CCT College Dublin

Module Title:	Data Exploration and Preparation
Assessment Title:	Individual
Lecturer Name:	Dr Muhammad lqbal
Student Full Name/ID:	Lucresse Pearle Tchatchoua Mbakop – 2021404
Assessment Due Date:	3rd December 2023 23:59
Date of Submission:	

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Table of content

GitHub:

- LucressePearle/DEP_CA1: Data Exploration and Preparation (github.com)

Introduction

In the past two decades, the amount of data being produce has drastically improve, the volume of data being recorded in all various erea of civilisation is impressing. In order to get the must out of this evolution we have to transform and clean the data to get insightful informations. In this project we are going to understand and prepare a data set in the aim of getting clair and useful informations about crimes by the use of R programming language. The objectif here will be to identify our variable types, the correlation group and the subgroup identification throug the use of method such as principal component analysis and many others.

Data Set

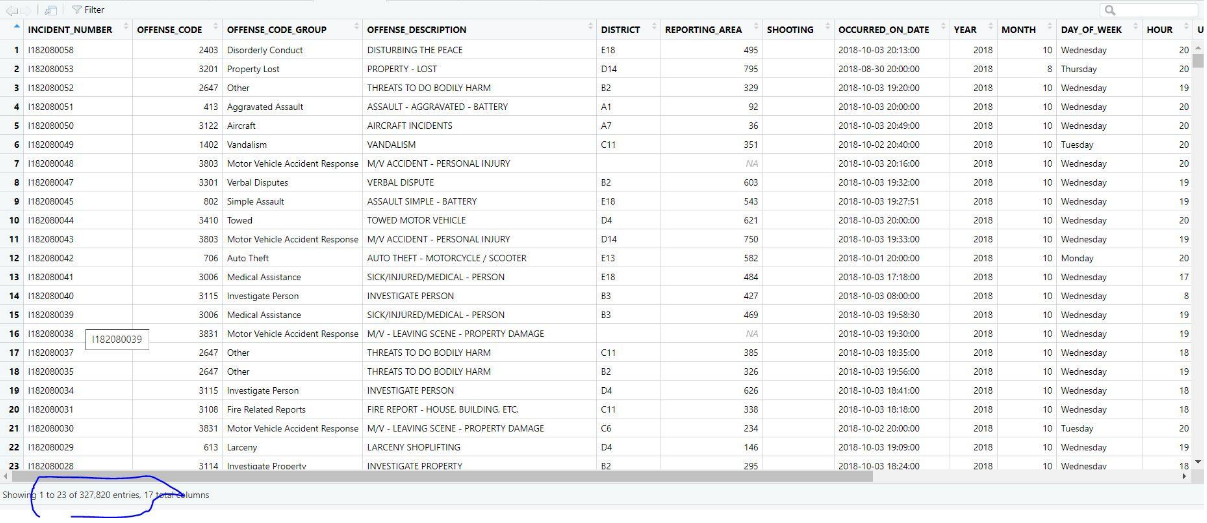
Data Set Introduction

For this project, I will be working with a crime data set called “crime” , this data set can be found at the Kaggle and deals with various crime observe in the United States(US). Our crime data set contains records of the crimes in the reposrt systems. This datasets has 17 columns and approximately 300,000 rows. Some of the variables are : Incident_Number, Offense_Code_Group, Offense_Description, Shooting , Year, Hour and many more.

The main idea on why I decided to work with a crime data set was to acknowledge how crimes have changed in the recent years and to have a better insights on what are the most commun crimes by using the data visualization techniques i have learn to get the most meaningful informations out of this data.

The first step consisted of importing the data set into our Rstudio environment and taking a grasp of it.

Image 1: A Display of the original data set



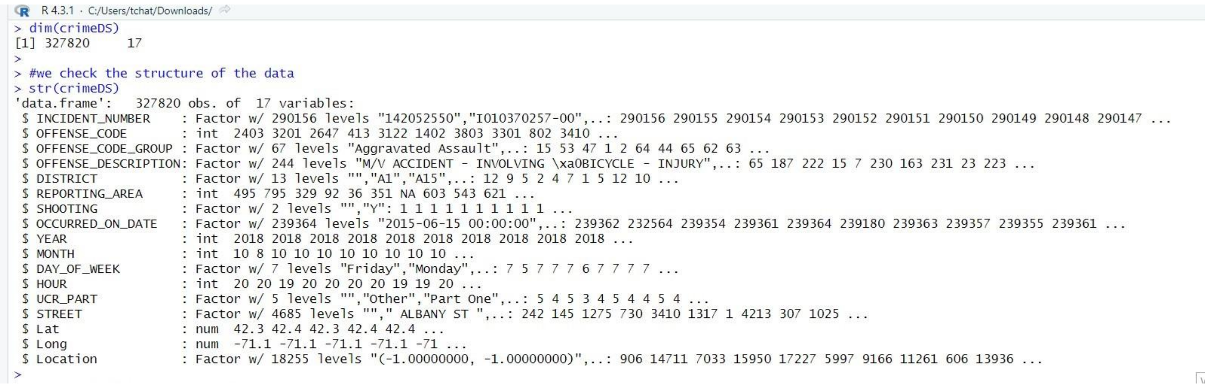
	INCIDENT_NUMBER	OFFENSE_CODE	OFFENSE_CODE_GROUP	OFFENSE_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR	U
1	I182080058	2403	Disorderly Conduct	DISTURBING THE PEACE	E18	495		2018-10-03 20:13:00	2018	10	Wednesday	20	
2	I182080053	3201	Property Lost	PROPERTY - LOST	D14	795		2018-08-30 20:00:00	2018	8	Thursday	20	
3	I182080052	2647	Other	THREATS TO DO BODILY HARM	B2	329		2018-10-03 19:20:00	2018	10	Wednesday	19	
4	I182080051	413	Aggravated Assault	ASSAULT - AGGRAVATED - BATTERY	A1	92		2018-10-03 20:00:00	2018	10	Wednesday	20	
5	I182080050	3122	Aircraft	AIRCRAFT INCIDENTS	A7	36		2018-10-03 20:49:00	2018	10	Wednesday	20	
6	I182080049	1402	Vandalism	VANDALISM	C11	351		2018-10-02 20:40:00	2018	10	Tuesday	20	
7	I182080048	3803	Motor Vehicle Accident Response	M/V ACCIDENT - PERSONAL INJURY		NA		2018-10-03 20:16:00	2018	10	Wednesday	20	
8	I182080047	3301	Verbal Disputes	VERBAL DISPUTE	B2	603		2018-10-03 19:32:00	2018	10	Wednesday	19	
9	I182080045	802	Simple Assault	ASSAULT SIMPLE - BATTERY	E18	543		2018-10-03 19:27:51	2018	10	Wednesday	19	
10	I182080044	3410	Towed	TOWED MOTOR VEHICLE	D4	621		2018-10-03 20:00:00	2018	10	Wednesday	20	
11	I182080043	3803	Motor Vehicle Accident Response	M/V ACCIDENT - PERSONAL INJURY	D14	750		2018-10-03 19:33:00	2018	10	Wednesday	19	
12	I182080042	706	Auto Theft	AUTO THEFT - MOTORCYCLE / SCOOTER	E13	582		2018-10-01 20:00:00	2018	10	Monday	20	
13	I182080041	3006	Medical Assistance	SICK/INJURED/MEDICAL - PERSON	E18	484		2018-10-03 17:18:00	2018	10	Wednesday	17	
14	I182080040	3115	Investigate Person	INVESTIGATE PERSON	B3	427		2018-10-03 08:00:00	2018	10	Wednesday	8	
15	I182080039	3006	Medical Assistance	SICK/INJURED/MEDICAL - PERSON	B3	469		2018-10-03 19:58:30	2018	10	Wednesday	19	
16	I182080038 I182080039	3831	Motor Vehicle Accident Response	M/V - LEAVING SCENE - PROPERTY DAMAGE		NA		2018-10-03 19:30:00	2018	10	Wednesday	19	
17	I182080037	2647	Other	THREATS TO DO BODILY HARM	C11	385		2018-10-03 18:35:00	2018	10	Wednesday	18	
18	I182080035	2647	Other	THREATS TO DO BODILY HARM	B2	326		2018-10-03 19:56:00	2018	10	Wednesday	19	
19	I182080034	3115	Investigate Person	INVESTIGATE PERSON	D4	626		2018-10-03 18:41:00	2018	10	Wednesday	18	
20	I182080031	3108	Fire Related Reports	FIRE REPORT - HOUSE, BUILDING, ETC.	C11	338		2018-10-03 18:18:00	2018	10	Wednesday	18	
21	I182080030	3831	Motor Vehicle Accident Response	M/V - LEAVING SCENE - PROPERTY DAMAGE	C6	234		2018-10-02 20:00:00	2018	10	Tuesday	20	
22	I182080029	613	Larceny	LARCENY SHOPLIFTING	D4	146		2018-10-03 19:09:00	2018	10	Wednesday	19	
23	I182080028	3114	Investigate Property	INVESTIGATE PROPERTY	B2	295		2018-10-03 18:24:00	2018	10	Wednesday	18	

Showing 1 to 23 of 327,820 entries, 17 total columns

Data Set Summary

As we observe our data, we move on to the data mining stage, where i mainly focus on clean the data. Before that i did a summary , this to have the number of rows and columns available in my dataset.

Image2: A brief summary

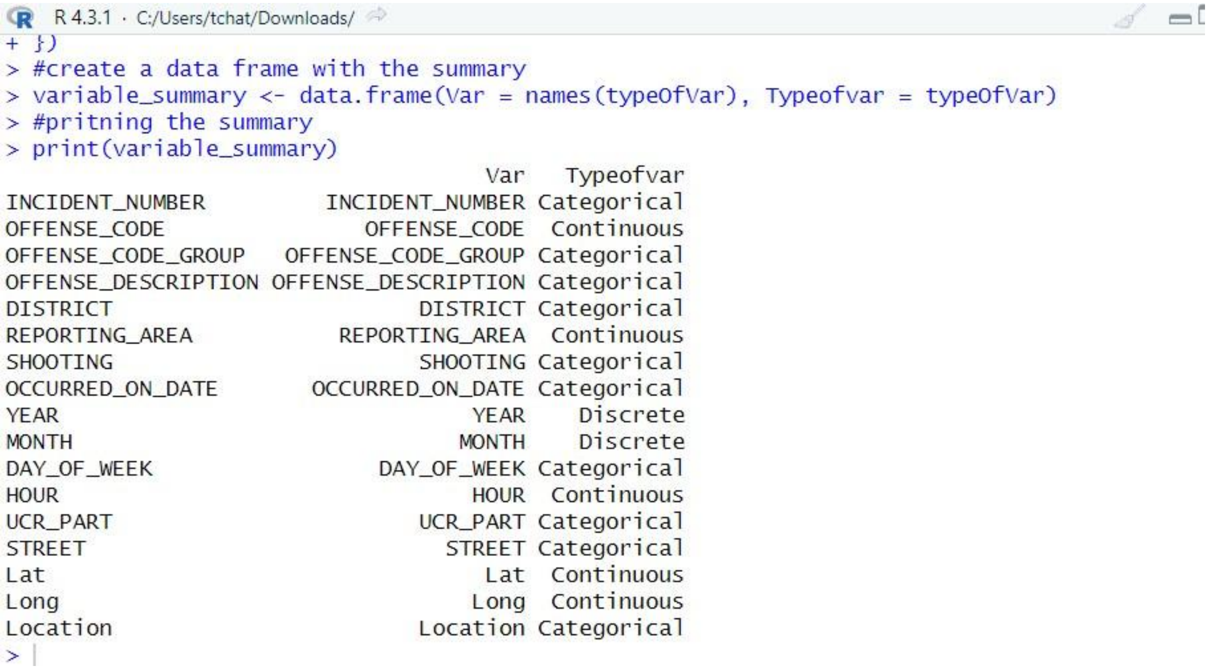


```
> dim(crimeDS)
[1] 327820     17
>
> #we check the structure of the data
> str(crimeDS)
'data.frame':	327820 obs. of  17 variables:
 $ INCIDENT_NUMBER    : Factor w/ 290156 levels "142052550","I010370257-00",..: 290156 290155 290154 290153 290152 290151 290150 290149 290148 290147 ...
 $ OFFENSE_CODE       : int  2403 3201 2647 413 3122 1402 3803 3301 802 3410 ...
 $ OFFENSE_CODE_GROUP : Factor w/ 67 levels "Aggravated Assault",..: 15 53 47 1 2 64 44 65 62 63 ...
 $ OFFENSE_DESCRIPTION: Factor w/ 244 levels "M/V ACCIDENT - INVOLVING \xa0BICYCLE - INJURY",..: 65 187 222 15 7 230 163 231 23 223 ...
 $ DISTRICT           : Factor w/ 13 levels "","A1","A15",..: 12 9 5 2 4 7 1 5 12 10 ...
 $ REPORTING_AREA     : int  495 795 329 92 36 351 NA 603 543 621 ...
 $ SHOOTING           : Factor w/ 2 levels "","Y": 1 1 1 1 1 1 1 1 1 1 ...
 $ OCCURRED_ON_DATE   : Factor w/ 239364 levels "2015-06-15 00:00:00",..: 239362 232564 239354 239361 239364 239180 239363 239357 239355 239361 ...
 $ YEAR               : int  2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
 $ MONTH              : int  10 8 10 10 10 10 10 10 10 10 ...
 $ DAY_OF_WEEK        : Factor w/ 7 levels "Friday","Monday",..: 7 5 7 7 7 6 7 7 7 7 ...
 $ HOUR               : int  20 20 19 20 20 20 20 19 19 20 ...
 $ UCR_PART           : Factor w/ 5 levels "","Other","Part One",..: 5 4 5 3 4 5 4 4 5 4 ...
 $ STREET             : Factor w/ 4685 levels ""," ALBANY ST ",..: 242 145 1275 730 3410 1317 1 4213 307 1025 ...
 $ Lat                : num  42.3 42.4 42.3 42.4 42.4 ...
 $ Long               : num  -71.1 -71.1 -71.1 -71.1 -71 ...
 $ Location           : Factor w/ 18255 levels "(-1.00000000, -1.00000000)",..: 906 14711 7033 15950 17227 5997 9166 11261 606 13936 ...
>
```

Variable type

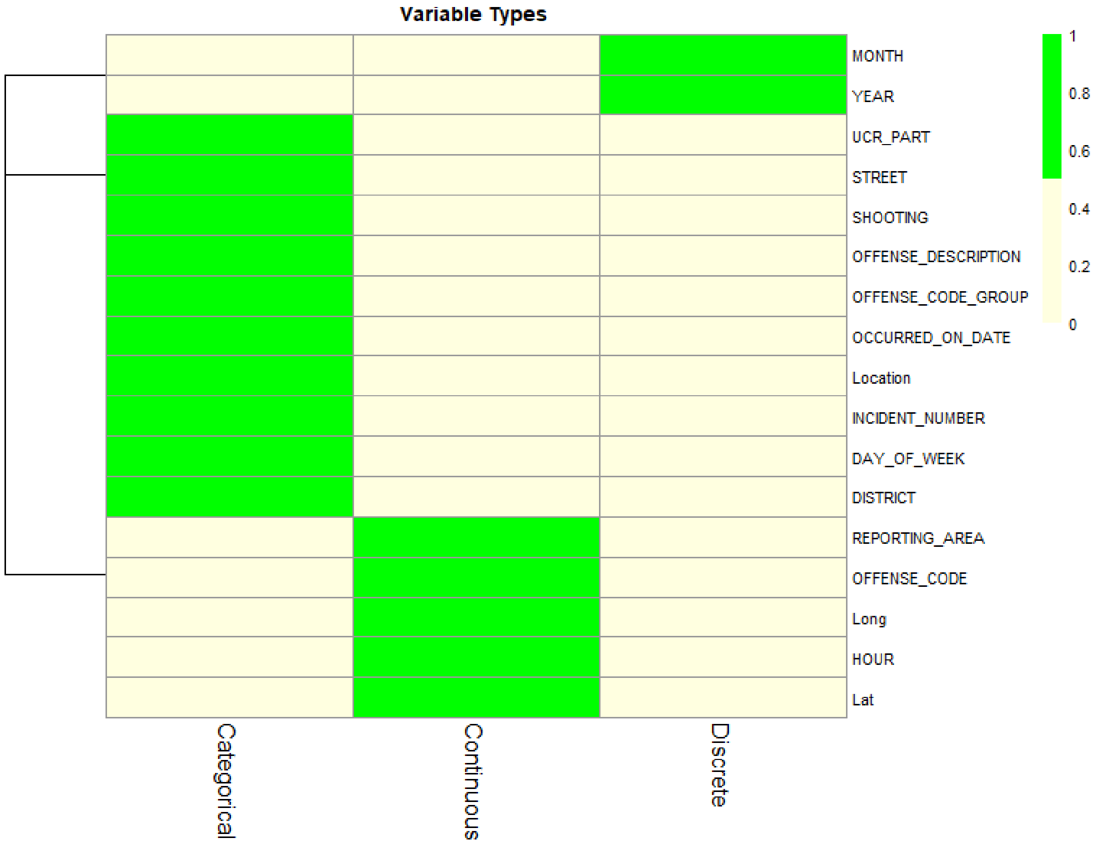
Still into the preparations of data, I explore the data set to understand the nature and variables type. This step is to make sure i have an insight in the data structure, knowing how to apply functions cause differents variable means implementing differents data analysis methods. Therefore this step helps me on the way i will handle and visualise my data.
To get the variable type, i created a function that goes through all the variable and check if they are categorical, discrete or continuous. In the following images, i want ran the functions and created a ggplot to show the distribution of variable type

Image 3:



```
+ })
> #create a data frame with the summary
> variable_summary <- data.frame(Var = names(typeOfVar), Typeofvar = typeOfVar)
> #pritning the summary
> print(variable_summary)
                                    Var   Typeofvar
INCIDENT_NUMBER         INCIDENT_NUMBER Categorical
OFFENSE_CODE               OFFENSE_CODE  Continuous
OFFENSE_CODE_GROUP   OFFENSE_CODE_GROUP Categorical
OFFENSE_DESCRIPTION OFFENSE_DESCRIPTION Categorical
DISTRICT                       DISTRICT Categorical
REPORTING_AREA           REPORTING_AREA  Continuous
SHOOTING                       SHOOTING Categorical
OCCURRED_ON_DATE       OCCURRED_ON_DATE Categorical
YEAR                               YEAR    Discrete
MONTH                             MONTH    Discrete
DAY_OF_WEEK                 DAY_OF_WEEK Categorical
HOUR                               HOUR  Continuous
UCR_PART                       UCR_PART Categorical
STREET                           STREET Categorical
Lat                                 Lat  Continuous
Long                               Long  Continuous
Location                       Location Categorical
> 
```

Image 4: Heat map to visualise the variable type

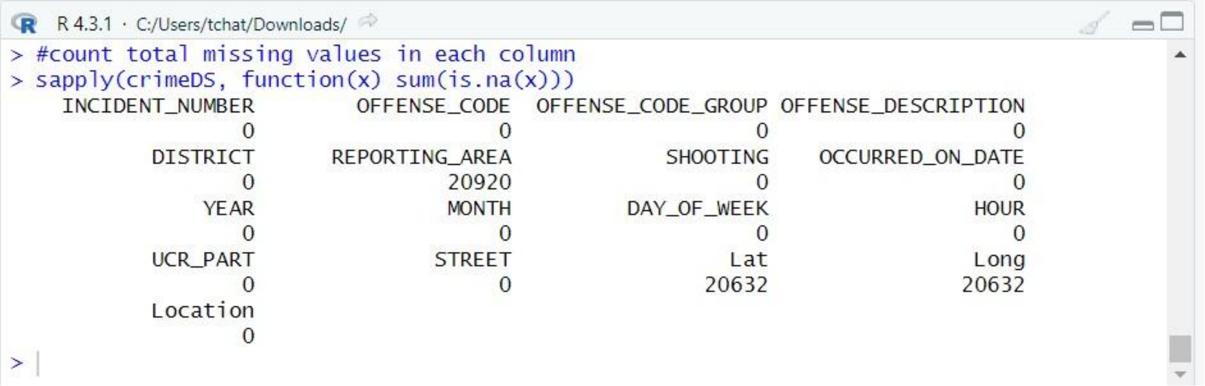


Data Cleaning

Data cleaning also called data mining is use in data exploration to improve the quality and fiability of the data. The main goal being to identify errors that could impact the computation of the statical parameters of standization such as mean, standard deviation, z-score and more.
Some reasons why its commun practice to clean data are:

- To avoid mistakes, it is commun to have missing values, mistake input or inaccurate data (outliers) therefore we use method like identifying NA, or empty space just to know the extend of this inaccurate

Image 5: NA



R 4.3.1 · C:/Users/tchat/Downloads/

```
> #count total missing values in each column
> sapply(crimeDS, function(x) sum(is.na(x)))
    INCIDENT_NUMBER        OFFENSE_CODE  OFFENSE_CODE_GROUP OFFENSE_DESCRIPTION 
                  0                   0                   0                   0 
           DISTRICT      REPORTING_AREA            SHOOTING    OCCURRED_ON_DATE 
                  0               20920                   0                   0 
               YEAR               MONTH         DAY_OF_WEEK                HOUR 
                  0                   0                   0                   0 
           UCR_PART              STREET                 Lat                Long 
                  0                   0               20632               20632 
           Location 
                  0 
> |
```

In the Image 5, we can see that our data sets contains 20920 Not Assign values in the Reporting_Area column, 20632 in Lat and 20632 in the Long.
To solve this inacuraty , i observe my data set. We have three possible ways to deal with the NA in the Reporting_Area which are:

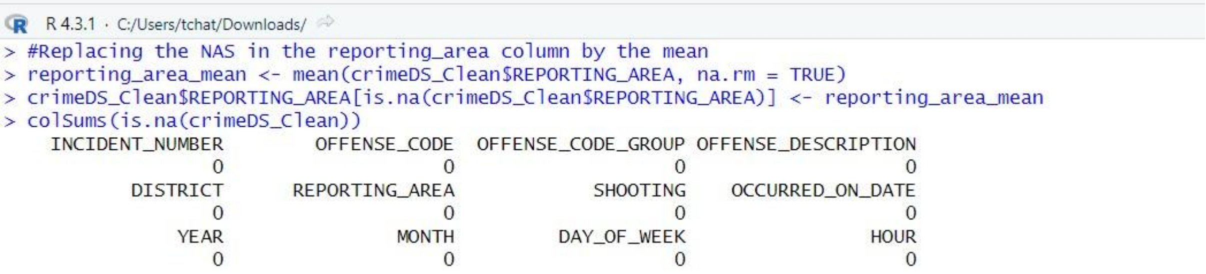
- Replace missing values with constant: this method consist of replacing the empty values by constant values (taking in consideration the data type of the variable)

- Replace missing values with Mode or Mean: it consist of replacing numerical missing values by the mean and categorical empty fields by the mode.

- Replace missing values with random values:this consist of replacing empty fields by random generated values.

In our case the technique use was to delete the LAT , Long, Ucr_part columns and to replace the missing valuesin the Reporting_Area column by the mean, like on the following image.

Image6: Handling missing values



```
> #Replacing the NAS in the reporting_area column by the mean
> reporting_area_mean <- mean(crimeDS_Clean$REPORTING_AREA, na.rm = TRUE)
> crimeDS_Clean$REPORTING_AREA[is.na(crimeDS_Clean$REPORTING_AREA)] <- reporting_area_mean
> colSums(is.na(crimeDS_Clean))
    INCIDENT_NUMBER        OFFENSE_CODE  OFFENSE_CODE_GROUP OFFENSE_DESCRIPTION 
                  0                   0                   0                   0 
           DISTRICT      REPORTING_AREA            SHOOTING    OCCURRED_ON_DATE 
                  0                   0                   0                   0 
               YEAR               MONTH         DAY_OF_WEEK                HOUR 
                  0                   0                   0                   0 
```

- I check the empty space in our dataset and we notice that SHOOTING column had empty spaces and after some analysis i notice that this column could be a binary column , “Y” to say if this incident was a shouting and “N” to say N. Therefore i replace all the empty space in this column with “N” to say that they were no shooting during this incident.
- I also implemented the outliers techniques, i check and remove all outliners from the data set .

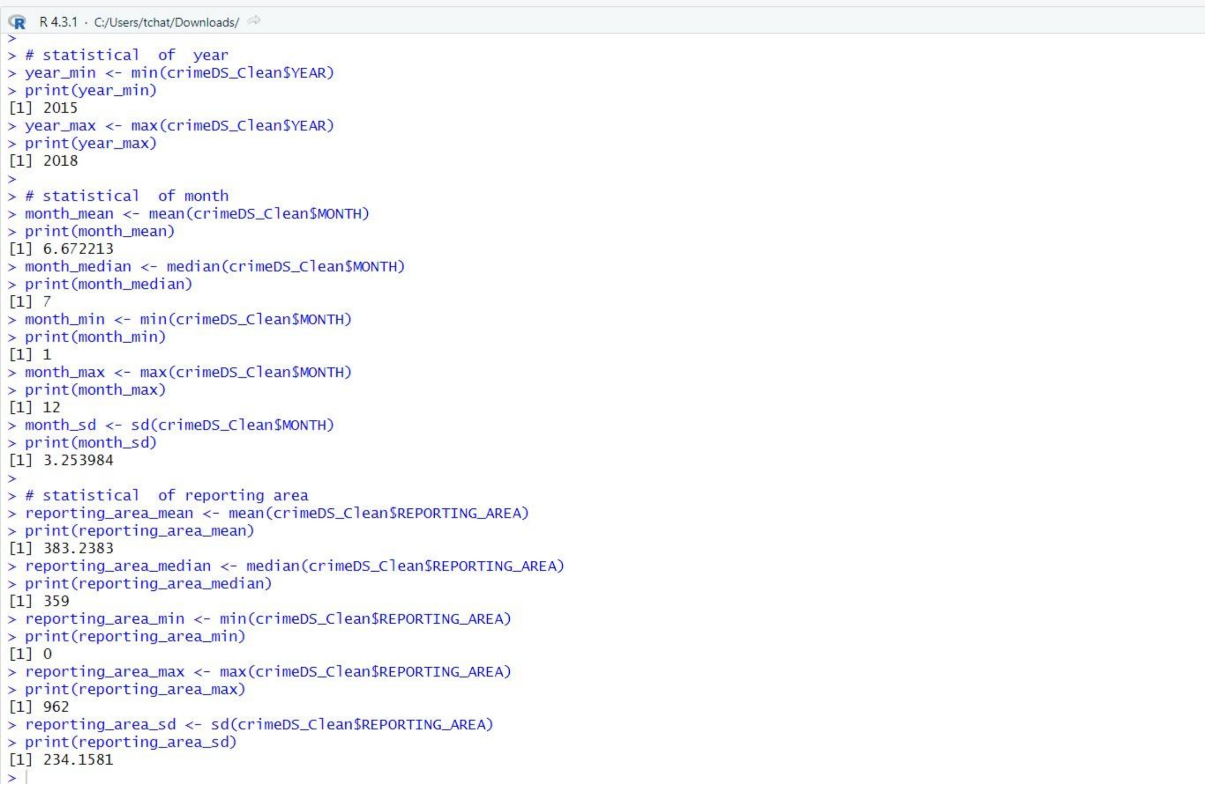
Exploratory Data Analysis

Statistical Exploration

The statistical exploration can be associated to the subsequent analyses of a data set, this is mainly applicable to numerical variables and consist of determining the mean, median, standard deviation, maximum and minimum of this variables. The statistical exploration is a summary of the variability of the data set, it helps understanding the distribution of the data before making any decisions therefore is a great decision support.

Before going deeply into ou data set exploratorey analysis, i conducted a statistical analys on the Year, Hour, Month and Reporting_Area where i respectfully determine the mean, median, standard deviation and mode (Year variable) of each.

Image 7: Statiscal Results



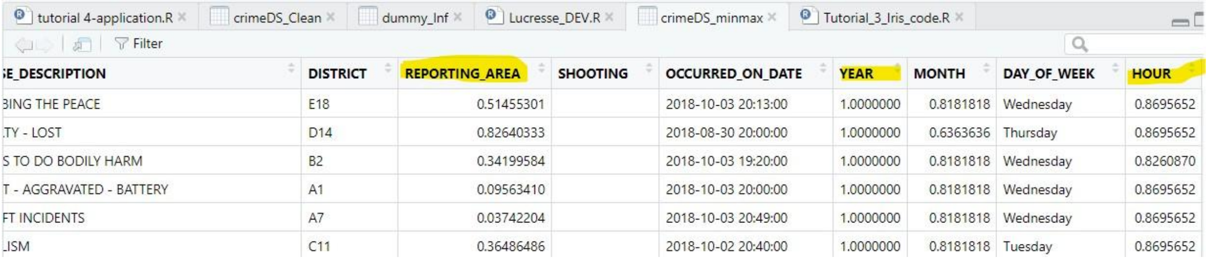
```
>
> # statistical  of  year
> year_min <- min(crimeDS_Clean$YEAR)
> print(year_min)
[1] 2015
> year_max <- max(crimeDS_Clean$YEAR)
> print(year_max)
[1] 2018
>
> # statistical  of month
> month_mean <- mean(crimeDS_Clean$MONTH)
> print(month_mean)
[1] 6.672213
> month_median <- median(crimeDS_Clean$MONTH)
> print(month_median)
[1] 7
> month_min <- min(crimeDS_Clean$MONTH)
> print(month_min)
[1] 1
> month_max <- max(crimeDS_Clean$MONTH)
> print(month_max)
[1] 12
> month_sd <- sd(crimeDS_Clean$MONTH)
> print(month_sd)
[1] 3.253984
>
> # statistical  of reporting area
> reporting_area_mean <- mean(crimeDS_Clean$REPORTING_AREA)
> print(reporting_area_mean)
[1] 383.2383
> reporting_area_median <- median(crimeDS_Clean$REPORTING_AREA)
> print(reporting_area_median)
[1] 359
> reporting_area_min <- min(crimeDS_Clean$REPORTING_AREA)
> print(reporting_area_min)
[1] 0
> reporting_area_max <- max(crimeDS_Clean$REPORTING_AREA)
> print(reporting_area_max)
[1] 962
> reporting_area_sd <- sd(crimeDS_Clean$REPORTING_AREA)
> print(reporting_area_sd)
[1] 234.1581
>
```

Data Normalization

Data normalization is a form of data exploration , it consist of scaling the numerical variable of a dataset . This is usually apply to data set to reduce the posible impact of outliers in order to do an approximate equal comparison between the features of this variable.
In this project , we applied three type of normalization

- Min - Max normalization: this takes each value subtrac it by the minimum of the values of the variable and divides its by the subtraction of the maximum value and the minimum value. After calculating the min- max this what our table looks like

Image 6: Min- Max normalization

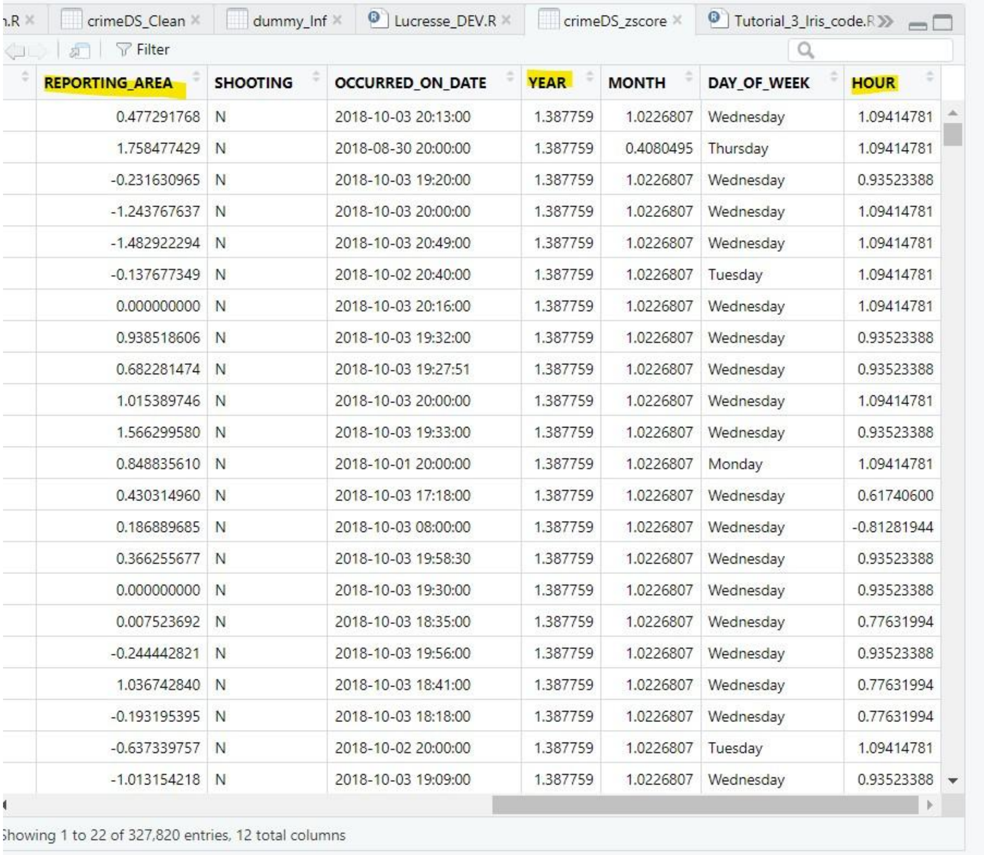


E_DESCRIPTION	DISTRICT	REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR
BING THE PEACE	E18	0.51455301		2018-10-03 20:13:00	1.0000000	0.8181818	Wednesday	0.8695652
TY - LOST	D14	0.82640333		2018-08-30 20:00:00	1.0000000	0.6363636	Thursday	0.8695652
S TO DO BODILY HARM	B2	0.34199584		2018-10-03 19:20:00	1.0000000	0.8181818	Wednesday	0.8260870
T - AGGRAVATED - BATTERY	A1	0.09563410		2018-10-03 20:00:00	1.0000000	0.8181818	Wednesday	0.8695652
FT INCIDENTS	A7	0.03742204		2018-10-03 20:49:00	1.0000000	0.8181818	Wednesday	0.8695652
ISM	C11	0.36486486		2018-10-02 20:40:00	1.0000000	0.8181818	Tuesday	0.8695652

- Robust scaling: compare to other normalization, robust scaling uses the median and interquartile range to scale.
- Z- score: we determine the z-score, the best way to interpret the zscore is dividing into categories. The z-score equal to 0 shows that the data is similar to the mean, 1 is equal to the standard deviation, if it is positive this indicates that above the mean

and if it is negative it is below the mean. This normalisation shows the relationship of the value to the mean .

Image 10 : Z-score



REPORTING_AREA	SHOOTING	OCCURRED_ON_DATE	YEAR	MONTH	DAY_OF_WEEK	HOUR
0.477291768	N	2018-10-03 20:13:00	1.387759	1.0226807	Wednesday	1.09414781
1.758477429	N	2018-08-30 20:00:00	1.387759	0.4080495	Thursday	1.09414781
-0.231630965	N	2018-10-03 19:20:00	1.387759	1.0226807	Wednesday	0.93523388
-1.243767637	N	2018-10-03 20:00:00	1.387759	1.0226807	Wednesday	1.09414781
-1.482922294	N	2018-10-03 20:49:00	1.387759	1.0226807	Wednesday	1.09414781
-0.137677349	N	2018-10-02 20:40:00	1.387759	1.0226807	Tuesday	1.09414781
0.000000000	N	2018-10-03 20:16:00	1.387759	1.0226807	Wednesday	1.09414781
0.938518606	N	2018-10-03 19:32:00	1.387759	1.0226807	Wednesday	0.93523388
0.682281474	N	2018-10-03 19:27:51	1.387759	1.0226807	Wednesday	0.93523388
1.015389746	N	2018-10-03 20:00:00	1.387759	1.0226807	Wednesday	1.09414781
1.566299580	N	2018-10-03 19:33:00	1.387759	1.0226807	Wednesday	0.93523388
0.848835610	N	2018-10-01 20:00:00	1.387759	1.0226807	Monday	1.09414781
0.430314960	N	2018-10-03 17:18:00	1.387759	1.0226807	Wednesday	0.61740600
0.186889685	N	2018-10-03 08:00:00	1.387759	1.0226807	Wednesday	-0.81281944
0.366255677	N	2018-10-03 19:58:30	1.387759	1.0226807	Wednesday	0.93523388
0.000000000	N	2018-10-03 19:30:00	1.387759	1.0226807	Wednesday	0.93523388
0.007523692	N	2018-10-03 18:35:00	1.387759	1.0226807	Wednesday	0.77631994
-0.244442821	N	2018-10-03 19:56:00	1.387759	1.0226807	Wednesday	0.93523388
1.036742840	N	2018-10-03 18:41:00	1.387759	1.0226807	Wednesday	0.77631994
-0.193195395	N	2018-10-03 18:18:00	1.387759	1.0226807	Wednesday	0.77631994
-0.637339757	N	2018-10-02 20:00:00	1.387759	1.0226807	Tuesday	1.09414781
-1.013154218	N	2018-10-03 19:09:00	1.387759	1.0226807	Wednesday	0.93523388

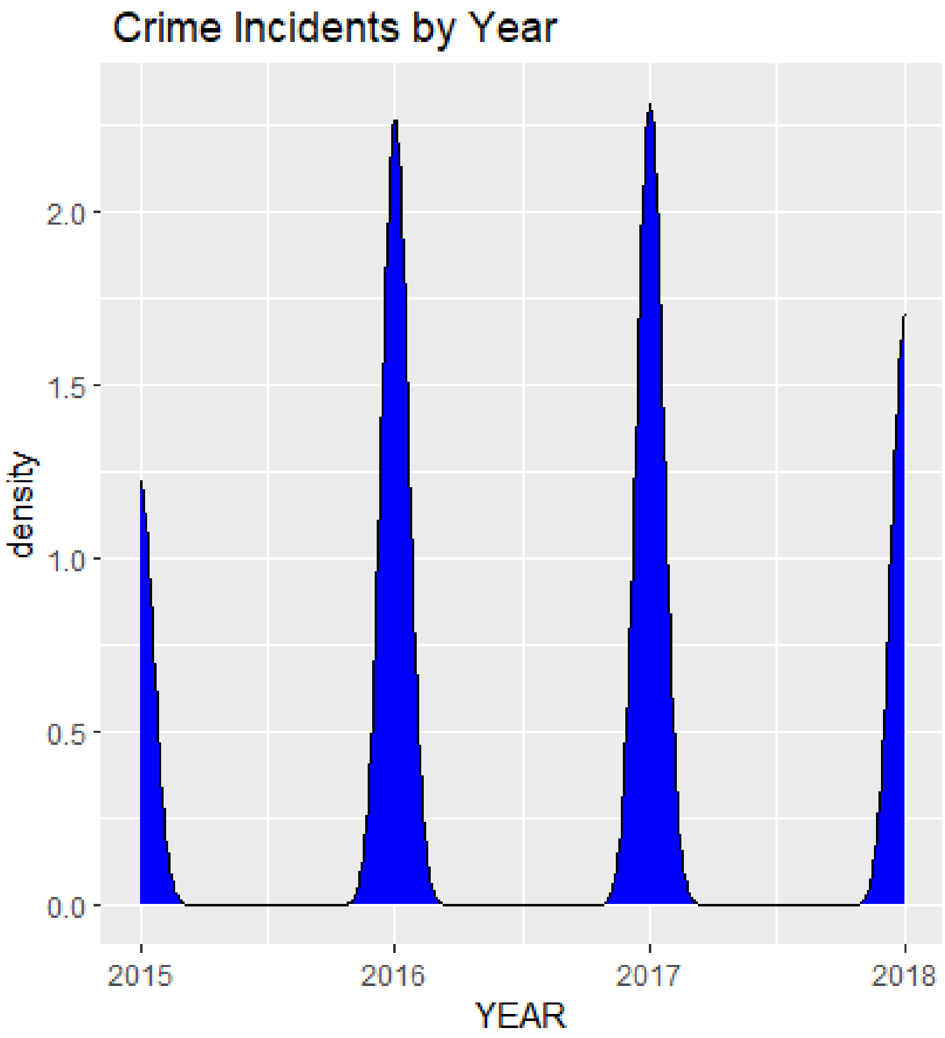
Showing 1 to 22 of 327,820 entries, 12 total columns

Data Features

Exploring the relationship between the different variable of the crime data set was a great way to understand in a visual and easy the possible informations that can be extracted from this data set. Throught the use of scatter plots, heatmaps, pie chart i got a different perspective of the features of the data set.

Case 1: Crime per Year

Question ask here was to know if they were spikes or drops in certain years? Are the variation in density drastic ?

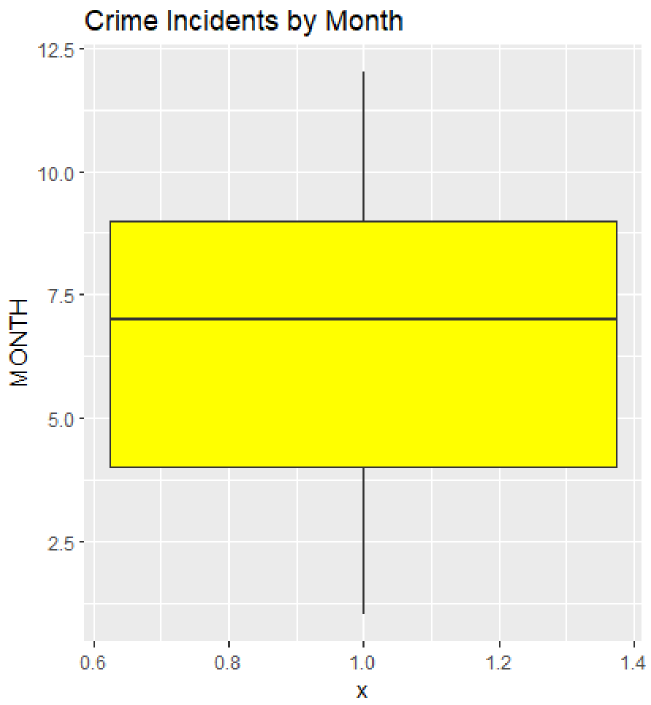


This density plot shows the distribution of crime over each year between 2015 and 1016. To obtain this we use the ggplot2 library. From the distribution of the density the information that is observe in this plot is that 2017 has had the highest number of crime amongst the four years present in our data set

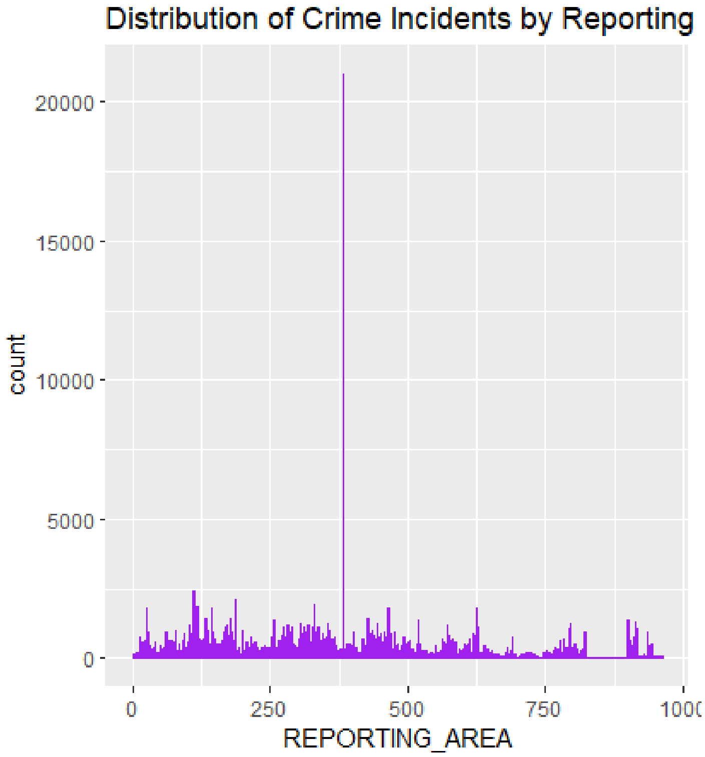
Case 2: Crime per Month

The question ask here was to know if they were any outliers? Where the crimes inconsistent?

The idea being very similar to the one for the years, this is a boxplot that shows the distribution of the crimes across. In this case we notice that crimes are the medium is around the seven month of the year.



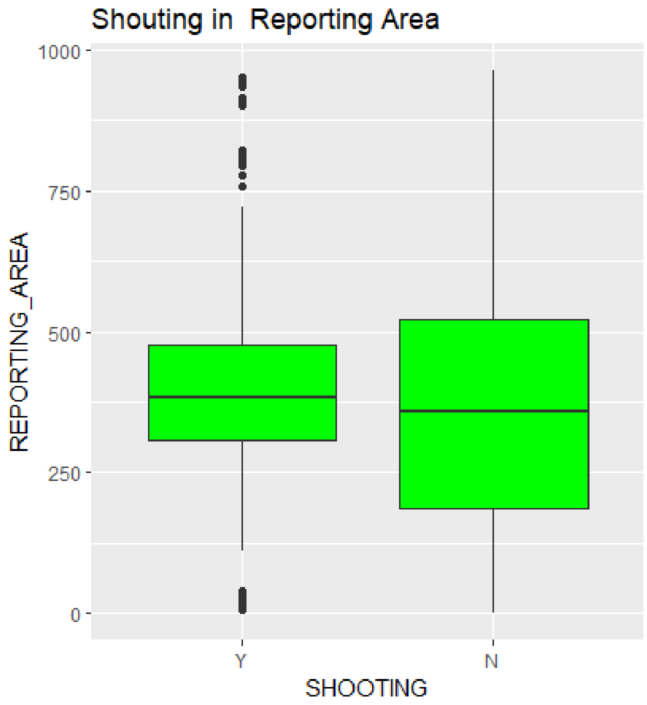
Case 3: Amount of crime reported per reporting area



This plot represent the total of crime reported in each Area , and the pick can be consider as a dangerous area cause is the area with the most report.

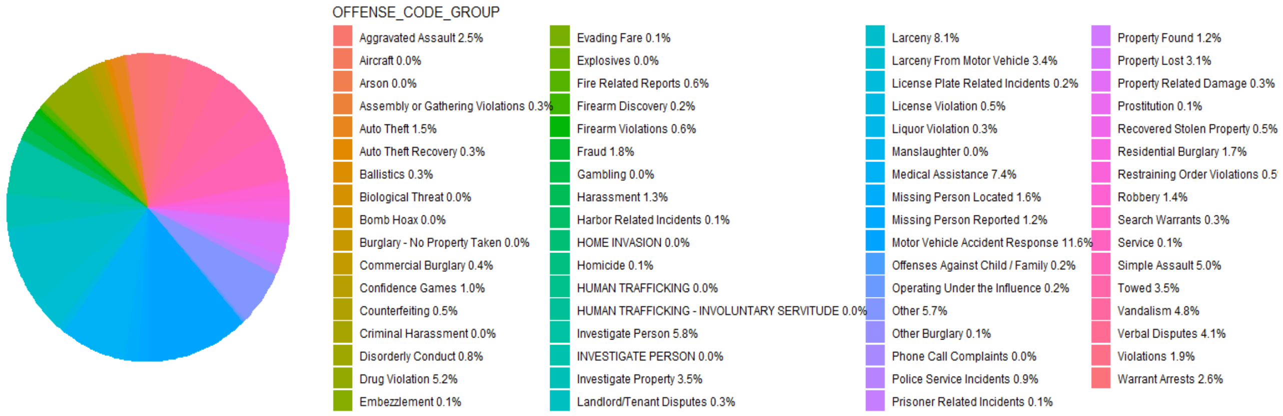
Case 4: Shooting in Reporting Area

The main question here was , are they any outliers? What do they represent ?



This plot is a representation of the shooting as an incident in the reported area, from this plot we can see that they are the proportion of shooting crime is smaller that the rest of the crime. In our next case we had a look in the distribution of the offense in order to know which are the crime usually reported.

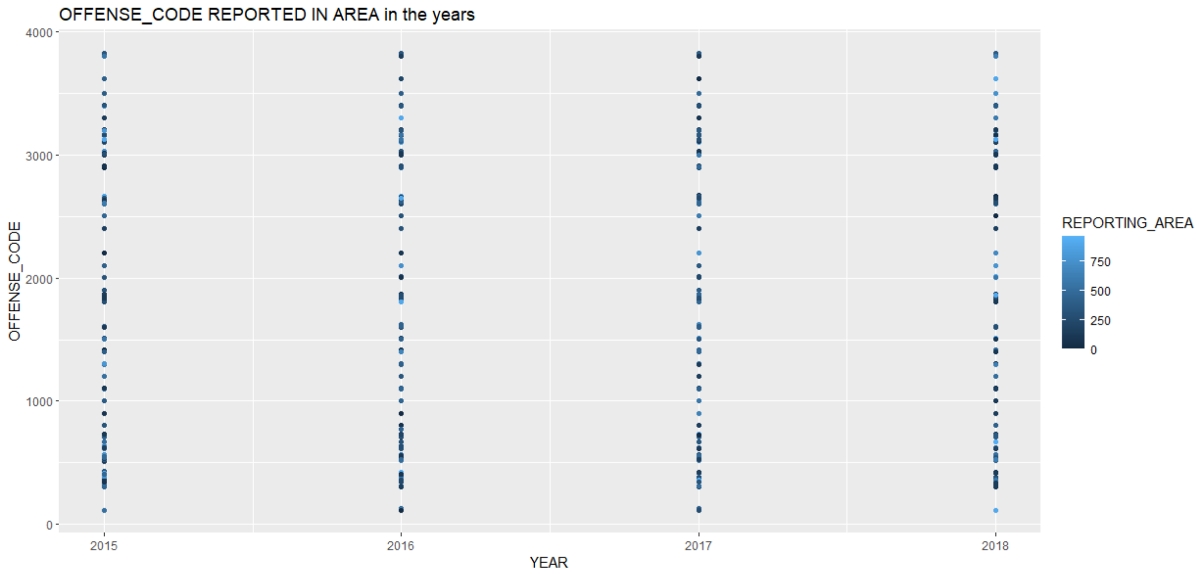
Case 5: Crime repartition in percentage



in this context i ask, are they any offense that should get more attention ?

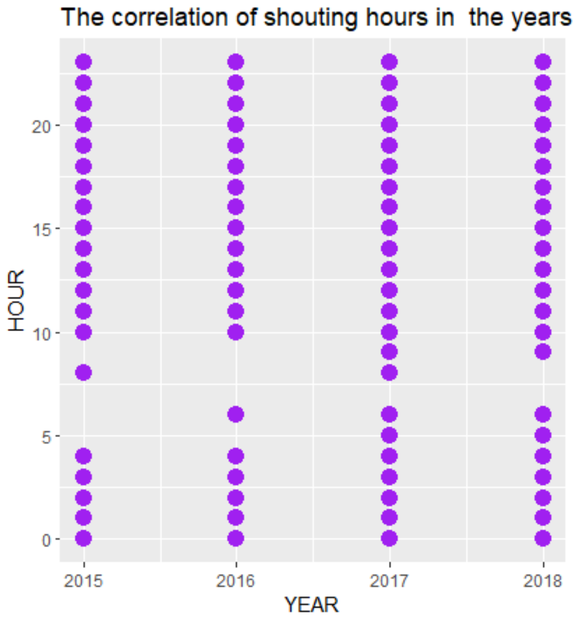
Base on the pie chart above, representing the different offense group, we can say “Motor Vehicle Accident Response” is the highest offense that appears in out data set with a percentage of 11.6 .

Case 6: Type of offense reported in the years



The question ask here is , are they any offense regular in a specific Area in each year ? This scatter plot show the distribution of the offense in the each year, given us a great insight on what offense where committed each year and the colors representing the Area in which this offense where committed. This plot is then a correlation between the year, the offense type the report area.

Case 7: Shooting hours in the each year

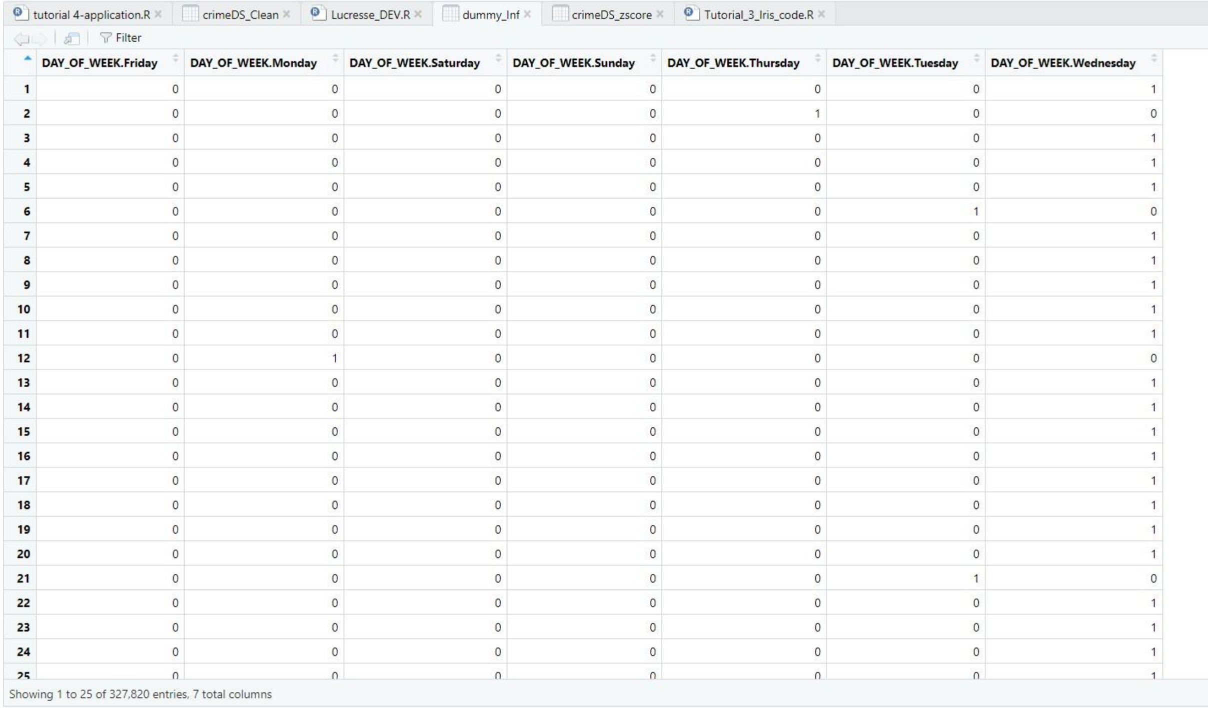


Exploring this features, we discover variations between the numerical and categorical variables of our data set.

Dummy Encoding -(One-hot)

Dummy encoding is offent use for categorical data, this method is use to represent categorical variables as numerical into order to be easily process for machine learning computation they are often transform into binary one being the positif and zero the negatif response. In our data set, I decided to encode the “Day_of_week” column, this is made of the seven of the week.
This is what our encoded data look like

Image 11: One- hot encoding



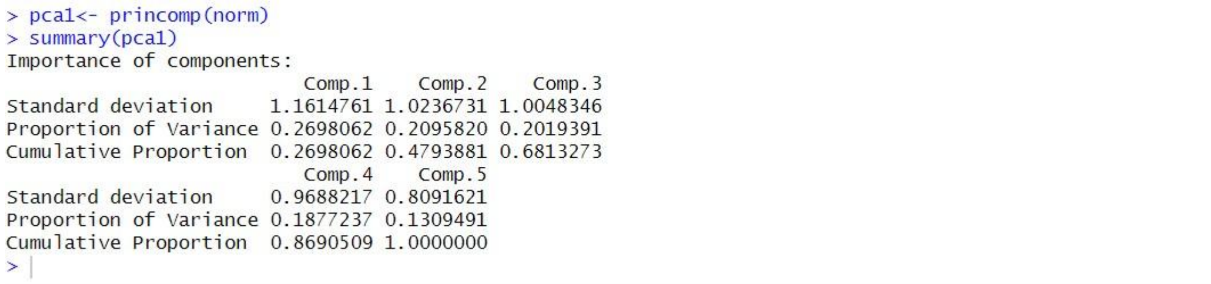
	DAY_OF_WEEK.Friday	DAY_OF_WEEK.Monday	DAY_OF_WEEK.Saturday	DAY_OF_WEEK.Sunday	DAY_OF_WEEK.Thursday	DAY_OF_WEEK.Tuesday	DAY_OF_WEEK.Wednesday
1	0	0	0	0	0	0	1
2	0	0	0	0	1	0	0
3	0	0	0	0	0	0	1
4	0	0	0	0	0	0	1
5	0	0	0	0	0	0	1
6	0	0	0	0	0	1	0
7	0	0	0	0	0	0	1
8	0	0	0	0	0	0	1
9	0	0	0	0	0	0	1
10	0	0	0	0	0	0	1
11	0	0	0	0	0	0	1
12	0	1	0	0	0	0	0
13	0	0	0	0	0	0	1
14	0	0	0	0	0	0	1
15	0	0	0	0	0	0	1
16	0	0	0	0	0	0	1
17	0	0	0	0	0	0	1
18	0	0	0	0	0	0	1
19	0	0	0	0	0	0	1
20	0	0	0	0	0	0	1
21	0	0	0	0	0	1	0
22	0	0	0	0	0	0	1
23	0	0	0	0	0	0	1
24	0	0	0	0	0	0	1
25	0	0	0	0	0	0	1

Showing 1 to 25 of 327,820 entries, 7 total columns

Principal Component Analysis(PCA)

PCA is use to reduce the dimension of a data set and turn it is features into uncorrelated variables.
In order to apply the PCA, I standardize the 5 numerical component of my data set to make sure they are equally scale. The outcome of this

Image 12: pca component norm



```
> pca1<- princomp(norm)
> summary(pca1)
Importance of components:
                          Comp.1    Comp.2    Comp.3
Standard deviation     1.1614761 1.0236731 1.0048346
Proportion of Variance 0.2698062 0.2095820 0.2019391
Cumulative Proportion  0.2698062 0.4793881 0.6813273
                          Comp.4    Comp.5
Standard deviation     0.9688217 0.8091621
Proportion of Variance 0.1877237 0.1309491
Cumulative Proportion  0.8690509 1.0000000
> |
```

An interpretation of this PCA is that component 1 having the highest standard deviation means it has the most variationin his data, also the cumulative proportion shos the cumulative data variance in each component and as shown above component 1 , 2,3 have cumulative proportion under 70 .

Image 13 : PCA

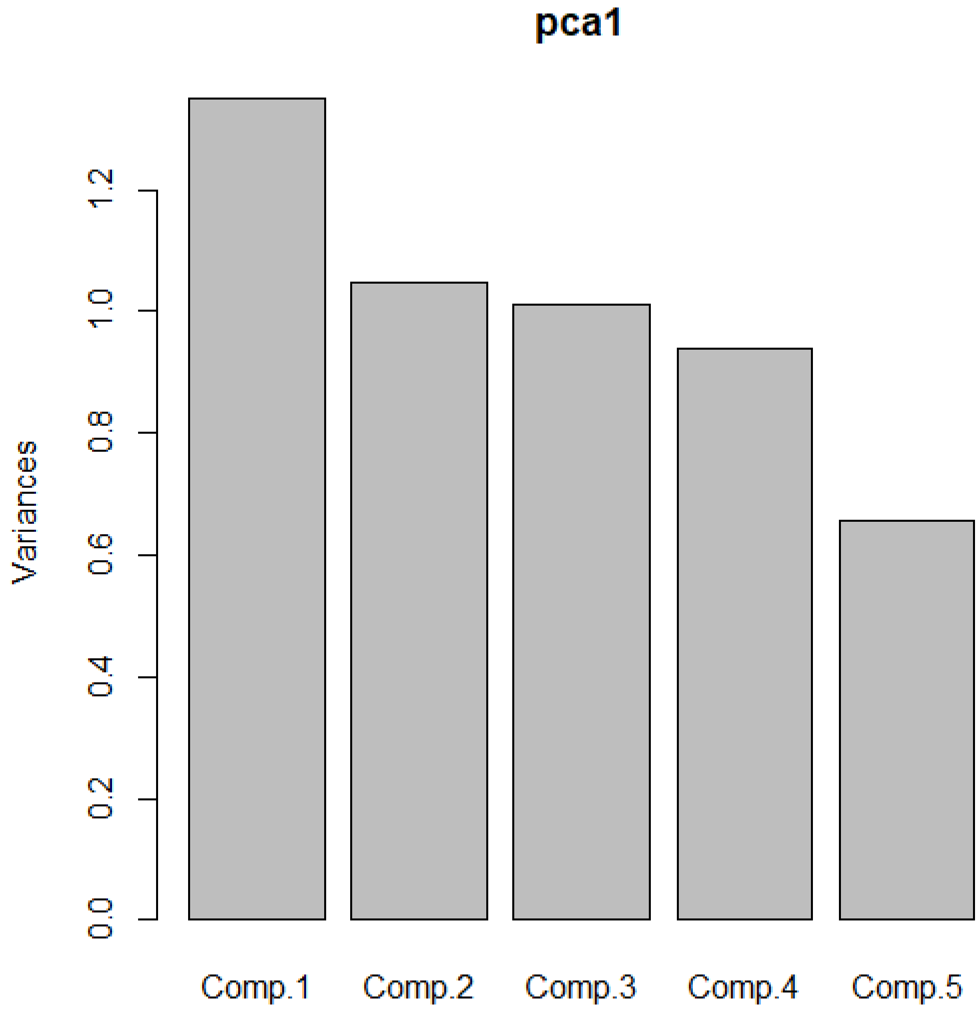
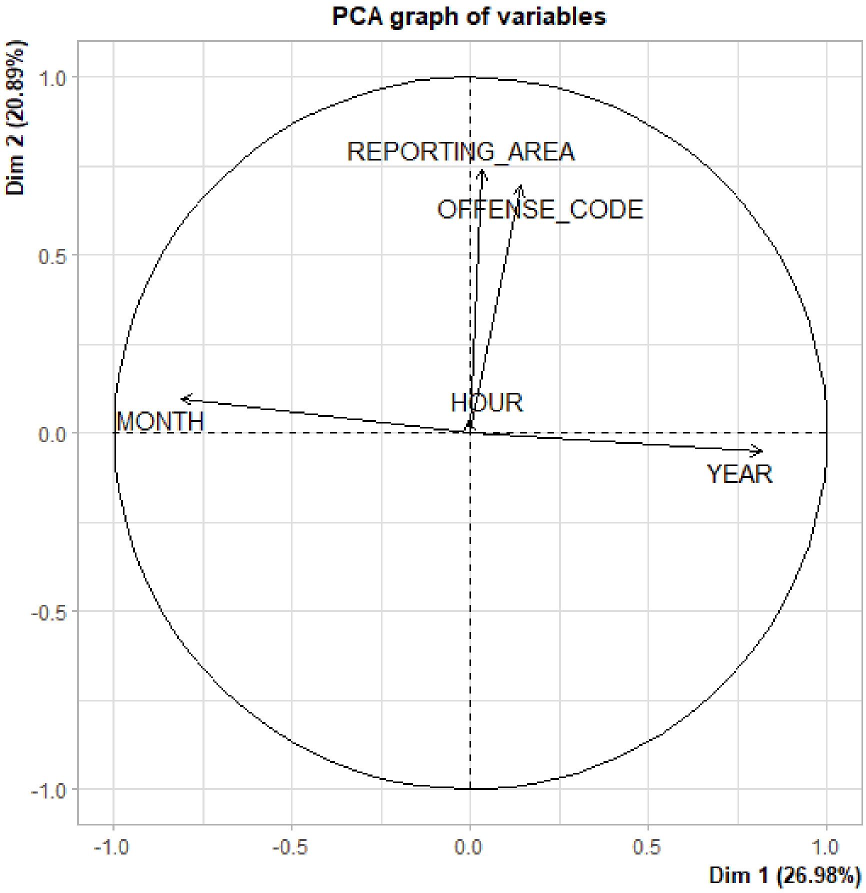


Image 14:



Dimensionality reduction

Dimensional reduction is often use in data visualization and is to reduce the number of variables in a data set, this is done to reduce redundant data that could after the interpretation . This process consist on going from high dimensional feature space to lower feature space. In our case of study, i decided to work with all the data features cause they correlated with each other and gave good information.

Challenges

While completing the various data cleaning and interpretation on this data set i encountered difficulties mainly in the programming part of the project, the data set choosen had a lot of missing values so it was difficult to make sure to clean the data without influencing the possible outcome. I also face difficulties in my time management, the workload required for this project help me improve my time management skills and keep focus on my objectifs.

Conclusion

In conclusion, the goal of this project was the most informations out of the crime dataset by the means of the different data exploration techniques.After identifying the variables types, cleaning the data and identifying possible outliers we they increase the accuracy of our set. We also made use of the EDA, a visualization method that helps us plot the different relationship between the features of our data set. We encountered some challenge but we mainly learn through research and implementation how data can be prepared before being use for machine learning, this project also enhance our understanding of our data set so that possible patterns could be use in the future to reduce the numbers of crimes.

Reference

Dataset:

- https://www.kaggle.com/datasets/ankkur13/boston-crime-data

Websites:

- prthmsh7. (2022). *Encoding Categorical Data in R*. [Online]. GEEKSFORGEEKS. Last Updated: 1 aug 2023. Available at: https://www.geeksforgeeks.org/encoding-categorical-data-in-r/ [Accessed .]
- Finnstats. (2021). *Data Normalization in R*. [Online]. R-bloggers. Last Updated: 17 oct 2021. Available at: https://www.r-bloggers.com/2021/10/data-normalization-in-r/
- Scott Nevil. (2023). *How to Calculate Z-Score and Its Meaning*. [Online]. Investopedia. Last Updated: 31 mars 2023. Available at: https://www.investopedia.com/terms/z/zscore.asp
- Oreilly. (NA). *Min–max normalization*. [Online]. oreilly. Last Updated: NA. Available at: https://www.oreilly.com/library/view/hands-on-machine-learning/9781788393485/fd5b8a44-e9d3-4c19-bebb
- The Startup Nikhita Singh. (2020). *Data Normalization With R*. [Online]. medium.com. Last Updated: Jul 5, 2020. Available at: https://medium.com/swlh/data-normalisation-with-r-6ef1d1947970
- Deepika Singh. (2019). *encodint data with r*. [Online]. pluralsight. Last Updated: 12 Nov 2019. Available at: https://www.pluralsight.com/guides/encoding-data-with-r
- Zoumana Keita. (2023). *Principal Component Analysis in R Tutorial*. [Online]. Data Camp. Last Updated: feb 2023. Available at: https://www.datacamp.com/tutorial/pca-analysis-r
- Nilesh barla. (2023). *Dimensionality Reduction for Machine Learning*. [Online]. neptune. Last Updated: auguest 2023. Available at: https://neptune.ai/blog/dimensionality-reduction [Accessed 30 November 2023].

Ebooks:

- Lingyun Zhan.. (.). *R tips: 16 HOWTO’s with examples for data analysts*. [bookdown]. .. Available at: https://bookdown.org/lyzhang10/lzhang_r_tips_book/preface.html [Accessed 22 November 2023].

GitHub:

- LucressePearle/DEP_CA1: Data Exploration and Preparation (github.com)

Powerpoint:

- In class tutorial on dummy encoding