORCE USED BY OFFICER - SUSPECT AGAINST WALL	Nominal	7	Left	A1	A1
pf_grnd	35	PHYSICAL FORCE USED BY OFFICER - SUSPECT ON GROUND	Nominal	7	Left	A1	A1
pf_drwep	36	PHYSICAL FORCE USED BY OFFICER - WEAPON DRAWN	Nominal	8	Left	A1	A1
pf_ptwep	37	PHYSICAL FORCE USED BY OFFICER - WEAPON POINTED	Nominal	8	Left	A1	A1
pf_baton	38	PHYSICAL FORCE USED BY OFFICER - BATON	Nominal	8	Left	A1	A1
pf_hcuff	39	PHYSICAL FORCE USED BY OFFICER - HANDCUFFS	Nominal	8	Left	A1	A1
pf_pepsp	40	PHYSICAL FORCE USED BY OFFICER - PEPPER SPRAY	Nominal	8	Left	A1	A1
pf_other	41	PHYSICAL FORCE USED BY OFFICER - OTHER	Nominal	8	Left	A1	A1
radio	42	RADIO RUN	Nominal	5	Left	A1	A1
ac_rept	43	ADDITIONAL CIRCUMSTANCES - REPORT BY VICTIM/WITNESS/OFFICER	Nominal	7	Left	A1	A1
ac_inves	44	ADDITIONAL CIRCUMSTANCES - ONGOING INVESTIGATION	Nominal	8	Left	A1	A1
rf_vcrim	45	REASON FOR FRISK - VIOLENT CRIME SUSPECTED	Nominal	8	Left	A1	A1
rf_othsw	46	REASON FOR FRISK - OTHER SUSPICION OF WEAPONS	Nominal	8	Left	A1	A1
ac_proxm	47	ADDITIONAL CIRCUMSTANCES - PROXIMITY TO SCENE OF OFFENSE	Nominal	8	Left	A1	A1
rf_attir	48	REASON FOR FRISK - INAPPROPRIATE ATTIRE FOR SEASON	Nominal	8	Left	A1	A1
cs_objcs	49	REASON FOR STOP - CARRYING SUSPICIOUS OBJECT	Nominal	8	Left	A1	A1
cs_descr	50	REASON FOR STOP - FITS A RELEVANT DESCRIPTION	Nominal	8	Left	A1	A1
cs_casng	51	REASON FOR STOP - CASING A VICTIM OR LOCATION	Nominal	8	Left	A1	A1
cs_lkout	52	REASON FOR STOP - SUSPECT ACTING AS A LOOKOUT	Nominal	8	Left	A1	A1
rf_vcact	53	REASON FOR FRISK- ACTIONS OF ENGAGING IN A VIOLENT CRIME	Nominal	8	Left	A1	A1
cs_cloth	54	REASON FOR STOP - WEARING CLOTHES COMMONLY USED IN A CRIME	Nominal	8	Left	A1	A1
cs_drgtr	55	REASON FOR STOP - ACTIONS INDICATIVE OF A DRUG TRANSACTION	Nominal	8	Left	A1	A1
ac_evasv	56	ADDITIONAL CIRCUMSTANCES - EVASIVE RESPONSE TO QUESTIONING	Nominal	8	Left	A1	A1

61	ac_assoc	57	ADDITIONAL CIRCUMSTANCES - ASSOCIATING WITH KNOWN CRIMINALS	Nominal	8	Left	A1	A1
62	cs_furtv	58	REASON FOR STOP - FURTIVE MOVEMENTS	Nominal	8	Left	A1	A1
63	rf_rfcmp	59	REASON FOR FRISK - REFUSE TO COMPLY W OFFICER'S DIRECTIONS	Nominal	8	Left	A1	A1
64	ac_cgdir	60	ADDITIONAL CIRCUMSTANCES - CHANGE DIRECTION AT SIGHT OF OFFICER	Nominal	8	Left	A1	A1
65	rf_verbl	61	REASON FOR FRISK - VERBAL THREATS BY SUSPECT	Nominal	8	Left	A1	A1
66	cs_vcrim	62	REASON FOR STOP - ACTIONS OF ENGAGING IN A VIOLENT CRIME	Nominal	8	Left	A1	A1
67	cs_bulge	63	REASON FOR STOP - SUSPICIOUS BULGE	Nominal	8	Left	A1	A1
68	cs_other	64	REASON FOR STOP - OTHER	Nominal	8	Left	A1	A1
69	ac_incid	65	ADDITIONAL CIRCUMSTANCES - AREA HAS HIGH CRIME INCIDENCE	Nominal	8	Left	A1	A1
70	ac_time	66	ADDITIONAL CIRCUMSTANCES - TIME OF DAY FITS CRIME INCIDENCE	Nominal	7	Left	A1	A1
71	rf_knowl	67	REASON FOR FRISK - KNOWLEDGE OF SUSPECT'S PRIOR CRIM BEHAV	Nominal	8	Left	A1	A1
72	ac_stsnd	68	ADDITIONAL CIRCUMSTANCES - SIGHTS OR SOUNDS OF CRIMINAL ACTIVITY	Nominal	8	Left	A1	A1
73	ac_other	69	ADDITIONAL CIRCUMSTANCES - OTHER	Nominal	8	Left	A1	A1
74	sb_hdobj	70	BASIS OF SEARCH - HARD OBJECT	Nominal	8	Left	A1	A1
75	sb_outln	71	BASIS OF SEARCH - OUTLINE OF WEAPON	Nominal	8	Left	A1	A1
76	sb_admis	72	BASIS OF SEARCH - ADMISSION BY SUSPECT	Nominal	8	Left	A1	A1
77	sb_other	73	BASIS OF SEARCH - OTHER	Nominal	8	Left	A1	A1
78	repcmd	74	REPORTING OFFICER'S COMMAND (1 TO 999)	Nominal	6	Left	A4	A4
79	revcmd	75	REVIEWING OFFICER'S COMMAND (1 TO 999)	Nominal	6	Left	A4	A4
80	rf_furt	76	REASON FOR FRISK - FURTIVE MOVEMENTS	Nominal	7	Left	A1	A1
81	rf_bulg	77	REASON FOR FRISK - SUSPICIOUS BULGE	Nominal	7	Left	A1	A1
82	offverb	78	VERBAL STATEMENT PROVIDED BY OFFICER (IF NOT IN UNIFORM)	Nominal	7	Left	A1	A1
83	offshld	79	SHIELD PROVIDED BY OFFICER (IF NOT IN UNIFORM)	Nominal	7	Left	A1	A1
84	sex	80	SUSPECT'S SEX	Nominal	3	Left	A1	A1
85	race	81	SUSPECT'S RACE	Nominal	4	Left	A1	A1
86	dob	82	SUSPECT'S DATE OF BIRTH (CCYY-MM-DD)	Nominal	8	Left	A8	A8
87	age	83	SUSPECT'S AGE	Nominal	6	Right	F6	F6
88	ht_feet	84	SUSPECT'S HEIGHT (FEET)	Scale	7	Left	A1	A1
89	ht_inch	85	SUSPECT'S HEIGHT (INCHES)	Nominal	7	Left	A2	A2
90	weight	86	SUSPECT'S WEIGHT	Nominal	6	Right	F6	F6
91	haircolr	87	SUSPECT'S HAIRCOLOR	Scale	8	Left	A2	A2
92	eyecolor	88	SUSPECT'S EYE COLOR	Nominal	8	Left	A2	A2
93	build	89	SUSPECT'S BUILD	Nominal	5	Left	A2	A2
94	othfeatr	90	SUSPECT'S OTHER FEATURES (SCARS, TATOOS ETC.)	Nominal	20	Left	A20	A20
95	addrtyp	91	LOCATION OF STOP ADDRESS TYPE	Nominal	9	Left	A1	A1
96	rescode	92	LOCATION OF STOP RESIDENT CODE	Nominal	7	Left	A1	A1
97	premtype	93	LOCATION OF STOP PREMISE TYPE	Nominal	8	Left	A2	A2
98	premname	94	LOCATION OF STOP PREMISE NAME	Nominal	24	Left	A30	A30
99	addrnum	95	LOCATION OF STOP ADDRESS NUMBER	Nominal	7	Left	A7	A7
100	stname	96	LOCATION OF STOP STREET NAME	Nominal	24	Left	A32	A32
101	stinter	97	LOCATION OF STOP INTERSECTION	Nominal	24	Left	A32	A32
102	crossst	98	LOCATION OF STOP CROSS STREET	Nominal	24	Left	A32	A32
103	aptnum	99	LOCATION OF STOP APT NUMBER	Nominal	6	Left	A6	A6
104	city	100	LOCATION OF STOP CITY	Nominal	20	Left	A20	A20
105	state	101	LOCATION OF STOP STATE	Nominal	5	Left	A2	A2
106	zip	102	LOCATION OF STOP ZIP CODE	Nominal	10	Left	A10	A10
107	addrpct	103	LOCATION OF STOP ADDRESS PRECINCT	Nominal	7	Left	A3	A3
108	sector	104	LOCATION OF STOP SECTOR	Nominal	6	Left	A30	A30
109	beat	105	LOCATION OF STOP BEAT	Nominal	16	Left	A7	A7
110	post	106	LOCATION OF STOP POST	Nominal	9	Left	A32	A32
111	xcoord	107	LOCATION OF STOP X COORD	Nominal	24	Left	A32	A32
112	ycoord	108	LOCATION OF STOP Y COORD	Nominal	24	Left	A6	A6
113	dettypCM	109	DETAILS TYPES CODE	Nominal	24	Left	A2	A2
114	lineCM	110	COUNT >1 ADDITIONAL DETAILS	Nominal	24	Left	A5	A5
115	detailCM	111	CRIME CODE DESCRIPTION	Nominal	6	Left	A78	A78

Fig 3: Attributes of the NYPD dataset from [3]

In this project, we are utilizing the NYPD dataset to get information about people in New York City (NYC) located at the area of Manhattan, Brooklyn, Bronx, Queens, and Staten Island to predict the frisked attribute of Stop, Question and Frisk (SQF) problems that the NYPD should be targeted to make final decision from the state various attributes, and information in crime status in future. Depend upon the primary target of the Big dataset, they can make result's future and reach out campaigns for their own duties. Analysis of the data shows to police officers who can work better in their own works including patrol, stop, question, and frisk and the data help to answer that why and how can apply SQF to people. Following of the predicted results, it can help police officers estimate their works that they should or not should deploy weapons or use several methods to govern to people. Police should understand about race attribute with radical different people, about age, city, crime background, gender, etc. to have consistent actions and avoid problems which do not allowed to use gun or strongly activities to people as well as police know they can

intensify to patrol at what are areas and how are important. Additionally, people who have been living at those areas will be received announcements about the status of crime, race, and other information from the NYPD by following of the Big Data through the statistics of many years, hence people will understand information, be careful and self- guard themselves or find the support from police.

The Big dataset that we are using is getting from the NYPD at **https://www1.nyc.gov.** The primary goal of this Analytics project go to find the answer of the questions about "***Upon different attributes from the datasets, could we report and predict to the juncture of crime in New York City***". In other words, we can ask the further questions to take predictions in the ***future*** as follows:

- ***Why the juncture of crime is different between areas, race, gender in NYC?***
- ***What are the reasons to increase or decrease crime in NYC?***
- ***How should the NYPD face in future?***

The questions above gave us a significance to consider the datasets and make decision-making results by choosing of the best target and modeling method. The best method will give us accurate answer from more questions and get along with the questions: "***What target could we predict on the dataset?***" and "***Which modeling method could we take to make a prediction?***" Since we got the topic based on SQF so our target is Frisked attribute which can accomplish the expectation of tasks. In many modeling methods have known why we can choose that method, also why we need modeling methods to discover and "dance" with them during we can look at on that dataset normally? May we not use the tools of Mathematical and Statistical methods? Why we can apply methods obviously to get correct answers. As prerequisites, we start this design and implementation project go along with analysis process to require the Big dataset get to "speak out".

The used software for:

Combining and Cleaning: MS-DOS (Command Prompt), Microsoft Excel, Access.

Storing, analyzing and predicting analytics: Tableau, IBM SPSS Modeler, R language with

Rattle tools.

List of Variables (Fig 4) [2]

The list below shows all attributes of the dataset with 112 attributes that contains detailed content of each attribute by variable, description, and values in each of them. For example, about Race attribute, we have six values of Black, Black Hispanic, White Hispanic, White, Asian/Pacific, Indian in the description of suspect's race of this race. The Stop, Question, and Frisked focus on criminal activities that descripted in the list such as criminal activities of using weapons, arrest, contraband, pistol, riflshot, assault weapon, knife or cutting instrument, machgun, othrweap, physical force used by officer, reason of stop, reason of frisked, basis of search, additional circumstances, location of stop. Otherwise, the list considers on the attributes as sex, haircolor, eyecolor, state, age, etc. In the crime code description variable of detailcm, we need to add more crime codes in a separately table as robbery, criminal activities, terrorism, fraud, killing, gambling, etc.

List of Variables

variable	description	values
year	year of stop	
pct	precinct of stop	1-123
ser_num	UF-250 serial number	
datestop	date of stop	mmddyyyy
timestop	time of stop	hhmm
city	location of stop city	1-Manhattan 2-Brooklyn 3-Bronx 4-Queens 5-Staten Island
sex	suspect's sex	0-female 1-male
race	suspect's race	1-black 2-black Hispanic 3-white Hispanic 4-white 5-Asian/Pacific Islander 6-Am. Indian/ Native Alaskan
dob	suspect's date of birth	mmddyyyy
age	suspect's age	
height	suspect's height in inches	
weight	suspect's weight in pounds	
haircolr	suspect's haircolor	1-black 2-brown 3-blonde 4-red 5-gray 6-white 7-bald 8-sandy 9-salt and pepper 10-dyed 11-frosted
eyecolor	suspect's eye color	1-black 2-brown 3-blue

eyecolor	suspect's eye color	1-black 2-brown 3-blue 4-green 5-hazel 6-gray 7-maroon, pink, violet 8-two different
build	suspect's build	1-heavy 2-muscular 3-medium 4-thin
othfeatr*	suspect's other features	
frisked	was suspect frisked?	
searched	was suspect searched?	
contrabn	was contraband found on suspect?	
pistol	was a pistol found on suspect?	
riflshot	was a rifle found on suspect?	
asltweap	was an assault weapon found on suspect?	
knifcuti	was a knife or cutting instrument found on suspect?	
machgun	was a machine gun found on suspect?	
othrweap	was another type of weapon found on suspect	
arstmade	was an arrest made?	
arstoffn*	offense suspect arrested for	
sumissue	was a summons issued?	
sumoffen*	offense suspect was summonsed for	
crimsusp*	crime suspected	
detailcm	crime code description	see attached
perobs	period of observation in minutes	

perstop	period of stop in minutes	
pf_hands	physical force used by officer - hands	
pf_wall	physical force used by officer - suspect on ground	
pf_grnd	physical force used by officer - suspect against wall	
pf_drwep	physical force used by officer - weapon drawn	
pf_ptwep	physical force used by officer - weapon pointed	

	pf_ptwep	physical force used by officer - weapon pointed	
	pf_baton	physical force used by officer - baton	
	pf_hcuff	physical force used by officer - handcuffs	
	pf_pepsp	physical force used by officer - pepper spray	
	pf_other	physical force used by officer - other	
	cs_objcs	reason for stop - carrying suspicious object	
	cs_descr	reason for stop - fits a relevant description	
	cs_casng	reason for stop - casing a victim or location	
	cs_lkout	reason for stop - suspect acting as a lookout	
	cs_cloth	reason for stop - wearing clothes commonly used in a crime	
	cs_drgtr	reason for stop - actions indicative of a drug transaction	
	cs_furtv	reason for stop - furtive movements	
	cs_vcrim	reason for stop - actions of engaging in a violent crime	
	cs_bulge	reason for stop - suspicious bulge	
	cs_other	reason for stop - other	
	rf_vcrim	reason for frisk - violent crime suspected	
	rf_othsw	reason for frisk - other suspicion of weapons	
	rf_attir	reason for frisk - inappropriate attire for season	
	rf_vcact	reason for frisk- actions of engaging in a violent crime	
	rf_rfcmp	reason for frisk - refuse to comply w officer's directions	
	rf_verbl	reason for frisk - verbal threats by suspect	
	rf_knowl	reason for frisk - knowledge of suspect's prior crim behav	
	rf_furt	reason for frisk - furtive movements	
	rf_bulg	reason for frisk - suspicious bulge	
	sb_hdobj	basis of search - hard object	
	sb_outln	basis of search - outline of weapon	
	sb_admis	basis of search - admission by suspect	
	sb_other	basis of search - other	
	ac_proxm	additional circumstances - proximity to scene of offense	
	ac_evasv	additional circumstances - evasive response to questioning	
	ac_assoc	additional circumstances - associating with known criminals	
	ac_cgdir	additional circumstances - change direction at sight of officer	
	ac_incid	additional circumstances - area has high crime incidence	
	ac_time	additional circumstances - time of day fits crime incidence	
	ac_stsnd	additional circumstances - sights or sounds of criminal activity	
	ac_other	additional circumstances - other	
	ac_rept	additional circumstances - report by victim/witness/officer	
	ac_inves	additional circumstances - ongoing investigation	
	forceuse	reason for force	1-defense of other 2-defense of self

ac_inves	additional circumstances - ongoing investigation	
forceuse	reason for force	1-defense of other 2-defense of self 3-overcome resistence 4-other 5-suspected flight 6-suspected weapon
inout	was stop inside or outside?	0-outside 1-inside
trhsloc	was location housing or transit authority?	0-neither 1-housing authority 2-transit authority
premname*	location of stop premise name	
addrnum*	location of stop address number	
stname*	location of stop street name	
stinter*	location of stop intersection	
crossst*	location of stop cross street	
addrpct*	location of stop address precinct	
sector*	location of stop sector	
beat*	location of stop beat	
post	location of stop post	
xcoord	location of stop x coord	
ycoord	location of stop y coord	

typeofid	stopped person's identification type	1-photo id 2-verbal id 3-refused to provide id
othpers	were other persons stopped, questioned or frisked ?	
explnstp	did officer explain reason for stop?	
repcmd	reporting officer's command	1-999
revcmd	reviewing officer's command	1-999
offunif	was officer in uniform?	
offverb	verbal statement provided by officer (if not in uniform)	
officrid	id card provided by officer (if not in uniform)	
offshld	shield provided by officer (if not in uniform)	

radio	radio run	
recstat*	record status	0-original value A 1-original value 1
linecm	count >1 additional details	

*string variables

Fig 4: List of Attributes from [2]

Below is the table of detailcm attribute about crime codes **(Fig 5).**

Over the following of the list of variable above via the detailcm attribute, we need to show a completely table of crime codes that has given by police officers in 113 criminal codes to recognize complex criminal activities in NYC as robbery, criminal activities, terrorism, sexual abuse, fraud, killing, gambling, etc.

Crime Codes

1 ABANDONMENT OF A CHILD
2 ABORTION
3 ABSCONDING
4 ADULTERY
5 AGGRAVATED ASSAULT
6 AGGRAVATED HARASSMENT
7 AGGRAVATED SEXUAL ABUSE
8 ARSON
9 ASSAULT
10 AUTO STRIPPING
11 BIGAMY
12 BRIBE RECEIVING
13 BRIBERY
14 BURGLARY
15 COERCION
16 COMPUTER TAMPERING
17 COMPUTER TRESPASS
18 COURSE OF SEXUAL CONDUCT
19 CPSP
20 CPW
21 CREATING A HAZARD
22 CRIMINAL CONTEMPT
23 CRIMINAL MISCHIEF
24 CRIMINAL POSSESION OF CONTROLLED SUBSTANCE
25 CRIMINAL POSSESSION OF COMPUTER MATERIAL
26 CRIMINAL POSSESSION OF FORGED INSTRUMENT
27 CRIMINAL POSSESSION OF MARIHUANA
28 CRIMINAL SALE OF CONTROLLED SUBSTANCE
29 CRIMINAL SALE OF MARIHUANA
30 CRIMINAL TAMPERING
31 CRIMINAL TRESPASS
32 CUSTODIAL INTERFERENCE
33 EAVES DROPPING
34 ENDANGER THE WELFARE OF A CHILD
35 ESCAPE
36 FALSIFY BUSINESS RECORDS
37 FORGERY
38 FORGERY OF A VIN
39 FORTUNE TELLING
40 FRAUD
41 FRAUDULENT ACCOSTING
42 FRAUDULENT MAKE ELECTRONIC ACCESS DEVICE
43 FRAUDULENT OBTAINING A SIGNATURE
44 GAMBLING
45 GRAND LARCENY
46 GRAND LARCENY AUTO
47 HARASSMENT
48 HAZING
49 HINDERING PROSECUTION
50 INCEST
51 INSURANCE FRAUD
52 ISSUE A FALSE CERTIFICATE
53 ISSUE A FALSE FINANCIAL STATEMENT
54 ISSUING ABORTION ARTICLES
55 JOSTLING
56 KIDNAPPING
57 KILLING OR INJURING A POILCE ANIMAL
58 LOITERING
59 MAKING GRAFFITI
60 MENACING
61 MISAPPLICATION OF PROPERTY
62 MURDER
63 OBSCENITY
64 OBSTRUCTING FIREFIGHTING OPERATIONS
65 OBSTRUCTING GOVERNMENTAL ADMINISTRATION
66 OFFERING A FALSE INSTRUMENT
67 OFFICIAL MISCONDUCT
68 PETIT LARCENY
69 POSSESSION OF BURGLAR TOOLS
70 POSSESSION OF EAVES DROPPING DEVICES
71 POSSESSION OF GRAFFITI INSTRUMENTS
72 PROHIBITED USE OF WEAPON
73 PROMOTING SUICIDE
74 PROSTITUTION
75 PUBLIC DISPLAY OF OFFENSIVE SEXUAL MATERIAL
76 PUBLIC LEWDNESS
77 RAPE
78 RECKLESS ENDANGERMENT
79 RECKLESS ENDANGERMENT PROPERTY
80 REFUSING TO AID A PEACE OR POLICE OFFICER
81 RENT GOUGING
82 RESISTING ARREST
83 REWARD OFFICIAL MISCONDUCT
84 RIOT
85 ROBBERY
86 SELF ABORTION
87 SEXUAL ABUSE
88 SEXUAL MISCONDUCT
89 SEXUAL PERFORMANCE BY A CHILD
90 SODOMY
91 SUBSTITUTION OF CHILDREN
92 TAMPERING WITH A PUBLIC RECORD
93 TAMPERING WITH CONSUMER PRODUCT
94 TAMPERING WITH PRIVATE COMMUNICATIONS
95 TERRORISM
96 THEFT OF SERVICES
97 TRADEMARK COUNTERFEITING
98 UNLAWFULLY DEALING WITH FIREWORKS
99 UNAUTHORIZED RECORDING
100 UNAUTHORIZED USE OF A VEHICLE
101 UNAUTHORIZED USE OF COMPUTER
102 UNLAWFUL ASSEMBLY
103 UNLAWFUL DUPLICATION OF COMPUTER MATERIAL
104 UNLAWFUL POSSESSION OF RADIO DEVICES
105 UNLAWFUL USE OF CREDIT CARD, DEBIT CARD
106 UNLAWFUL USE OF SECRET SCIENTIFIC MATERIAL
107 UNLAWFUL WEARING A BODY VEST
108 UNLAWFULL IMPRISONMENT
109 UNLAWFULLY DEALING WITH A CHILD
110 UNLAWFULLY USE SLUGS
111 VEHICULAR ASSAULT
112 OTHER
113 FORCIBLE TOUCHING

Fig 5: List of Crime codes from [2]

2.4 Cleaning dataset

About the problem of Remove Duplicates data in the dataset (Fig 6), we work on Microsoft Excel to remove duplicate values may be happened somewhere around four million rows and shows unique values that help to clean the dataset. The dataset has created by NYPD in variables, description, and values exactly, so there are no many duplicate values in rows and columns, though the dataset has many variables, values in a complex dataset but they are not duplicable and we only have a found duplicate values to remove besides 1,048,574 unique values remain exists in the data. This removing duplicate values found out in each dataset of *.csv file after that we would combined them from10 file into a file. The figure below shows the duplicate data found in the dataset of 2007 as the following figure:

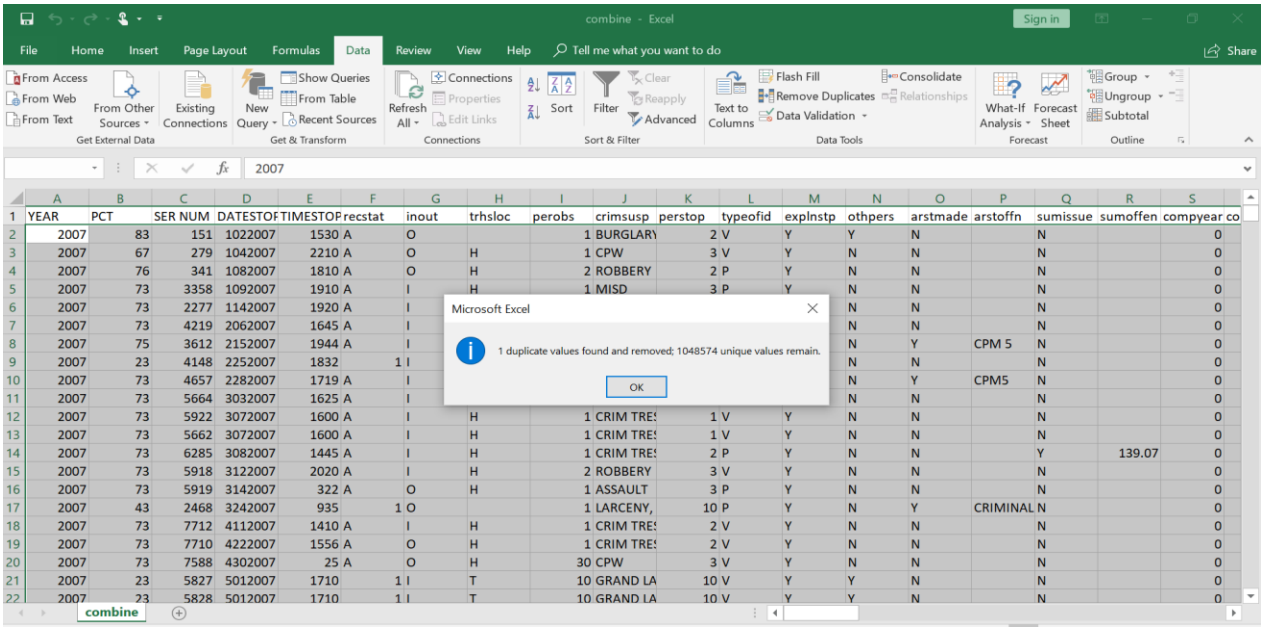


Fig 6: Remove Duplicates data

Since we had a combined *.csv file from the ten *.csv file, we need using of Microsoft Access to process the combined file from changing form *.csv file into text format *.txt file via the export tool. Why we do not keep *.csv file to utilize to our works? Because, as a result, the text format file can run to be faster than *.csv file by trying with software and apply them to works, although text file has the volume is larger the *.csv format file but we need to get the dataset with a better pattern over transformation.

Firstly, we choose the *.csv combined file to import to Microsoft Access for transformation process, and then generate them and view and check all rows for clean dataset.

Second, we use tools of Microsoft Access to export the combined file of *.csv into *.txt file of text file depends on the transformation process, hence we will have a text file to ready for analyzing and predicting works. Transformation process in Access tools has been shown in the figure (**Fig** 7):

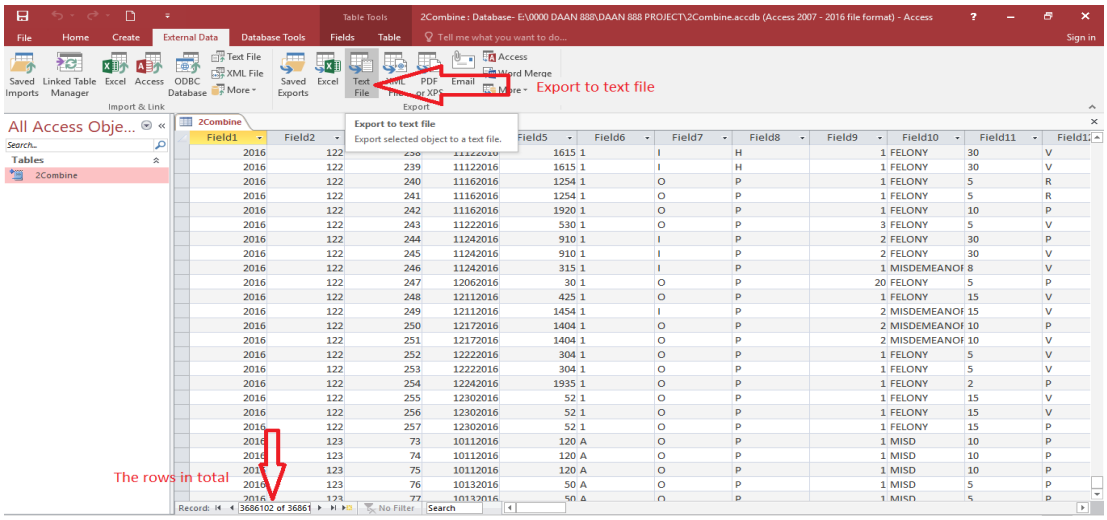


Fig 7: Transformation of *.csv file into *.txt file of the dataset

As the generated result above, there is the text dataset to prepare for next steps, we would import it to a showing values table in IBM SPSS Modeler to look and check general values. Totally, the dataset has 112 columns and 3,686,101 rows but we would reduce them to around 100 attributes to be easy for analyzing and predicting process. The dataset has shown by the text file in SPSS Modeler software completely of attributes, variables and values. During we execute the dataset with *.csv file then data will be missed values and not be completed in data so that is reason why we need to transfer the *.csv file to text file that is shown in the data table of IBM SPSS Modeler. In the final decision, we should recommend to take the text dataset for the data from NYPD for whole process of evaluation, calculation, and prediction to get exactly results as good as running faster in the workflows of nodes in the SPSS Modeler software. The dataset is now always available for considering as follows (Fig 8):

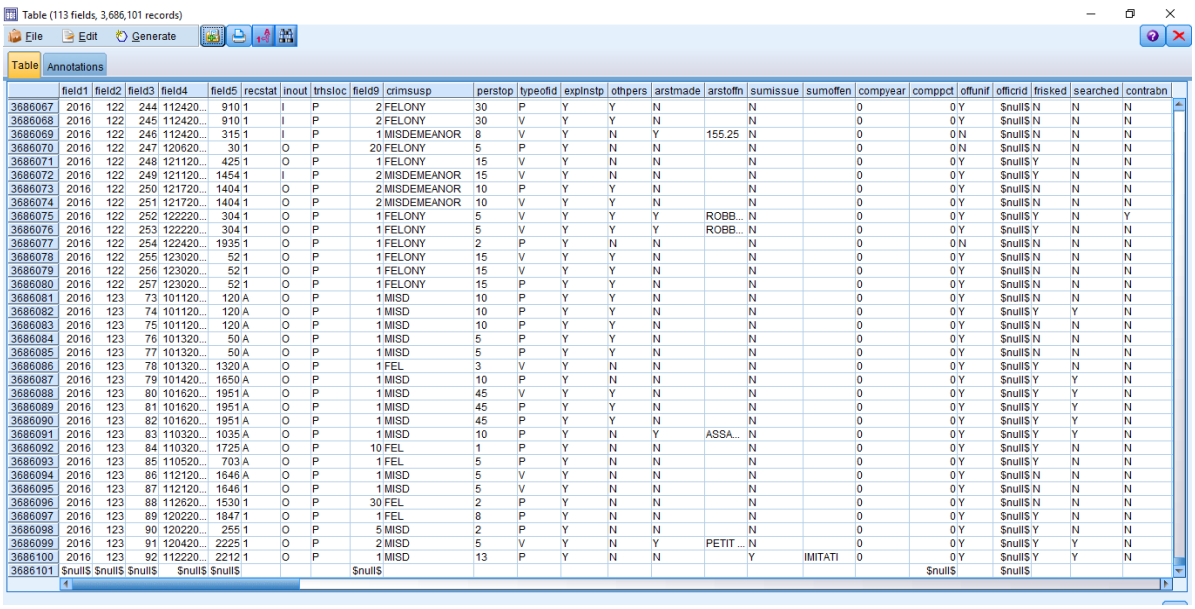
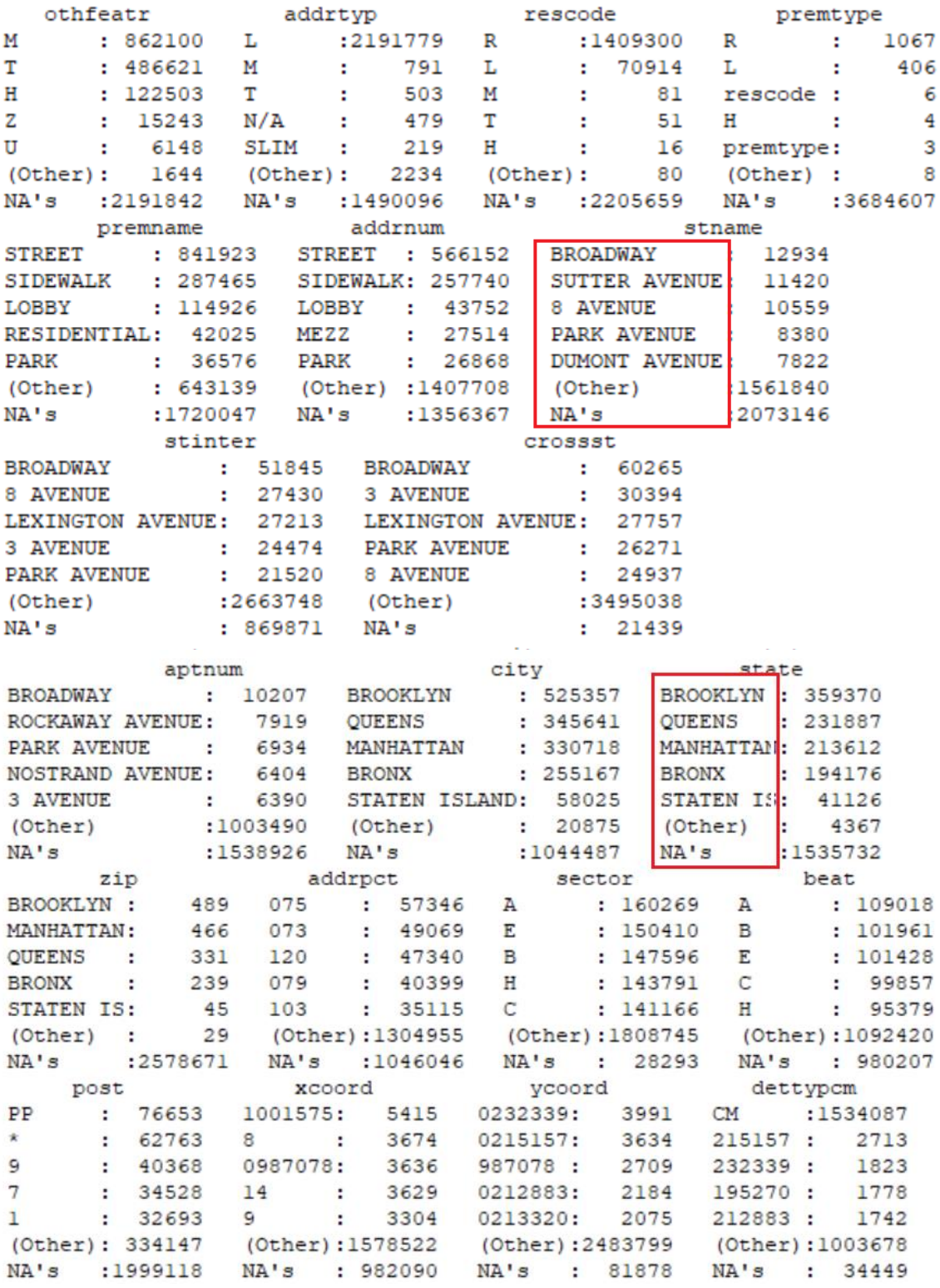


Fig 8: The cleaned dataset with 112 attributes and 3,686,101 rows

Summary table of some attributes (Fig 8A)

```
          sumoffen          compyear          comppct          offunif
OPEN CONTAINER:  15821   0       :3681736   0       :3685661   Y       :2659553
DISCON         :  14924   N       :    346   N       :     24   N       :1021739
DIS CON        :   7769   "       :     33   comppct:      9   0       :   4322
240.20         :   6288   Y       :     20   "       :      5   M       :    444
TRESPASS       :   6240   DISCON  :     19   240.20  :      3   offunif:      9
(Other)        : 167551   (Other):    568   (Other):     51   (Other):      7
NA's           :3467508   NA's    :   3379   NA's    :    348   NA's    :     27
    officrid          frisked          searched          contrabn
0       :   4999   0       :     34   0       :      3   contrabn:      9
401"    :      1   frisked:      9   I       :      8   I       :      1
I       :  56722   I       :     70   N       :3352165   N       :3611214
N       :    859   N       :1634799   searched:      9   Y       :  69640
officrid:      9   Y       :2047334   Y       : 333526   NA's    :   5237
Y       :   3066   NA's    :   3855   NA's    :    390
NA's    :3620445
    adtlrept          pistol          riflshot          asltweap
adtlrept:      9   1       :     54   1       :      1   1       :      1
N       :3671873   N       :3670126   N       :3675003   asltweap:      9
Y       :   3359   pistol:      9   riflshot:      9   N       :3674941
NA's    :  10860   Y       :   5105   Y       :    228   Y       :    290
                   NA's    :  10807   NA's    :  10860   NA's    :  10860
```

```
   offverb           offshld             sex               race
V       : 785684   S       :1008572   M       :1996431   M       :1366734
0       :   4129   0       :   3350   F       : 154671   B       :1143097
N       :   3543   V       :    675   DS      :  49806   Q       : 544949
Y       :    768   N       :    343   Z       :  40738   W       : 218718
offverb:      9   Y       :     84   SF      :  25224   P       : 134079
(Other):     34   (Other):     12   (Other):  40155   (Other): 277170
NA's   :2891934   NA's   :2673065   NA's   :1379076   NA's   :   1354
      dob                age              ht_feet           ht_inch
B       : 783164   12311900: 534847   5       :1765484   5       :1349076
Q       : 365082   19      : 133854   6       : 417833   6       : 497901
12311900: 204118   18      : 131654   20      :  87167   8       : 291572
W       : 142326   20      : 130942   18      :  84769   9       : 283147
P       : 101603   17      : 118051   19      :  83378   10      : 261612
(Other) :2089683   (Other) :2410205   (Other):1247183   (Other):1002758
NA's    :    125   NA's    : 226548   NA's   :    287   NA's   :     35
    weight           haircolr          eyecolor           build
160     : 277927   BK      :1630585   BR      :2233761   BR      :1340025
180     : 243847   BR      : 442201   BK      :1276803   M       :1315478
150     : 235382   160     : 193426   BL      :  56453   T       : 661169
170     : 227004   150     : 167291   BA      :  26400   H       : 183617
8       : 204130   180     : 166021   XX      :  21177   BK      :  96689
(Other):2497807   (Other):1086575   (Other):  71505   (Other):  89121
NA's    :      4   NA's    :      2   NA's    :      2   NA's    :      2
      dob                age              ht_feet           ht_inch
B       : 783164   12311900: 534847   5       :1765484   5       :1349076
Q       : 365082   19      : 133854   6       : 417833   6       : 497901
12311900: 204118   18      : 131654   20      :  87167   8       : 291572
W       : 142326   20      : 130942   18      :  84769   9       : 283147
P       : 101603   17      : 118051   19      :  83378   10      : 261612
(Other) :2089683   (Other) :2410205   (Other):1247183   (Other):1002758
NA's    :    125   NA's    : 226548   NA's   :    287   NA's   :     35
    weight           haircolr          eyecolor           build
160     : 277927   BK      :1630585   BR      :2233761   BR      :1340025
180     : 243847   BR      : 442201   BK      :1276803   M       :1315478
150     : 235382   160     : 193426   BL      :  56453   T       : 661169
170     : 227004   150     : 167291   BA      :  26400   H       : 183617
8       : 204130   180     : 166021   XX      :  21177   BK      :  96689
(Other):2497807   (Other):1086575   (Other):  71505   (Other):  89121
NA's    :      4   NA's    :      2   NA's    :      2   NA's    :      2
   othfeatr          addrtyp           rescode           premtype
M       : 862100   L       :2191779   R       :1409300   R       :   1067
T       : 486621   M       :    791   L       :  70914   L       :    406
H       : 122503   T       :    503   M       :     81   rescode :      6
Z       :  15243   N/A     :    479   T       :     51   H       :      4
U       :   6148   SLIM    :    219   H       :     16   premtype:      3
(Other):   1644   (Other):   2234   (Other):     80   (Other) :      8
NA's    :2191842   NA's    :1490096   NA's    :2205659   NA's    :3684607
```



```
   othfeatr          addrtyp          rescode          premtype
M      : 862100   L      :2191779   R      :1409300   R       :   1067
T      : 486621   M      :    791   L      :  70914   L       :    406
H      : 122503   T      :    503   M      :     81   rescode :      6
Z      :  15243   N/A    :    479   T      :     51   H       :      4
U      :   6148   SLIM   :    219   H      :     16   premtype:      3
(Other):   1644   (Other):   2234   (Other):     80   (Other) :      8
NA's   :2191842   NA's   :1490096   NA's   :2205659   NA's    :3684607
       premname          addrnum                stname
STREET      : 841923   STREET  : 566152   BROADWAY       :  12934
SIDEWALK    : 287465   SIDEWALK: 257740   SUTTER AVENUE  :  11420
LOBBY       : 114926   LOBBY   :  43752   8 AVENUE       :  10559
RESIDENTIAL:  42025   MEZZ    :  27514   PARK AVENUE    :   8380
PARK        :  36576   PARK    :  26868   DUMONT AVENUE  :   7822
(Other)     : 643139   (Other) :1407708   (Other)        :1561840
NA's        :1720047   NA's    :1356367   NA's           :2073146
          stinter                    crossst
BROADWAY         :  51845   BROADWAY         :  60265
8 AVENUE         :  27430   3 AVENUE         :  30394
LEXINGTON AVENUE:  27213   LEXINGTON AVENUE:  27757
3 AVENUE         :  24474   PARK AVENUE      :  26271
PARK AVENUE      :  21520   8 AVENUE         :  24937
(Other)          :2663748   (Other)          :3495038
NA's             : 869871   NA's             :  21439
          aptnum                    city                 state
BROADWAY         :  10207   BROOKLYN     : 525357   BROOKLYN : 359370
ROCKAWAY AVENUE:   7919   QUEENS       : 345641   QUEENS   : 231887
PARK AVENUE      :   6934   MANHATTAN    : 330718   MANHATTAN: 213612
NOSTRAND AVENUE:   6404   BRONX        : 255167   BRONX    : 194176
3 AVENUE         :   6390   STATEN ISLAND:  58025   STATEN IS:  41126
(Other)          :1003490   (Other)      :  20875   (Other)  :   4367
NA's             :1538926   NA's         :1044487   NA's     :1535732
       zip              addrpct           sector           beat
BROOKLYN :    489   075    :  57346   A      : 160269   A      : 109018
MANHATTAN:    466   073    :  49069   E      : 150410   B      : 101961
QUEENS   :    331   120    :  47340   B      : 147596   E      : 101428
BRONX    :    239   079    :  40399   H      : 143791   C      :  99857
STATEN IS:     45   103    :  35115   C      : 141166   H      :  95379
(Other)  :     29   (Other):1304955   (Other):1808745   (Other):1092420
NA's     :2578671   NA's   :1046046   NA's   :  28293   NA's   : 980207
     post             xcoord            ycoord           dettypcm
PP      :  76653   1001575:   5415   0232339:   3991   CM     :1534087
*       :  62763   8      :   3674   0215157:   3634   215157 :   2713
9       :  40368   0987078:   3636   987078 :   2709   232339 :   1823
7       :  34528   14     :   3629   0212883:   2184   195270 :   1778
1       :  32693   9      :   3304   0213320:   2075   212883 :   1742
(Other): 334147   (Other):1578522   (Other):2483799   (Other):1003678
NA's    :1999118   NA's   : 982090   NA's   :  81878   NA's   :  34449
```

Fig 8A: Summary of some attributes with number and values of variables

The figure in Fig 8A have shown in summary of attributes with detailed variables comprising each value among each variable to have understood insight the dataset. In the red rectangles above we present some attributes to see details of them. The figure 8A indicates 112 attributes of the dataset in summary of clear details by variables and values to us for planning of analyzing works since it was not easy to look on all data in a data table due to a vast of amount of data over our observation. Although, the summary data generated all variables and values but they had main variables which had mostly values to become predictors in progress process. Furthermore, we will also waive attributes having many missing values, typeless, duplicate values by automatic choosing process by IBM SPSS Modeler 18 software to modeling algorithm in predicting to execute better accuracies and strongly results. As a recommendation to the features of dataset, we could apply fitting machine learning to evaluate and predict to a purpose of target component, and overcomes some limitations of attributes.

The tables above describe the insight details of attributes as:

Outliers: We use boxplot to recognize outliers of attributes by concentrating values around years so we can find out outliers to consider and prepare predictors of the dataset.
Outliers only are verified as further component parts. It does not have much influence to the results of dataset but also need to verify data for insight understanding, as well as doing among analyzing data.

By the observation of the qualify table data of boxplot below, the primary two attributes are Y and N between 2008 and 2011 are most on the light blue and light purple box, besides the pink one of 0 value of 2007, 2008, and 2009; the light green boxplot is from 2008 to 2011 with I attribute. As our observations, there are three attributes with outlier areas from points on the dataset of the attributes Y, N and 0 from 2013 to 2016 but they are small as in the boxplot. Totally, we have the gray color is a summary boxplot of entire variables as below (Fig 9).

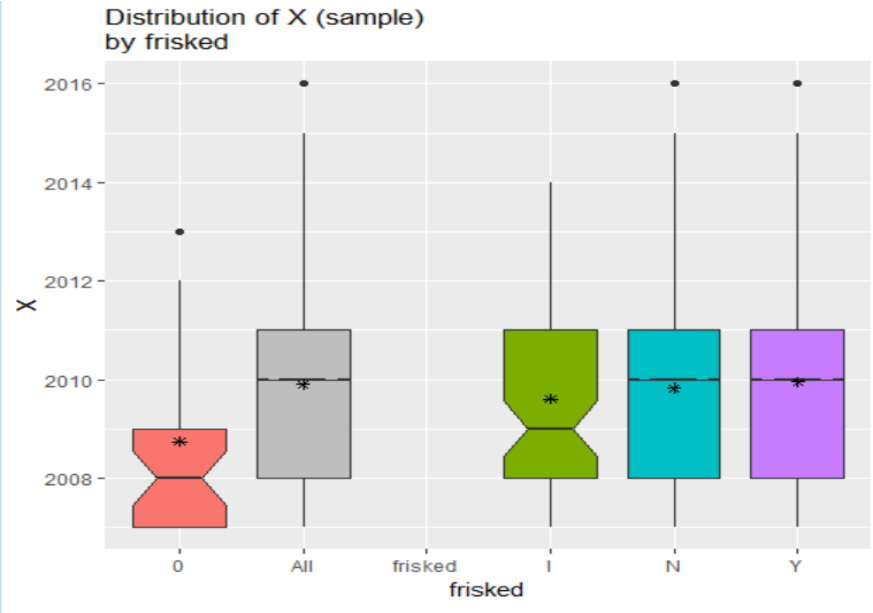


Fig 9: Boxplot shows the three outliers area.

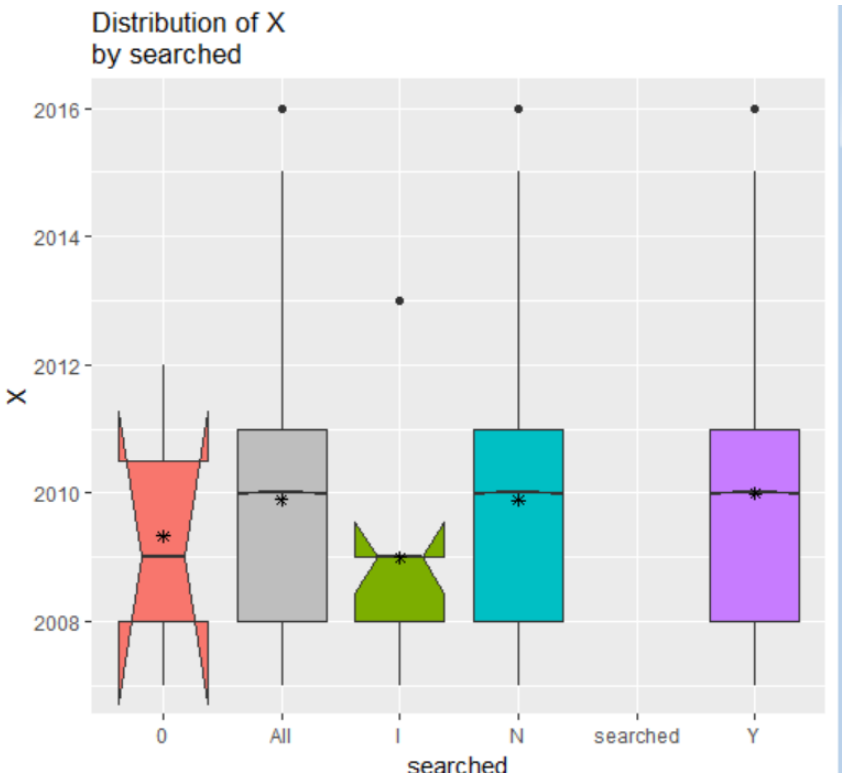


Fig 9A: Boxplot of searched with four variables.

Next of the observation on the boxplot of searched attribute, similarly as the frisked attribute, the primary two attributes are also Y and N between 2008 and 2011 are mostly values on the light blue and light purple box, besides the pink one of 0 value of 2007 to 2011; The light green boxplot is from 2008 to 2009 with I attribute. As our observations, there are three attributes with outlier areas from points on the dataset of the attributes Y, N and I from 2011 for I and to 2016 for Y and N but they are small as in the boxplot. Under observations of the two attributes of frisked and searched, we have generated results in different features via years of variables, however the main variables of Y and N are the same throughout years of 2008 to 2011 but they are different of values, and then having change of the two attributes of 0 and I by influence of years

Boxplot of searched attribute with the main attributes of Y and N are among 2008 and 2011 are mostly values but other years are so small values (9A).

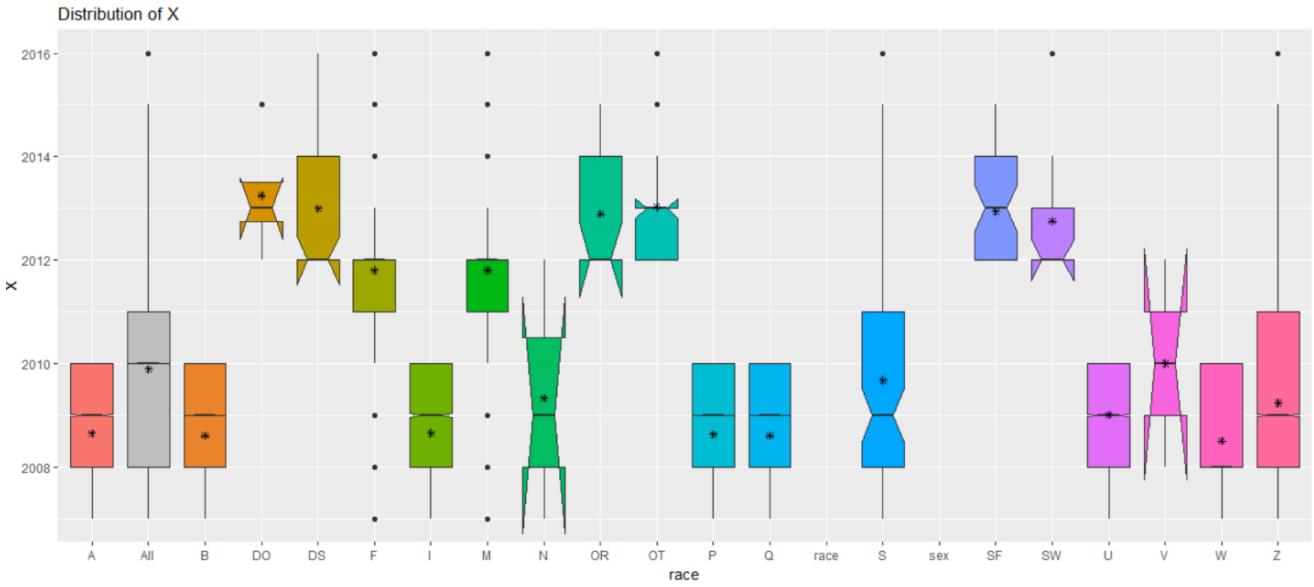


Fig 9B: Boxplot of Race attribute by 10 years.

As a completely results, the boxplot of Race attributes between 2007 to 2016 with at most of attributes of A, B variables from 2008 to 2010 in the pink and orange color boxplot on the left of the boxplot; S attribute from 2008 to 2011 in the light blue boxplot; DS, OR, SF are from 2012 to 2014 in the light green boxplot; U, W, P, Q attributes are in the year of 2008 to 2010, ect. DO, F, M, OT, S, SW, Z attributes have outliers through years from 2007 to 2016 but they are concentrated at the two years of 2015 and 2016, and F, M attributes have also outliers in years of 2007 to 2009 and 2014 that visualized by dark points on the boxplot. Following of observed boxplot, we can figure out outliers easily to understand, observe, and analyze any dataset on the predicting of plans for our common purposes because we need to overcome limitations of dataset, differences to associate them in a clean sourcing data. The boxplot used many varietal color in the boxplot to support well for visualizations and to make conclusions for related outlier problems. As a result, the gray color exhibits totally values that focus on the years of 2008 to 20111 as the figure (9B).

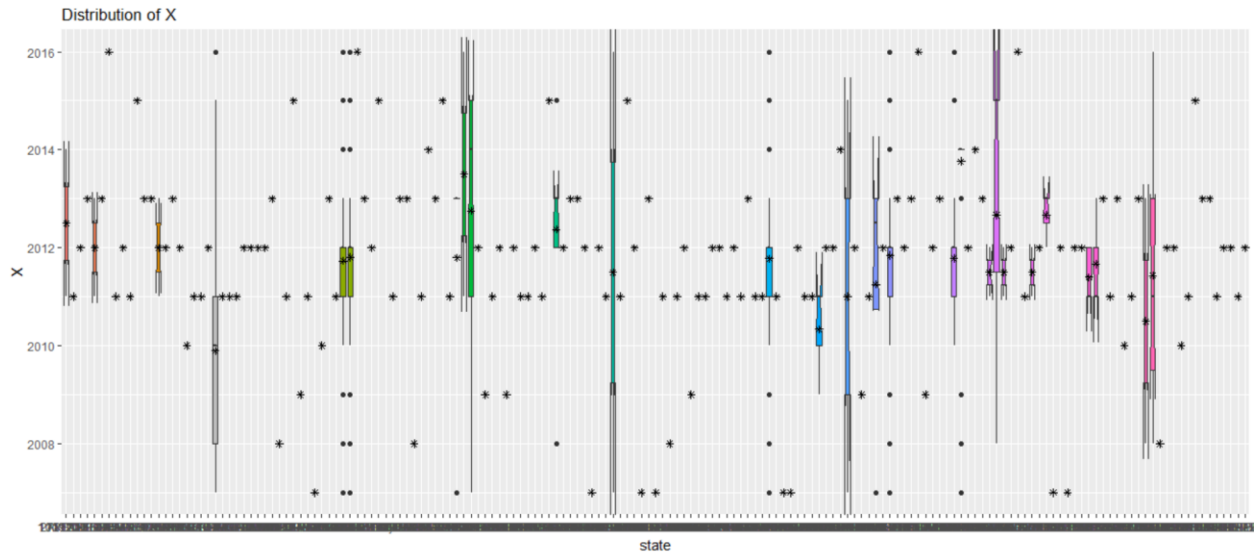


Fig 9C: Boxplot of State attribute by 10 years.

The boxplot of State attributes in the figure of (9C) follow between 2007 to 2016 with at most of attributes of Manhattan, Brooklyn areas, and then from Bronx, Queen, and Staten Island through years. Similarly, from the five color boxplot indicated values of years on the five colors with variables of them that look like the boxplot above, and they are summary in the gray color for total of all attributes.

The boxplot exhibits too many outliers in ten years and distributes to all attributes of the state attribute.

The gray color carries out of data in total values in the years of 2008 to 2011 as the considered attributes above. The brown boxplot color shows data between 2002 and 2003 with many outliers; the light green boxplot color of 2011 and 2012 with many outliers from 2007 to 2016; the green boxplot color of 2011 and 2012 with many outliers from 2011 to 2015; and other boxplot colors along values exhibit as the illustrated boxplot graphs in many outliers in the related attributes.

Other qualify attributes data: We consider the dataset and generate to have results on extremes, missing data, whitespace, null values, blank values, % complete of the attributes depend on the cleaning process as results below

Missing data: Attributes have missing data on all variables but mostly are little and we can waive attributes that have missing values more than two million values, each attribute presented missing values and distinct data. Describing of each attribute focuses on each variable of 112 attributes that contained value, frequency, proportion, and evaluated on each variable of each attribute. This figure helps us to determine and prepare for analyzing and predicting works in the arranging of data into input, target, risk, ident, ignore, weigh, and comment (Fig 10 & 11).

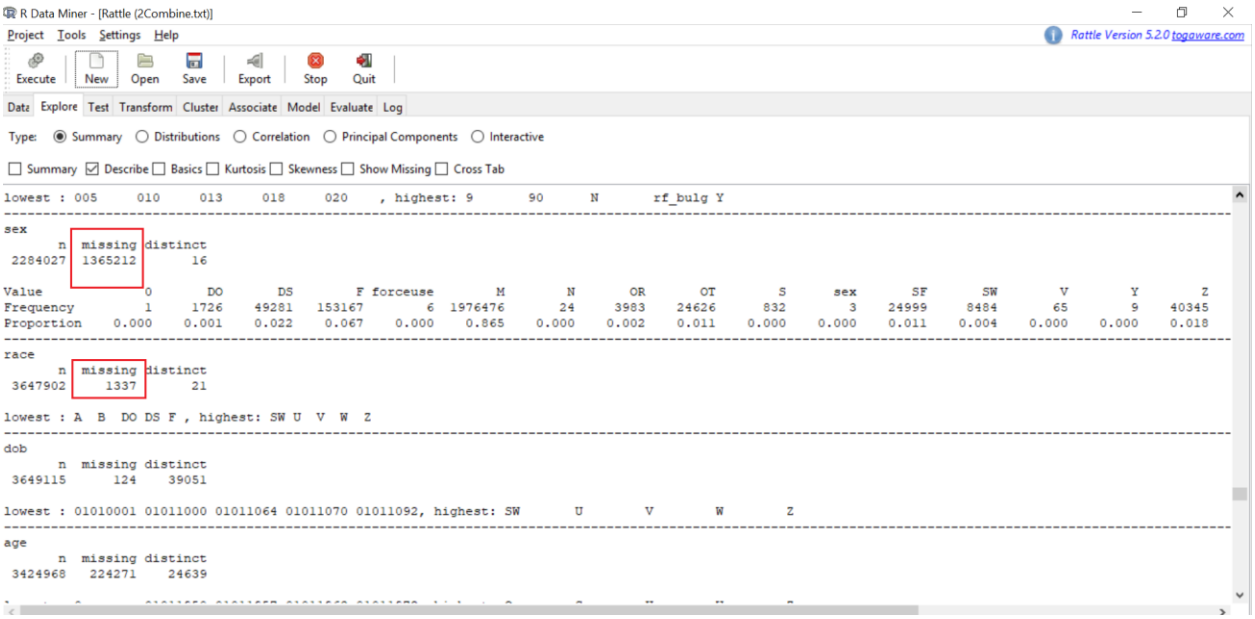


Fig 10: Missing, distinct values of the dataset

Extremes data: Having some attributes have extremes data but they are small and are to use to predict on the dataset. It is having of duty as outlier data and we can use it to analyze to the dataset, although it does not have much influence to the results of predicting process.

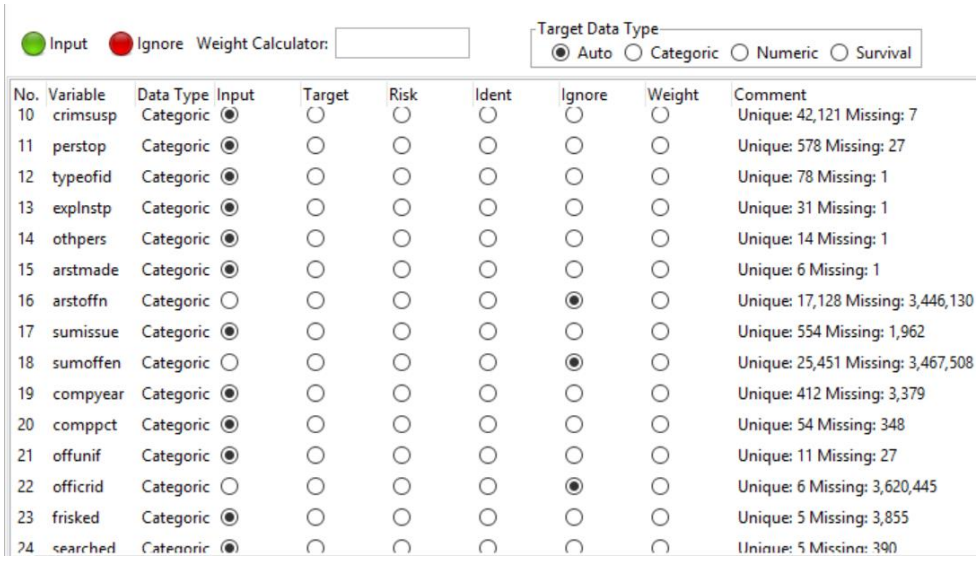
% Complete of column data: At most attributes have completion of 100% unless columns have to many lacks or empty of data. We will use the data with at least 40% of completely data by the column looking on each attribute for our works on the exploratory data analysis, modeling methods and visualizations. Because of completion of dataset, we plan to handle tools, models, and algorithm for necessary progress process (Fig 12).

Empty String of dataset: As in the figure of 12 & 13 have shown many empty string values of columns, this means that have not values because they were not recorded due to there are other actions occurred or nobody is suspected or is mistook on those attributes, thus they are a part of real life dataset and we need to consider and works on them (Fig 12&13).

White space of dataset: By the way above, the number of white spaces are similar the number of empty strings by comparing of the figure of 12 and 13, we arrive and know them in depth as an observation. If empty strings are an empty positions of values, and not have any data then white spaces are no something or no data stay in white spaces.

Blank data: As the results in the dataset, we can see the datasets do not have any blank data. Blank data on columns or rows do not have any ones, this definition is to do not have blank rows or columns. They really are different to white space and empty string, thought they mean to be empty but empty string and white space are blank on some cell but blank is on whole column or whole rows (Fig 13).

Unique data: In order to prevent duplicate data, we could utilize tools to remove them by software. Usually, we can work on the MS – Excel for this responsibility. That works only clean duplicate data but also create clearly considering data. Pursuant to generated unique data from the dataset, we understand to the meaning of variables of data to setup plans of forwarding statements (Fig 11). In the figure from R & Rattle tool expressed to missing and unique values by feature of comments on the detailed components.



No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
10	crimsusp	Categoric	◉	○	○	○	○	○	Unique: 42,121 Missing: 7
11	perstop	Categoric	◉	○	○	○	○	○	Unique: 578 Missing: 27
12	typeofid	Categoric	◉	○	○	○	○	○	Unique: 78 Missing: 1
13	explnstp	Categoric	◉	○	○	○	○	○	Unique: 31 Missing: 1
14	othpers	Categoric	◉	○	○	○	○	○	Unique: 14 Missing: 1
15	arstmade	Categoric	◉	○	○	○	○	○	Unique: 6 Missing: 1
16	arstoffn	Categoric	○	○	○	○	◉	○	Unique: 17,128 Missing: 3,446,130
17	sumissue	Categoric	◉	○	○	○	○	○	Unique: 554 Missing: 1,962
18	sumoffen	Categoric	○	○	○	○	◉	○	Unique: 25,451 Missing: 3,467,508
19	compyear	Categoric	◉	○	○	○	○	○	Unique: 412 Missing: 3,379
20	comppct	Categoric	◉	○	○	○	○	○	Unique: 54 Missing: 348
21	offunif	Categoric	◉	○	○	○	○	○	Unique: 11 Missing: 27
22	officrid	Categoric	○	○	○	○	◉	○	Unique: 6 Missing: 3,620,445
23	frisked	Categoric	◉	○	○	○	○	○	Unique: 5 Missing: 3,855
24	searched	Categoric	◉	○	○	○	○	○	Unique: 5 Missing: 390

Fig 11: Missing, Unique values of the dataset

Measurement and Type data of attributes: 18 attributes of Continuous values and 100 attributes of Categorical/Nominal values.

Type data: Obviously, the analyzing software need to determine automatically the type of attributes for any task as using of IBM SPSS Modeler or we need to setup manually measurement as working on Rattle software. Typically, attributes in numerical values will be putted as a continuous value, besides attributes in categorical values will be shown as nominal/categorical value which are not numerical parameters. Furthermore, we also know types of attributes as typeless that cannot use for any calculating of dataset or the type of flag that will apply to a chosen target to find ROC curve of Gain curve by its numerical values during getting on the duties of predicted works. For example, based on the two variables of Yes and No, we will work on categorical values to do any its calculable ability on the duty of project but they lie in an allowed limitation, however, it is not the same in mind of numerical values, they have widely calculable ability of attributes so they will have many results in many different formulas (see Fig 11, 12&13).

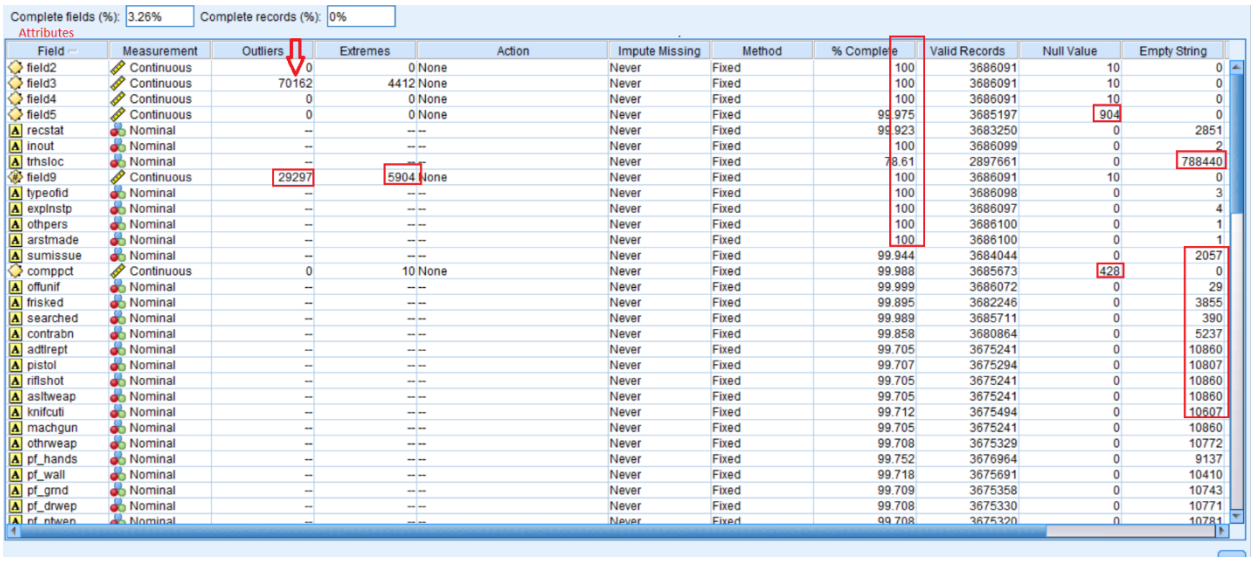


Fig 12: Continuous and Categorical/Nominal values

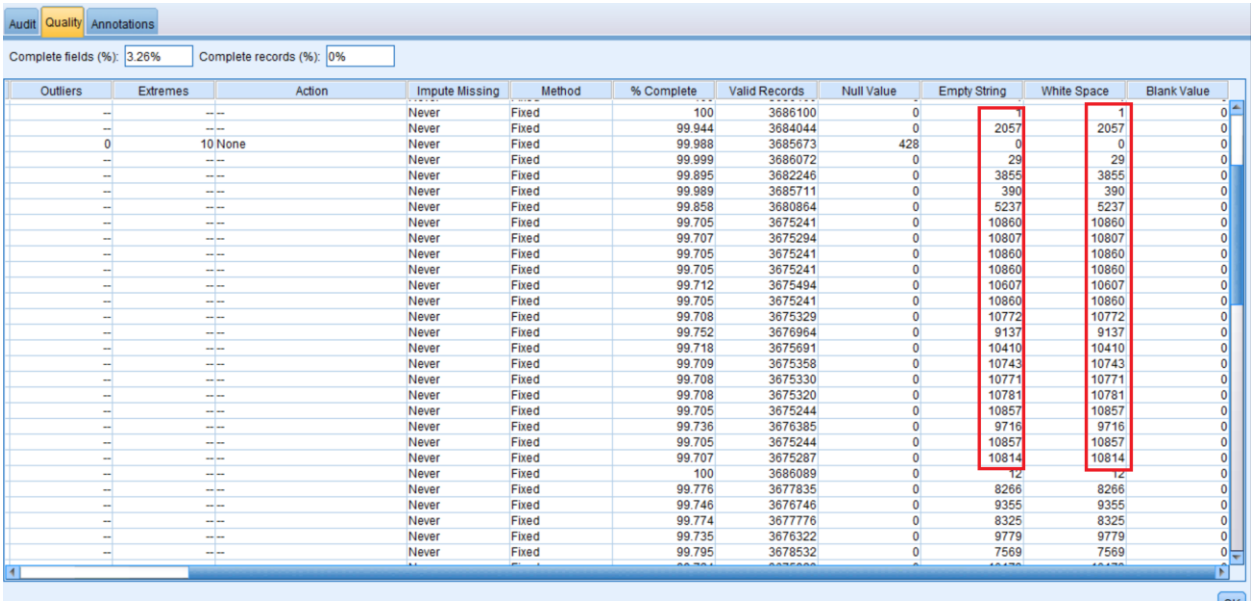
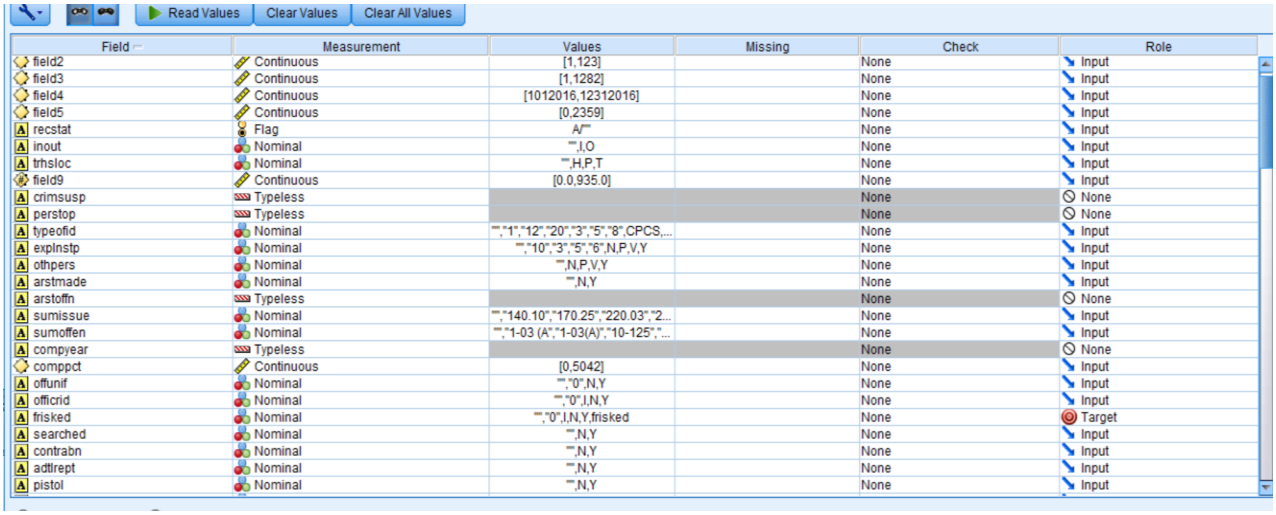


Fig 13: Empty, White Space, Blank, %Complete values

Interestingly and deeply, we were discovering, cleaning, analyzing, and understanding about attributes of the dataset by a combination of analysis type of data as above to do certain research of insight data as was going to find obviously answers in the dataset and having the best preparing for report project so far.

In most case, after we had ten *.csv files separately by downloading, and then combining them in a file, we removed duplicate data in Microsoft Excel tool and get 100 attributes on the table below. Mostly, the dataset has more than 90% of data completely, this is a good signification to predict and analyze our purposes. By strictly ways as the looking above, we finally result of the cleaned dataset has attributed columns with the 100 attributes (99 inputs and 1 targets in Role) in the figure

below after we waive all attributes that have bad effect to targeted prediction as well as make increasing of accuracy.
Data type now is clear to use in the type of Nominal (Categorical), Typeless, Flag, and Continuous (Numeric).
As a result and work way automatically, we have received a variety type of attributes that service in whole processing duty since we have waived 10 attributes of the **typeless** type by analyzing works and setting a type with the name of None on the Role column (Fig 14).



Field	Measurement	Values	Missing	Check	Role
field2	Continuous	[1,123]		None	Input
field3	Continuous	[1,1282]		None	Input
field4	Continuous	[1012016,12312016]		None	Input
field5	Continuous	[0,2359]		None	Input
recstat	Flag	A/"		None	Input
inout	Nominal	"",I,O		None	Input
trhsloc	Nominal	"",H,P,T		None	Input
field9	Continuous	[0.0,935.0]		None	Input
crimsusp	Typeless			None	None
perstop	Typeless			None	None
typeofid	Nominal	"","1","12","20","3","5","8",CPCS,...		None	Input
explnstp	Nominal	"","10","3","5","6",N,P,V,Y		None	Input
othpers	Nominal	"",N,P,V,Y		None	Input
arstmade	Nominal	"",N,Y		None	Input
arstoffn	Typeless			None	None
sumissue	Nominal	"","140.10","170.25","220.03","2...		None	Input
sumoffen	Nominal	"","1-03 (A)","1-03(A)","10-125","...		None	Input
compyear	Typeless			None	None
comppct	Continuous	[0,5042]		None	Input
offunif	Nominal	"","0",N,Y		None	Input
officrid	Nominal	"","0",I,N,Y		None	Input
frisked	Nominal	"","0",I,N,Y,frisked		None	Target
searched	Nominal	"",N,Y		None	Input
contrabn	Nominal	"",N,Y		None	Input
adtlrept	Nominal	"",N,Y		None	Input
pistol	Nominal	"",N,Y		None	Input

Fig 14: Some attributes of typeless values waived.

3. Understanding and General Analyzing Dataset

3.1 Histogram and Dotplot of Distribution of Frisked by Years (X variable)

Distribution below (Fig15) is the statistical result of Frisked attribute on the double of variable of Yes and No of them comments from police officers of 10 years between 2007 and 2016. We comprise that processes of Stop, Question, and then is Frisk on suspected people as using weapons, robbers, contraband goods, and other reasons. Someone can be required to Stop and answer Questions from police officer, during of asking and exhibiting on the problems related to suspected criminal activities, police will frisk or will not frisked to current those people, so we have many columns in this project which were asked about the status of Yes or No. Under doing way so, Frisk involves two main attributes of Yes or No, this means people are or are not frisked by police officers. The records of frisked was Yes status with 2,047,334 cases (55.6%) but No status was 1,634,799 (44.4%) cases of 10 years, and the highest cases in years of 2009, 2010, and 2011 with the values of each, this status decreases between 2012 and 2016 with the number of case, Additionally, we also have the total of number of Yes and No of 10 years by two pink bars, each year of number of frisked plotted by colors to recognize in visualizing results with the numbers on the top bars since we have dispensed the values of I and O attributes because they are not considerable as follows.

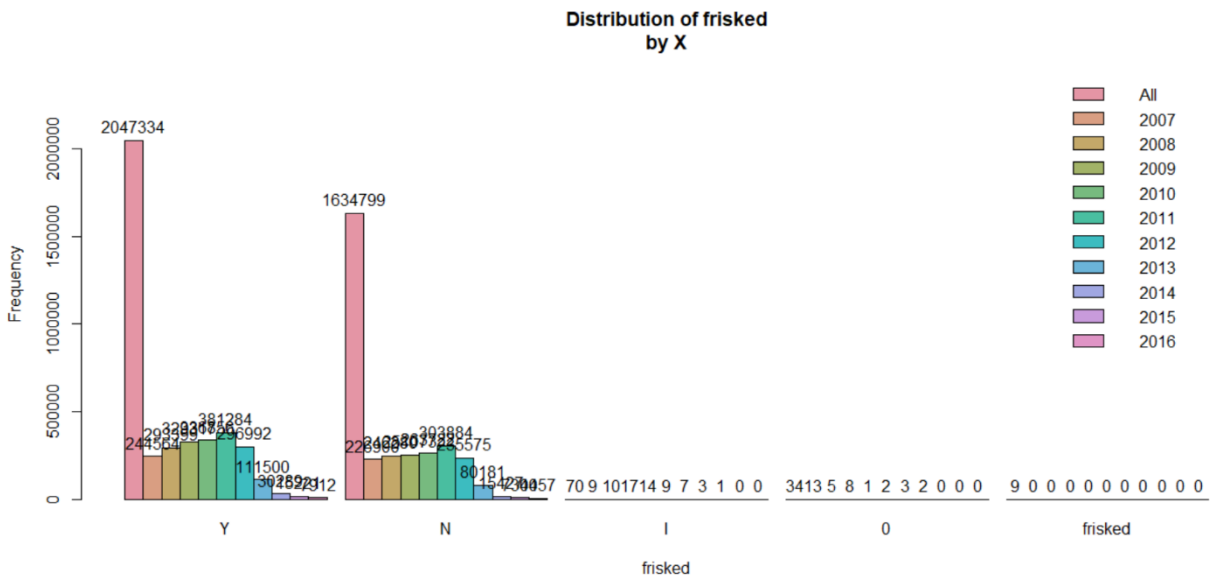


Fig15: Distribution of Frisked by Years

Continuously, as the discussed task in the figure 15, the histogram below (Fig 16) is the same in details but they had been presented in the two variable Yes and No in each year of 10 years in the straight line of the x- axis that helps us to explore the number of cases of frisked in each year easier on Yes (Pink) or No (Yellow) by clicking and moving the mouse on the bar chart via those bar-years.

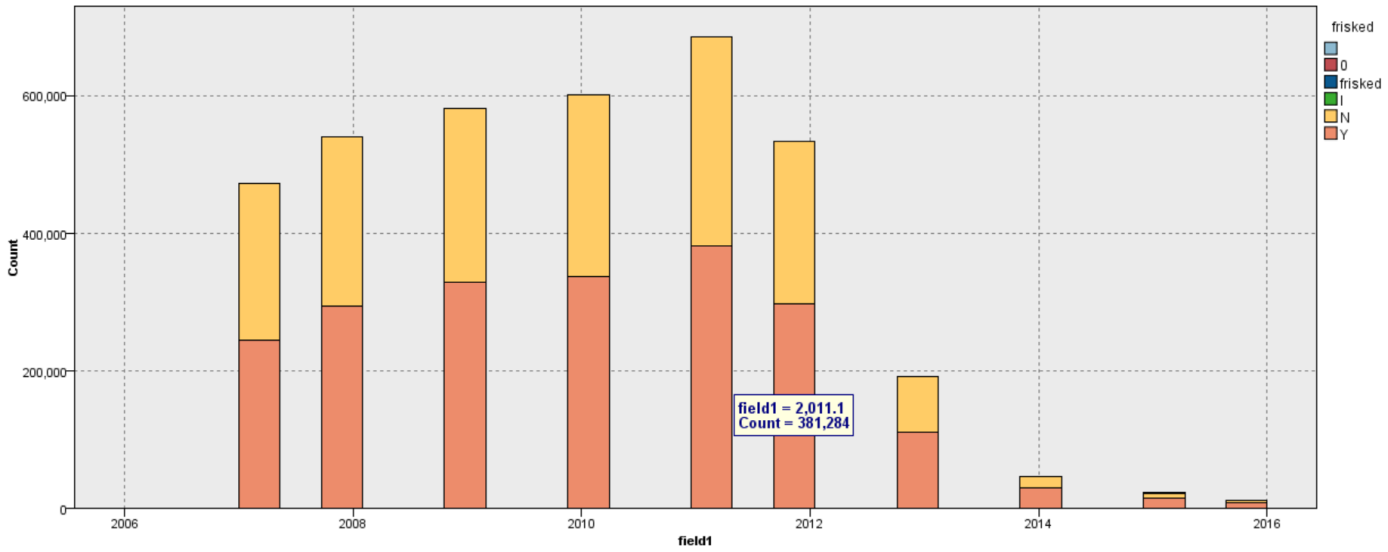


Fig16: The number of Yes and No variables by each year of 10 years.

In summary, we have the statistical frisked attribute through 10 years as the table result (Fig 17)

As the result table, it has given us the highest values of the years of 2009, 2010, 2011 of the bold black numbers by the number of frisked (Yes variable have already ticked on the record) by police officers. Obviously, the number of frisked has been increasing via 2007 to 2011 and has been strongly decreasing by then years of 2013 to 2016 while the values of 2008 and 2012 are the same of the number of frisked values. This indicates criminal cases was decreased by the duties and responsibilities of NYPD as training of police, criminal people activity occurred, and security, educating on people and areas, and by the experience from the information of the prior of dataset. Under the conclusion of the comparing of final results, the number of frisked is 55.6% while no frisk is 44.4% in 100% of the responsibilities of police officers on the task of Stop and Questions as shown in the summary table below.

Years	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	**Total %**
YES (Frisked)	244,564	293,399	**329,317**	**336,856**	**381,284**	296,992	111,500	30,289	15,221	7,912	**55.6%**
NO (Frisk)	226,900	246,280	251,073	263,722	303,884	235,575	80,181	15,427	7,300	4,457	**44.4%**

Fig17: The summary table of the number of Yes and No variables via each year.

Dotplot of Frisk by Year

Next, we work on dotplot of distribution of frisked by year that gave us the colored points through years in Yes and No status of frisked and the most values of 2009 to 2011 of the start, triangle, and rectangle. This visualization in the same mind of the bar chart above, however, as a new progress step, we would like to focus on a new graphics to understand the frisk attribute in depth(Fig18).

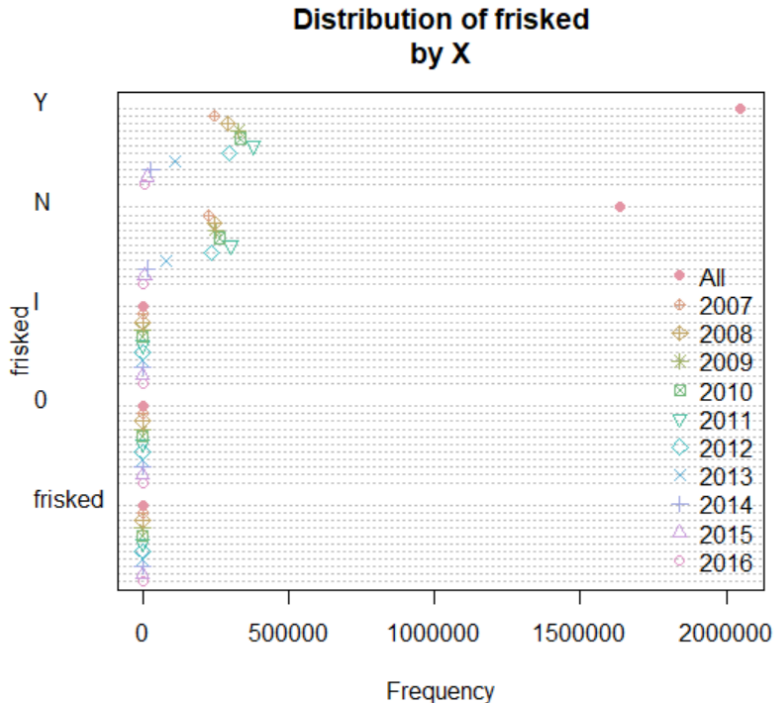


Fig18: Dotplot of Frisked by years

In another way, the graphic has shown the consideration of frisked in a new way to see their values in the range of 0 to 500,000, and see the highest values of the best ones is in 2008 and 2011, also look fast in all of a fast looking of 10 years with a variety of fashions of each year, and we can fast look in total of Yes and No variables on the graphic. The two red points on the dotplot that indicates exactly values of these variables.

3.2 Histogram and Mosaic plot of Frisk by Arrest made

Over using the histogram from Rattle software, we discover the frisked attribute in the 10 years with the two variables of Yes and No in the clearly figure as a looking at the year of 2012 with 381,284 times of Yes. This bar char gives the information about frisked by arrest made in 10 years, because of this graphic, we know people were arrested after they were Stopped, asked some Questions, and Frisked along with the reason of using weapons, or relating to criminal activities.

Regarding of the variable of Yes and No based upon Frisked, we have two statuses for Arrest made including total bar in pink one, No of arrest made in blue bar, and the purple one of Yes in Arrest made actions. In order to have deeply understanding of this graphic, we would arrange them in the comparing table as a good looking. For instance, getting with the Yes one in the frisked with 2,047,334 cases in pink bar but they have been divided to two bars of No one in blue bar of Arrest made is 1,848,849 cases, and the purple bar of Yes of Arrest made is 198,485 case respectively. Similarly, we also can determine on No variable of Frisked for Arrest made.

(Fig 19).

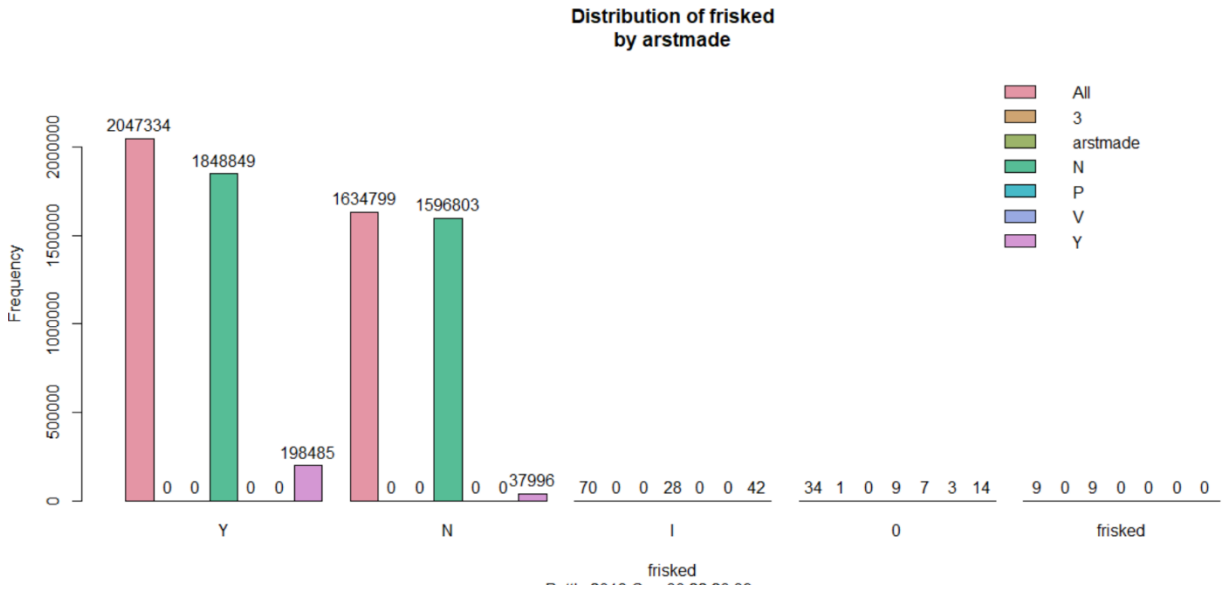


Fig 19: The bar chart of Frisked by Arrest made.

Lastly of this case, we have the result table that present the number of variable comparing between the two attributes as follows (Fig 20).

Frisked			**Note**
YES	Frisked	1,848,849	
	Arrest Made	198,485	Arrest People
NO	Frisked	1,596,803	
	Arrest Made	37,996	Arrest People

Fig 20: The results of Arstmade by Frisked

With conducting of the table above of distribution of frisked by arrest made, we would compare and determine the cases involving Yes or No in the frisked and the arrest made. There are 198,485 (9.7%) people in arrest made and frisked (Yes-Yes), with 1,848,849 (90.3%) are not in arrest made but are frisked (No-Yes). Contrarily, there are 1,596,803 (97.7%) people are in not arrest made and are not frisked (No-No) but 37,996 (2.3%) are in arrest made, and are not frisked (Yes-No).

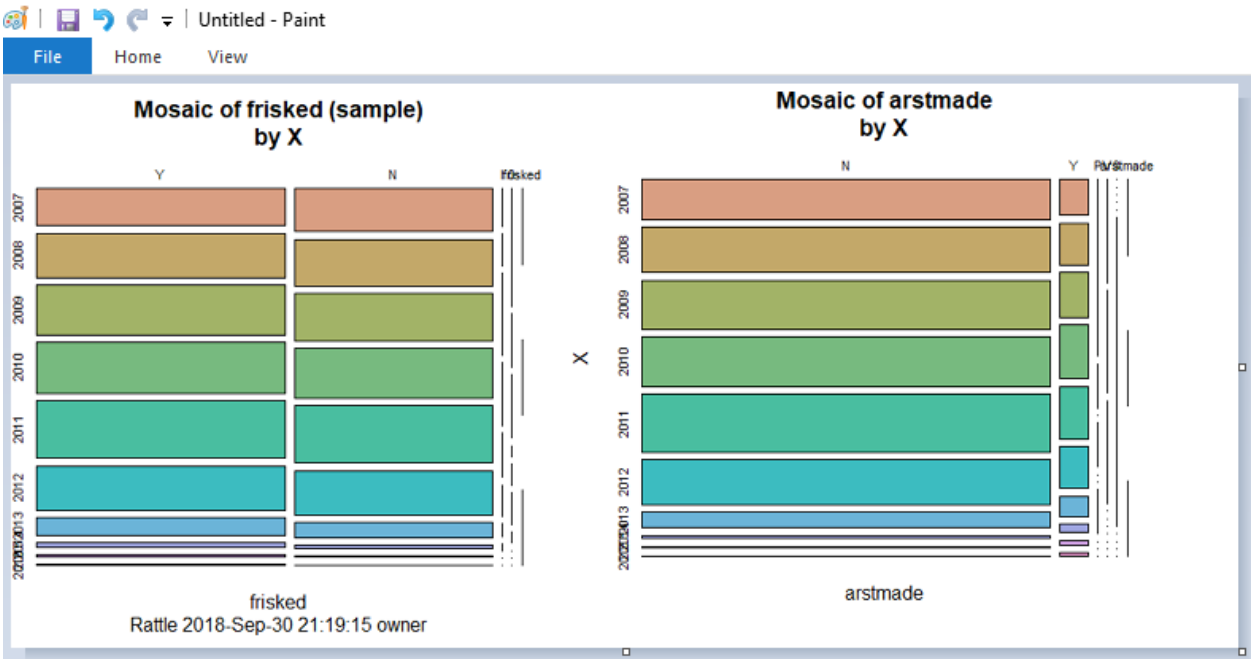


Fig 21: Mosaic plot of Frisked and Arstmade attribute.

In the left of Mosaic graphic of Frisked of 10 along with the Yes-No of this attribute, this was giving us new looking way of Mosaic type as detailed in the different color in the most values of from 2007 to 2012 of both variables, but the values decreased in 2013 to 2016 on both of Yes and No, and we can take a look faster with the Yes values are bigger than about 10% respect to No values. As prerequisite of the problem, we used Mosaic graphic as a looking fast which is similar

to dotplot plotted graphics. Moreover, we also can see very small values of I and O variables above as very small black lines (Fig 21).

In an additional state of part in the left of Mosaic graphic of Arrest made in 10 years following with the Yes-No variable, this was giving us a new insight way of Mosaic type as detailed in the different color in the most values of from 2007 to 2012, but the values decreased in 2013 to 2016 on both of Yes and No, and we can take a look faster with the No values are smaller than about 90% respect to Yes values. Actually, Mosaic helps us to see the graphic as a looking fast. Besides, we also know very small other values of variables above in very small black lines (Fig 21).

3.3 Histogram plot of Frisk by Record status

Record status attribute is to know as the state of someone who recorded on own their actions by police officers such as variables of 1, A, and 9. The Record paper has been offered to someone by police officers based upon the note from Stop, Question, and Frisked works.

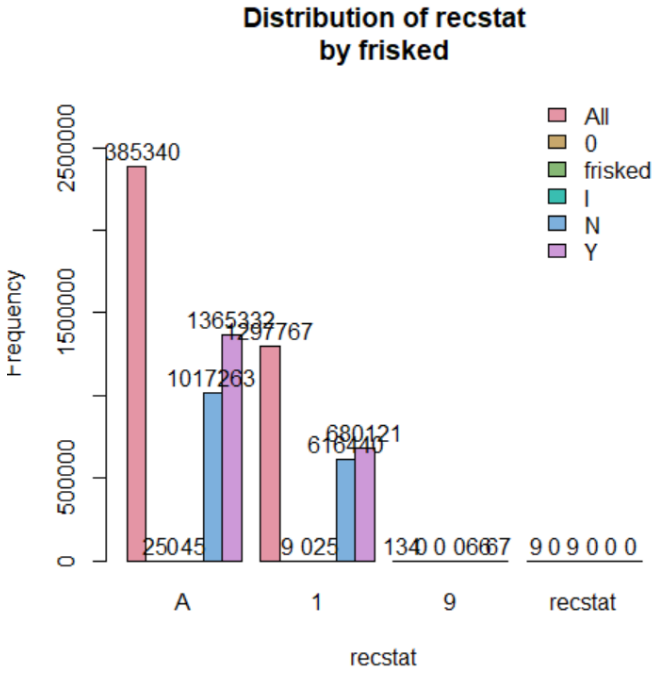


Fig 22: Distribution of Recstat by Frisked

In order to look the Frisked of people in NYC by Record status based on A (No suspect crime) and 1(suspect crime) variable; 9 is another variable but very small (Fig 22):

For A variable, in total of 2,385,340 of Record status in which there are 1,365,332 cases in purple bar with frisked actions by police officers (57.3% of Yes and 42.7% of No frisked status with 1,017,263 cases in blue bar).

For 1 variable, in total we have 1,297,767 case in pink bar in which there are 680,121 cases with frisked actions by police officers in purple bar (52.4% of Yes status and 47.6% of No status in blue bar). In this suspect crime, we recorded people had been frisked are higher than other ones with

52.4% - 47.6%= 4.8%. As following this variable, we recorded 100% of suspect criminal in record status by the paper form which offered by police officers to people but police officers frisked with 52.4% people and 47.6% people who have not been frisked of using weapons, criminal tools. 0 and I variable of Record status are very small so we cannot check the state of Frisked.

3.4 Histogram plot of Race by Year

Race status or Racial attribute is an important component in more than 100 attributes, it indicates the influence to Stop, Question, and Frisk problem as interest as the criminal state. This attribute also gives us to recognize about racial people such as Black, Hispanic, White Hispanic, White, Asian-Pacific, Indian, etc. with the letters of Race column as M, F, B, Q, W, P, Z, A, U, I, etc.

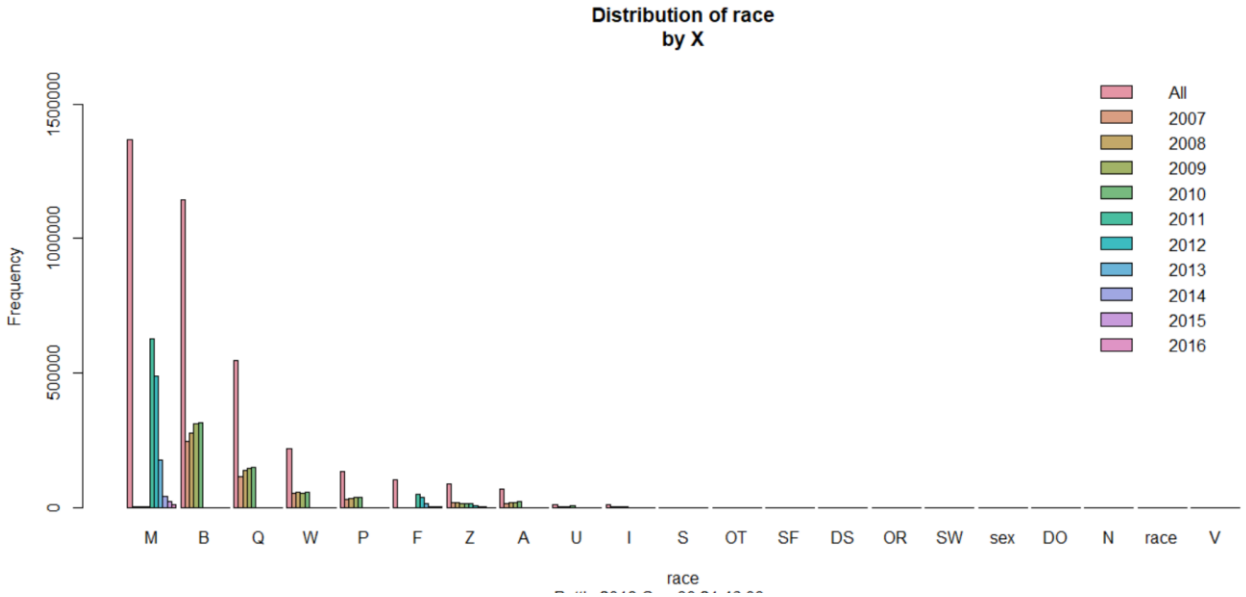


Fig 23: Distribution of Race by Years

The figure shows Race (Fig 23) by year from 2007 to 2016, we can see the most values in Black, Q, White, Pacific people in the highest values of years 2009 to 2013. The statistical values give us the generous state of racial activities in NCY. This histogram shows variables completely to consider via each year of all related attributes to getting eight looking sub-bar charts on the figure to estimate to y- axis for values of frequency.

To continue upon Race attribute in this research works, we arrive a new significance as below is the comparing of race by frisked with the highest values of race such as B, Q, W, and P racial variables that stayed with Yes of Frisked to see relationship between both attributes and work on the impaction of race to frisk work (Fig 24).

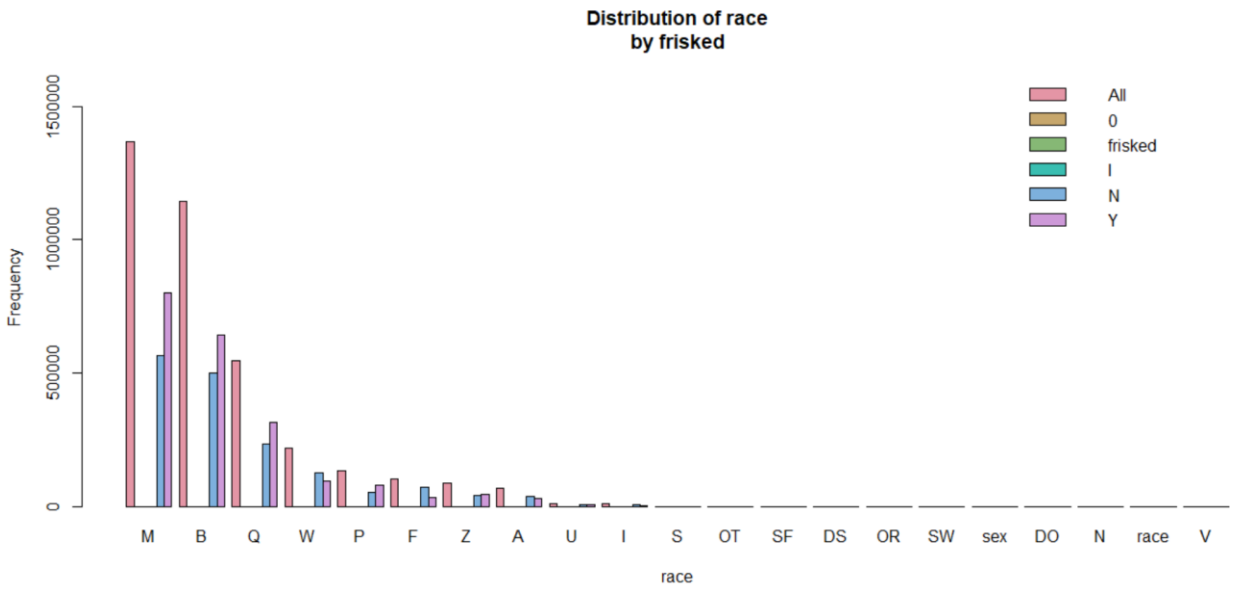


Fig 24: Distribution of Frisked by Race.

For M, B, Q, W variable are most values of the Race attribute with the highest values around 1,400,000 as the longest bar. In purple color bar, we have the frisked case of Yes variable, and in blue bar, we have No variable of not be frisked of actions. Mostly, the number of frisk occurred is greater than the number of No variable in each racial variable of Race attribute. In pink one as we know, they are sum of purple of Yes variable and blue of No variable.

For P, F, Z, A variable are low values of the Race attribute with the highest values around 100,000 as the showing of bar chars. In purple color bar, we have the frisked case of Yes variable, and in blue bar. As the figure, we exhibited the decreasing of bars from highest bar of M values to the smallest variable of V. Due to this arranging, we have understood the properties of each variable and had interesting to the results what is important or not. The remain of variables as from U to V by following the bar charts of Fig 24 are unnoticeable to forward to report data.

3.5 Distribution of sex, age, haircolor, and eyecolor by frisked (Fig 25)

Sex: As the histogram of the top left og the figure Fig 25, there are mostly people are **male** in total of more than 2,000,00 people of both who had been frisked with 94% by police offices during female had small number of frisked with 6%. In addition, the number of frisked cases in purple bars are higher the number of not frisked case in blue bars

Age: The top right bar char shows people had been frisked were increase from 40 to **less than 17 years** old with the greater values are around 700,000 people, criminal people are mostly young people in other variable, and a variety range from 14 to 70 years old in NYC. However, this histogram is only summary of statistical values, we can entire looking on distribution of age by frisked in a separately bigger single bar char by choosing on age attribute to discover it.

Haircolor: Focus on the most haircolor of who had been frisked are **Black** color (52%) to determine Racial people via haircolor with the highest people of 1,700,000 cases in total in which around 800,000 of Black color of both Yes and No of Frisked. We can conclude racial people by following this result with Brown color, Blond hair, etc.

Eyecolor: Similarly, as analyzing haircolor, we arrive with eyecolor attribute in the bottom right of the figure that had been frisked are **Brown** color (58%) focus on Black, besides Green, Brown eyecolor people. This is an interesting and important problem to make conclusions of evaluating and analyzing tasks in the population communities. We can also deduce racial attribute people by following this result to widely discover forward analyzed problems.

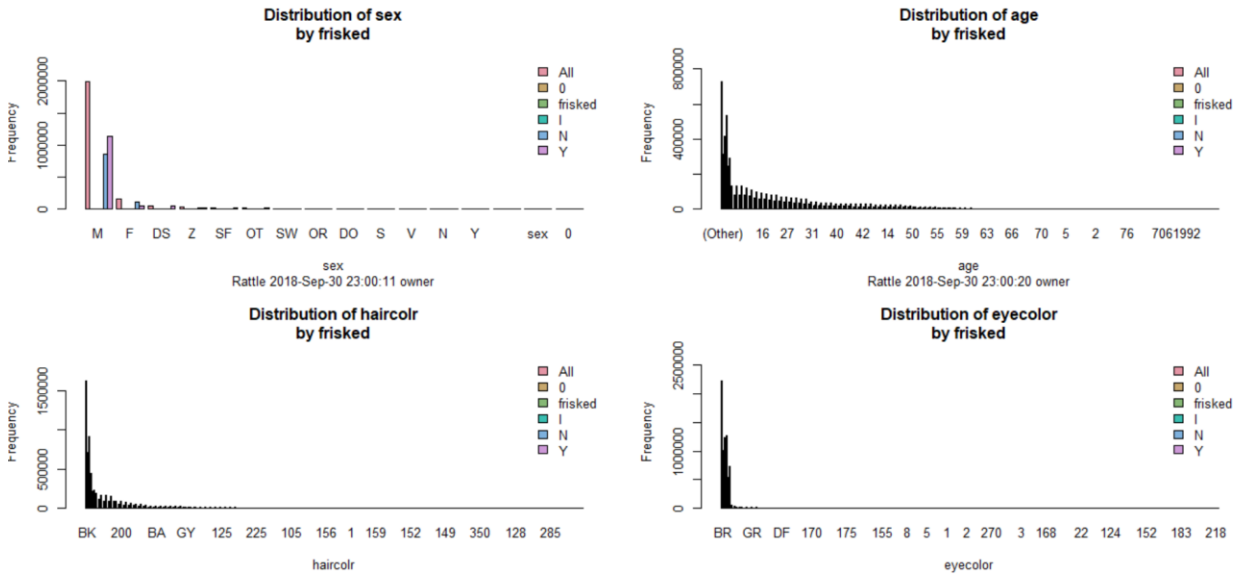


Fig 25: Distribution of Sex, Age, Haircolor, and Eyecolor by Frisked

3.6 Distribution of weapons and reasons by frisked (Fig 26)

Weapons by Frisked

Pistol: From the statistical of figure below, we have 55.4% in 3,670,126 are frisked cases but they are to do not have using pistol during we have 8% of frisked and they have used pistol as weapons with 294,771 cases. 0 and I are small values as the figure. Based on the result, we see in 2,035,240 case of frisked but we have 14.5% with 294,771 cases used pistol, otherwise, used knives and did not use any weapon.

Riflshot: The statistical of the figure, we have 55.5% in 3,675,003 are frisked cases and not frisked 44.5% with 1,631,637 cases, but we have the case of used riflshot with 18,178 cases in frisked with 0.9% and used riflshot but not frisked with 22,811 cases.

Machgun: We have 2,039,969 frisked cases with 55.5% in 3,675,008 cases. There are 3642 either machgun and frisked cases. And 22,300 cases of used machgun but the cases are in frisked. By analyzing of weapons, we know the kinds of used weapons including guns and other ones.

Knifcuti: There are 2,010,426 (55.1%) in total 3,643,631 frisked and 1,629,263 cases with not frisked and all of 3,643,631 case did not use knife during and there are 0.81% of frisked and have use of kinifcuti.

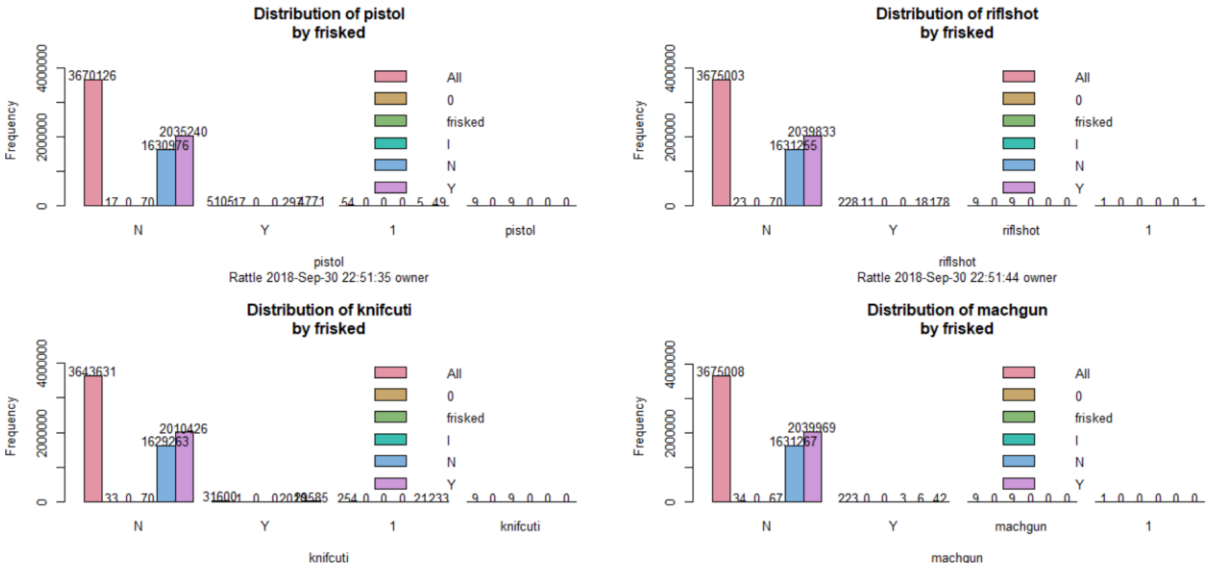


Fig 26: Distribution of weapons by Frisked.

Regarding of using weapon problem as the detailed presentation above, we can summate them in a result table:

Frisked	Pistol		Riflshot		Knifcuti		Machgun	
	YES	NO	YES	NO	YES	NO	YES	NO
YES (Frisked)	294,771	2,035,240	18,178	2,039,833	20,585	2,010,426	3,642	2,039,969
NO (Frisk)	510,517	1,630,976	22,811	1,631,256	31,600	1,629,263	223	1,631,267

We proceed analyzing works on the other attributes such as is in/out, crimsusp, contrabn, detailcm. This means that criminal activities occurred in or out house, subway, markets, etc. Crime suspect means police officers require someone Stop, Questions, and Frisk since they have suspected criminal problems (Fig 27):

For Out status: There are 1,682,799 cases (60%) of Frisked that occurred at outside and 1,159,391 with no Frisked (40%) of at outside. By the percent results, we can understand the actions of crime frisked are bigger than outside not frisked so NYDP can plan to work on future duties.

For In status: We continue to discover that are 364,535 cases of Frisked cases occurred inside and 475,407 with no Frisked. Comparing of Out/In status, the outside Frisked (79.4%) cases are bigger than inside Frisked (21.6%) cases. In conclusion, many outside cases are occurred in outside as on streets, road, outside house, subways, markets, etc.

For Contraband status: There are 1,986,782 cases of Frisked but there is no contraband. There are 58,224 cases of Frisked and Yes of contraband in the left bottom of the figure. Of course, someone who were contraband then required to Stop, ask some questions and Frisked to get something which follow to them by police officers.

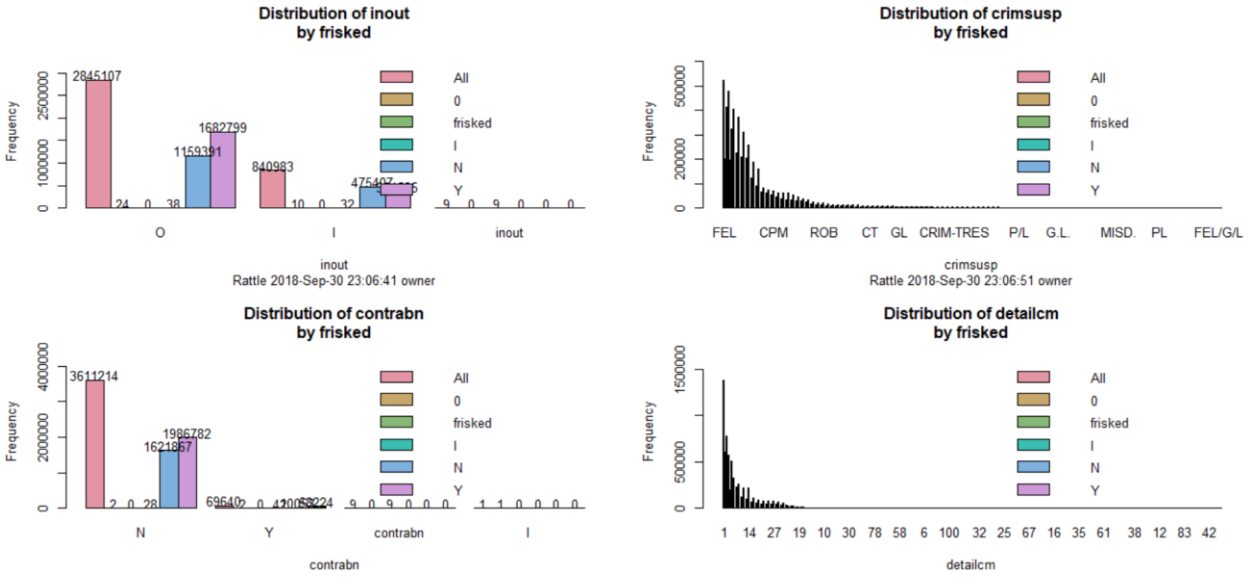


Fig 27: Distribution of in/out, crimsusp, contrabn, detailcm by frisked

Criminal Reasons by Frisked

We proceed to focus on the reasons of frisked, we can take a look on the attributes as violent crime suspected (Rt-vcrim), actions of engaging in a violent crime (Rf-vcact), knowledge of suspect's prior crime behaviour (Rf-knowl), other suspicion of weapons (Rf-othsw) to have a commonly signification to causes to create criminal activities. The classification helps us to know reasons of Stop, Question, Frisked and understand insight of criminal states (Fig 28).

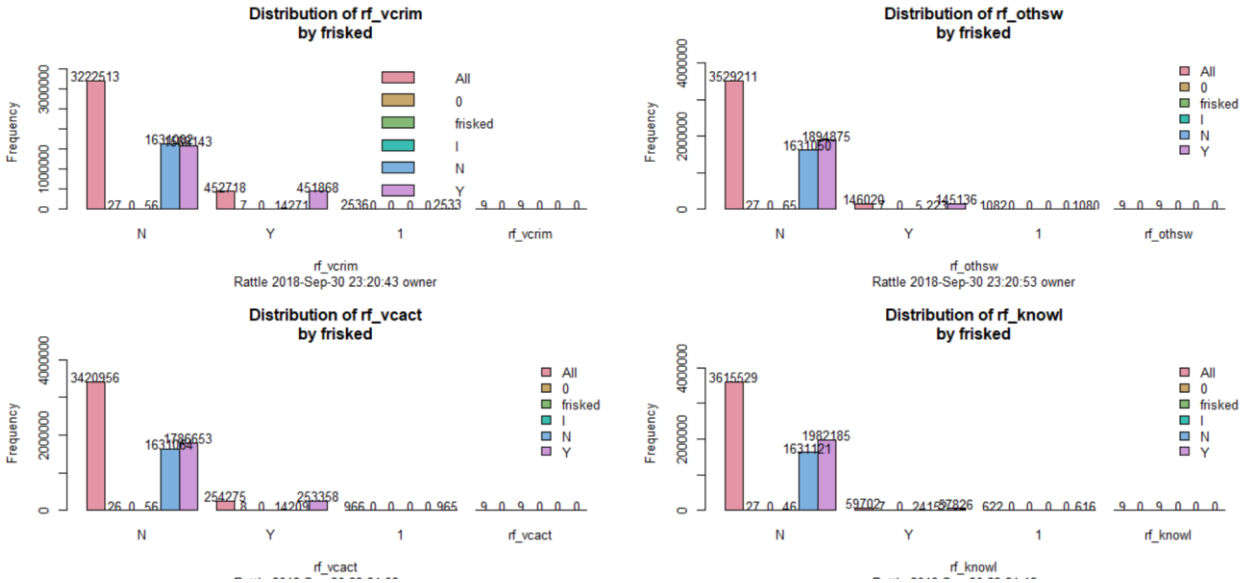


Fig 28: Distribution of the Rt-vcrim, Rf-vcact, Rf-knowl, Rf-othsw attributes by Frisked

As the generated histogram, we need to indicate the final results of the bar charts above in the following table (Fig 29). The showing table indicates causes of criminal actions to someone who were frisked by the kind of criminal reasons including Yes or No have reasons based on variables. The dataset recorded completely of all real life works of police officers in NYC. By the dataset, we have analysis in general of many attributes which can apply to real life with the highest values belong to Rt-vcrim, Rf-vcact variable as many as number of frisked on four variables.

Frisked	Rt-vcrim		Rf-vcact		Rf-othsw		Rf-knowl	
	YES	NO	YES	NO	YES	NO	YES	NO
YES (Frisked)	451,868	1,588,143	253,358	1,786,653	145,136	1,894,875	57,826	1,982,185
NO (Frisk)	14,271	1,631,002	14,209	1,631,064	5,223	1,631,050	2415	1,631,121

Fig 29: The result distribution table of attributes

Physical force used by Frisked

Of course by continuous, distribution of physical force used by officer – **hands, pf_wall** physical force used by officer - suspect on ground, **pf_grnd** physical force used by officer - suspect against wall – **pf drwep**-as follows the figure of Fig 30. This is physical force which used to someone by police officers to master someone with the highest values are at pf-hands 650,371 of frisked by physical, next is pf-wall with 77,486 of frisked and physical force based on in total of frisked are more than two millions cases (Fig 30)

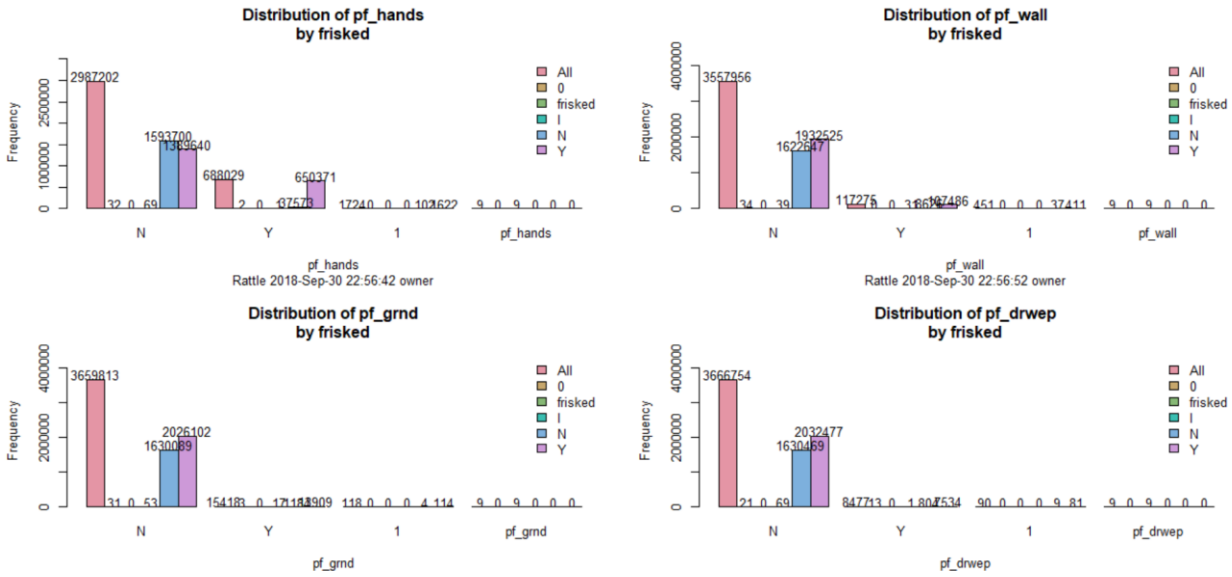


Fig 30: Distribution of pf – hands, pf_wall, pf_grnd, pf_drwep by Frisked7486

Focusing of using physical force as the detailed presentation above, we can summate them in a results table:

Frisked	Pf-hands		Pf-wall		Pf-grnd		Pf-drwep	
	YES	NO	YES	NO	YES	NO	YES	NO
YES (Frisked)	**650,371**	1,593,700	77,486	1,932,525	17,118	2,026,102	7,477	2,032,477
NO (Frisk)	37,578	1,369,640	34,234	1,622,647	134,090	1,630,089	1805	1,630,469

3.7 Analysis of some attributes by 10 years

Analysis of perstop at areas by 10 years

The Big Dataset was collected information on the five areas in New York City as Manhattan, Brooklyn, Bronx, Queens, Staten Island as the map shows in Tableau software. Our responsibility are works on the five areas in NYC to discover each them on the map of initial considering for locations of the dataset in different colors (Fig 31)

Next, using of the map to shows to perstop attributes by years between 2007 and 2016, the figure indicates number of stop of people in 10 years at the areas above. At each area of data, we have values in total of 10 years so it is easy to see them following by location.

The map gives us the number of perstop at Manhattan is 2,723,559 case, Brooklyn is 3,956,204 cases, Queens is 3,042,996 cases, Staten Island is 487,992 cases, and Bronx is 1,725,202 cases in 10 years. As a statistical result, we can understand the juncture of crime at those areas with the

highest value at Brooklyn area as the most complex location in comparing to other ones, next, we also have Queens and Manhattan respectively as a complex area in criminal problem. Finally, we have the smaller values of Bronx and Staten Island area and it is feel safer than other ones

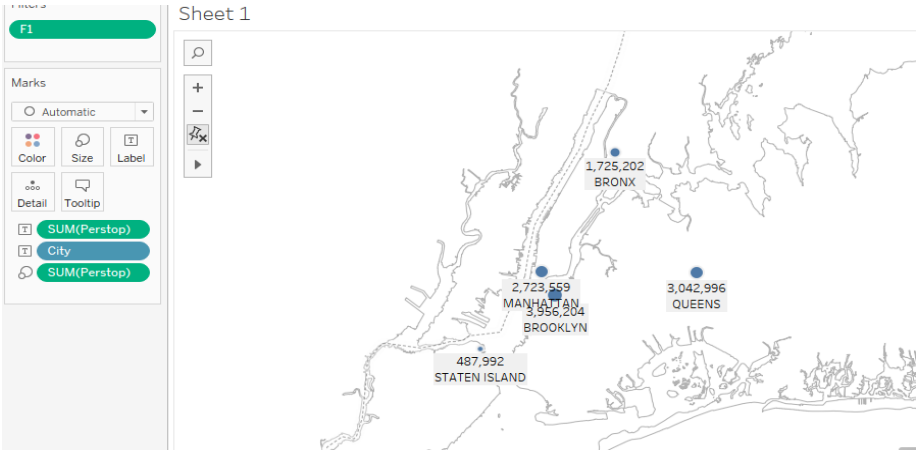


Fig 31: Analysis of perstop by 10 years

Applying of perstop attribute by mapping area as the detailed presentation above, we can summate them in a results table (Fig 32):

Areas	Brooklyn	Queens	Manhattan	Bronx	Staten Island
Values	3,956,204	3,042,996	2,723,559	1,725,202	487,992

Fig 32: Analysis Table of perstop by 10 years

3.8 Distribution of Weight, Age, Height, and Observation

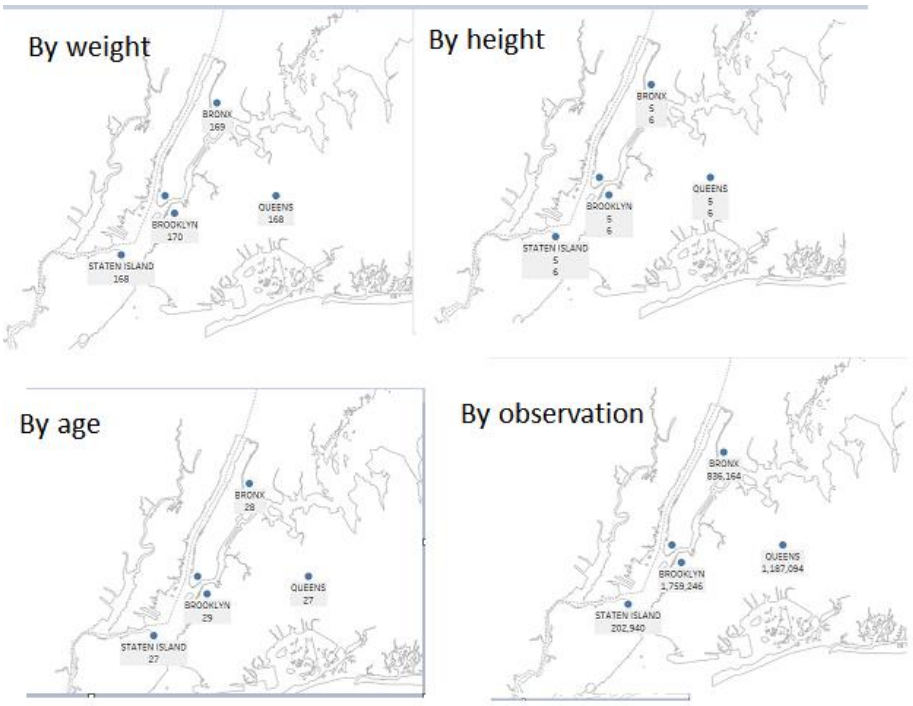


Fig 33: Distribution of the average values of Weight, Age, Height, and Observation.

Weight: The weight of people in five areas with their average weight are 171 pounds for Manhattan, 170 pounds for Brooklyn, 168 for Queens and Staten Island, 169 pounds for Bronx area. Based on the weight of people, police officers know to prepare methods to govern felon by physical forge as the useful actions.

Height: As the histogram, we have the height of people in five areas with their average height are 5 feet 6 on people analyzed. This is also a portion to check to the ability of police offices because they need to have insight of related problems to forward the actual statuses. Probability, the height attribute had much appreciation to the strategy action of the NYPD.

Age: Because of the age of people in five areas, we can deduce the strategies to come near some and understand population community along felon. Typically, their average age are 32 years old for Manhattan, 29 for Brooklyn, 27 for Queens and Staten Island, 28 for Bronx area.

Period of observation: This portion is number of observation of someone by police officer in each area. This helps officers to observe someone in their areas. Totally, the number of observation of people in five areas are 1,645,574 for Manhattan; 1,769,246 for Brooklyn; 1,187,094 for Queens; 202,949 for Staten Island; 867,164 for Bronx area.
To proceed to evaluate, we can get an average value for an observation day of10 years will be 3.48 times for Manhattan; 2.45 for Staten Island; 2.41 for Queens; 2.34 for Brooklyn; 2.30 for Bronx as the showing figure (Fig 34):

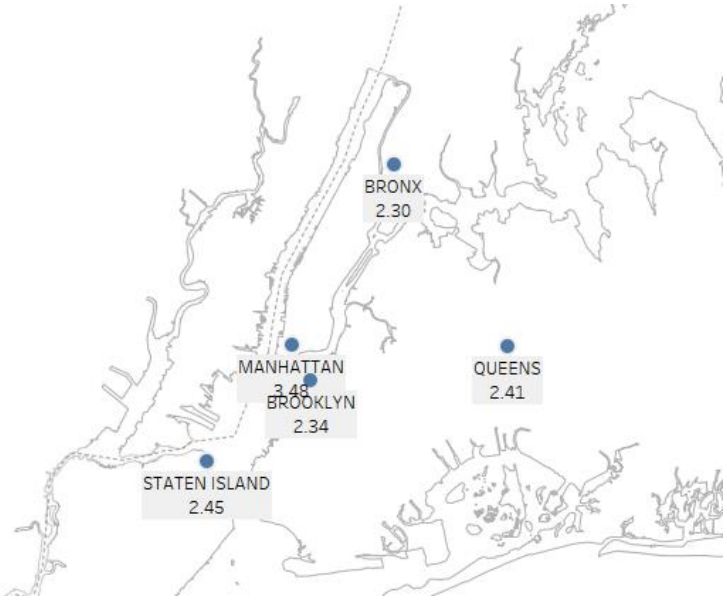


Fig 34: Average observation in 10 years

We can find the average of period of observation in a day is 3.48 times for Manhattan, more 2 times for other ones.

Determined result data, we had analytics of the areas in NYC, where is important with high parameter record, where is complex problems to have preparations in the interested works.

3.9 Combining of attributes on some charts

Analysis of Record Status of 10 years by bar chart in each year

Analysis of restat by 10 years

Considering of the restat attribute includes four variables are A, 1, 9, and NA but there are two main variables of A and 1 variable to analyze on the distribution of restat by 10 years (Fig 33). This is the status of people who have recorded about their state based on Stop, Q and F by police officers.

The distribution with highest values through the years of 2010 -2012 with pink and blue color of main variables, the variable of 1 decrease via years of 2010 t0 2016 and A variable decreased from 2011 to 2016. The 9 variable is the highest in 2011 and have appear in 2015 but not in other years.

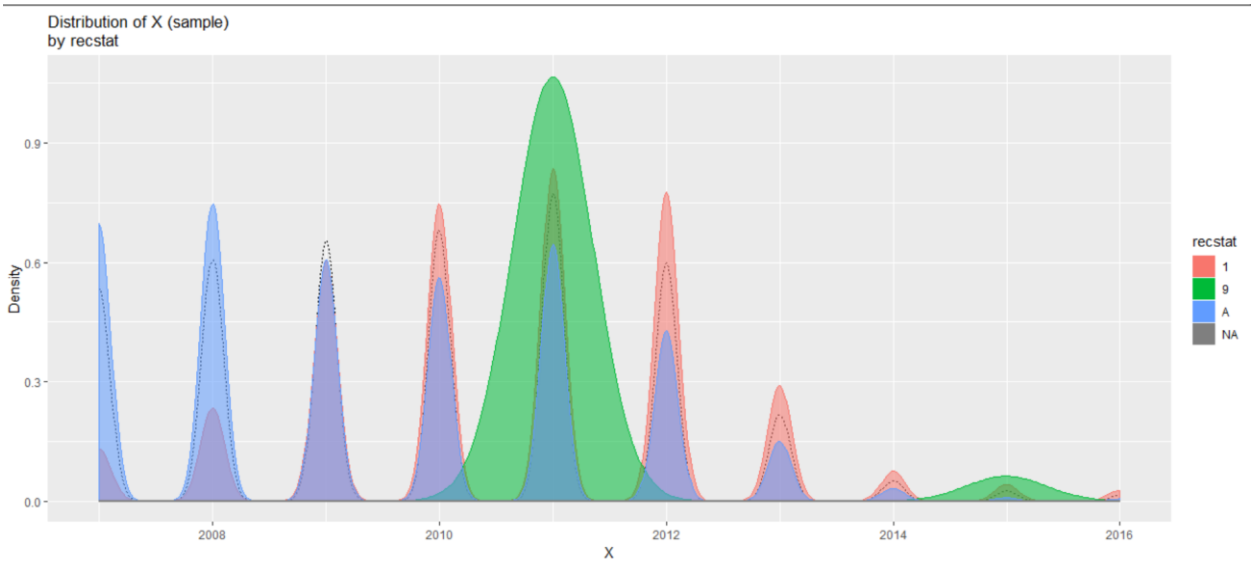


Fig 33: Analysis of restat by 10 years

Analysis of restat by each year and main attribute

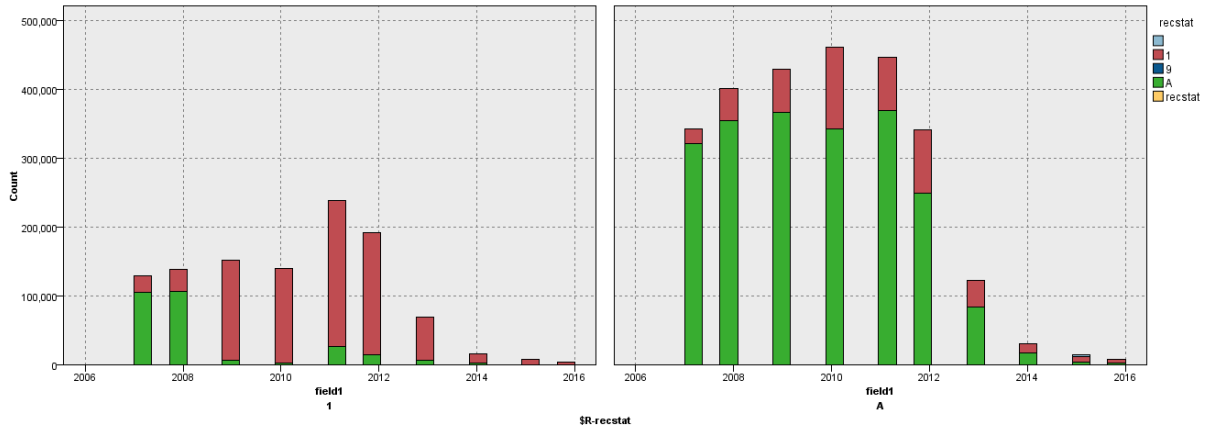


Fig 36: Analysis of Record Status by 10 years

Record Status has 4 values including 1, 9, A and NA but we focus on 1 is original value 1, 0 is original value A this means that 1 is Yes, 0 is No for crime status recorded by police officers.

Blue bar color shows the original of A values equal to 0 that is No for crime record status by years in 10 years.

Red bar color shows the original of 1 values equal to 1 that is Yes for crime record status by years in 10 years. We would make a comparison between years as the histogram indicates the highest values in 2010 and next are in 2012, and then are in 2011, 2009, 2008, the smallest values are in 2016, 2015 and 2014. By this observation, we could finalize the junctures of crime are decreasing throughout years from 2012 to 2016 but are increasing from 2007 to 2010.

Analysis of PerStop of 10 years by line chart (Fig 36)

In a new way, we plotted in the line char which gives us the number of perstop of whole five areas in a year but in the average of 10 years, we have Manhattan is 2,723,559 case, Brooklyn is 3,956,204 cases, Queens is 3,042,996 cases, Staten Island is 487,992 cases, and Bronx is 1,725,202 cases. As a statistical result, we can understand the juncture of crime at those areas with the highest value at Brooklyn area as the most complex location in comparing to other ones, next, we also have Queens and Manhattan respectively as a complex area in criminal problem. Finally, we have the smaller values of Bronx and Staten Island area and it is feel safer than other ones

Stops though each of all 10 years by police offices. Obviously, there have many more the number of stops that increase is from 2007 to 2011 and then the line char decrease from 2012 to 2016 but there are the highest values in 2011 with 3.817,800 cases by the figure:

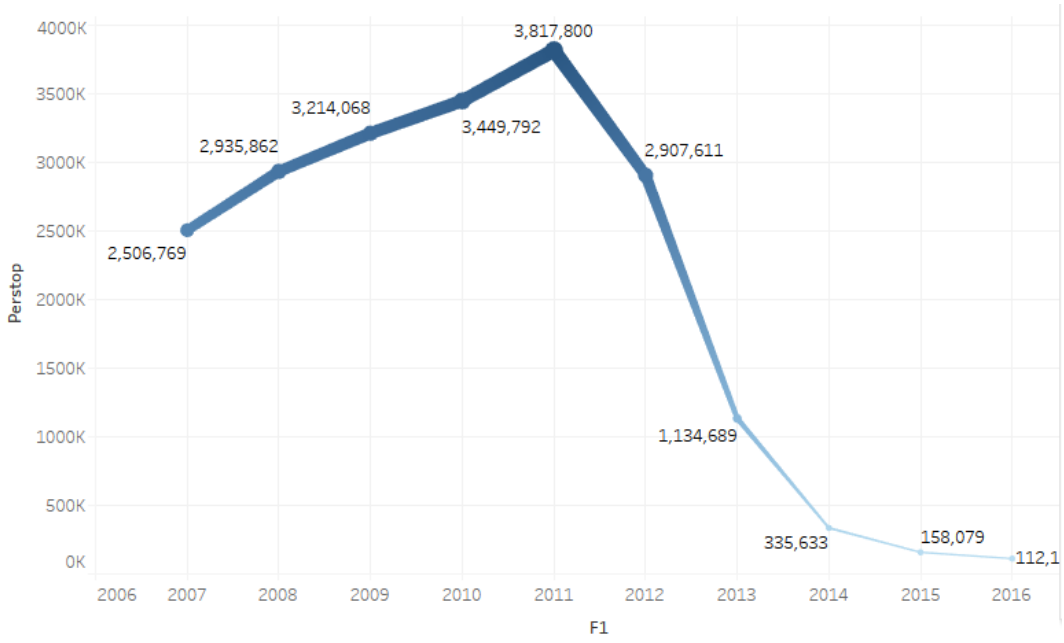


Fig 36: Distribution of PerStop of 10 years by line chart

We can also use average feature to find the average stop in a day of each year from the graph above by the table (Fig 37):

Years	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Total in a year	2,506,769	2,935,862	**3,214,068**	**3,449,792**	**3,817,800**	2,907,611	1,134,689	335,633	158,079	112,150
a day	6867	8043	8805	9451	10459	7966	3108	919	433	307

Fig 37: Distribution of perstop in 10 years and the average values in a day of each year

Analysis of Search of 10 years

By doing so, there is clear understanding of what are they as frisked attribute, we can see searched in the creasing through 2007 to 2011 and decreasing from 2011 to 2016 (Fig 38) with green of No variable and in pink one of Yes variable and we get number of variables of Yes and No by moving the mouse on the histogram of each year. This bar chart is easy to look each values on each bar of each year.

In 2011 is the biggest value year of Stop, Question and Frisk with more than 700,000 cases as many as having searched people by police offices. Searched values are the same between 2007 to 2012, and then they strongly reduced from 2012 to 2016 as the surprise problems of decreasing values due to strategies from NYPD to master criminal people. The variables of 0, 1 and searched are very small so they have been represented in the bar charts for useful analyzing process, hence we always follow the two primary variables of Yes and No to accomplish problems.

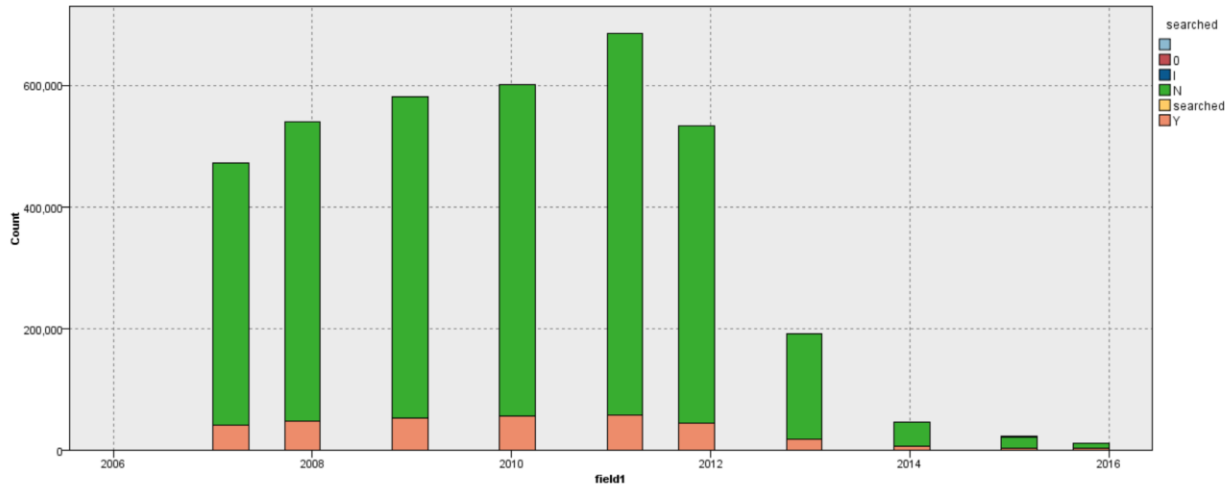


Fig 38: Distribution of Searched of 10 years

Distribution and Comparison of Arrest made and Searched of 10 years

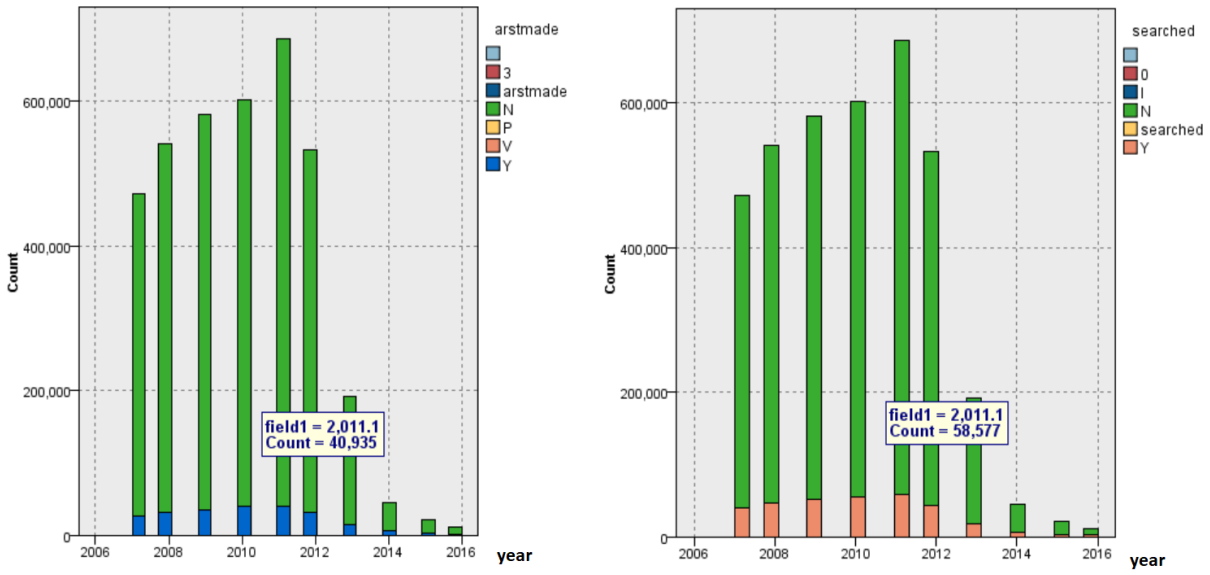


Fig 39& 40: Distribution and Comparison of Arrest made and Searched of 10 years

In order to discover the dataset, we directly focus and compare on the arrest made attribute (Fig 39&40) and searched attribute of the dataset as an additional research.

Arrest made is an action which has been made by police offices to someone who have really criminal problems.

As the above presentation, searched actions have been made to search something from someone in the SQF.

For the histogram of arrest made on the left, we need to use in the main attributes, there are the Yes variable increases each year of 2007 to 2011, and then decrease from 2011 to 2016. Number of arrest made problem depend on the activities of people as frisked and searched by police officers. Values on two bar chart can be found on each bar of each histogram by moving of the mouse. For example, we can see the number of arrest made is 40,935 in 2011 by discovering at position of 2011 bar.

Because of relationship between arrest made and searched, we can see in the same in their variation since number of arrest made increased then number of searched were to increase and contrarily, both are decreasing in the same way, although they had different values in the same year.

By looking on two histograms, we can the variations between blue-green bar on the left and green-pink on the right are similar. No matter what the results of frisked and searched actions, then arrest made actions are follow them, this means, someone have criminal problems, thus police officers will stop, ask question, search, and frisk to forward to arrest them in the necessary time.

3.10 **Distribution of related other separate attributes that follow analyzed attributes above as an additional analysis**

Typically, we discover on each attribute as the figure has two statuses of Yes and No attributes in Frisked and Searched with number of each attributes.

For Distribution of Frisked: There were 2,026,915 frisked people (55.6%) in 3,645,307 people in 10 years at all areas of NYC that include Stop and Question. By finding of average value in each year then we can see 202,691 frisked cases per each year in all five areas, and 555 frisked cases per a day by calculating. Additionally, there are with 70 cases of I, 34 cases of 0, and 9 frisked cases (Fig 41).

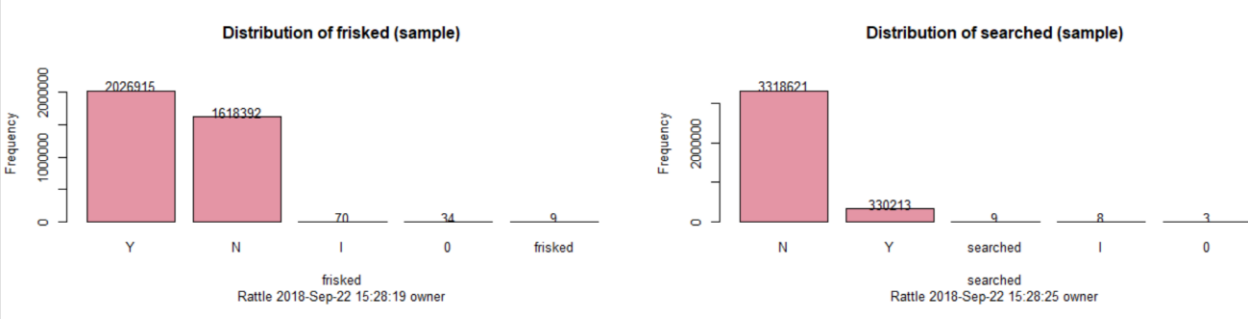


Fig 41: Distribution of Frisked and Searched in 10 years

For Distribution of Searched: There were 330,213 searched people (9.95%) and were not searched 3,318,621 people in 10 years at all areas of NYC. By finding of average value in each year then we can see 33,021 frisked cases per each year in all five areas, and 91 frisked cases per a day by calculating. During distribution of typeofid with 1,989,170 cases of P, next one is 1,536,873 (Fig 41).

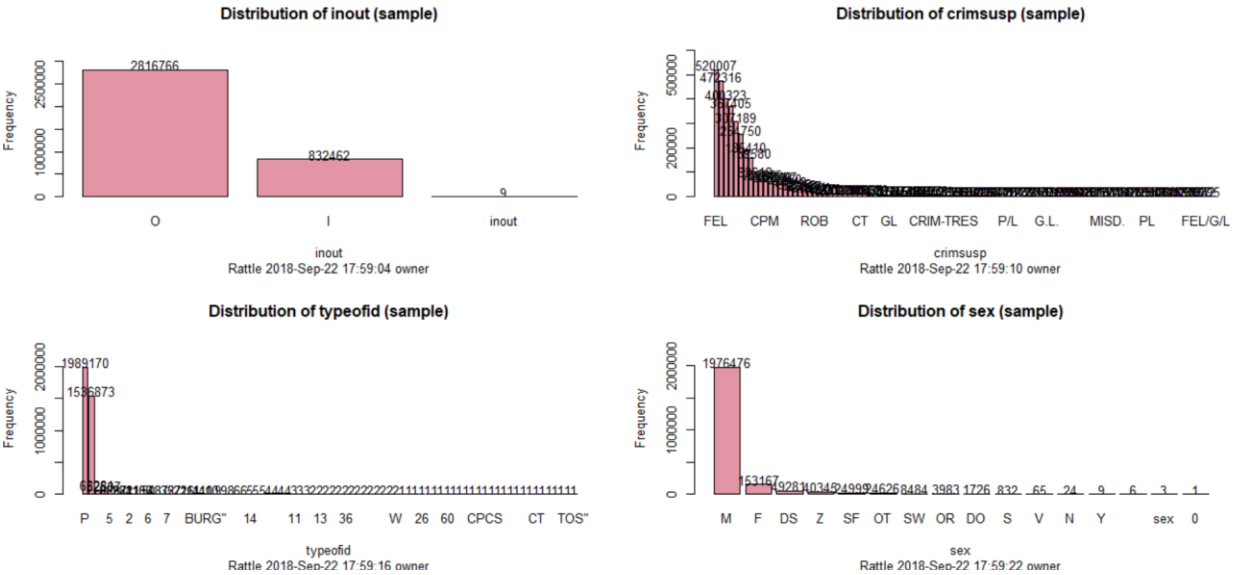


Fig 42: Distribution of in/out, typeofid, crimsusp, and sex attribute.

The statistical plotted histograms above figure out detailed results as in/out of stop people by police with many outraced data of stop out 2,816,766 but stop in is 832,462, this shows criminal activities occurred from out of houses, offices, buildings, subways, etc. Such as that are actions on streets, areas, outlets We also know about the gender of people with 1,976,476 (92.25%) for male and 153,167 (7.75%) for female in at most of Felony, Criminal Possession of Marihuana (CPM), Robbery activities. As the right top of the figure, we can see with 520,007 cases of felon; 472,316 of CPM, etc. On the sex attribute, we have the male with 1,976,476 people and 153,167 in female.

As the out/in attribute to demonstrate the analyzed results above (Fig 42).

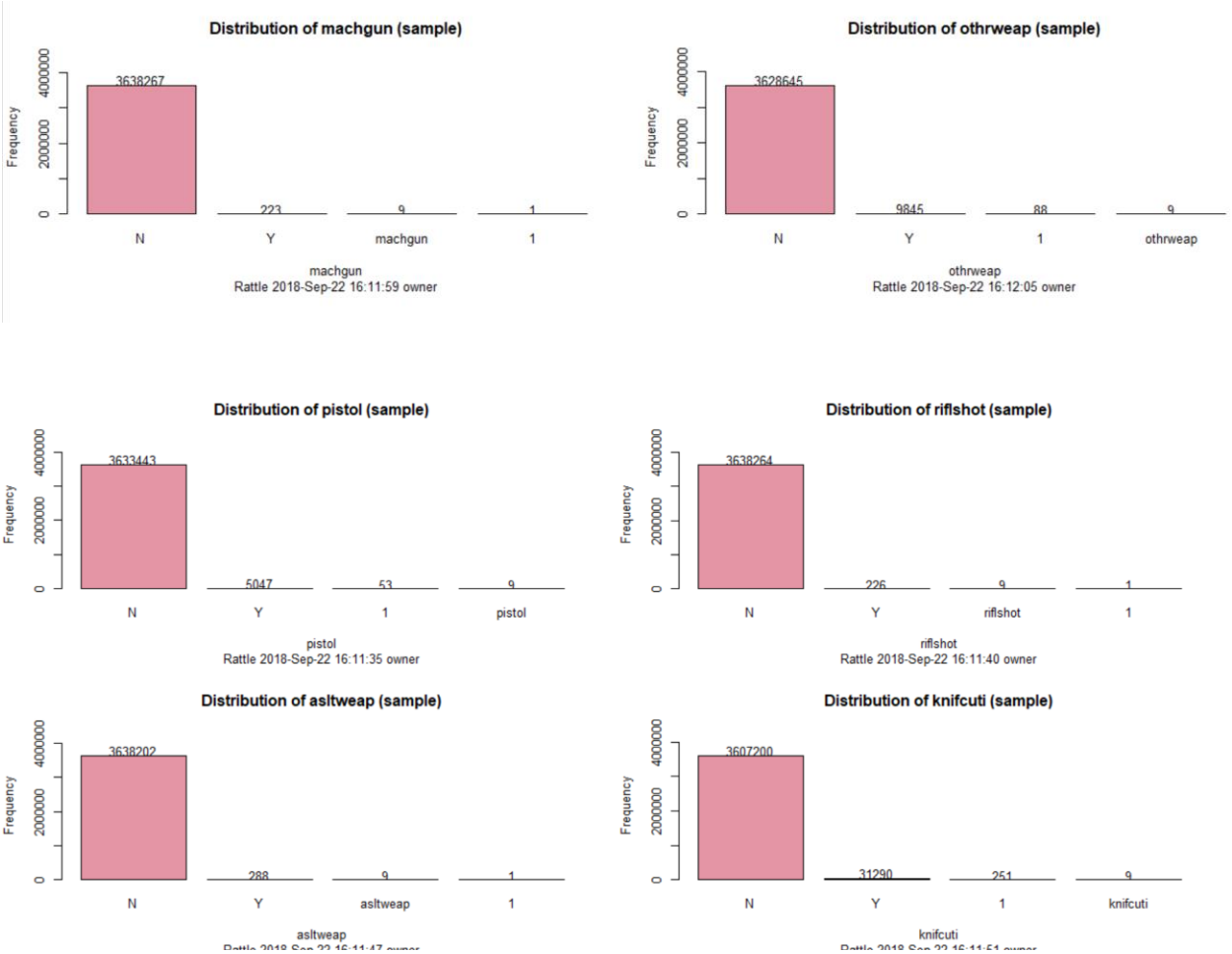


Fig 43: Distribution of weapons of 10 years.

As we have written in the part above, in addition about the use of weapons in detail namely machgun (0.0061%), other weapons (0.271%), pistol (0.138%), rifilshot (0.0062%), assault weapon (0.0079%), knife (0.867%) showed on the histogram in total of 10 years. In other hands, the attributes related to weapon with two choices of Yes and No which we determined in 10 years to understand criminal people, who used knives more than use of guns so we can see and predict dangerous statuses at areas in NYC (Fig 43).

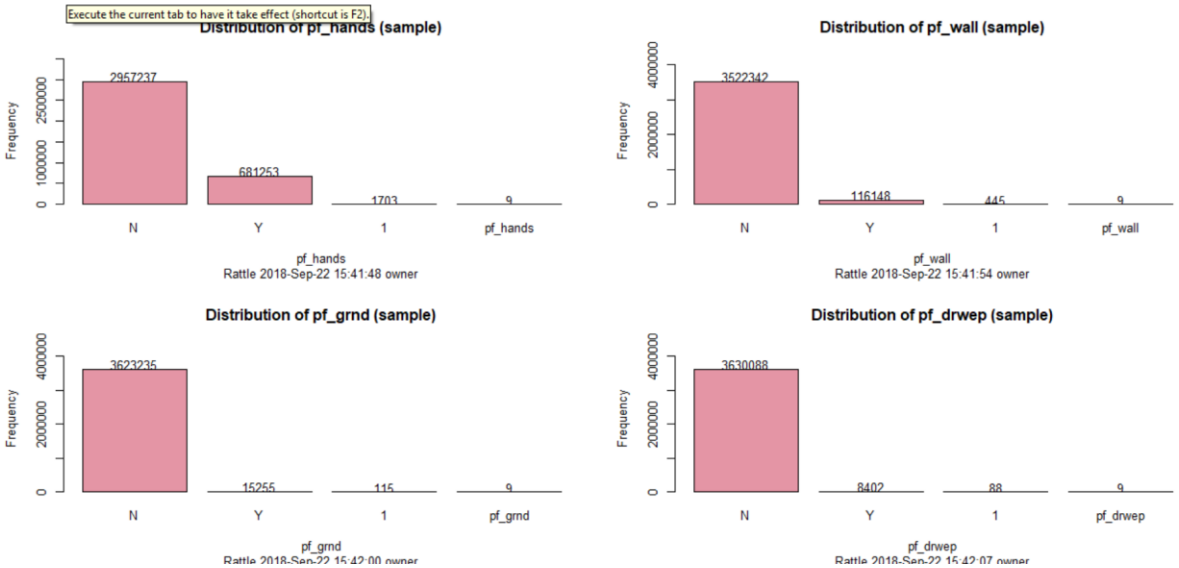


Fig 44: Distribution of pf_hand, pf_wall, pf_grnd, and pf_drwep of 10 years

The results table (Fig 44) gave the physical force used by officer as hands (18.7%), wall, against wall (3.2%), weapon drawn (0.23%) with at most of hands and against wall that applied to people, and they have small effectiveness to target attribute. This figure as is detailed attributes to analyze clearly by number as the showing figure and by evaluating of percent of each.

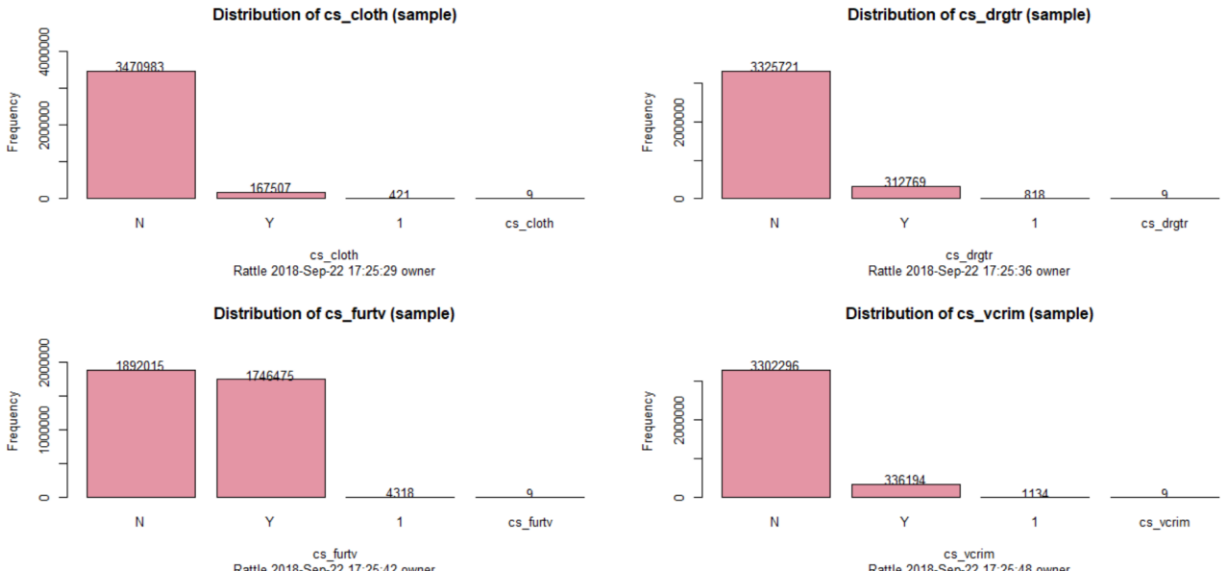


Fig 45: Distribution of cs_cloth, cs_drgtr, cs_furtv, and cs_vcrim attribute of 10 years.

The results of histogram table gave the reason of stop as reason for stop - wearing clothes commonly used in a crime (4.6 % of Yes), reason for stop - actions indicative of a drug transaction (8.6% of Yes), reason for stop - furtive movements (48% of Yes), and reason for stop - actions of

engaging in a violent crime (9.23% of Yes), used by officer applied to people, and they the most values of furtive movement in the reason for stop from police (Fig 45).

Distribution of race by recstat

Again analyzing of the attribute, we can take a look more on the attribute but of race by recstat due to we would like to look an important attribute of race to consider more than in detail. Recstat was recorded to see the state of people by SQF problems but not in frisked also by race on the record paper with primary variables are 1 and A to accomplish on racial people.

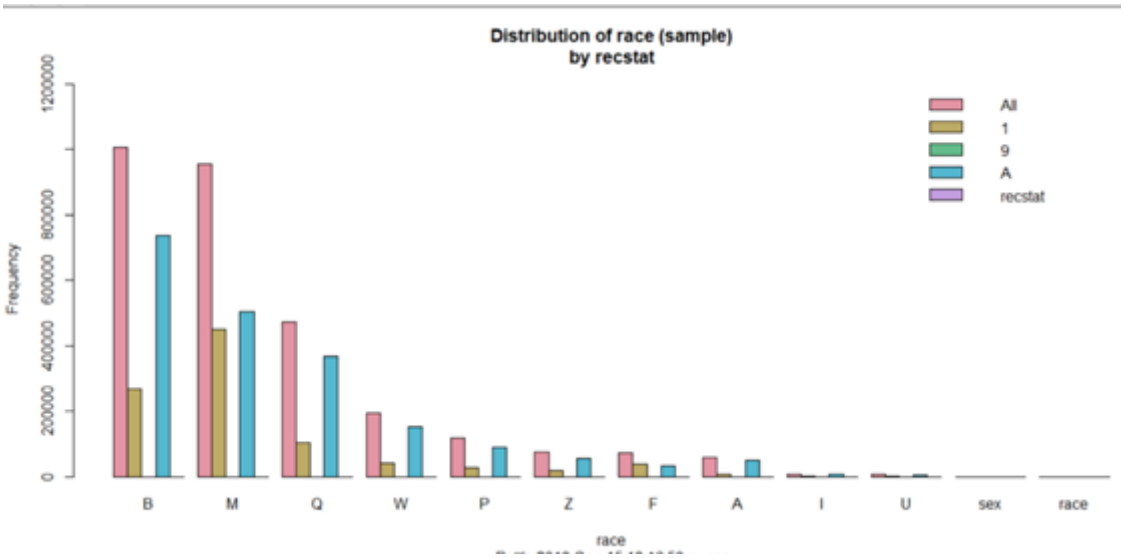


Fig 46: Distribution of race by recstat.

In order to make an analysis of race by recstat but only focus on 1 and A variable of recstat with the highest values of Black people at 1,000,000 totally cases with the 1 variable of recstat is close to 300,00 case and the number of A variable in blue color, next to M variable in total of around 900,000 cases but in 1 case are around 500,000 case, and other ones of Q, W, P, Z, F, A racial variable including of A and 1 variable but number values of A is bigger than number values of 1 all of them are from 1,000 cases to 100,000 cases.(Fig 46)

The detailed race by recstat

So far in a way, the histogram of race by record in another bar chart, the status containing the attributes of 1 and A for recstat on each variable as Black, Black-Hispanic, White Hispanic, White, Asian/Pacific Islander, Indian/ Native Alaskan in two histograms of Yes and No. The figure shows each racial attribute for number of 1 and A status, respectively Yes and No of recstat in each variable of race in many colors to see each portion of race on each of Yes and No (Fig 47).

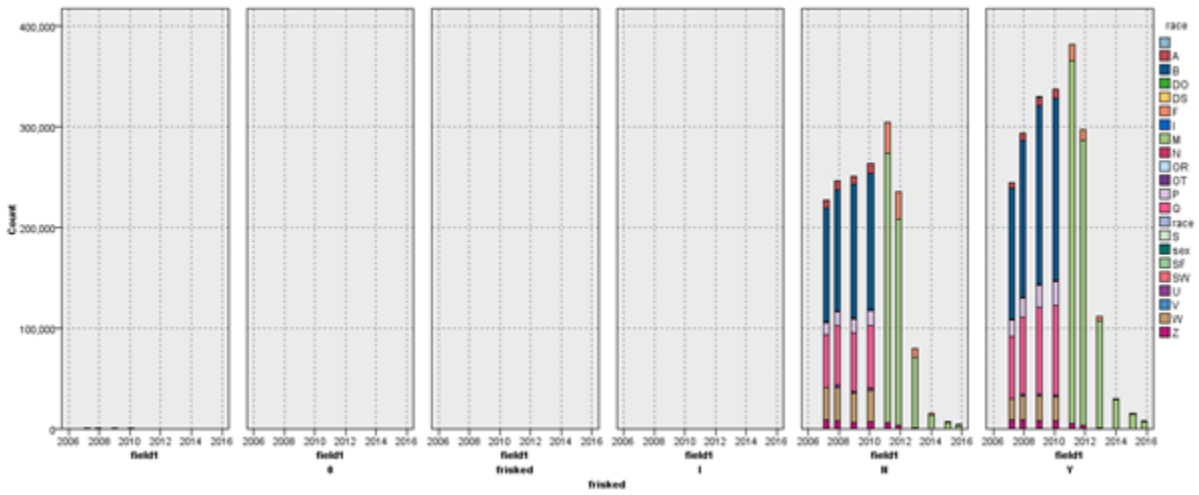


Fig 47: Distribution of race by recstat.

The histogram indicates racial people by frisk with two variables of Yes and No with the attributes of race as Black, White, Hispanic, etc. and these racial attributes show in each year of 2007 to 2016 to number of racial people in each year with each racial people. Besides, focusing on the main variable of Yes and No, we also have other variables such as I, 0, and frisked but those are small so we cannot to see showing values. In other datasets, we can see all variable which can be available on a combining histogram. The division of each portion on race attribute as is an easy kind to analyze problems in depth and visual works in many part of interested statistical bars.

IV. Choosing of Predictive Attributes of Dataset for Plans of Prediction

4.1 Choosing testing and training data

After we have analyzed and accomplished about typically & mainly attributes by the part above. Now we are using a Big Dataset of 10 years between 2007 and 2016 related to the New York Police Department dataset based on Stop, Question, and Frisk practices to predict on a chosen target and predictors.

The purpose of this project would determine and predict with **why people were stopped, asked question, frisked and arrested in NYC?** We have various attributes to support cause a person to be frisked or arrested by Police officers, as well as we also can predict on other people who can be stopped, asked, frisked or arrested depending on their own activities occurred.

For this prediction on analyzed data works, we would use **100 attributes** of columns and **3,686,101 rows** of the Big Dataset with the Target of **Frisked attribute**. In order to predict the Frisked target, we need getting of percentage on the datasets with the rate as:

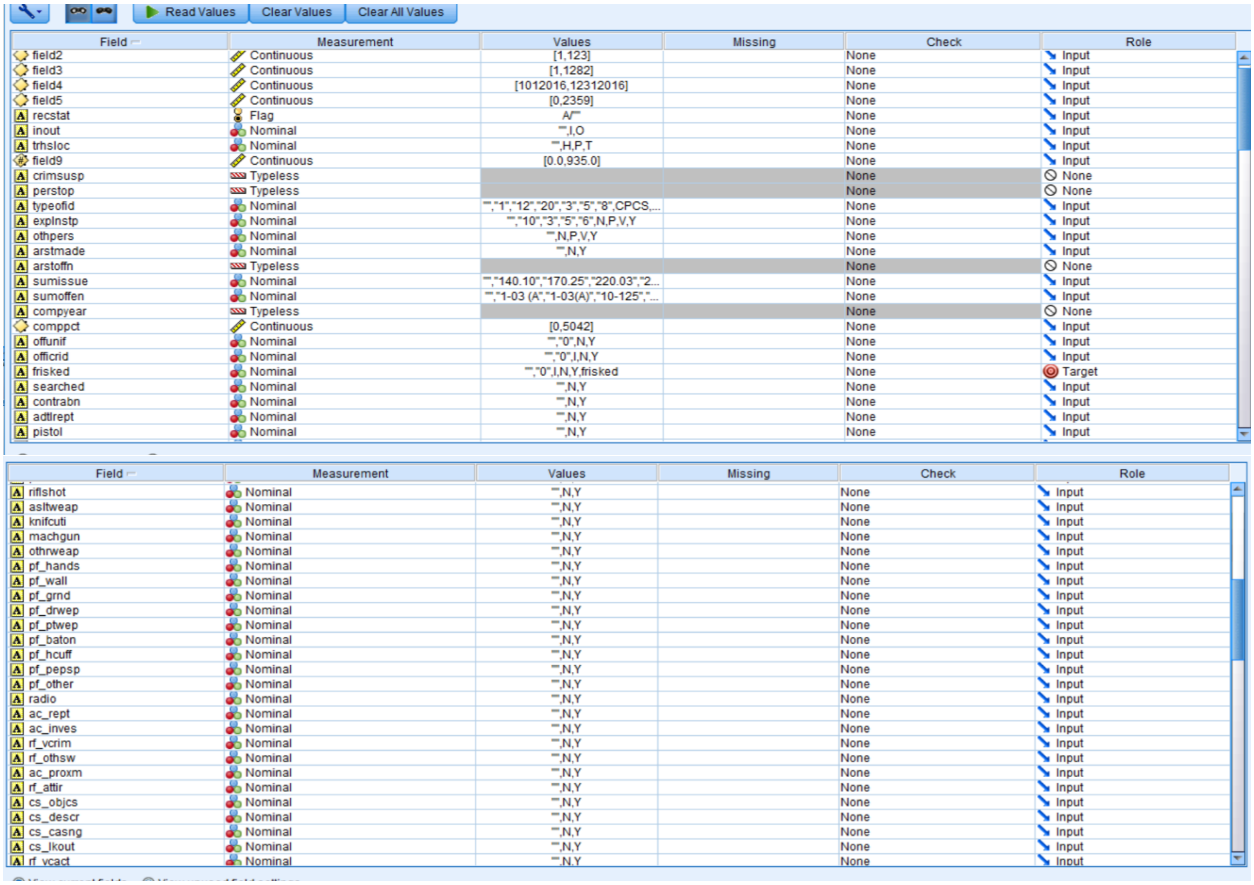
Training dataset: **50%** of the Big Dataset.
Testing dataset: **50%** of the Big Dataset.

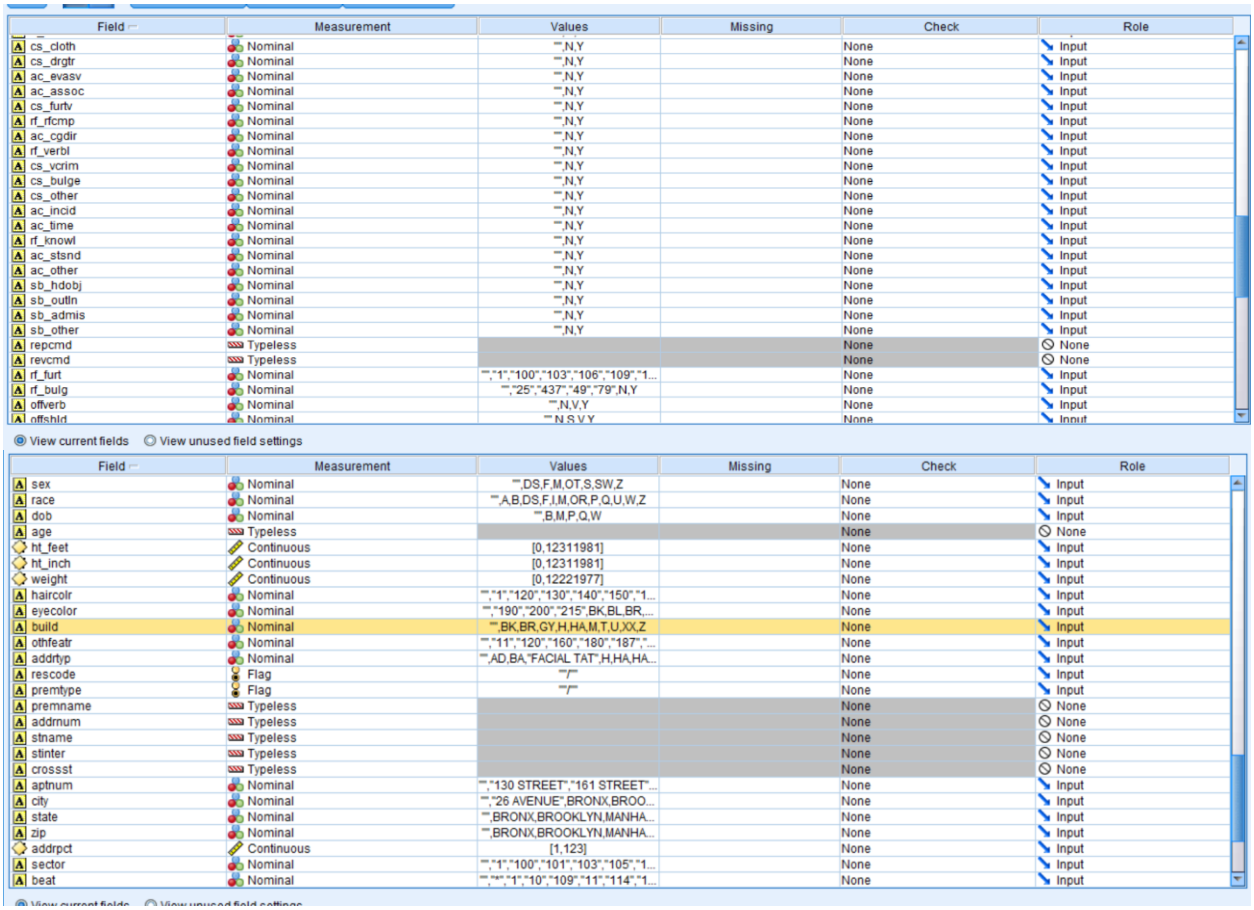
As the setting of the training and testing dataset, we can make to getting ready on the divided dataset to find better analyzing results in predicting and evaluating since we tried on the dataset with the rate of 70% for training and 30% for testing data but results are not as our expectation.

4.2 Choosing predictor and target attributes

Attributed columns with the **100 attributes (99 inputs and 1 targets in Role)** in the figure below after we waive all attributes do not have effect to targeted prediction. We will waive the attributes **crimsusp, perstop, age, arsfoffn, repcmd, revcmd, premtype, addnum, stname, stinten,crosst.** To be successful on the predicting tasks, we would approach to datatype includes Nominal (Categorical), Typeless, Flag, and Continuous (Numeric). This tasks, we told as the representation above (Fig 48).

Field column:





Field	Measurement	Values	Missing	Check	Role
cs_cloth	Nominal	"",N,Y		None	Input
cs_drgtr	Nominal	"",N,Y		None	Input
ac_evasv	Nominal	"",N,Y		None	Input
ac_assoc	Nominal	"",N,Y		None	Input
cs_furtv	Nominal	"",N,Y		None	Input
rf_rfcmp	Nominal	"",N,Y		None	Input
ac_cgdir	Nominal	"",N,Y		None	Input
rf_verbl	Nominal	"",N,Y		None	Input
cs_vcrim	Nominal	"",N,Y		None	Input
cs_bulge	Nominal	"",N,Y		None	Input
cs_other	Nominal	"",N,Y		None	Input
ac_incid	Nominal	"",N,Y		None	Input
ac_time	Nominal	"",N,Y		None	Input
rf_knowl	Nominal	"",N,Y		None	Input
ac_stsnd	Nominal	"",N,Y		None	Input
ac_other	Nominal	"",N,Y		None	Input
sb_hdobj	Nominal	"",N,Y		None	Input
sb_outln	Nominal	"",N,Y		None	Input
sb_admis	Nominal	"",N,Y		None	Input
sb_other	Nominal	"",N,Y		None	Input
repcmd	Typeless			None	None
revcmd	Typeless			None	None
rf_furt	Nominal	"","1","100","103","106","109","1...		None	Input
rf_bulg	Nominal	"","25","437","49","79",N,Y		None	Input
offverb	Nominal	"",N,V,Y		None	Input
offshld	Nominal	"",N,S,V,Y		None	Input

View current fields View unused field settings

Field	Measurement	Values	Missing	Check	Role
sex	Nominal	"",DS,F,M,OT,S,SW,Z		None	Input
race	Nominal	"",A,B,DS,F,I,M,OR,P,Q,U,W,Z		None	Input
dob	Nominal	"",B,M,P,Q,W		None	Input
age	Typeless			None	None
ht_feet	Continuous	[0,12311981]		None	Input
ht_inch	Continuous	[0,12311981]		None	Input
weight	Continuous	[0,12221977]		None	Input
haircolr	Nominal	"","1","120","130","140","150","1...		None	Input
eyecolor	Nominal	"","190","200","215",BK,BL,BR,...		None	Input
build	Nominal	"",BK,BR,GY,H,HA,M,T,U,XX,Z		None	Input
othfeatr	Nominal	"","11","120","160","180","187","...		None	Input
addrtyp	Nominal	"",AD,BA,"FACIAL TAT",H,HA,HA...		None	Input
rescode	Flag	"/"		None	Input
premtype	Flag	"/"		None	Input
premname	Typeless			None	None
addrnum	Typeless			None	None
stname	Typeless			None	None
stinter	Typeless			None	None
crossst	Typeless			None	None
aptnum	Nominal	"","130 STREET","161 STREET"...		None	Input
city	Nominal	"","26 AVENUE",BRONX,BROO...		None	Input
state	Nominal	"",BRONX,BROOKLYN,MANHA...		None	Input
zip	Nominal	"",BRONX,BROOKLYN,MANHA...		None	Input
addrpct	Continuous	[1,123]		None	Input
sector	Nominal	"","1","100","101","103","105","1...		None	Input
beat	Nominal	"","*","1","10","109","11","114","1...		None	Input

Fig 48: Attributes of the columns for predicting dataset

In which we can explain some attributes in the black rows as the figures above since we have waived 10 attributes of the **typeless** measurement attributes by analyzing works automatically as the light black rows.

Target attributes: frisked with the two attributes are Yes and No that mean that people were frisked or not, and they will be recoded to numerical values of 1 and 0 to predict to final decision and also give how to satisfied people who can be frisked.
year: Including 10 years, this value gives time in 10 years between 2007 and 2016.
pct: This is the **precinct of stop** in NYC with the values of 1 to 123.
race: This attribute is **racial people** in NYC with the races as:
1-black.
2-black Hispanic.
3-white Hispanic.
4-white.
5-Asian/Pacific Islander.
6-Am. Indian/ Native Alaskan.
sex: This is **gender of people** in the dataset that has two attributes of Male for 1 and Female for 0.

age: This gives the **age of people** that was recorded by officers.
knifcute: This was **a knife or cutting instrument found on** suspect that provide information to police with two values of Yes and No that will be recoded to 1 and 0.
searched: This was to give **suspect searched** of people with two values of Yes and No that will be recoded to 1 and 0.
contrabn: This was **contraband found** on suspect of people with two values of Yes and No that will be recoded to 1 and 0.
pistol: This was a **pistol found** on suspect of people with two values of Yes and No that will be recoded to 1 and 0.
riflshot: This was a **rifle found** on suspect of people with two values of Yes and No that will be recoded to 1 and 0.
asltweap: This was an **assault weapon found** on suspect of people with two values of Yes and No that will be recoded to 1 and 0.
machgun: This was a **machine gun found** on suspect of people with two values of Yes and No that will be recoded to 1 and 0.
othrweap: This was another **type of weapon found** on suspect of people with two values of Yes and No that will be recoded to 1 and 0.
arstmade: This was an **arrest made of people** with two values of Yes and No that will be recoded to 1 and 0.
crimsusp: This is **crime suspected** with the name of crime attributes.
detailcm: This includes **crime codes of people** with the crime codes above.
perstop: This is **period of stop** in minute.
height: This is **suspect's height** in inches.
recstat: This is **record status**, there are 0-original value A, 1-original value 1.
Attributes of physical force, reason for stop, reason for frisk, additional circumstances.

And **other attributes** with a total of 100 attributes as the figures above and in the list of variables.

5. Analyzing and Predicting on the Big Dataset in total of 10 years

In general, we would build models analytics of the dataset on IBM SSPS Modeler workflow and predict by **C&R Tree Method (The Classification and Regression Trees)** in total of 10 years, and then we would make predictions on each year by the following the workflow in IBM SPS Modeler 18.

Use of Modeling Method of Classification with Regression Trees (C&RT node) node generates via **decision tree** mothed that allows you to predict or classify future observations. We need to consider carefully based on the Frisk target and 100 attribute fields before predicting such a decision.

To begin for analyzing and predicting process, we build to express to a workflow in IBM SPSS Modeler with the addressing of worked nodes involved we firstly input the text combined file to File node which will run exactly and faster than *csv file, next Table node, Audi, Partition, Type, three nodes of Modeling Tree methods as **CRT, CHAID**, and **AS Tree** (Fig 49) that generated to three churns of results and we can proceed to analyzing, visualizing, and statistical progress from these churns as adding more Analysis node, Statistic, Graph, Audi node to get statistical final results.

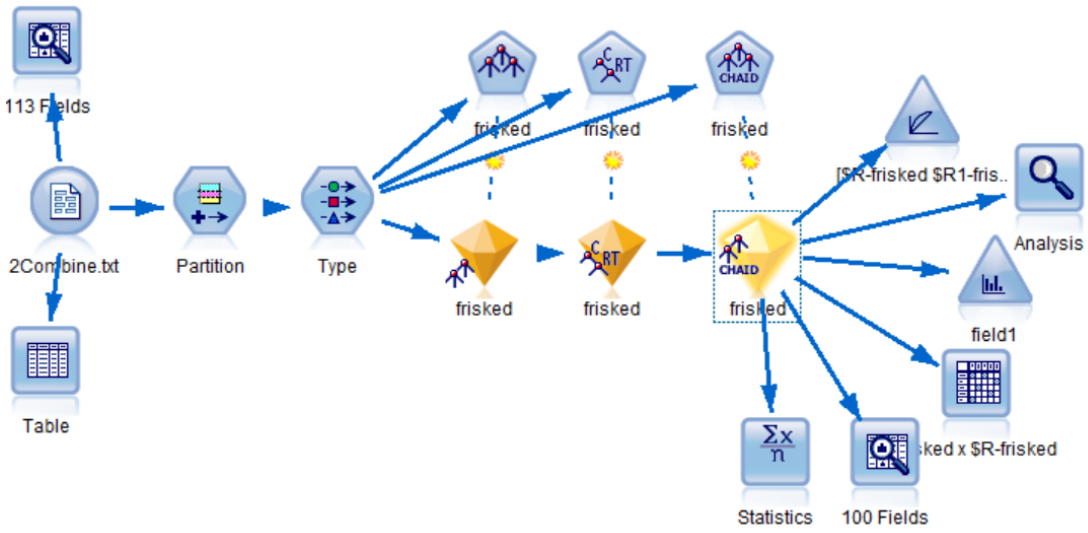


Fig 49: Workflow of Decision trees for analyzing and predicting.

As the schema of workflow above of frisked, we can recall this bar chart as below to acknowledge to number of frisked times by each year, and to compare to total times of 10 years. This chart brings a whole view on the frisked target before we can address to analytics process. At most at 2011 with around 700,000 cases and at least values at 2016, the number of frisks are increasing from 2007 to 2011, and then they are decreasing between 2012 and 2016 with much different values. As we see the histogram, in total, we have **3,686,101** cases including of Yes and No on the frisk, hence we need to predict on all rows to get final results (Fig 50).

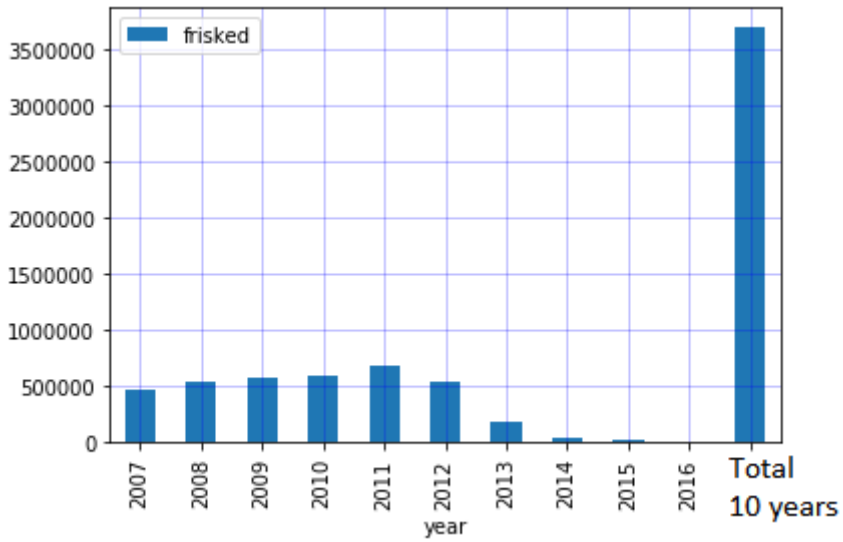


Fig 50: Histogram of Frisked attribute throughout 10 years.

Regarding of the map trees as below, we had map tree generated the importance of attributes by frisked target with nodes of workflow by reason for frisk reason for frisk-furtive movements

rf_furt (0.48), reason for frisk - knowledge of suspect's prior crim behavour (**rf_knowl**), reason, reason for frisk - violent crime suspected **(rf_vcrim),** crime code description **(detailCM),** reason for frisk - inappropriate attire for season **(rf_attir),** reason for frisk - other suspicion of weapons (**rf_othsw**), reason for frisk - suspicious bulge (**rf_bulg**), additional circumstances - area has high crime incidence **(ac_incid),** suspect's weight in pounds **(weight),** radio run **(radio)**, precinct of stop **(pct)**. These attributes influenced to results of target and fetch out the importance of related attributes (Fig 51).

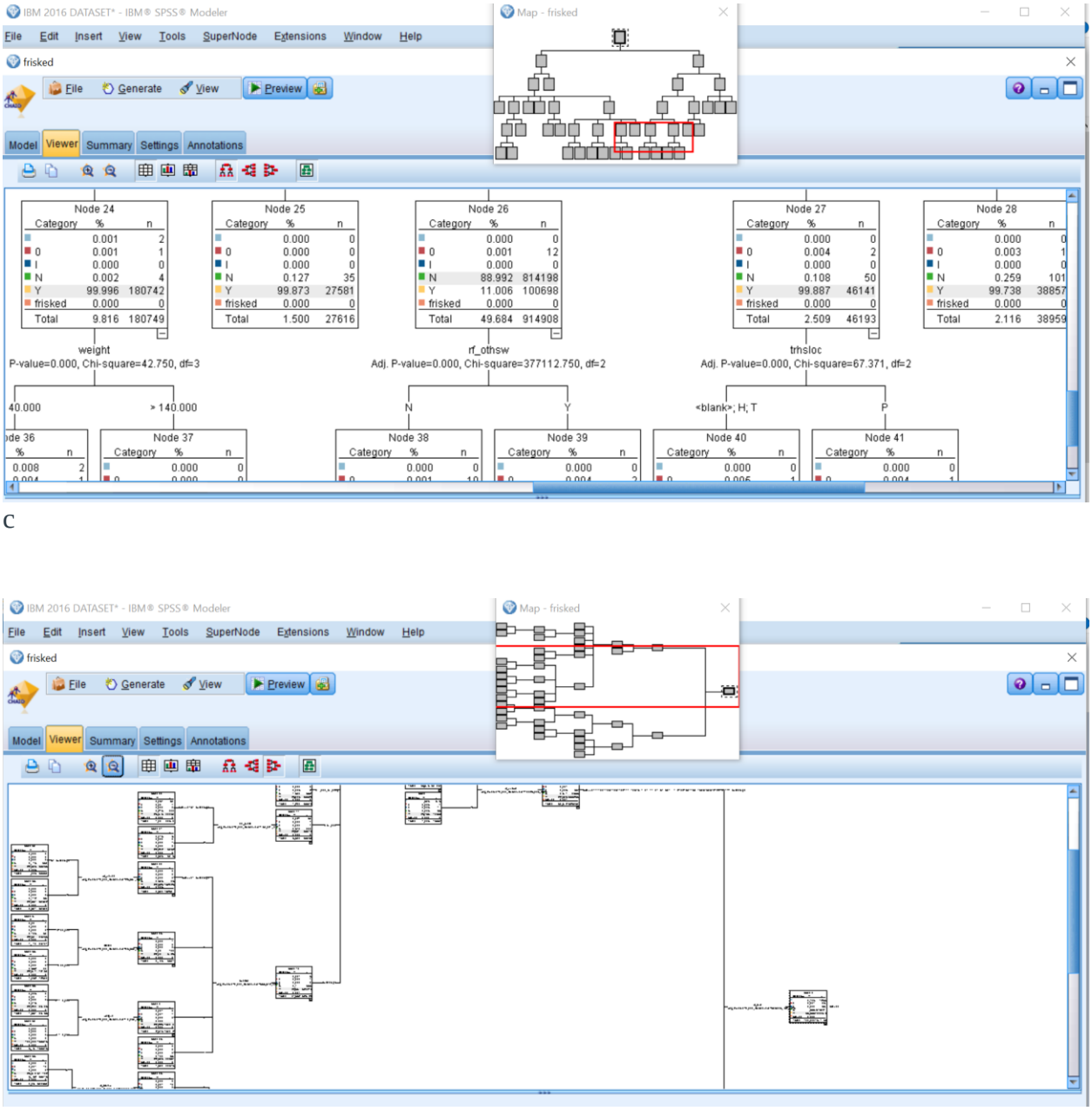


Fig 51: The map trees of important attributes.

In order to predict the importance of attributes, we would like to add to a node with the best accuracy which is optimal solution of the workflow as CHAID Tree node to find predictor importance (Fig 52) as **rf_furt (0.54), rf_vcrim, rf_knowl, detailCM, rf_atti**, **rf_othsw**, **rf_bulg**, **radio**, **ac_incid, weight, pct**. With such a generating result for this strategy, these are the most attributes have influences to frisked target.

As a special result, we conceived the related attributes to Frisked target as the obvious attributes of predictor because they include necessary reasons for frisk by popular problems as furtive movements, crime behavior, violent crime suspected, crime code description, precinct of stop.

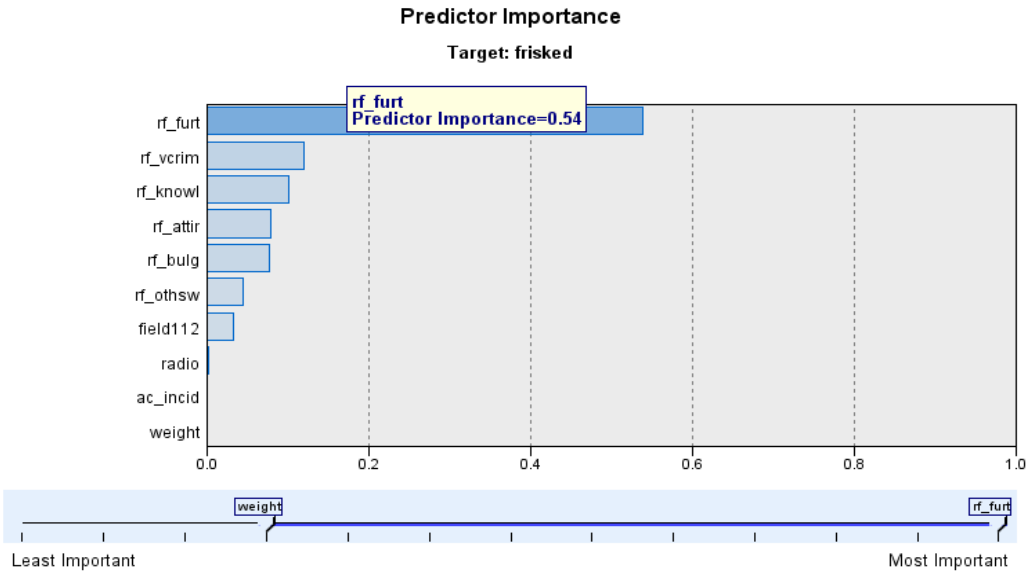


Fig 52: Predictor Importance attributes in predicting process

Obviously, we continue to innovate to get the important attributes for the frisked target to answer for predictive process is Yes based on the attributes as the figure below (Fig 53).

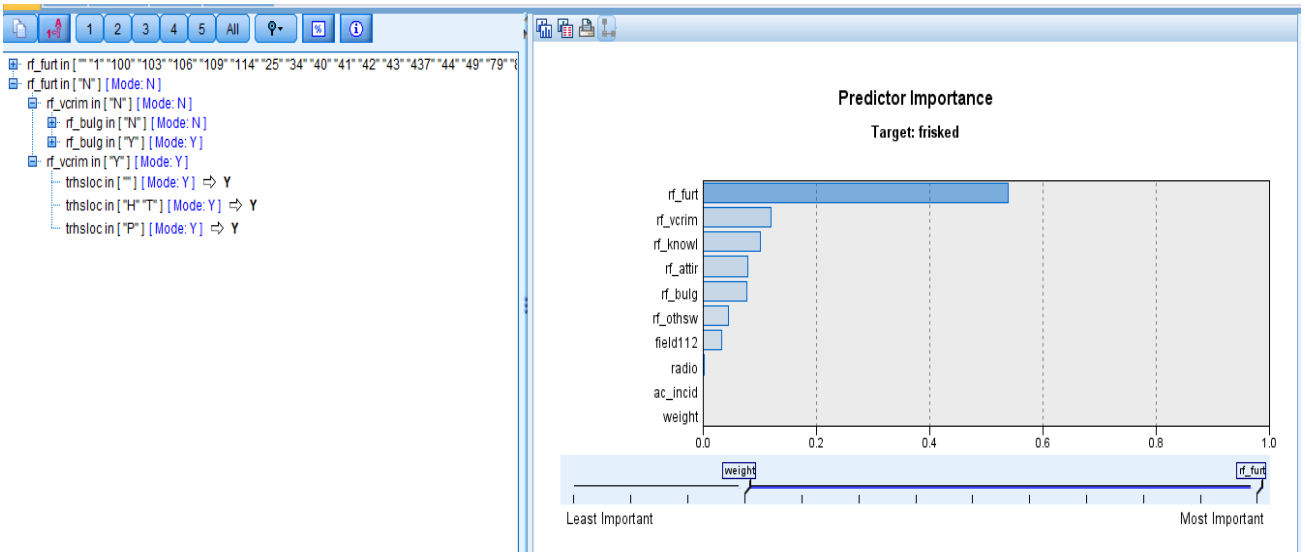
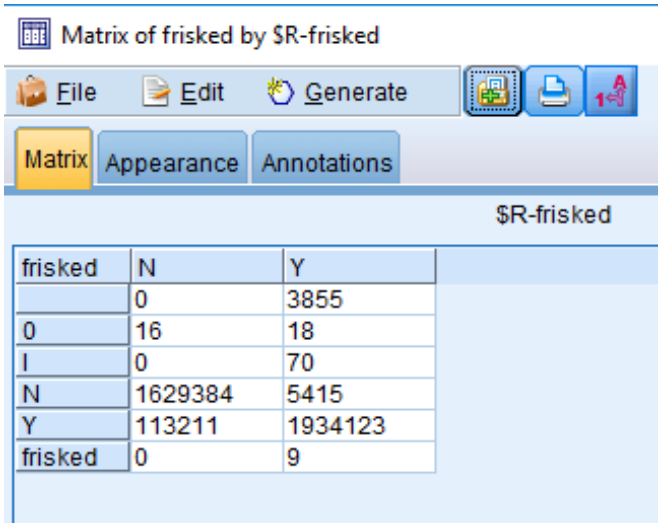


Fig 53: Predictor Important attributes table with variables of Yes or No.

Within area of analytics responsibilities, we have generated the results by frisked target to interpret the statistical particularly values of True Positive (TP), True Negative (TN), False Negative (FN), Positive Negative (PN) as in No, Yes below (Fig 54).



frisked	N	Y
	0	3855
0	16	18
I	0	70
N	1629384	5415
Y	113211	1934123
frisked	0	9

Fig 54: Confusion matrix of frisked attribute

Recall the definitions of predicted parameters as the following:

Accuracy: What percentage of your **predictions** were correct?
Recall: What percentage of the **positive cases** did you **catch**?
Precision: What percentage of the **positive predictions** were **correct**?

Accuracy = (TN+TP)/(TN+FP+FN+TP) = 96.7%.

Recall (r) = TP/(TP+FN) = 94.47%.

Precision (p) = TP/(TP+FP) = 99.7%.

F-Measure = 2 x (p x r)/(p + r) = 97%, which is close to the smaller one of r and p.

The ***accuracy*** can be better to avoid the accuracy in likely of other metrics as ***precision*** and ***recall.***

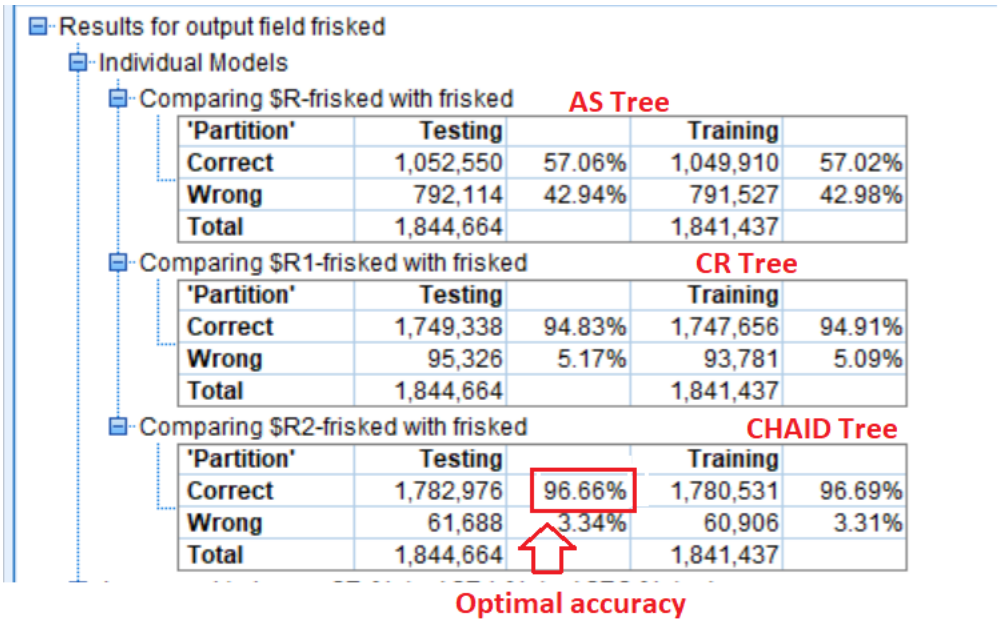


Fig 55: The results of Accuracy of Decision Tree methods of frisked target.

By this result of the table for testing data, we had the accuracy value on Testing data for the Frisked target is **96.66**% (Fig 55).

Gini index

Gini index is the divergences between the probabilities distributions of the target attribute of values. The Gini index is given by the formula as:

$$Gini\ (S, A) = 1 - \sum_{v \epsilon A} \left(\frac{|S_v|}{|S|}\right)^2$$

GinaGains are defined as the proportion of total hits which occurs in each quantile.

$$GiniGain\ (S, A) = Gini(S) - \sum_{v \epsilon A} \frac{|S_v|}{|S|} Gini(S_v)$$

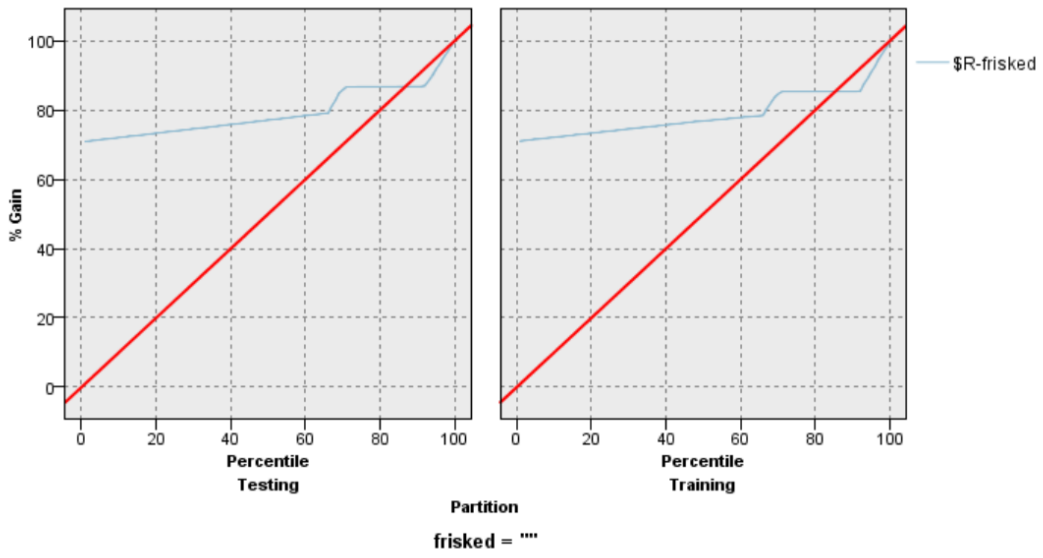


Fig 56: Gain curve of frisked target

Below is the ROC curve of the Frisk target with accuracy above.
In statistics, a **receiver operating characteristic** (**ROC**) is a graphical plot that has change of **True Positive** rate vs. **False Positive** rate**.** We would find the **ROC curve** by the function y = f(x) from x = a to x = b, this means the **integration of** y = f(x) between the limits of a and b. Areas under the x-axis is negative and **areas** above the x-axis is positive. The Area Under Curve (AUC) of frisked target is **0.9444** as the figure of (Fig 57).

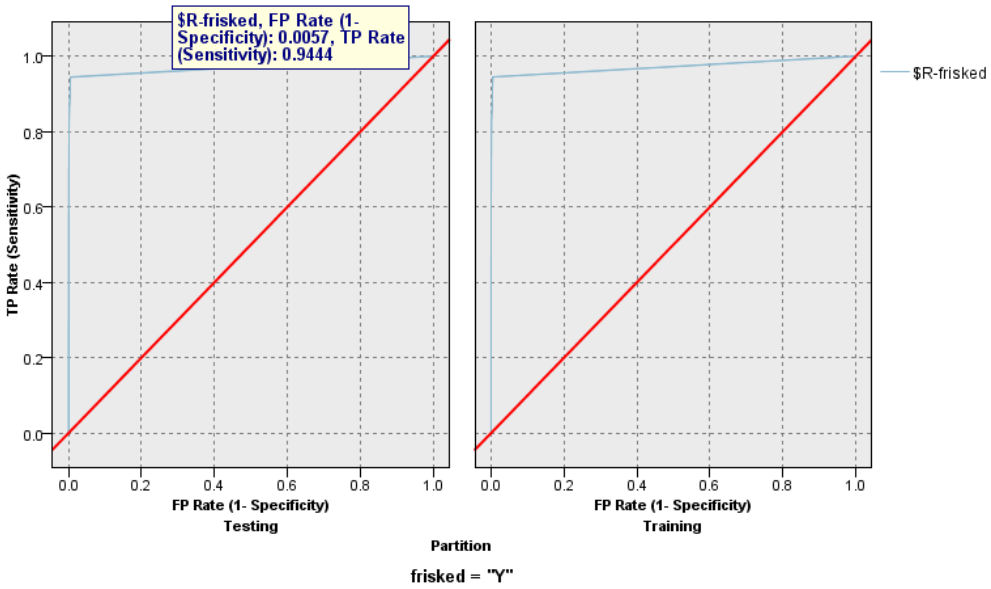


Fig 57: The AUC curve of the frisked target

Generated results on the dataset with histogram of each attribute that indicate Sample Graph, Measurement, Skewness, Min/Max/Mean, Standard Dev, Skewness, Unique, Valid as the table (Fig 58).

Audit | Quality | Annotations

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
field1		Continuous	2007	2016	2009.897	1.895	0.200	--	3686091
field2		Continuous	1	123	67.822	33.149	-0.093	--	3686091
field3		Continuous	1	31763	5138.256	4842.739	1.733	--	3686091
field4		Continuous	1012007	12312016	6176127.350	3468143.835	0.166	--	3686091
field5		Continuous	0	2959	1415.487	751.916	-0.650	--	3685197
recstat		Flag	--	--	--	--	--	5	3683250
inout		Nominal	--	--	--	--	--	4	3686099
trhsloc		Nominal	--	--	--	--	--	5	2897661
field9		Continuous	0.000	999.000	2.537	5.764	63.953	--	3686091

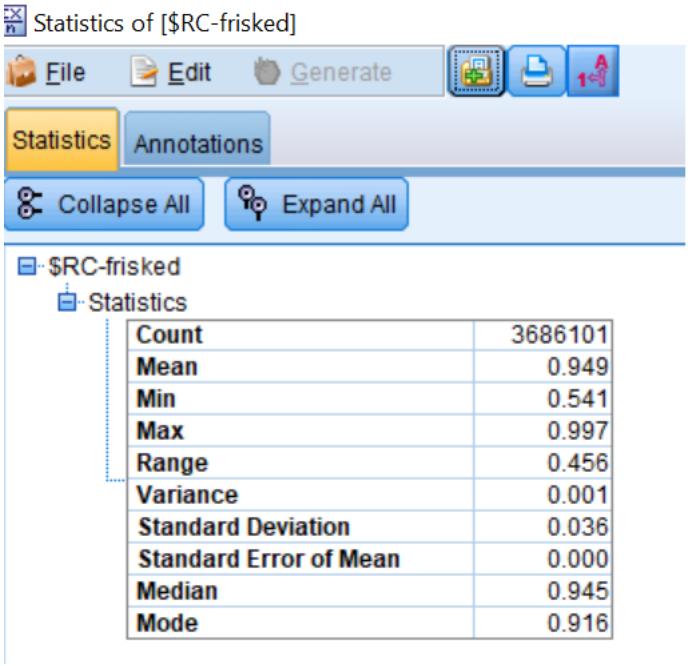
Audit | Quality | Annotations

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
post		Continuous	1	122	12.826	15.162	2.525	--	627706
xcoord		Continuous	1	1067249	937022.605	255085.027	-3.376	--	2281812
ycoord		Continuous	1	1067256	528967.978	394835.109	0.383	--	3569330
linecm		Continuous	1	1058778	294.167	11492.612	64.481	--	2193444
field112		Continuous	0	1008694	43.877	1097.009	606.574	--	1494718
field113		Continuous	0	179350	106.685	3622.824	49.494	--	2450
Partition		Nominal	--	--	--	--	--	2	3686101
$R-frisked		Nominal	--	--	--	--	--	2	3686101
$RC-frisk...		Continuous	0.541	0.997	0.949	0.036	0.362	--	3686101

[1] Indicates a multimode result. [3] Indicates a sampled result.

Fig 58: The quality and statistical results of attributes.

Statistical results based on frisked target with Count, Min/Max/Mean, Range, Variance, Standard Deviation, Standard Error of Mean, Median, Mode (Fig 59).



Count	3686101
Mean	0.949
Min	0.541
Max	0.997
Range	0.456
Variance	0.001
Standard Deviation	0.036
Standard Error of Mean	0.000
Median	0.945
Mode	0.916

Fig 59: The statistical results of frisked target

Separately, we can find accuracies of each year of 10 years between 2007 and 2016 by the Frisked target to evaluate, compare, and predict to results and important attributes. Looking on the changed accuracies, we accomplished the differences of values from 51.75% to 67.52%, this have given that had chance of the difference of predictor importance in each year and the differences of each dataset of each year. In this case, we should respect to the accuracies were followed and depended the datasets and the predictor importance (Fig 60).

Results for output field frisked

Comparing $R-frisked with frisked

2007

'Partition'	Testing		Training	
Correct	122,355	51.75%	122,209	51.86%
Wrong	114,098	48.25%	113,434	48.14%
Total	236,453		235,643	

Results for output field frisked

Comparing $R-frisked with frisked

2008

'Partition'	Testing		Training	
Correct	146,928	54.32%	146,471	54.28%
Wrong	123,534	45.68%	123,369	45.72%
Total	270,462		269,840	

Results for output field frisked

Comparing $R-frisked with frisked

2009

'Partition'	Testing		Training	
Correct	168,633	56%	168,223	56.05%
Wrong	132,505	44%	131,924	43.95%
Total	301,138		300,147	

Results for output field frisked

Comparing $R-frisked with frisked

2010

'Partition'	Testing		Training	
Correct	168,633	56%	168,223	56.05%
Wrong	132,505	44%	131,924	43.95%
Total	301,138		300,147	

Results for output field frisked

Comparing $R-frisked with frisked

2011

'Partition'	Testing		Training	
Correct	190,992	55.59%	190,292	55.62%
Wrong	152,582	44.41%	151,858	44.38%
Total	343,574		342,150	

Results for output field frisked

Comparing $R-frisked with frisked

2012

'Partition'	Testing		Training	
Correct	148,614	55.71%	148,378	55.75%
Wrong	118,137	44.29%	117,782	44.25%
Total	266,751		266,160	

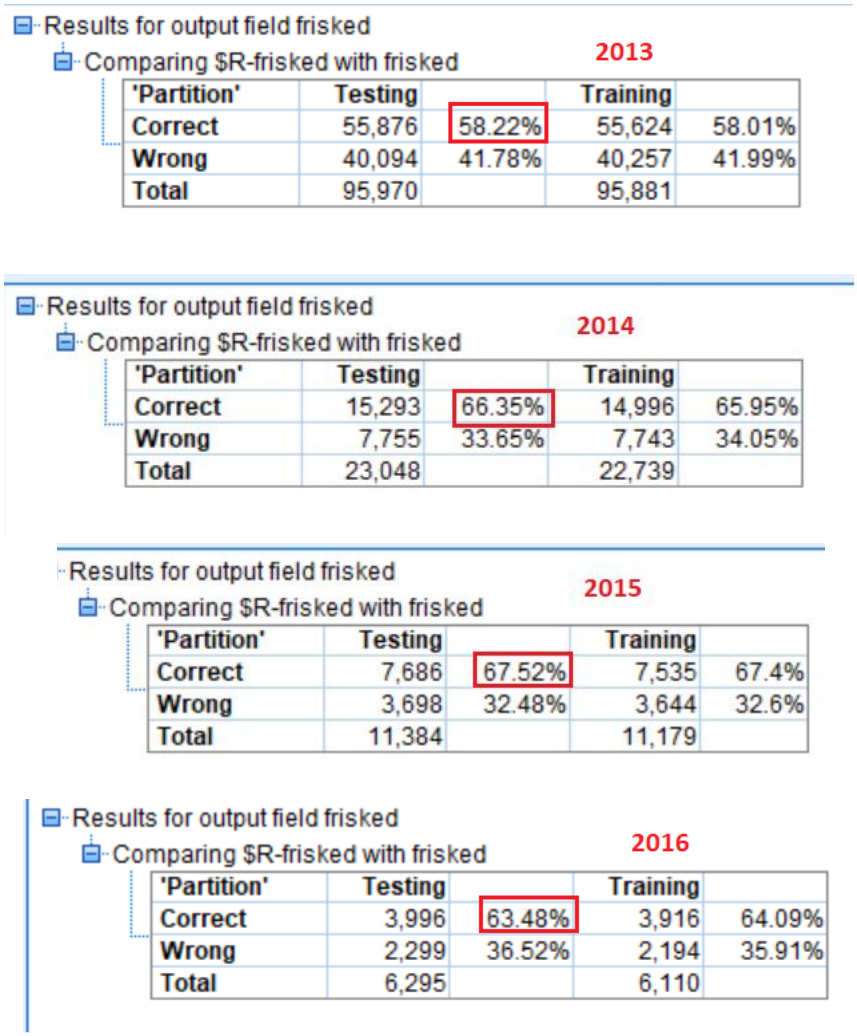
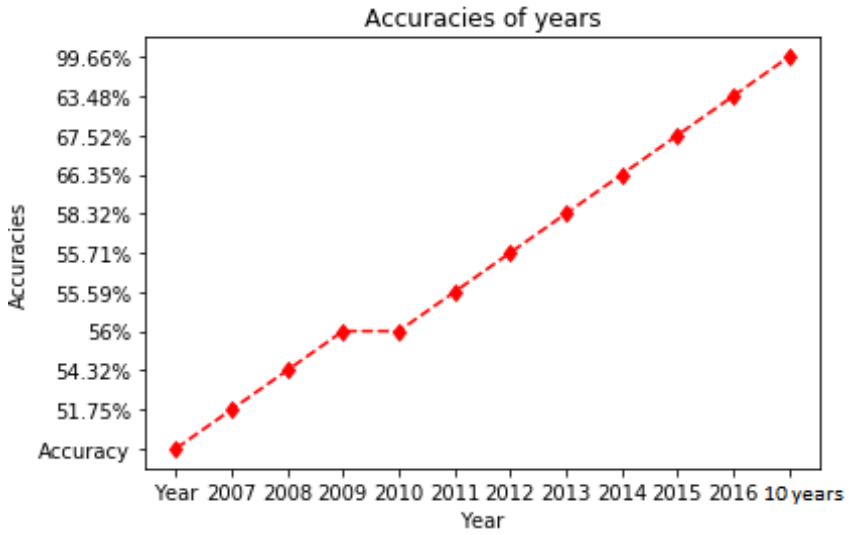
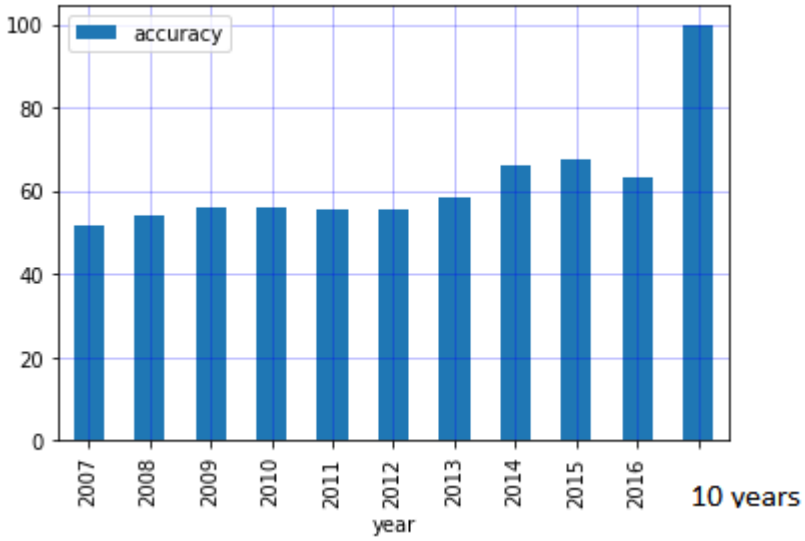


Fig 60: Accuracies of each year between 2007 and 206

Years	Method	Accuracy	Number of Rows	Number of Attributes	RECALL	F1	Pre
2007	CHAID Tree	51.75%	472,097	100			
2008	CHAID	54.32%	540,303	100			
2009	CHAID	56%	581,169	100			
2010	CHAID	56%	601,286	100			
2011	CHAID	55.59%	685,725	100			
2012	CHAID	55.71%	532,912	100			
2013	CHAID	58.32%	191,852	100			
2014	CHAID	66.35%	45,788	100			
2015	CHAID	67.52%	22,564	100			

2016	CHAID	63.48%	12,495	100			
	CHAID	99.66%	3,686,101	100			
	CR Tree	94.83%	3,686,101	100			
10 years	AS Tree	57.06%	3,686,101	100			





Conclusion: We are making conclusions from the predicted results for optimal decision making purposes. As anticipated, the Decision Tree models has been choosing for this project due to it is Big Datasets with many attributes related to nominal values. It is not exact in predicting if we transform them to categorical or numeric or continuous values, therefore, the best choice of this

project is Decision Tree models in three kinds of Tree including CRTree, AS Tree, and CHAID Tree but the generated optimal results belong to the CHAID Tree results with the highest accuracy is 96.66%. The causes of CHAID Tree has been chosen in the best accuracy on the Frisked target for final decision that the predictions of the Stop, Question and Frisk process realized. Throughout of this analyzed work, we have clearly found the important attributes influence to the target in an efficient and timely manner on the column of Frisked attribute since we followed it on the two variables of Yes and No, at this mind that NYPD should know who, what, how to crime activities in NYC and they really are able to manage a variety of training and preparing programs as well as they can ensure practicing objectives are achieved to decrease number of criminal people in future. No matter what to NYPD work, they always need to match to data analytics to get out their own decisions for a bright future in the city. The Big Dataset has spoken out all information, intentions, attributes, predictions to real life and future.

Upon the result of prediction process of the Frisked attribute and by AUC and accuracy value, we were able to be connected to real life NYPD Dataset that endeavoured to manage top crime activires and predicted to searching, frisking and arresting people by police oficers with many kinds of crime. The result assessed the levels of dangerous man depend on Firsked or consider that how about the status of someone. While it is not easy to determine the state of components that ask to why someone are felonious, but NYPD has establishment to know they need or do not need to frisk or arrest someone, this gave many knowledge to understand their works. As a helpful result, data predicting has shown police need to frisked anyone once they have required stoping and asking to ensure that someone do not have any menace by acting of bring weapons, of thievery, of contraband, ect. Hence, the series of action by Stop, Question, and Frisk must be continued to be sure the safety of everyone.

Anyone who has been stoped and answered the low and bad questions or had wrong questions to police officer can be viewed as strong candidates to be frisked. Among the idea point police officers joined with another deparment to check background and control the use of weapon to someone, for example, someone who have concluded to frisked and arrested so they must be prohibited to posssess any weapon and need to observe their actions in the future.

References:

[1] http://shukka.com/exploratory-analysis-services.php

[2] http://michael.hahsler.net/research/arules_RUG_2015/demo/SQF_Codebook.pdf

[3] https://www1.nyc.gov/site/nypd/stats/reports-analysis/stopfrisk.page

[4] https://www.usccr.gov/pubs/nypolice/ch5.htm

[5] Delores Jones-Brown, Brett G. Stoudt Brian Johnston Kevin Moran, Stop, Question, And Frisk Policing Practices In New York City, Center on Race, Crime and Justice John, Jay College of Criminal Justice, July, 2013.

[6] AdrienneN.Milner, BrandonJ.George, DavidB.Allison, Black and Hispanic Men Perceived to Be Large Areat Increased Risk for Police Frisk, Search, and Force.