AIT-580: Big Data Analytics Project

FINAL PROJECT REPORT ON

Washington DC Crime Data

GEORGE MASON UNIVERSITY

FAIRFAX, VA

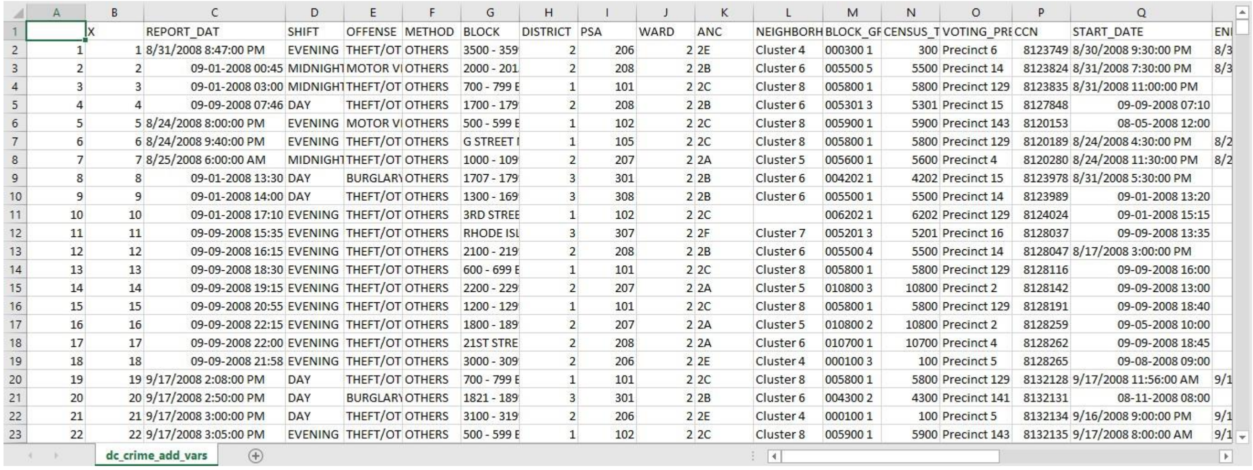
BY

SHOURYASIMHA ADDEPALLI

G01156040

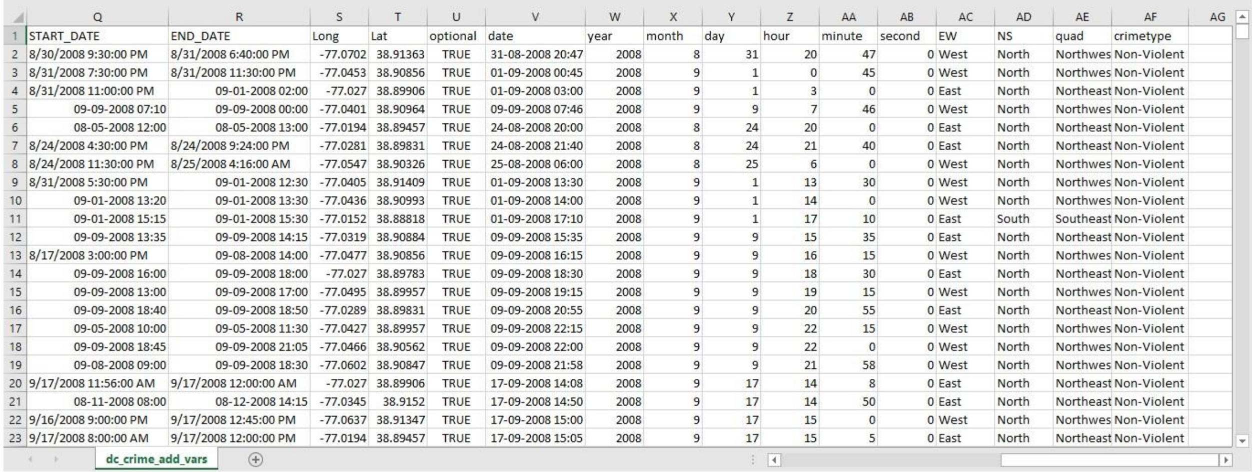
INTRODUCTION

I picked this informational collection like this could be useful in building certain prescient models to examine and foresee future crimes. The datasets include certain vital properties like offense, neighborhood bunches, scope, longitude and a lot increasingly that are exceptionally basic. This informational collection gives result in kinds of violations, sorts of crimes submitted by individuals in seven days, month or hourly premise. This dataset would be useful in reaching an appropriate determination to get a decent result.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1		X	REPORT_DAT	SHIFT	OFFENSE	METHOD	BLOCK	DISTRICT	PSA	WARD	ANC	NEIGHBORH	BLOCK_GF	CENSUS_T	VOTING_PRE	CCN	START_DATE	ENI
2	1	1	8/31/2008 8:47:00 PM	EVENING	THEFT/OT	OTHERS	3500 - 359	2	206	2	2E	Cluster 4	000300 1	300	Precinct 6	8123749	8/30/2008 9:30:00 PM	8/3
3	2	2	09-01-2008 00:45	MIDNIGHT	MOTOR VI	OTHERS	2000 - 201	2	208	2	2B	Cluster 6	005500 5	5500	Precinct 14	8123824	8/31/2008 7:30:00 PM	8/3
4	3	3	09-01-2008 03:00	MIDNIGHT	THEFT/OT	OTHERS	700 - 799 E	1	101	2	2C	Cluster 8	005800 1	5800	Precinct 129	8123835	8/31/2008 11:00:00 PM	
5	4	4	09-09-2008 07:46	DAY	THEFT/OT	OTHERS	1700 - 179	2	208	2	2B	Cluster 6	005301 3	5301	Precinct 15	8127848	09-09-2008 07:10	
6	5	5	8/24/2008 8:00:00 PM	EVENING	MOTOR VI	OTHERS	500 - 599 E	1	102	2	2C	Cluster 8	005900 1	5900	Precinct 143	8120153	08-05-2008 12:00	
7	6	6	8/24/2008 9:40:00 PM	EVENING	THEFT/OT	OTHERS	G STREET I	1	105	2	2C	Cluster 8	005800 1	5800	Precinct 129	8120189	8/24/2008 4:30:00 PM	8/2
8	7	7	8/25/2008 6:00:00 AM	MIDNIGHT	THEFT/OT	OTHERS	1000 - 109	2	207	2	2A	Cluster 5	005600 1	5600	Precinct 4	8120280	8/24/2008 11:30:00 PM	8/2
9	8	8	09-01-2008 13:30	DAY	BURGLARY	OTHERS	1707 - 179	3	301	2	2B	Cluster 6	004202 1	4202	Precinct 15	8123978	8/31/2008 5:30:00 PM	
10	9	9	09-01-2008 14:00	DAY	THEFT/OT	OTHERS	1300 - 169	3	308	2	2B	Cluster 6	005500 1	5500	Precinct 14	8123989	09-01-2008 13:20	
11	10	10	09-01-2008 17:10	EVENING	THEFT/OT	OTHERS	3RD STREE	1	102	2	2C		006202 1	6202	Precinct 129	8124024	09-01-2008 15:15	
12	11	11	09-09-2008 15:35	EVENING	THEFT/OT	OTHERS	RHODE ISL	3	307	2	2F	Cluster 7	005201 3	5201	Precinct 16	8128037	09-09-2008 13:35	
13	12	12	09-09-2008 16:15	EVENING	THEFT/OT	OTHERS	2100 - 219	2	208	2	2B	Cluster 6	005500 4	5500	Precinct 14	8128047	8/17/2008 3:00:00 PM	
14	13	13	09-09-2008 18:30	EVENING	THEFT/OT	OTHERS	600 - 699 E	1	101	2	2C	Cluster 8	005800 1	5800	Precinct 129	8128116	09-09-2008 16:00	
15	14	14	09-09-2008 19:15	EVENING	THEFT/OT	OTHERS	2200 - 229	2	207	2	2A	Cluster 5	010800 3	10800	Precinct 2	8128142	09-09-2008 13:00	
16	15	15	09-09-2008 20:55	EVENING	THEFT/OT	OTHERS	1200 - 129	1	101	2	2C	Cluster 8	005800 1	5800	Precinct 129	8128191	09-09-2008 18:40	
17	16	16	09-09-2008 22:15	EVENING	THEFT/OT	OTHERS	1800 - 189	2	207	2	2A	Cluster 5	010800 2	10800	Precinct 2	8128259	09-05-2008 10:00	
18	17	17	09-09-2008 22:00	EVENING	THEFT/OT	OTHERS	21ST STRE	2	208	2	2A	Cluster 6	010700 1	10700	Precinct 4	8128262	09-09-2008 18:45	
19	18	18	09-09-2008 21:58	EVENING	THEFT/OT	OTHERS	3000 - 309	2	206	2	2E	Cluster 4	000100 3	100	Precinct 5	8128265	09-08-2008 09:00	
20	19	19	9/17/2008 2:08:00 PM	DAY	THEFT/OT	OTHERS	700 - 799 E	1	101	2	2C	Cluster 8	005800 1	5800	Precinct 129	8132128	9/17/2008 11:56:00 AM	9/1
21	20	20	9/17/2008 2:50:00 PM	DAY	BURGLARY	OTHERS	1821 - 189	3	301	2	2B	Cluster 6	004300 2	4300	Precinct 141	8132131	08-11-2008 08:00	
22	21	21	9/17/2008 3:00:00 PM	DAY	THEFT/OT	OTHERS	3100 - 319	2	206	2	2E	Cluster 4	000100 1	100	Precinct 5	8132134	9/16/2008 9:00:00 PM	9/1
23	22	22	9/17/2008 3:05:00 PM	EVENING	THEFT/OT	OTHERS	500 - 599 E	1	102	2	2C	Cluster 8	005900 1	5900	Precinct 143	8132135	9/17/2008 8:00:00 AM	9/1

Fig 1: Dataset



	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	START_DATE	END_DATE	Long	Lat	optional	date	year	month	day	hour	minute	second	EW	NS	quad	crimetype	
2	8/30/2008 9:30:00 PM	8/31/2008 6:40:00 PM	-77.0702	38.91363	TRUE	31-08-2008 20:47	2008	8	31	20	47	0	West	North	Northwes	Non-Violent	
3	8/31/2008 7:30:00 PM	8/31/2008 11:30:00 PM	-77.0453	38.90856	TRUE	01-09-2008 00:45	2008	9	1	0	45	0	West	North	Northwes	Non-Violent	
4	8/31/2008 11:00:00 PM	09-01-2008 02:00	-77.027	38.89906	TRUE	01-09-2008 03:00	2008	9	1	3	0	0	East	North	Northeast	Non-Violent	
5	09-09-2008 07:10	09-09-2008 00:00	-77.0401	38.90964	TRUE	09-09-2008 07:46	2008	9	9	7	46	0	West	North	Northwes	Non-Violent	
6	08-05-2008 12:00	08-05-2008 13:00	-77.0194	38.89457	TRUE	24-08-2008 20:00	2008	8	24	20	0	0	East	North	Northeast	Non-Violent	
7	8/24/2008 4:30:00 PM	8/24/2008 9:24:00 PM	-77.0281	38.89831	TRUE	24-08-2008 21:40	2008	8	24	21	40	0	East	North	Northeast	Non-Violent	
8	8/24/2008 11:30:00 PM	8/25/2008 4:16:00 AM	-77.0547	38.90326	TRUE	25-08-2008 06:00	2008	8	25	6	0	0	West	North	Northwes	Non-Violent	
9	8/31/2008 5:30:00 PM	09-01-2008 12:30	-77.0405	38.91409	TRUE	01-09-2008 13:30	2008	9	1	13	30	0	West	North	Northwes	Non-Violent	
10	09-01-2008 13:20	09-01-2008 13:30	-77.0436	38.90993	TRUE	01-09-2008 14:00	2008	9	1	14	0	0	West	North	Northwes	Non-Violent	
11	09-01-2008 15:15	09-01-2008 15:30	-77.0152	38.88818	TRUE	01-09-2008 17:10	2008	9	1	17	10	0	East	South	Southeast	Non-Violent	
12	09-09-2008 13:35	09-09-2008 14:15	-77.0319	38.90884	TRUE	09-09-2008 15:35	2008	9	9	15	35	0	East	North	Northeast	Non-Violent	
13	8/17/2008 3:00:00 PM	09-08-2008 14:00	-77.0477	38.90856	TRUE	09-09-2008 16:15	2008	9	9	16	15	0	West	North	Northwes	Non-Violent	
14	09-09-2008 16:00	09-09-2008 18:00	-77.027	38.89783	TRUE	09-09-2008 18:30	2008	9	9	18	30	0	East	North	Northeast	Non-Violent	
15	09-09-2008 13:00	09-09-2008 17:00	-77.0495	38.89957	TRUE	09-09-2008 19:15	2008	9	9	19	15	0	West	North	Northwes	Non-Violent	
16	09-09-2008 18:40	09-09-2008 18:50	-77.0289	38.89831	TRUE	09-09-2008 20:55	2008	9	9	20	55	0	East	North	Northeast	Non-Violent	
17	09-05-2008 10:00	09-05-2008 11:30	-77.0427	38.89957	TRUE	09-09-2008 22:15	2008	9	9	22	15	0	West	North	Northwes	Non-Violent	
18	09-09-2008 18:45	09-09-2008 21:05	-77.0466	38.90562	TRUE	09-09-2008 22:00	2008	9	9	22	0	0	West	North	Northwes	Non-Violent	
19	09-08-2008 09:00	09-09-2008 18:30	-77.0602	38.90847	TRUE	09-09-2008 21:58	2008	9	9	21	58	0	West	North	Northwes	Non-Violent	
20	9/17/2008 11:56:00 AM	9/17/2008 12:00:00 AM	-77.027	38.89906	TRUE	17-09-2008 14:08	2008	9	17	14	8	0	East	North	Northeast	Non-Violent	
21	08-11-2008 08:00	08-12-2008 14:15	-77.0345	38.9152	TRUE	17-09-2008 14:50	2008	9	17	14	50	0	East	North	Northeast	Non-Violent	
22	9/16/2008 9:00:00 PM	9/17/2008 12:45:00 PM	-77.0637	38.91347	TRUE	17-09-2008 15:00	2008	9	17	15	0	0	West	North	Northwes	Non-Violent	
23	9/17/2008 8:00:00 AM	9/17/2008 12:00:00 PM	-77.0194	38.89457	TRUE	17-09-2008 15:05	2008	9	17	15	5	0	East	North	Northeast	Non-Violent	

Fig 2: Dataset

Description: Dataset of all the crimes in the DC metro police system ranging from Theft, Arson, Assault, Homicide, Sex Abuse, Robbery, and Burglary. Data can be easily geocoded and mapped, trends can be extracted, and predictions can be made. Would be interesting to combine with other datasets, i.e. changes in housing prices, history of construction sites etc. An informal hypothesis would be: If the local government invests in fixing the sidewalks in a neighborhood, how much would the investment decrease crime levels on a block by block basis. (Vinze, 2017)

WHO: The data is collected by Metropolitan Police Department, which is known to be one of the largest police agency in United States of America.
MPT is one of the primary law enforcement agency for the Colombia district. (Metro Police Department, n.d.)

The dataset has been published in Kaggle.com, platform for predictive modelling and analytics competitions in which companies and researchers post data and statisticians and data miners compete to produce the best models for predicting and describing the data. (google, n.d.)

This data is very helpful in analyzing the main attributes like the time and location affecting the crime rates, and to also use certain predictive models that could be helpful in analyzing the crime patterns in an ability to predict the future crimes.

NEED: The data collected is helpful in giving hope to the police officer to respond immediately to different types of crimes in advance with the criminals in a suitable way. This could be helpful in reducing the crime rate further. (Metro Police Department, n.d.)

Requirements Resources- To study and analyze this dataset, the resources needed are Tableau, R, python and SQL. The Visualizations are done using Tableau, R and python.

Data Curation- The data set had missing values and null values so as this missing data and null values could produce incorrect or inappropriate results/conclusions, data curation needs to be done which is very essential to produce good output. Subsequent to expelling the missing information by supplanting it with a most plausible value, I considered the vital qualities from the given informational indexes to make another one.

POTENTIAL SET OF QUESTIONS:

- **Geographically, which part of the Washington area tends to represent higher crime rate?**
- **From the dataset provided and analyses of the dataset, which year could provide a higher crime rate?**
- **Among the different types of crimes, which type of crime has the highest crime record?**
- **From the analyses of the data set which could be the most suitable time for the criminals to commit crime.**

I have examined the dataset by taking the fundamental traits, which are basic to imagine the information utilizing R, python, and scene as this could be useful in noting the above inquiries with a decent portrayal of information and could make a helpful determination and appropriate comprehension of the information

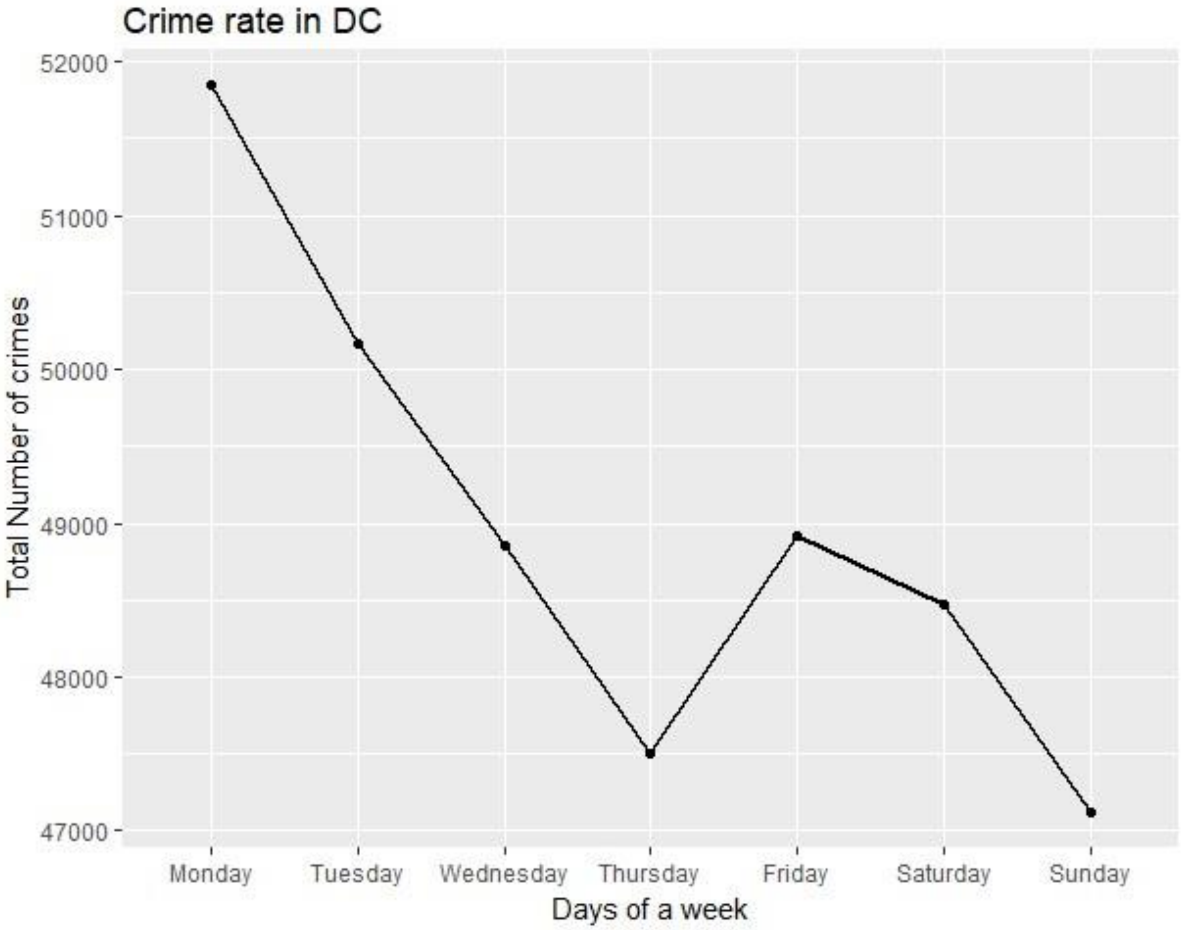


Fig 3: The Total number of crimes Vs Days of a week

I generally had an illusion that most of the crimes carried out by lawbreakers happen as a rule toward the week's end might resemble Saturdays or Sundays as a large portion of the general population are free normally at ends of the week and can have more opportunities to perpetrate a Crime. Yet, shockingly after a point by point investigation and comprehension from the chart delineated, it has been realized that most of the crimes will, in general, happen for the most part in the weekdays. The chart which I depicted utilizing R portrays that crime rate in DC is more on Mondays where almost there is a possibility of 52000 violations perpetrated by law breakers, the crime rate has diminished consistently by Thursdays about to 47500 violations and before the week's over, the crime rate has diminished to an incredible degree.

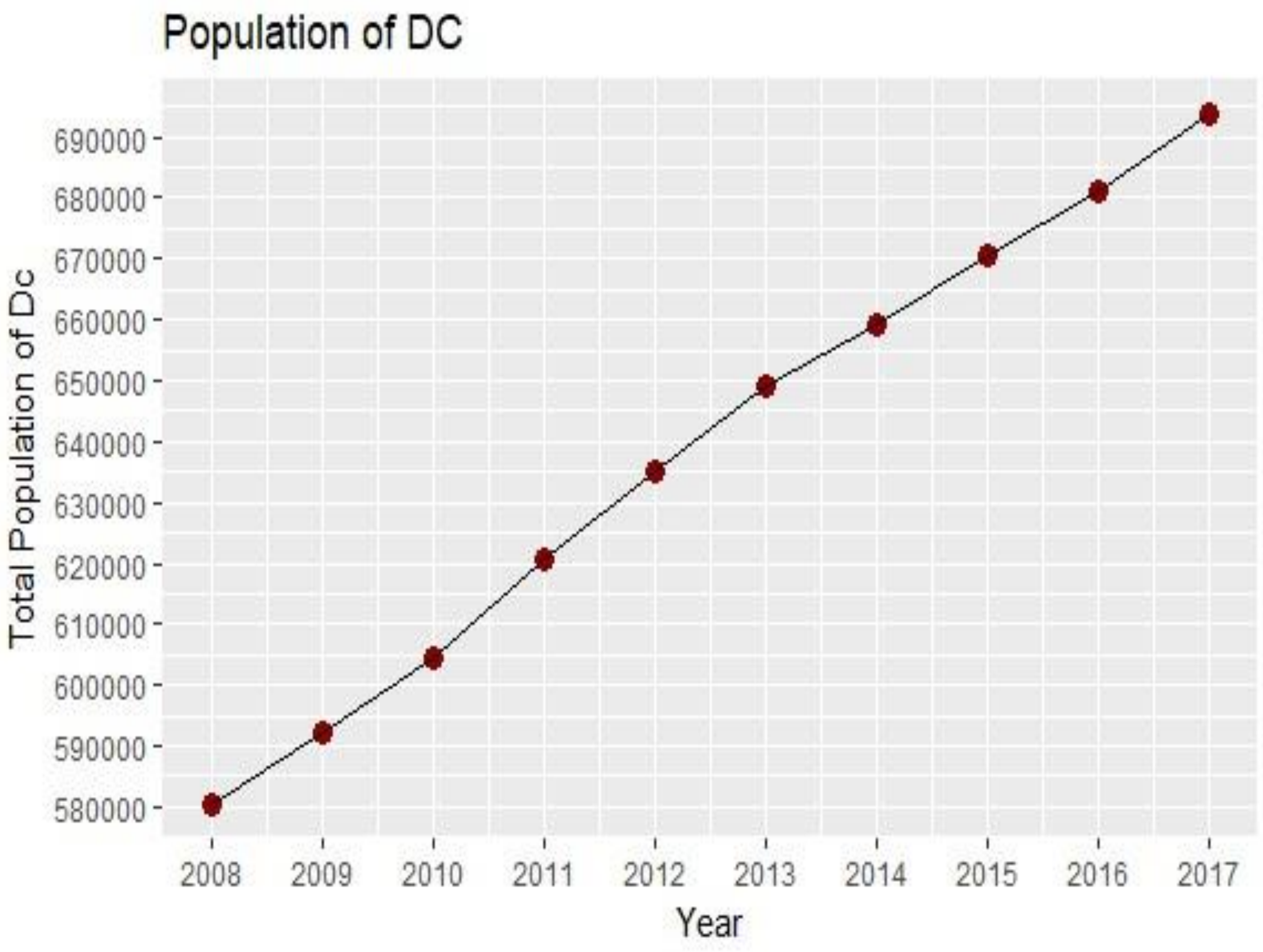


Fig 4: total population of DC from 2008-2017

The chart, which envisioned in R, portrays the aggregate populace of DC, which has been consistently expanding throughout the years from the year 2008 to 2017. It is seen from the chart that the aggregate populace in DC was 58000 in the year 2008 and has expanded to 69000 in the year 2017. In this way, the increment in populace could be one reason that could build the wrongdoing rate directly as because of more populace more could be the Crime rate.

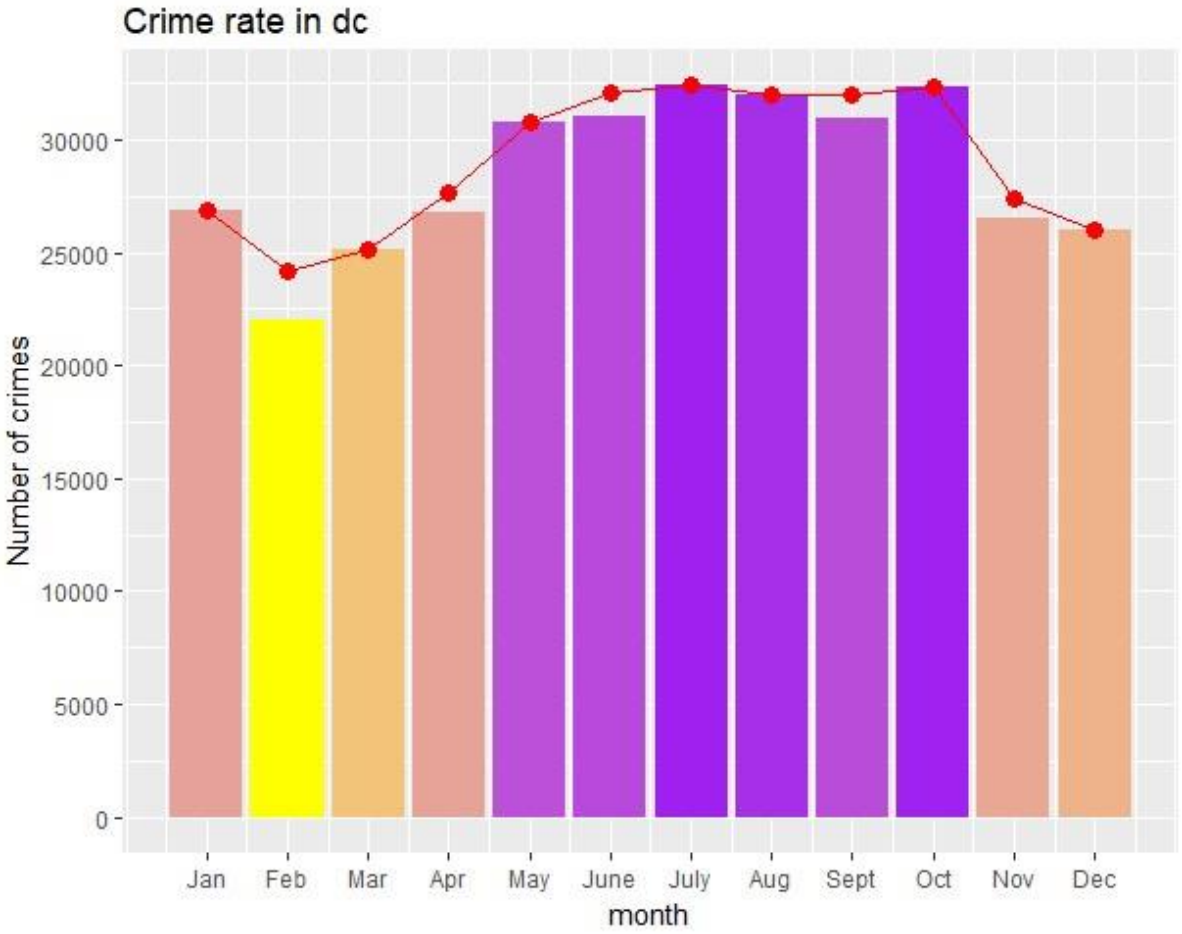


Fig 5: Total crimes Vs Months

This diagram delineates the aggregate number of crimes in DC from January to December. After a point by point investigation. I comprehended that the criminals had carried out a more prominent number of violations that are around 32500 crimes in the long stretch of October, which is featured by utilizing purple shading to separate the highest noteworthy wrongdoing rate and lower crime rate. All as the year progressed, relatively the crime rate is brought down in the period of February that is the start of the year, which is featured in yellow shading the crime rate is almost 22000 which is bring down in the number of violations when contrasted with rest of the months all as the year progressed.

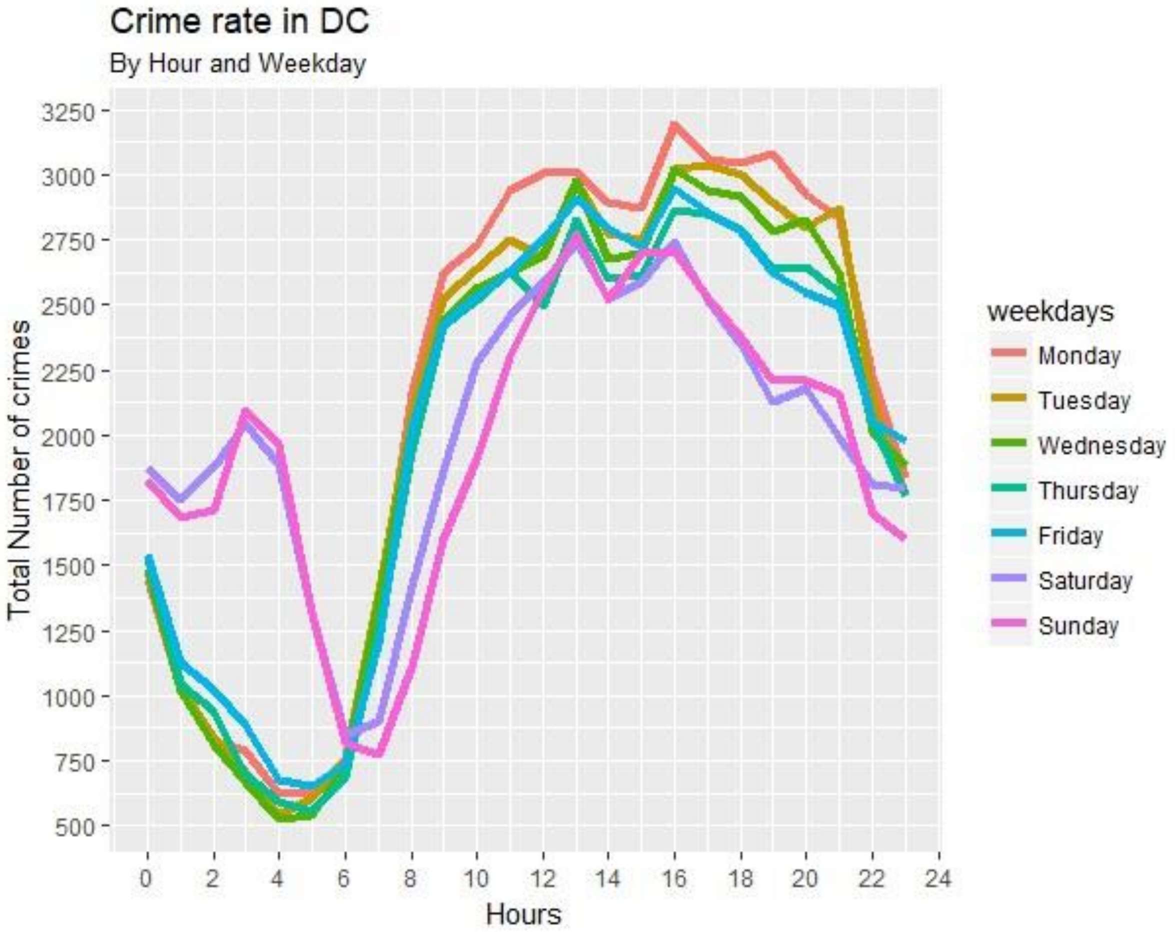


Fig 6: Number of Crimes Vs Hours

The diagram portrays the aggregate number of violations in DC that is the number of crimes that are appropriated all for the duration of the day, which is contrasted and the times of the week. From the graph, it is unmistakably comprehended that on Saturdays and Sundays generally amid the early hours of the day it is seen that the crime rate is higher when contrasted with whatever is left of the week. From the chart, it is likewise expressed that crime rates for the most part in ends of the week are lessened at around 6a.m toward the beginning of the day. It is closed plainly from the graph that lawbreakers who will, in general, carry out the crimes in the weekdays incline toward evening or early night through the crime rates by culprits in ends of the week are more in the murkiness of early mornings.

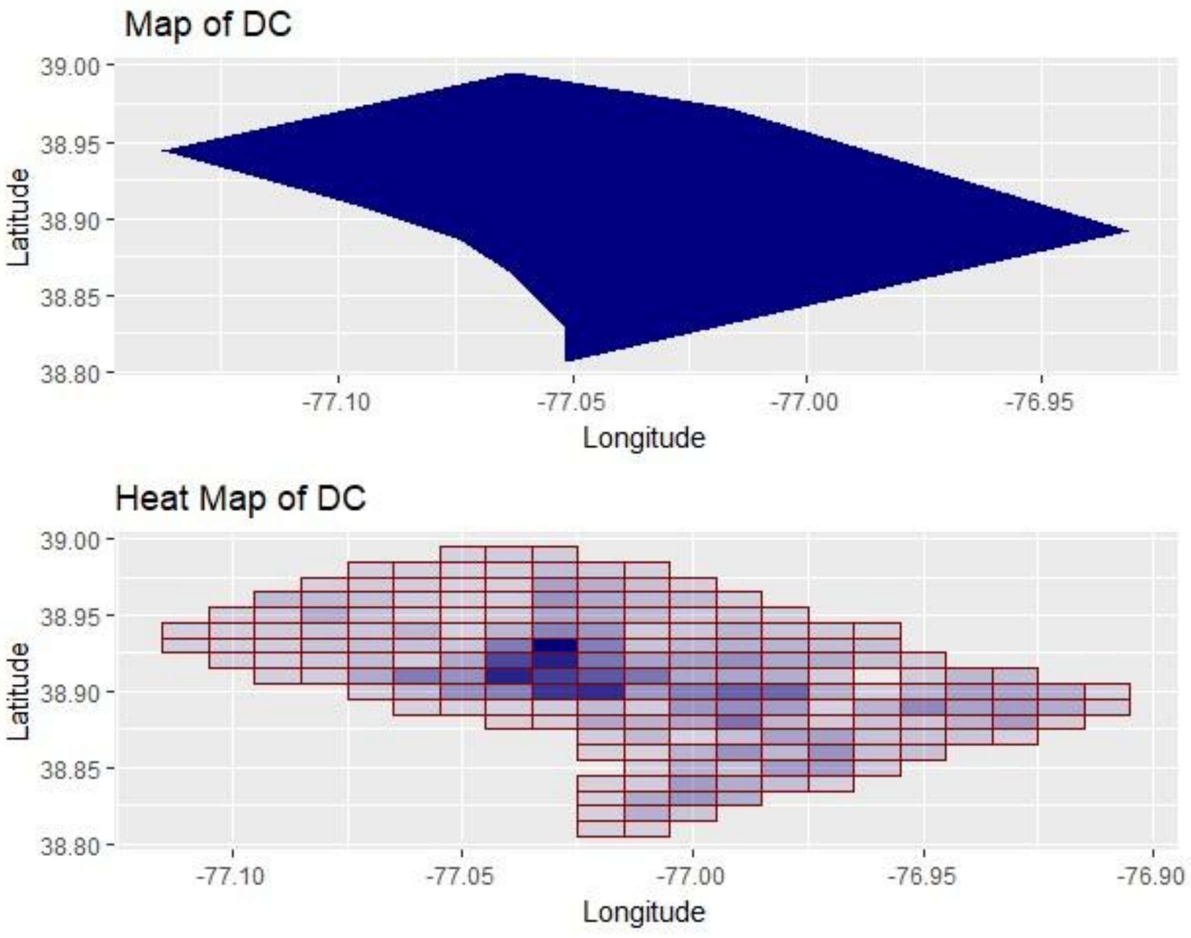
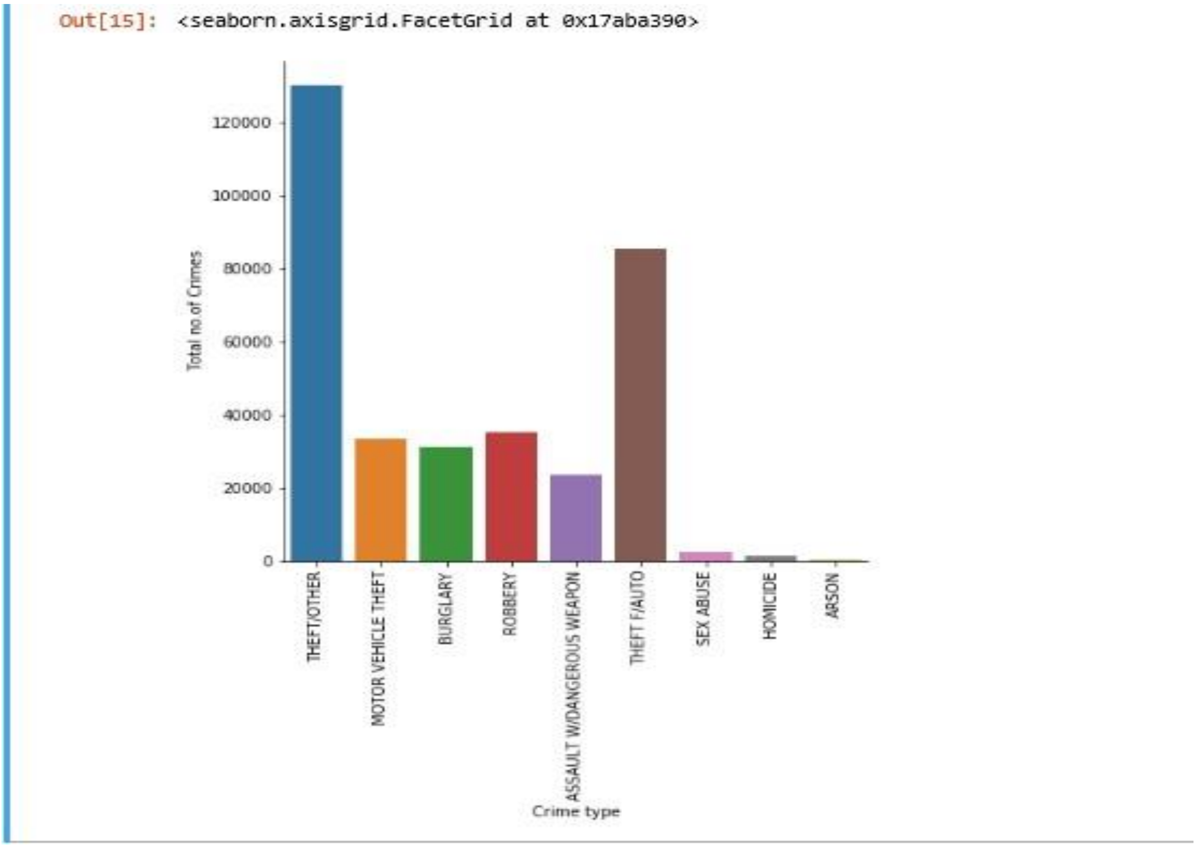


Fig 7: geographical view

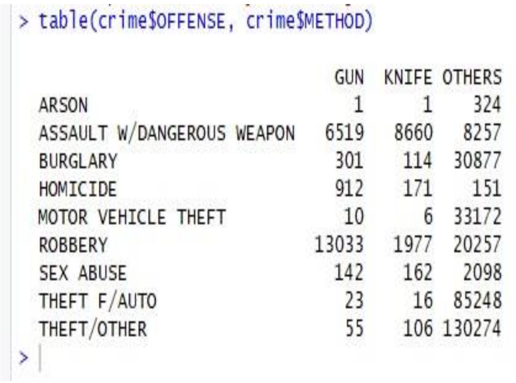
The grid view chart delineates the conveyance of crime rate all through most of the Washington territory. This warmth delineate demonstrates that the crime rate is brought down in the edge of the Washington DC and a greater amount of the crime rate is moved in a focal region which is plainly comprehended from the guide. The purple shading in the inside that is 38.90 N-77.03 W unmistakably portrays various crimes perpetrated by the lawbreakers were contrasted with the edge of the territory. Perhaps because of the nearness of the most essential spots like the white house more crimes could occur in the focal point of the zone.



```
Out[15]: <seaborn.axisgrid.FacetGrid at 0x17aba390>
```

Fig 8: Total crimes Vs types

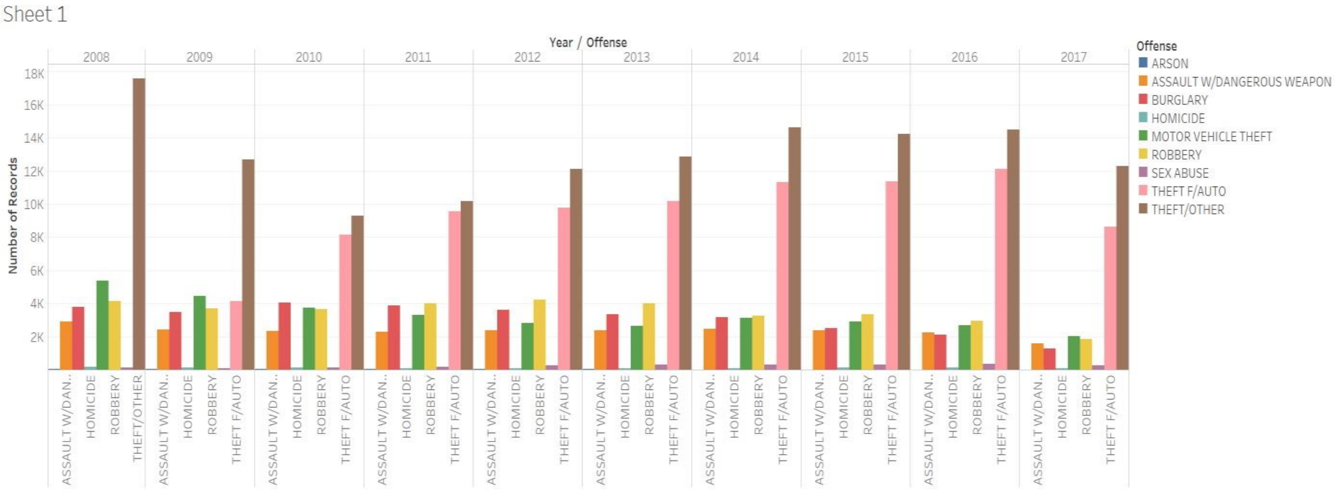
The above reference chart portrays the most fundamental thing that could be useful in giving enough data about the aggregate number of violations and kinds of crimes. I visualized this graph using python. They are a shifted number of crimes perpetrated by offenders of burglary, thievery, theft and some more. From examination and picturing a chart, it is comprehended that burglary records to most astounding number of violations that is it records to in excess of 12000 crimes while they are likewise sure violations, which record to about zero percent crime rate like the sex abuse, homicide, and arson. Motor vehicle theft, burglary likewise records to an about same crime rate that is around 43000 violations are perpetrated by the culprits. Along these lines, from the graph, it is reasoned that diverse kinds of crimes could prompt a more noteworthy increment in crime rate all as the year progressed.



```
> table(crime$OFFENSE, crime$METHOD)

                            GUN  KNIFE OTHERS
  ARSON                       1      1    324
  ASSAULT W/DANGEROUS WEAPON 6519   8660   8257
  BURGLARY                  301    114  30877
  HOMICIDE                  912    171    151
  MOTOR VEHICLE THEFT        10      6  33172
  ROBBERY                 13033   1977  20257
  SEX ABUSE                 142    162   2098
  THEFT F/AUTO               23     16  85248
  THEFT/OTHER                55    106 130274
> |
```

The table describes the types of crimes and specific crimes committed by the criminals . From the table it is comprehended that theft accounts to higher crime rate than other crimes including than gun and knife.



Sum of Number of Records for each Offense broken down by Year. Color shows details about Offense.

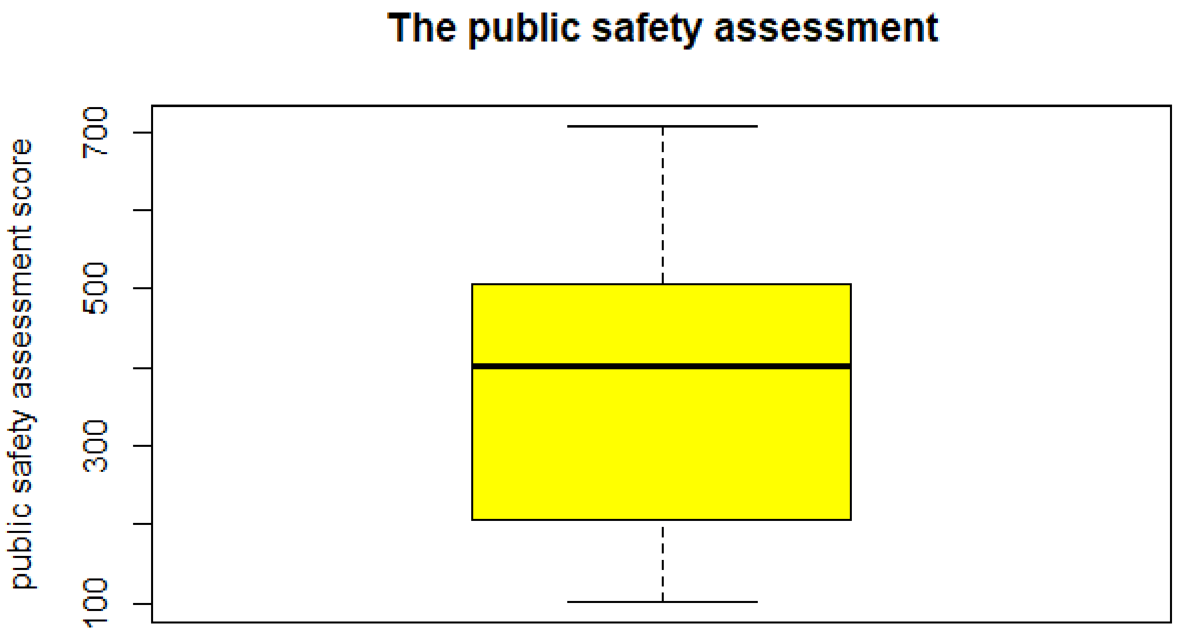
Figure 9: Total crimes/ offense

The above bar plot portrays different kinds of crimes and the number of violations carried out by the lawbreakers from a time of 2008-2017. I visualized the graph utilizing scene. In the graph, theft which is in darker shading records to the most astounding number of crimes in the year 2008 that is around 18,000 violations, from the plot it is comprehended that robbery vacillates all through the period. From the plot, it is comprehended that burglary which is in yellow shading is more in the year 2012 and 2013 and the minimum in the year 2017. It is finished up from the plot that Arson records to about to zero percent number of violations all through the period 2008 to 2018.

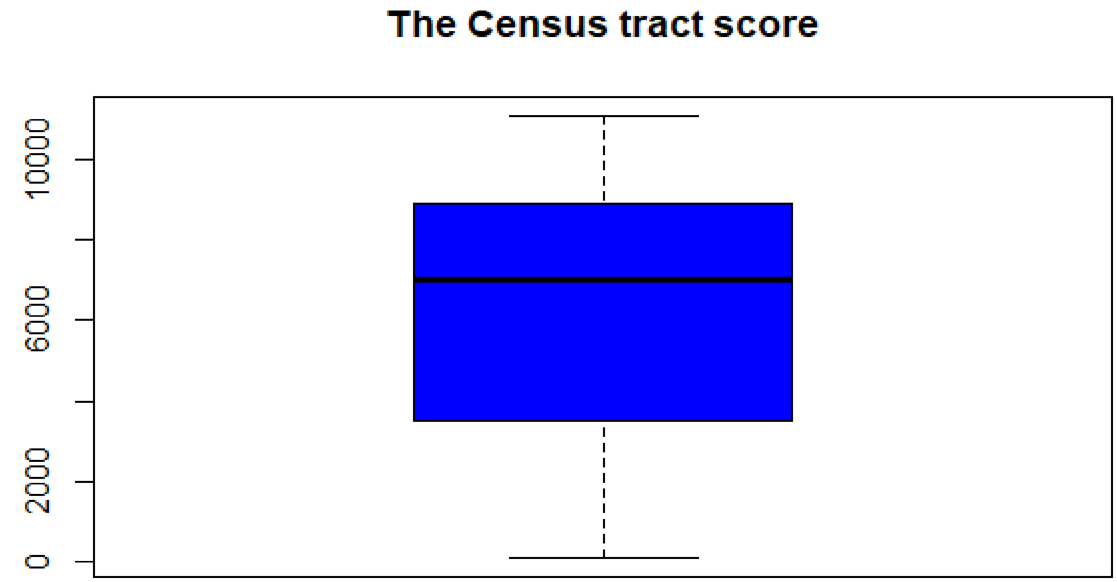
BOXPLOT

A boxplot of a variable is a graphical representation based on its quartiles as well as its smallest and largest values of the variable. It helps to provide a visual shape of the data distribution. (Mathur, n.d.)

The “PSA” or Public safety assessment is to help the judges gauge the risk that the defendant or criminal poses, i.e., the likelihood of an individual will commit a new crime if released before trial and to predict the likelihood that he will fail to return to future court hearing.



The average public safety assessment score approximately close to 400, with minimum as 100 and maximum approximately 700. The upper quartile is close to 500,i.e. 25% of the data is greater than this value and the lower quartile is close to 200 i.e. 25% of the data is less than this value.



Census tracts (CTs) are small, relatively stable geographic areas that usually have a population between 2,500 and 8,000 persons. They are in census metropolitan areas and in census agglomerations that had a core population of 50,000 or more in the previous census . (Census Tract Detailed Definition, 2018)

The minimum census tract score is close to 100 and the maximum census tract score is approximately 11100.The mean i.e. is the average census tract score is just above 6000 and the upper quartile is close to 9000 i.e. the 25% of the data is greater than this value ,lower quartile just below 4000 which states that 25% of the data is lesser than this value.

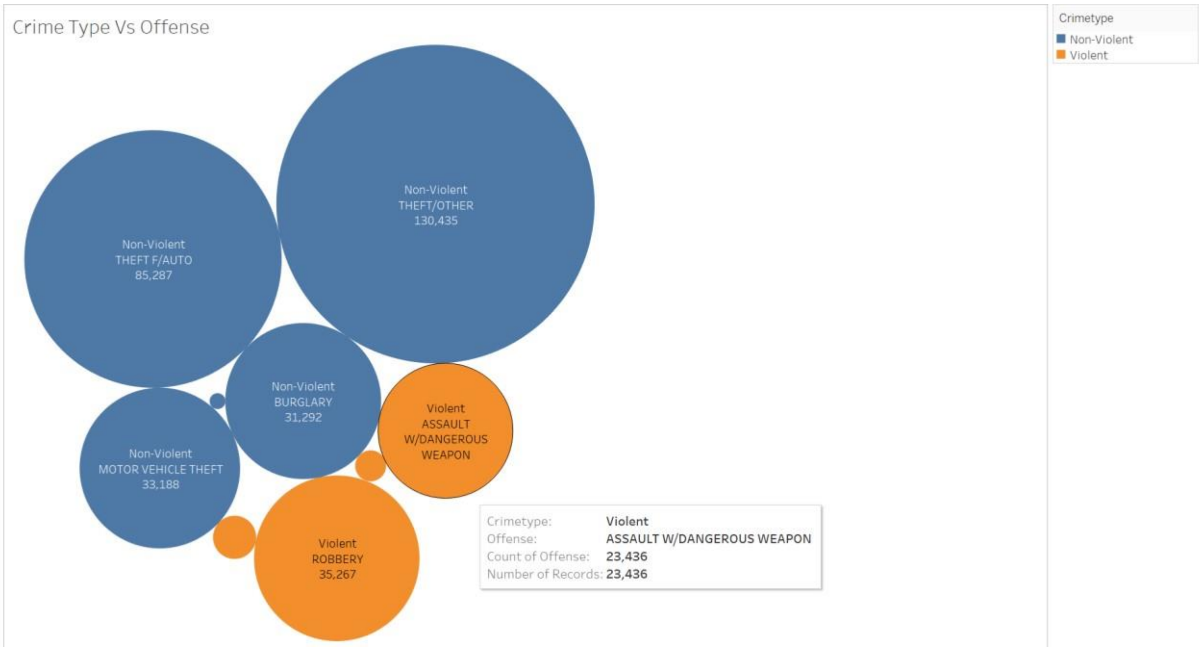


Fig 10: Bubble plot for crime type vs offense

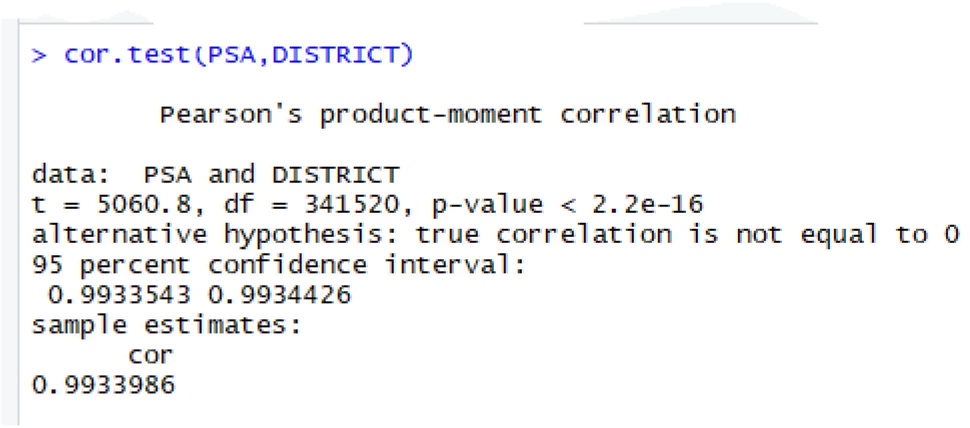
This bubble plot clearly displays the types of crimes classified as violent crime and non-violent crime. Violent crimes such as robbery and assault records to 32267 number of violations. From the graph depicted non violent acts such as theft accounts to the most astounding number of violations that is around 130435 crimes are committed by the criminals.

Correlation Test

The correlation test is used to evaluate the association between two or more variables.
I have done a correlation test using the variables “PSA” and “District”.
The Null Hypothesis is denoted as ‘H0’ and the Alternative Hypothesis is denoted as ‘H1’.

H0: The Public safety Assignment score and the districts are not significantly associated.

H1: The Public safety Assignment score and the districts are significantly associated.



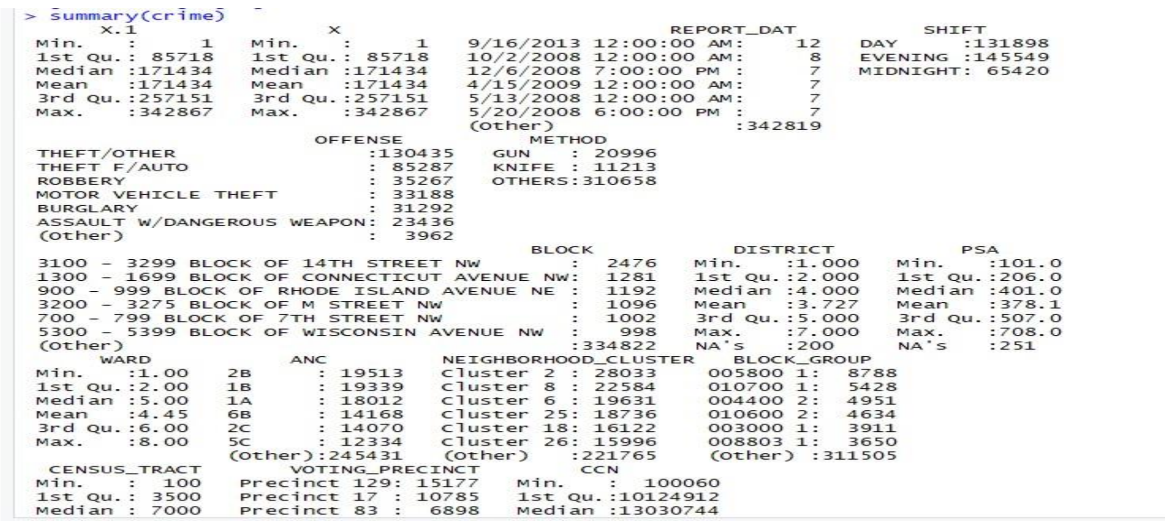
```
> cor.test(PSA,DISTRICT)

        Pearson's product-moment correlation

data:  PSA and DISTRICT
t = 5060.8, df = 341520, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9933543 0.9934426
sample estimates:
      cor
0.9933986
```

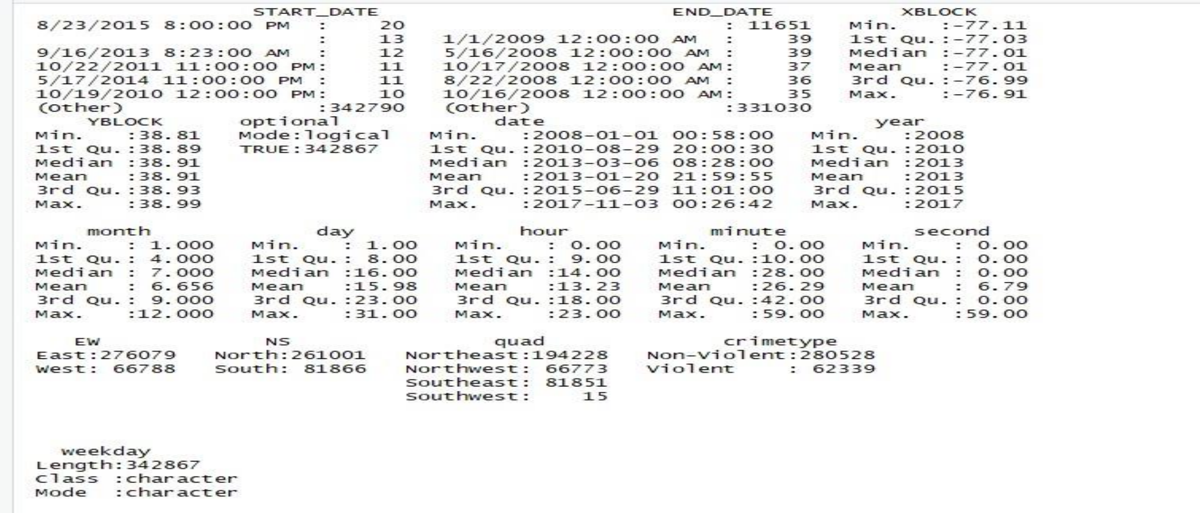
The P-value from the Pearson’s Product moment Correlational test is less 2.2e-16 i.e. the P-value is very small which rejects the Null Hypothesis H0 and proves the Alternative Hypothesis is true which means that there is a stronger relationship between the public safety assessment score and districts and are highly significantly correlated.

Descriptive statistics



```
> summary(crime)
      X.1              X                          REPORT_DAT           SHIFT
 Min.   :     1   Min.   :     1   9/16/2013 12:00:00 AM:     12   DAY     :131898
 1st Qu.: 85718   1st Qu.: 85718   10/2/2008 12:00:00 AM:      8   EVENING :145549
 Median :171434   Median :171434   12/6/2008 7:00:00 PM :      7   MIDNIGHT: 65420
 Mean   :171434   Mean   :171434   4/15/2009 12:00:00 AM:      7
 3rd Qu.:257151   3rd Qu.:257151   5/13/2008 12:00:00 AM:      7
 Max.   :342867   Max.   :342867   5/20/2008 6:00:00 PM :      7
                                   (Other)              :342819
                         OFFENSE             METHOD
 THEFT/OTHER                :130435   GUN    : 20996
 THEFT F/AUTO               : 85287   KNIFE  : 11213
 ROBBERY                    : 35267   OTHERS:310658
 MOTOR VEHICLE THEFT        : 33188
 BURGLARY                   : 31292
 ASSAULT W/DANGEROUS WEAPON: 23436
 (Other)                    :  3962
                                          BLOCK            DISTRICT           PSA
 3100 - 3299 BLOCK OF 14TH STREET NW        :  2476   Min.   :1.000   Min.   :101.0
 1300 - 1699 BLOCK OF CONNECTICUT AVENUE NW:  1281   1st Qu.:2.000   1st Qu.:206.0
 900 - 999 BLOCK OF RHODE ISLAND AVENUE NE :  1192   Median :4.000   Median :401.0
 3200 - 3275 BLOCK OF M STREET NW           :  1096   Mean   :3.727   Mean   :378.1
 700 - 799 BLOCK OF 7TH STREET NW           :  1002   3rd Qu.:5.000   3rd Qu.:507.0
 5300 - 5399 BLOCK OF WISCONSIN AVENUE NW   :   998   Max.   :7.000   Max.   :708.0
 (Other)                                    :334822   NA's   :200     NA's   :251
      WARD             ANC          NEIGHBORHOOD_CLUSTER    BLOCK_GROUP
 Min.   :1.00   2B       : 19513   Cluster 2 : 28033    005800 1:   8788
 1st Qu.:2.00   1B       : 19339   Cluster 8 : 22584    010700 1:   5428
 Median :5.00   1A       : 18012   Cluster 6 : 19631    004400 2:   4951
 Mean   :4.45   6B       : 14168   Cluster 25: 18736    010600 2:   4634
 3rd Qu.:6.00   2C       : 14070   Cluster 18: 16122    003000 1:   3911
 Max.   :8.00   5C       : 12334   Cluster 26: 15996    008803 1:   3650
                (Other):245431   (Other)   :221765    (Other) :311505
  CENSUS_TRACT      VOTING_PRECINCT          CCN
 Min.   :  100   Precinct 129: 15177   Min.   :  100060
 1st Qu.: 3500   Precinct 17 : 10785   1st Qu.:10124912
 Median : 7000   Precinct 83 :  6898   Median :13030744
```

Fig 11:Summary Statistical values

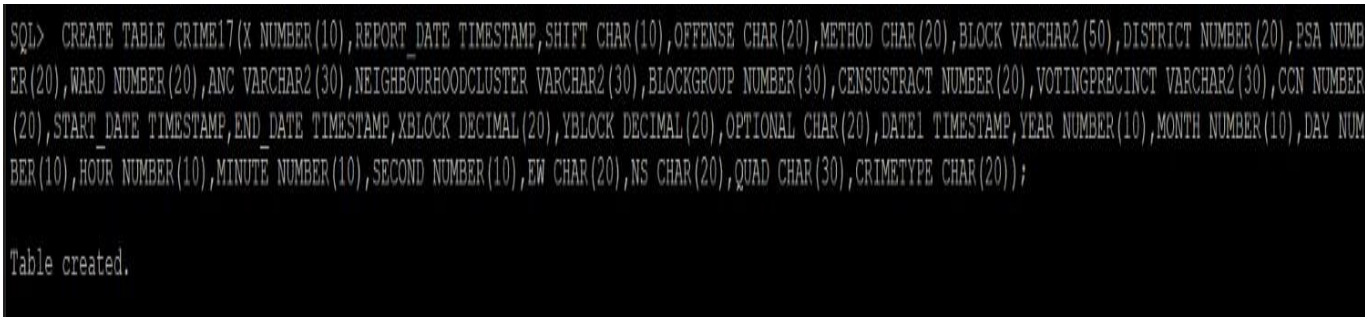


```
                START_DATE                                 END_DATE         XBLOCK      
 8/23/2015 8:00:00 PM  :     20                                  : 11651   Min.   :-77.11  
                       :     13   1/1/2009 12:00:00 AM  :     39   1st Qu.:-77.03  
 9/16/2013 8:23:00 AM  :     12   5/16/2008 12:00:00 AM :     39   Median :-77.01  
 10/22/2011 11:00:00 PM:     11   10/17/2008 12:00:00 AM:     37   Mean   :-77.01  
 5/17/2014 11:00:00 PM :     11   8/22/2008 12:00:00 AM :     36   3rd Qu.:-76.99  
 10/19/2010 12:00:00 PM:     10   10/16/2008 12:00:00 AM:     35   Max.   :-76.91  
 (Other)               :342790   (Other)                :331030                   
     YBLOCK       optional            date                          year     
 Min.   :38.81   Mode:logical   Min.   :2008-01-01 00:58:00   Min.   :2008  
 1st Qu.:38.89   TRUE:342867    1st Qu.:2010-08-29 20:00:30   1st Qu.:2010  
 Median :38.91                  Median :2013-03-06 08:28:00   Median :2013  
 Mean   :38.91                  Mean   :2013-01-20 21:59:55   Mean   :2013  
 3rd Qu.:38.93                  3rd Qu.:2015-06-29 11:01:00   3rd Qu.:2015  
 Max.   :38.99                  Max.   :2017-11-03 00:26:42   Max.   :2017  
                                                                            
     month             day             hour           minute          second     
 Min.   : 1.000   Min.   : 1.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00  
 1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.:10.00   1st Qu.: 0.00  
 Median : 7.000   Median :16.00   Median :14.00   Median :28.00   Median : 0.00  
 Mean   : 6.656   Mean   :15.98   Mean   :13.23   Mean   :26.29   Mean   : 6.79  
 3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:18.00   3rd Qu.:42.00   3rd Qu.: 0.00  
 Max.   :12.000   Max.   :31.00   Max.   :23.00   Max.   :59.00   Max.   :59.00  
                                                                                 
    EW              NS                 quad               crimetype     
 East:276079   North:261001   Northeast:194228   Non-Violent:280528  
 West: 66788   South: 81866   Northwest: 66773   Violent    : 62339  
                              Southeast: 81851                       
                              Southwest:    15                       
                                                                     
                                                                     
   weekday         
 Length:342867     
 Class :character  
 Mode  :character  
```

Fig 12: Summary Statistical values

SQL Schema and SQL based data exploration:

Creating the table

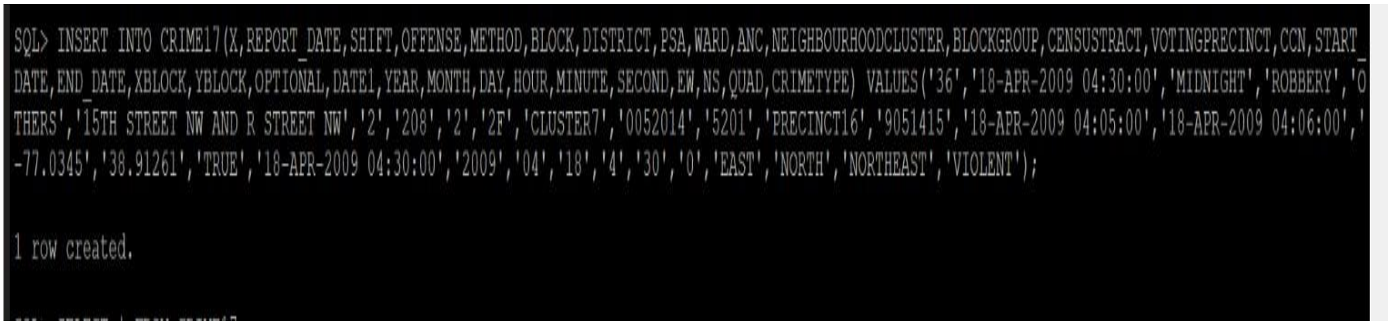


```
SQL>  CREATE TABLE CRIME17(X NUMBER(10),REPORT_DATE TIMESTAMP,SHIFT CHAR(10),OFFENSE CHAR(20),METHOD CHAR(20),BLOCK VARCHAR2(50),DISTRICT NUMBER(20),PSA NUMB
ER(20),WARD NUMBER(20),ANC VARCHAR2(30),NEIGHBOURHOODCLUSTER VARCHAR2(30),BLOCKGROUP NUMBER(30),CENSUSTRACT NUMBER(20),VOTINGPRECINCT VARCHAR2(30),CCN NUMBER
(20),START_DATE TIMESTAMP,END_DATE TIMESTAMP,XBLOCK DECIMAL(20),YBLOCK DECIMAL(20),OPTIONAL CHAR(20),DATE1 TIMESTAMP,YEAR NUMBER(10),MONTH NUMBER(10),DAY NUM
BER(10),HOUR NUMBER(10),MINUTE NUMBER(10),SECOND NUMBER(10),EW CHAR(20),NS CHAR(20),QUAD CHAR(30),CRIMETYPE CHAR(20));


Table created.
```

Fig 13:Creation of Table

Inserting the values into table

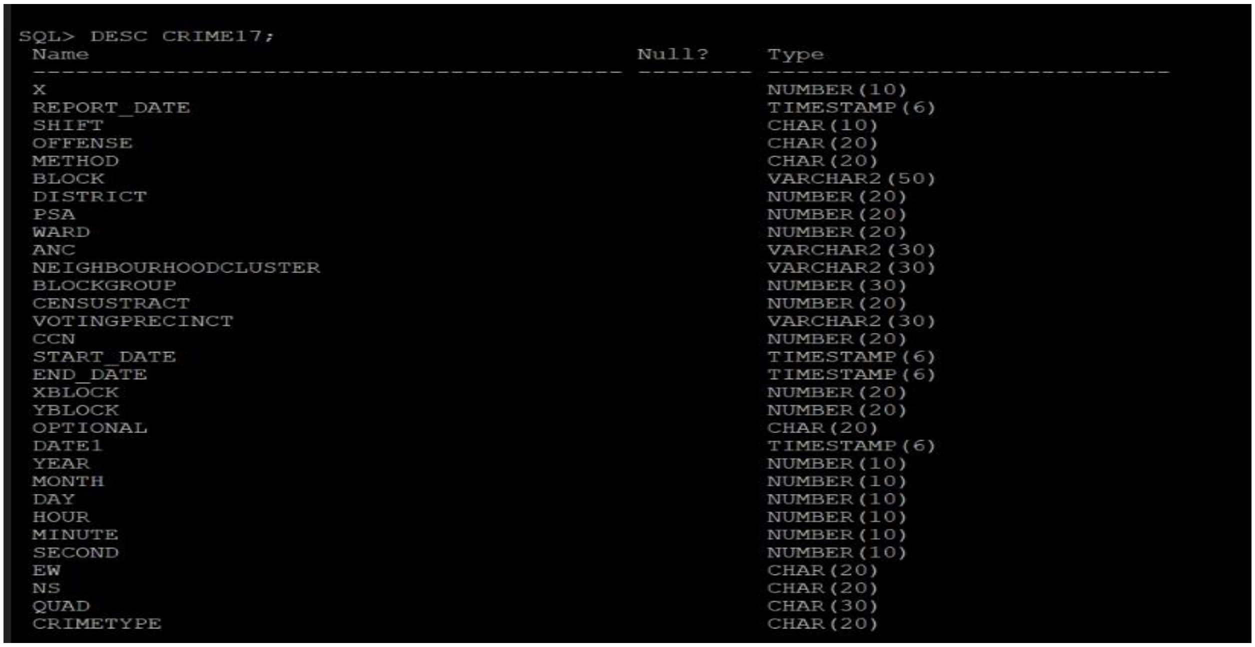


```
SQL> INSERT INTO CRIME17(X,REPORT_DATE,SHIFT,OFFENSE,METHOD,BLOCK,DISTRICT,PSA,WARD,ANC,NEIGHBOURHOODCLUSTER,BLOCKGROUP,CENSUSTRACT,VOTINGPRECINCT,CCN,START_
DATE,END_DATE,XBLOCK,YBLOCK,OPTIONAL,DATE1,YEAR,MONTH,DAY,HOUR,MINUTE,SECOND,EW,NS,QUAD,CRIMETYPE) VALUES('36','18-APR-2009 04:30:00','MIDNIGHT','ROBBERY','O
THERS','15TH STREET NW AND R STREET NW','2','208','2','2F','CLUSTER7','0052014','5201','PRECINCT16','9051415','18-APR-2009 04:05:00','18-APR-2009 04:06:00','
-77.0345','38.91261','TRUE','18-APR-2009 04:30:00','2009','04','18','4','30','0','EAST','NORTH','NORTHEAST','VIOLENT');

1 row created.
```

Fig 14: Insertion of Values

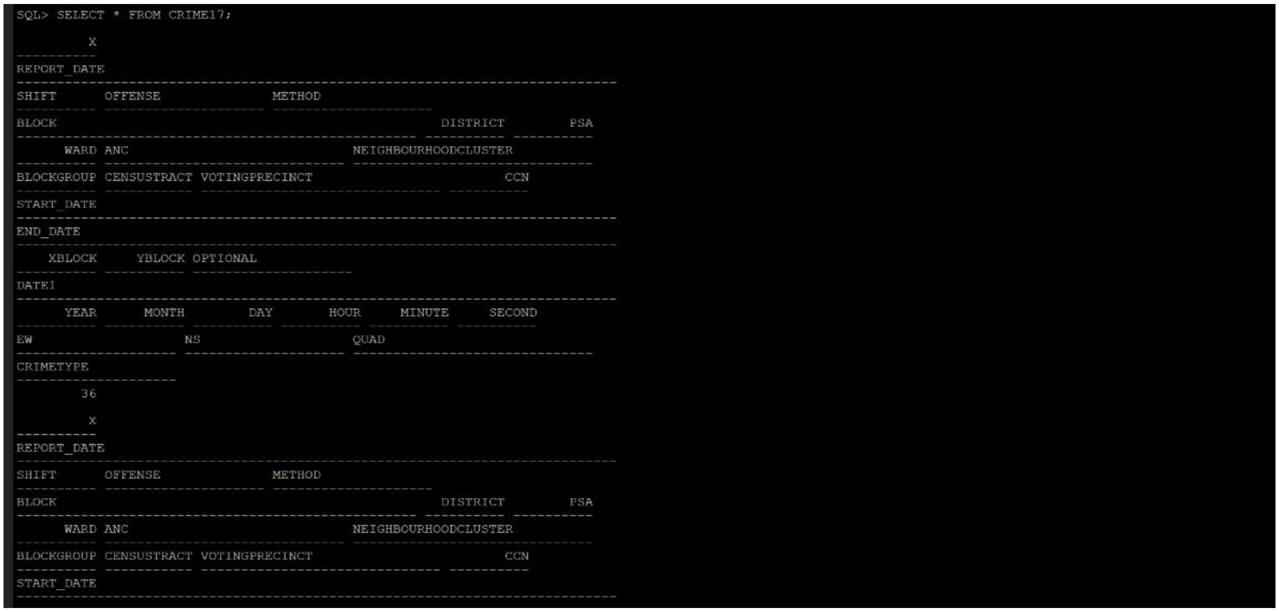
Describing the table



```
SQL> DESC CRIME17;
 Name                                      Null?    Type
 ----------------------------------------- -------- ----------------------------
 X                                                  NUMBER(10)
 REPORT_DATE                                        TIMESTAMP(6)
 SHIFT                                              CHAR(10)
 OFFENSE                                            CHAR(20)
 METHOD                                             CHAR(20)
 BLOCK                                              VARCHAR2(50)
 DISTRICT                                           NUMBER(20)
 PSA                                                NUMBER(20)
 WARD                                               NUMBER(20)
 ANC                                                VARCHAR2(30)
 NEIGHBOURHOODCLUSTER                               VARCHAR2(30)
 BLOCKGROUP                                         NUMBER(30)
 CENSUSTRACT                                        NUMBER(20)
 VOTINGPRECINCT                                     VARCHAR2(30)
 CCN                                                NUMBER(20)
 START_DATE                                         TIMESTAMP(6)
 END_DATE                                           TIMESTAMP(6)
 XBLOCK                                             NUMBER(20)
 YBLOCK                                             NUMBER(20)
 OPTIONAL                                           CHAR(20)
 DATE1                                              TIMESTAMP(6)
 YEAR                                               NUMBER(10)
 MONTH                                              NUMBER(10)
 DAY                                                NUMBER(10)
 HOUR                                               NUMBER(10)
 MINUTE                                             NUMBER(10)
 SECOND                                             NUMBER(10)
 EW                                                 CHAR(20)
 NS                                                 CHAR(20)
 QUAD                                               CHAR(30)
 CRIMETYPE                                          CHAR(20)
```

Fig 15: Description of Table

Selection of Variables



```
SQL> SELECT * FROM CRIME17;

         X
----------
REPORT_DATE
---------------------------------------------------------------------------
SHIFT      OFFENSE              METHOD
---------- -------------------- --------------------
BLOCK                                                 DISTRICT        PSA
--------------------------------------------------- ---------- ----------
      WARD ANC                            NEIGHBOURHOODCLUSTER
---------- ------------------------------ -------------------------------
BLOCKGROUP CENSUSTRACT VOTINGPRECINCT                        CCN
---------- ----------- ------------------------------ ----------
START_DATE
---------------------------------------------------------------------------
END_DATE
---------------------------------------------------------------------------
    XBLOCK     YBLOCK OPTIONAL
---------- ---------- --------------------
DATE1
---------------------------------------------------------------------------
      YEAR      MONTH        DAY       HOUR     MINUTE     SECOND
---------- ---------- ---------- ---------- ---------- ----------
EW                   NS                   QUAD
-------------------- -------------------- -------------------------------
CRIMETYPE
--------------------
        36

         X
----------
REPORT_DATE
---------------------------------------------------------------------------
SHIFT      OFFENSE              METHOD
---------- -------------------- --------------------
BLOCK                                                 DISTRICT        PSA
--------------------------------------------------- ---------- ----------
      WARD ANC                            NEIGHBOURHOODCLUSTER
---------- ------------------------------ -------------------------------
BLOCKGROUP CENSUSTRACT VOTINGPRECINCT                        CCN
---------- ----------- ------------------------------ ----------
START_DATE
---------------------------------------------------------------------------
```

Fig 16: Attribute Selection

CONCLUSION:

After a total understanding and more profound examinations of various fields in the dataset. This dataset and investigations of the dataset could be useful to assemble a prescient model which could anticipate the future violations to a palatable level. Over, the entire range of the task that has engaged with different stages like information cleaning, analyzing, and visualizing with probably the best instruments that could deliver good visualizations which could be envisioned legitimately. I have significant information on every one of tools, technologies used to investigate, translate and envision the information.

Through this project, I have figured out how to get a more profound learning of the considerable number of tools utilized alongside the most critical resources that are utilized. It is additionally plainly comprehended about the working of the scene. Python and R that could be extremely helpful in visualizing the information appropriately to get the correct output.

Technical terms

CURATION- This is one of techniques which involves cleaning the data completely in order to remove missing data, null values if present they could be removed by replacing the missing data with a most probable value, or using a mean, median value. On proper data curation, it could retain the quality of data that could be helpful in getting a good outcome.

Predictive model- This is one of kind of model which uses data mining techniques to predict future outcomes with a defined success criterion.

References

1. References

(n.d.). Retrieved from google: https://www.google.com/search?biw=1093&bih=501&ei=hEjyW7aQG8GyggeH5JOgDA&q=kaggle&oq=kaggle&gs_l=psy-ab.3..0l10.62888293.62889012..62889212...0.0..0.111.625.2j4......0....1..gws-wiz.......0i131j0i67j0i131i67j0i10.XMTwqnL1Vak

Census Tract Detailed Definition. (2018, 09 17). Retrieved from symbol os statistics canada: https://www.google.com/search?ei=888GXKGWIY7z5gK8iYmgDw&q=census+tract+meaning+&oq=census+tract+meaning+&gs_l=psy-ab.3..0i22i30l2.19147.24944..25866...1.0..0.258.466.4j0j1......0....1..gws-wiz.......0j0i71j0i67.G1leDE7-794

Mathur, P. S. (n.d.). *Single variable visualization*. Retrieved from Udemy: https://www.udemy.com/draft/1159488/learn/v4/t/lecture/6849526?start=27

Metro Police Department. (n.d.). Retrieved from Dc.gov: https://mpdc.dc.gov/

Vinze, L. (2017, nov 11). *DC Metro Crime Data*. Retrieved from kaggle: https://www.kaggle.com/vinchinzu/dc-metro-crime-data/home