Module Title:	Data Exploration & Preparation
Assessment Title:	CA1 Project
Lecturer Name:	Dr. Muhammad Iqbal
Student Full Name:	Eliabe Baliero De Moura
Student Number:	2022474
Assessment Due Date:	03/December /2023
Date of Submission:	03/December /2023

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.

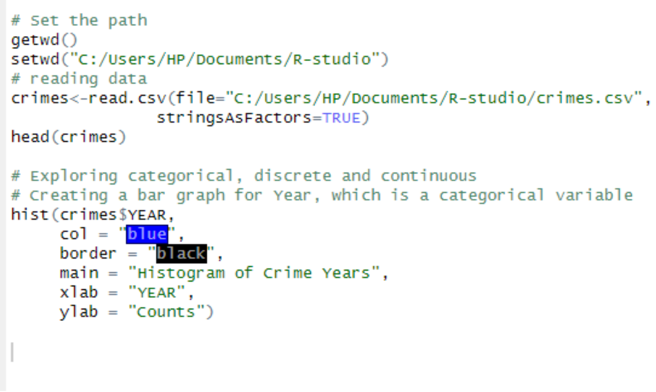
Introduction:

First CA for Data Exploration and Preparation module. In this short introduction I would like to point and summarise some tasks involved for this CA1. As students we are required to provide a dataset, report and our code in the R studio. For this assement, I am using a dataset which contains 18 columns and 9018 rows. By using this dataset csv file, I will be able to explore it on R studio and plot some graphs for results such as PCA, Statical parameters, and so on and so forth. I planned my question from A to H, put some graphs and pictures of code and results and a brief explanation.

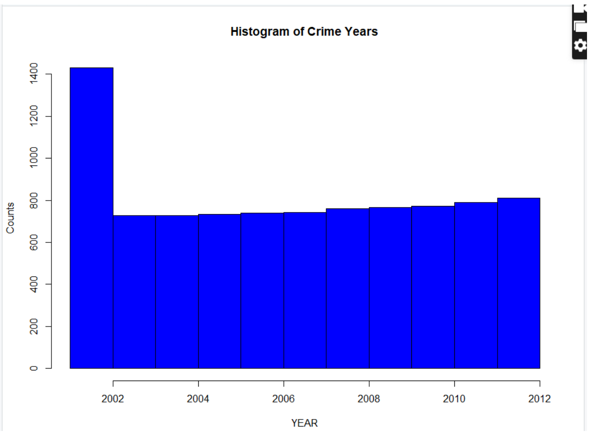
Question A

This is how I start my code in question "a". In this question, we are required to Identify which variables are categorical, discrete and continuous, then we need to explore those variables through using some graphs or visualization. In my case I am using a dataset which is a criminal one. I chose 3 variables, one is categorical, other is discrete and one is continuous. The 3 variable columns are: "YEAR", TIME", and "RAPE". Once it is done, I will check whether there are missing values for any of the variables. So below in the picture I will start exploring my categorical variable "YEAR" and use a simple bar graph to plot it. This basically has a blue bar colour, and black border.

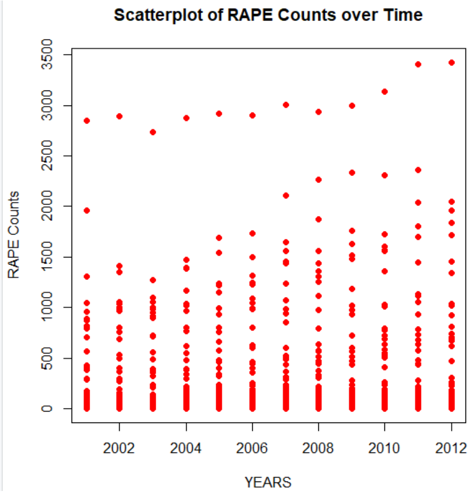
Categorical variable "YEAR".



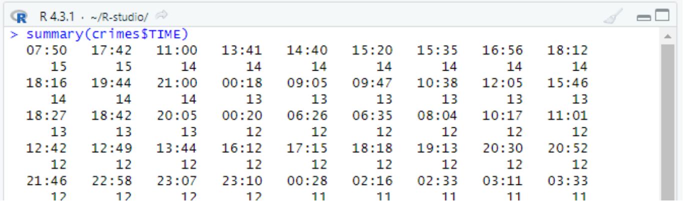
```
# Set the path
getwd()
setwd("C:/Users/HP/Documents/R-studio")
# reading data
crimes<-read.csv(file="C:/Users/HP/Documents/R-studio/crimes.csv",
               stringsAsFactors=TRUE)
head(crimes)

# Exploring categorical, discrete and continuous
# Creating a bar graph for Year, which is a categorical variable
hist(crimes$YEAR,
     col = "blue",
     border = "black",
     main = "Histogram of Crime Years",
     xlab = "YEAR",
     ylab = "Counts")
```



The graph below, is representing my discrete variable e” RAPE”. I set up a graph where shows the rape in years, using a Scatterplot and dots in red colour as show.

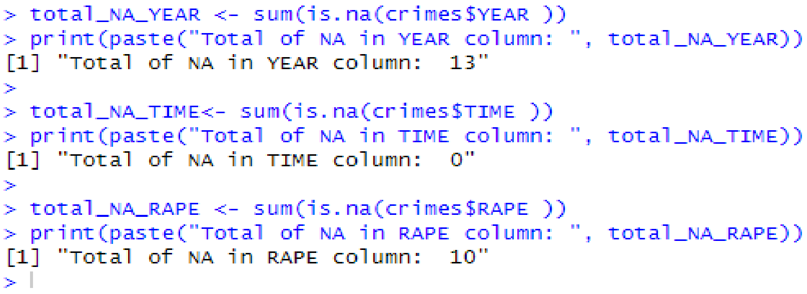


The next picture is about my “TIME” variable, this variable I take as my continuous variable. In the example below, I am using “summary” function to show the content inside of my column.



```
> summary(crimes$TIME)
  07:50   17:42   11:00   13:41   14:40   15:20   15:35   16:56   18:12 
     15      15      14      14      14      14      14      14      14 
  18:16   19:44   21:00   00:18   09:05   09:47   10:38   12:05   15:46 
     14      14      14      13      13      13      13      13      13 
  18:27   18:42   20:05   00:20   06:26   06:35   08:04   10:17   11:01 
     13      13      13      12      12      12      12      12      12 
  12:42   12:49   13:44   16:12   17:15   18:18   19:13   20:30   20:52 
     12      12      12      12      12      12      12      12      12 
  21:46   22:58   23:07   23:10   00:28   02:16   02:33   03:11   03:33 
     12      12      12      12      11      11      11      11      11 
```

The second part below is showing the code and the output that I got when I check if there are any missing “NA” or missing values in my variable. The output shows that in my “YEAR” variable I have 13 missing, in the “TIME” we do not have missing and in “RAPE we have 10

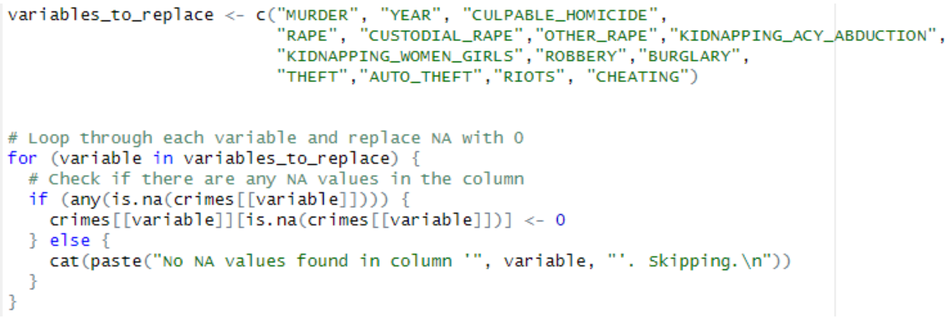


```
> total_NA_YEAR <- sum(is.na(crimes$YEAR ))
> print(paste("Total of NA in YEAR column: ", total_NA_YEAR))
[1] "Total of NA in YEAR column:  13"
> 
> total_NA_TIME<- sum(is.na(crimes$TIME ))
> print(paste("Total of NA in TIME column: ", total_NA_TIME))
[1] "Total of NA in TIME column:  0"
> 
> total_NA_RAPE <- sum(is.na(crimes$RAPE ))
> print(paste("Total of NA in RAPE column: ", total_NA_RAPE))
[1] "Total of NA in RAPE column:  10"
> 
```

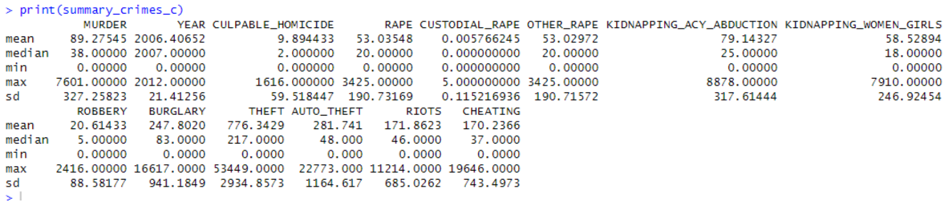
Question B

Calculate the statistical parameters (mean, median, minimum, maximum, and standard deviation) for each of the numerical variables.

The picture below shows how I planned my data. First I select my variables from MURDER to CHEATING and then I create a loop to fill up NA place with 0.



```
variables_to_replace <- c("MURDER", "YEAR", "CULPABLE_HOMICIDE",
                          "RAPE",  "CUSTODIAL_RAPE","OTHER_RAPE","KIDNAPPING_ACY_ABDUCTION",
                          "KIDNAPPING_WOMEN_GIRLS","ROBBERY","BURGLARY",
                          "THEFT","AUTO_THEFT","RIOTS",  "CHEATING")


# Loop through each variable and replace NA with 0
for (variable in variables_to_replace) {
  # Check if there are any NA values in the column
  if (any(is.na(crimes[[variable]]))) {
    crimes[[variable]][is.na(crimes[[variable]])] <- 0
  } else {
    cat(paste("No NA values found in column '", variable, "'. Skipping.\n"))
  }
}
```



```
> print(summary_crimes_c)
          MURDER        YEAR CULPABLE_HOMICIDE       RAPE CUSTODIAL_RAPE OTHER_RAPE KIDNAPPING_ACY_ABDUCTION KIDNAPPING_WOMEN_GIRLS
mean    89.27545 2006.40652          9.894433   53.03548    0.005766245   53.02972                 79.14327               58.52894
median  38.00000 2007.00000          2.000000   20.00000    0.000000000   20.00000                 25.00000               18.00000
min      0.00000    0.00000          0.000000    0.00000    0.000000000    0.00000                  0.00000                0.00000
max   7601.00000 2012.00000       1616.000000 3425.00000    5.000000000 3425.00000               8878.00000             7910.00000
sd     327.25823   21.41256         59.518447  190.73169    0.115216936  190.71572                317.61444              246.92454
         ROBBERY   BURGLARY      THEFT AUTO_THEFT      RIOTS    CHEATING
mean    20.61433   247.8020   776.3429    281.741   171.8623   170.2366
median   5.00000    83.0000   217.0000     48.000    46.0000    37.0000
min      0.00000     0.0000     0.0000      0.000     0.0000     0.0000
max   2416.00000 16617.0000 53449.0000  22773.000 11214.0000 19646.0000
sd      88.58177   941.1849  2934.8573   1164.617   685.0262   743.4973
> |
```

As it shows above the results for mean, median, minimum, maximum and standard deviation, I will go through it and explain my results for the CHEATING variable.

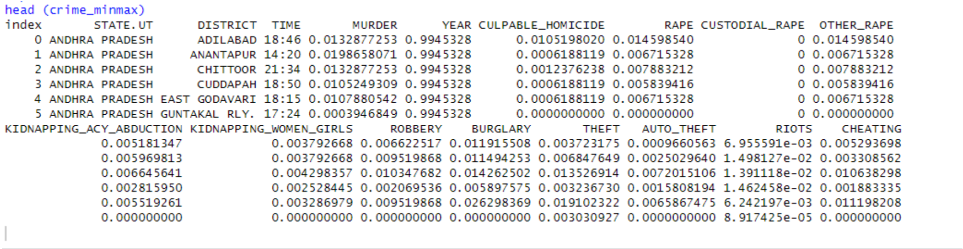
- **Mean:**
  The mean of 170.2366 indicates the average of the "CHEATING" values; this is the sum of all values divided by the number of values.
- **Median:**
  My median for variable CHEATING is 37.0000,"variable fall below this value, and half fall above it.
- **Minimum:**
  The minimum is 0.0000 is the smallest observed value in the "CHEATING" dataset.
- **Maximum:**
  19646.0000 is my largest value in the "CHEATING" dataset. This is the potential existence of outliers or extreme values.
- **Standard Deviation:**
  My standard is 743.4973. This is the amount of variability or dispersion in the "CHEATING" variable. If in case I would have higher standard deviation, it would be a greater variability among the values in my variable.

Question C

Min-Max Normalization, Z-score Standardization and Robust scalar:

Min-Max Normalization results:

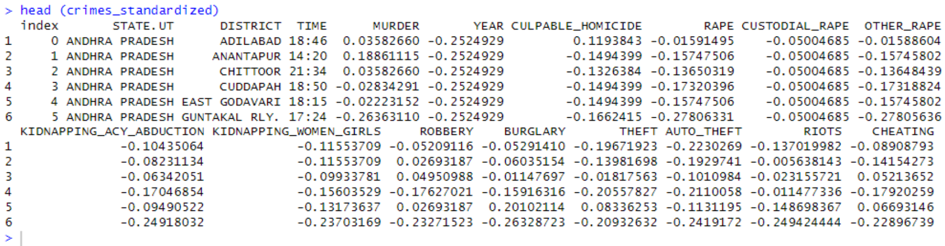
As shown in question B, I select my variable which are numerical ones to apply Min-Max Normalization, this referred to as feature scaling, this method employed in data manipulation has a role or purpose to rescale and normalize the numerical values of a feature in my data crimes.



```
head (crime_minmax)
index        STATE.UT       DISTRICT  TIME       MURDER      YEAR CULPABLE_HOMICIDE        RAPE CUSTODIAL_RAPE  OTHER_RAPE
    0 ANDHRA PRADESH       ADILABAD 18:46 0.0132877253 0.9945328      0.0105198020 0.014598540              0 0.014598540
    1 ANDHRA PRADESH      ANANTAPUR 14:20 0.0198658071 0.9945328      0.0006188119 0.006715328              0 0.006715328
    2 ANDHRA PRADESH       CHITTOOR 21:34 0.0132877253 0.9945328      0.0012376238 0.007883212              0 0.007883212
    3 ANDHRA PRADESH       CUDDAPAH 18:50 0.0105249309 0.9945328      0.0006188119 0.005839416              0 0.005839416
    4 ANDHRA PRADESH EAST GODAVARI 18:15 0.0107880542 0.9945328      0.0006188119 0.006715328              0 0.006715328
    5 ANDHRA PRADESH GUNTAKAL RLY. 17:24 0.0003946849 0.9945328      0.0000000000 0.000000000              0 0.000000000
KIDNAPPING_ACY_ABDUCTION KIDNAPPING_WOMEN_GIRLS     ROBBERY    BURGLARY       THEFT   AUTO_THEFT        RIOTS    CHEATING
             0.005181347            0.003792668 0.006622517 0.011915508 0.003723175 0.0009660563 6.955591e-03 0.005293698
             0.005969813            0.003792668 0.009519868 0.011494253 0.006847649 0.0025029640 1.498127e-02 0.003308562
             0.006645641            0.004298357 0.010347682 0.014262502 0.013526914 0.0072015106 1.391118e-02 0.010638298
             0.002815950            0.002528445 0.002069536 0.005897575 0.003236730 0.0015808194 1.462458e-02 0.001883335
             0.005519261            0.003286979 0.009519868 0.026298369 0.019102322 0.0065867475 6.242197e-03 0.011198208
             0.000000000            0.000000000 0.000000000 0.000000000 0.003030927 0.0000000000 8.917425e-05 0.000000000
|
```

My "MURDER" column for example in the crimes data table, was applied to Z-score standardization. This involves converting the initial murder counts into standardized scores. Positive z-scores signify counts surpassing the mean, while negative values indicate counts falling below the mean, and basically Z-score also take and convert the values of a numerical variable to have a mean of 0 and a standard deviation of 1

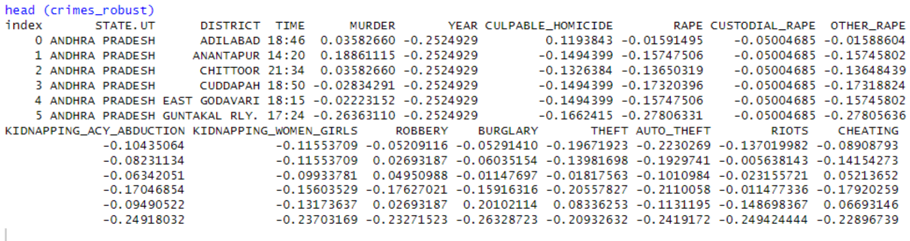
Z-score Standardization results:



```
> head (crimes_standardized)
  index        STATE.UT       DISTRICT  TIME      MURDER       YEAR CULPABLE_HOMICIDE        RAPE CUSTODIAL_RAPE  OTHER_RAPE
1     0 ANDHRA PRADESH      ADILABAD 18:46  0.03582660 -0.2524929         0.1193843 -0.01591495    -0.05004685 -0.01588604
2     1 ANDHRA PRADESH     ANANTAPUR 14:20  0.18861115 -0.2524929        -0.1494399 -0.15747506    -0.05004685 -0.15745802
3     2 ANDHRA PRADESH      CHITTOOR 21:34  0.03582660 -0.2524929        -0.1326384 -0.13650319    -0.05004685 -0.13648439
4     3 ANDHRA PRADESH      CUDDAPAH 18:50 -0.02834291 -0.2524929        -0.1494399 -0.17320396    -0.05004685 -0.17318824
5     4 ANDHRA PRADESH EAST GODAVARI 18:15 -0.02223152 -0.2524929        -0.1494399 -0.15747506    -0.05004685 -0.15745802
6     5 ANDHRA PRADESH GUNTAKAL RLY. 17:24 -0.26363110 -0.2524929        -0.1662415 -0.27806331    -0.05004685 -0.27805636
  KIDNAPPING_ACY_ABDUCTION KIDNAPPING_WOMEN_GIRLS     ROBBERY    BURGLARY       THEFT AUTO_THEFT        RIOTS    CHEATING
1              -0.10435064            -0.11553709 -0.05209116 -0.05291410 -0.19671923 -0.2230269 -0.137019982 -0.08908793
2              -0.08231134            -0.11553709  0.02693187 -0.06035154 -0.13981698 -0.1929741 -0.005638143 -0.14154273
3              -0.06342051            -0.09933781  0.04950988 -0.01147697 -0.01817563 -0.1010984 -0.023155721  0.05213652
4              -0.17046854            -0.15603529 -0.17627021 -0.15916316 -0.20557827 -0.2110058 -0.011477336 -0.17920259
5              -0.09490522            -0.13173637  0.02693187  0.20102114  0.08336253 -0.1131195 -0.148698367  0.06693146
6              -0.24918032            -0.23703169 -0.23271523 -0.26328723 -0.20932632 -0.2419172 -0.249424444 -0.22896739
> |
```

Robust results:

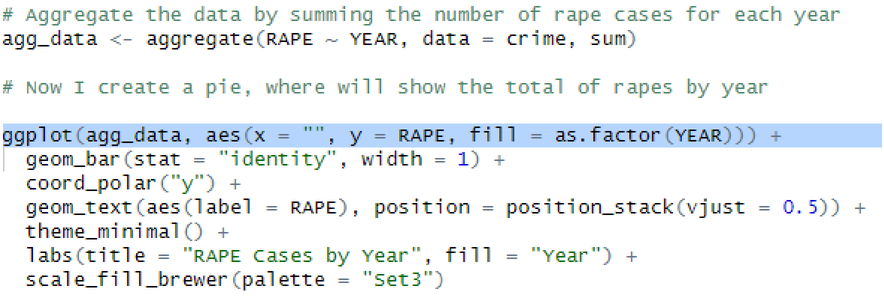
In the last one picture in this section, I am applying the “robust” function, this will give stability for my data analysis, where we have situations that we can contain outliers in our data



```
head (crimes_robust)
index        STATE.UT       DISTRICT  TIME      MURDER       YEAR CULPABLE_HOMICIDE        RAPE CUSTODIAL_RAPE  OTHER_RAPE
    0 ANDHRA PRADESH      ADILABAD 18:46  0.03582660 -0.2524929         0.1193843 -0.01591495    -0.05004685 -0.01588604
    1 ANDHRA PRADESH     ANANTAPUR 14:20  0.18861115 -0.2524929        -0.1494399 -0.15747506    -0.05004685 -0.15745802
    2 ANDHRA PRADESH      CHITTOOR 21:34  0.03582660 -0.2524929        -0.1326384 -0.13650319    -0.05004685 -0.13648439
    3 ANDHRA PRADESH      CUDDAPAH 18:50 -0.02834291 -0.2524929        -0.1494399 -0.17320396    -0.05004685 -0.17318824
    4 ANDHRA PRADESH EAST GODAVARI 18:15 -0.02223152 -0.2524929        -0.1494399 -0.15747506    -0.05004685 -0.15745802
    5 ANDHRA PRADESH GUNTAKAL RLY. 17:24 -0.26363110 -0.2524929        -0.1662415 -0.27806331    -0.05004685 -0.27805636
KIDNAPPING_ACY_ABDUCTION KIDNAPPING_WOMEN_GIRLS     ROBBERY    BURGLARY       THEFT AUTO_THEFT        RIOTS    CHEATING
             -0.10435064            -0.11553709 -0.05209116 -0.05291410 -0.19671923 -0.2230269 -0.137019982 -0.08908793
             -0.08231134            -0.11553709  0.02693187 -0.06035154 -0.13981698 -0.1929741 -0.005638143 -0.14154273
             -0.06342051            -0.09933781  0.04950988 -0.01147697 -0.01817563 -0.1010984 -0.023155721  0.05213652
             -0.17046854            -0.15603529 -0.17627021 -0.15916316 -0.20557827 -0.2110058 -0.011477336 -0.17920259
             -0.09490522            -0.13173637  0.02693187  0.20102114  0.08336253 -0.1131195 -0.148698367  0.06693146
             -0.24918032            -0.23703169 -0.23271523 -0.26328723 -0.20932632 -0.2419172 -0.249424444 -0.22896739
|
```

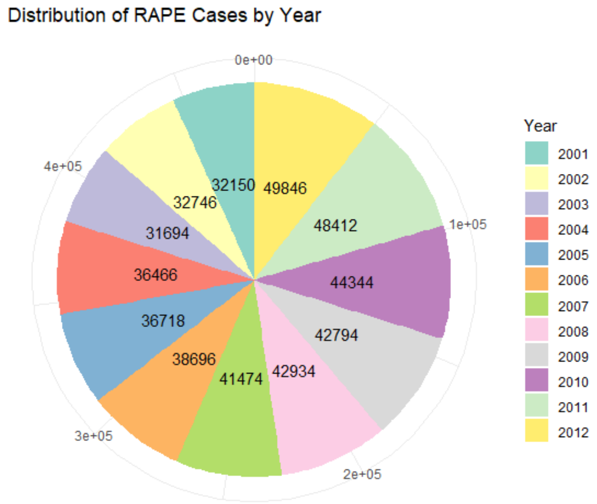
Question D / Question E

The code below is showing how I create my pie graph. I started by choosing the columns that I want to explore, which are YEAR and RAPE, after I use aggregate function to aggregate YEAR and RAPE variables and then we sum as I want the total for RAPE. As it is showing, I used a pie graph and chose the title as “Distributions of RAPE cases by YEAR”

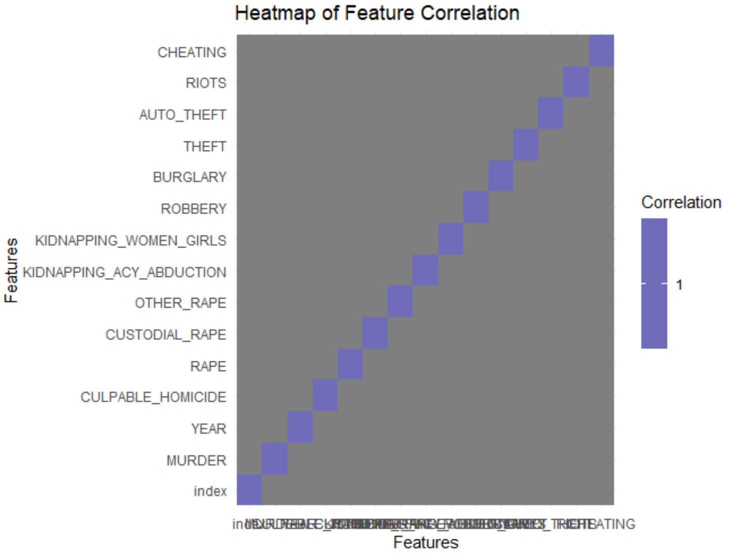


```
# Aggregate the data by summing the number of rape cases for each year
agg_data <- aggregate(RAPE ~ YEAR, data = crime, sum)

# Now I create a pie, where will show the total of rapes by year

ggplot(agg_data, aes(x = "", y = RAPE, fill = as.factor(YEAR))) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  geom_text(aes(label = RAPE), position = position_stack(vjust = 0.5)) +
  theme_minimal() +
  labs(title = "RAPE Cases by Year", fill = "Year") +
  scale_fill_brewer(palette = "Set3")
```

This is how the pie graph looks like to represent the total of rape by year, in the middle we can see the total of RAPE, and beside we see the YEAR represented by colours.

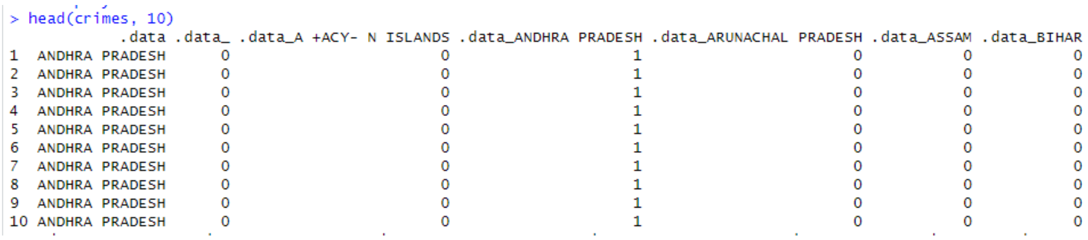


Here I am using heatmap to show how different features in a dataset relate to each other. The heatmap has squares, and each one of those squares tell us how much two features Var1 are and Var2 are connected. The colours in the squares go from light blue to dark blue, making it easy to see how features are related. The graph is named "Heatmap Correlation graph," and it has labels on the sides to show which features are being compared. The colours in the legend help us understand how strong the connections are.



Question F

Apply dummy encoding to categorical variables:



```
> head(crimes, 10)
          .data .data_ .data_A +ACY- N ISLANDS .data_ANDHRA PRADESH .data_ARUNACHAL PRADESH .data_ASSAM .data_BIHAR
1  ANDHRA PRADESH      0                    0                    1                       0           0           0
2  ANDHRA PRADESH      0                    0                    1                       0           0           0
3  ANDHRA PRADESH      0                    0                    1                       0           0           0
4  ANDHRA PRADESH      0                    0                    1                       0           0           0
5  ANDHRA PRADESH      0                    0                    1                       0           0           0
6  ANDHRA PRADESH      0                    0                    1                       0           0           0
7  ANDHRA PRADESH      0                    0                    1                       0           0           0
8  ANDHRA PRADESH      0                    0                    1                       0           0           0
9  ANDHRA PRADESH      0                    0                    1                       0           0           0
10 ANDHRA PRADESH      0                    0                    1                       0           0           0
```

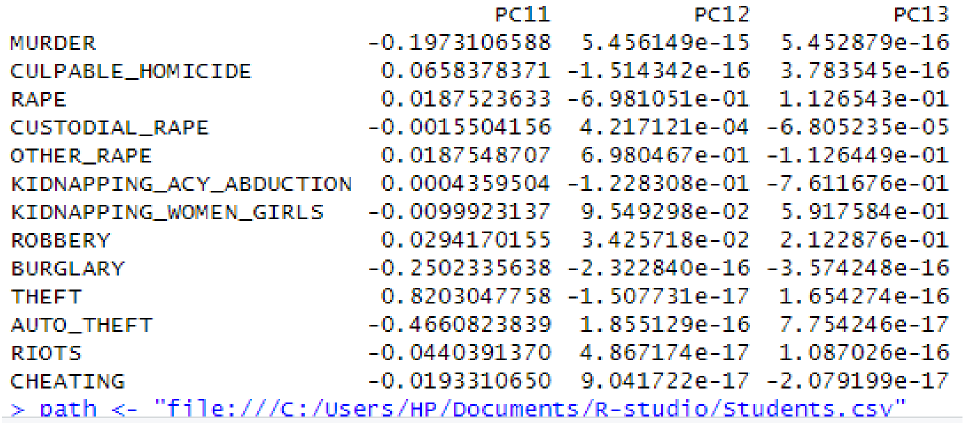
Results:

In the picture above is the result of my code for dummy applied. C**olumns**: The columns in the result represent each unique category in the original **STATE.UT** column. For example, there is a column named. Data **ANDHRA PRADESH** which indicates whether the original **STATE.IT** was "ANDHRA PRADESH." Each unique category gets its own dummy variable column.

Values: The values in my code, in these dummy variable columns are binary (0 or 1) and indicate the presence or absence of the corresponding category for each row. For example, in the first row, the value for the column **.data ANDHRA PRADESH** is 1, indicating that the original **STATE.UT** for that row was "ANDHRA PRADESH."

In summary, I would say that this creates a binary representation of the categorical variable **STATE.UT** by transforming it into a set of dummy variables, this istypes of statistical analyses and machine learning models that require numerical input.

Question G



```
                                   PC11          PC12          PC13
MURDER                    -0.1973106588  5.456149e-15  5.452879e-16
CULPABLE_HOMICIDE          0.0658378371 -1.514342e-16  3.783545e-16
RAPE                       0.0187523633 -6.981051e-01  1.126543e-01
CUSTODIAL_RAPE            -0.0015504156  4.217121e-04 -6.805235e-05
OTHER_RAPE                 0.0187548707  6.980467e-01 -1.126449e-01
KIDNAPPING_ACY_ABDUCTION   0.0004359504 -1.228308e-01 -7.611676e-01
KIDNAPPING_WOMEN_GIRLS    -0.0099923137  9.549298e-02  5.917584e-01
ROBBERY                    0.0294170155  3.425718e-02  2.122876e-01
BURGLARY                  -0.2502335638 -2.322840e-16 -3.574248e-16
THEFT                      0.8203047758 -1.507731e-17  1.654274e-16
AUTO_THEFT                -0.4660823839  1.855129e-16  7.754246e-17
RIOTS                     -0.0440391370  4.867174e-17  1.087026e-16
CHEATING                  -0.0193310650  9.041722e-17 -2.079199e-17
> path <- "file:///C:/Users/HP/Documents/R-studio/Students.csv"
```

The picture above is representing my 3 chosen components (PC11, PC12, PC13) derived from my crimes data set, often acquired through techniques that I used such as Principal Component Analysis (PCA) for dimensionality reduction. These weightings signify the impact of each crime category on its respective principal component. The second part be

Principal Component **11 (PC11):**

- PC11 exhibits negative weightings for crimes like Murder, Culpable Homicide, Burglary, Auto Theft, Riots, and Cheating.

- It demonstrates a substantial positive weighting for Theft, underscoring a robust correlation between this crime category and PC11.

Principal Component 12 (PC12):

- PC12 is characterized by a prominent negative weighting for Rape, coupled with positive weightings for Kidnapping & Abduction, Kidnapping Women & Girls, and Robbery.
- This suggests that PC12 encapsulates variations linked to sexual offenses (Rape) as well as crimes involving abduction and robbery.

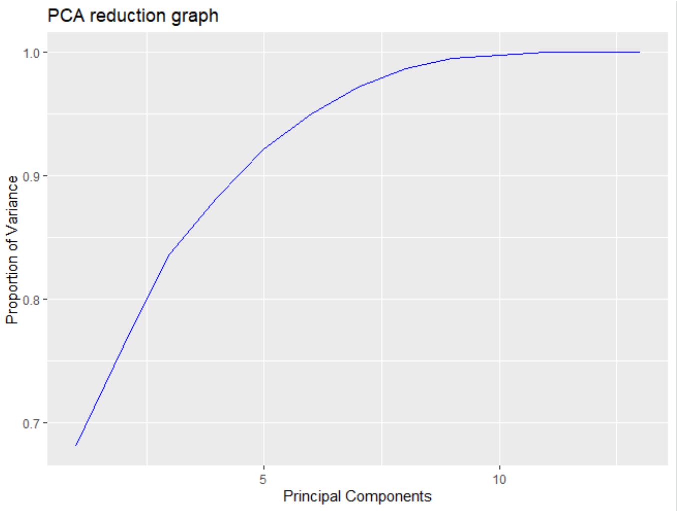
Principal Component 13 (PC13):

- PC13 exhibits positive weightings for crimes such as Rape, Kidnapping & Abduction, Kidnapping Women & Girls, Robbery, and Theft.
- This implies that PC13 is linked to crimes involving violence (Rape, Robbery) and abduction, particularly targeting women and girls.

Question H

In the example below, I took the summary of my three components PCA and I applied the reduction, then I plotted a graph where the x-axis represents the principal components in ascending order, while the y-axis indicates the cumulative proportion of variance. The declining line helps identify an optimal point.

Dimensionality reduction has an essential role to take challenges and improve the efficiency, interpretability, and performance of both machine learning and data analysis.



Noise:

In datasets with a high number of dimensions, we often have presence of noise or irrelevant features. Dimensionality reduction will eliminate noise and homing in on the most pertinent information. This process will result in a more refined and interpretable representation of the dataset.

Enhanced Computational Efficiency:

When we deal with datasets that boast numerous dimensions can be both computationally demanding and time intensive. By employing dimensionality reduction techniques to trim the number of features, subsequent analyses and model training procedures are expedited, contributing to increased computational efficiency.

Facilitating Visualization:

I would say the most important and satisfying one is Visualizing. Dimensionality reduction addresses the issue by projecting the data onto a lower-dimensional space, typically spanning two or three dimensions. This will transform and simplify the visualization process and make it more accessible for exploring.

Summary and Challenges:

At this point, I would say and include here in my challenges that Leading with Data Analysis (EDA) is satisfying, on the other hand it includes a lot of preparation and of course time is crucial in my view. Some challenges for this project I could include first of all selecting a data crime that could attend to the requirements that were asked, leading with outliers' values, missing data, some large data as well. All those problems or challenges cited we can manage and overcome in (EDA), I could find some error in my code by some research and by redoing my code multiple times to reach my goal and the results. As an example of a challenge that I faced, Ican cite the question (g). This is where I where required to find the (PCA), in this question get some error such as filter numeric columns, missing data and infinite values, so some update in my code were required, I need to filter my numeric values, fill up missing values, replace infinite values, then I could select my three component and overcome it.

Git hub link and Email:

Eliabe2022474/Eliabe_2022474CA1_dataEXP (github.com)

Email: 2022474@student.cct.ie

Links of websites that I used to study and help to build graphs and code:

R Graphics - Pie (w3schools.com)

Handling Errors in R Programming - GeeksforGeeks

https://stackoverflow.comn