Assignment1

code chunk to hide warnings or any extra message .

```
knitr::opts_chunk$set(warning = FALSE, message = FALSE)
```

Assignment 1

This file explains the steps and logic behind the solution provided to each questions given in Assignment 1.

Task 1 : Manipulation

1 : Load the data from excel file

I am loading **readxl** package which is needed to read data from excel file in R.

I have used range attribute to select rows starting from row 9 till row 50

```
library(readxl)
ds_a1 <- read_excel("crim_off_cat_2022.xlsx",
                    col_names = TRUE,na=":",range = cell_rows(9:50))
```

2 What is the size (number of rows and columns) and structure of dataset ?

I will use **str** function to get this information, this provides with data size and structure as output

The `echo: false` option disables the printing of code (only output is displayed).

```
#|ndefined echo: false

str(ds_a1, width = 60, strict.width = "wrap")
```

```
tibble [41 x 22] (S3: tbl_df/tbl/data.frame)
$ ICCS (Labels) : chr [1:41] "Belgium" "Bulgaria" "Czechia"
   "Denmark" ...
$ Intentional homicide : num [1:41] 1.54 1.11 0.75 1 0.74
   1.35 0.87 0.76 0.69 1.21 ...
$ Attempted intentional homicide : num [1:41] 10.29 0.67
   0.6 2.38 2.07 ...
$ Serious assault : num [1:41] 560.2 45.8 39.7 36.7 173.8
   ...
$ Kidnapping : num [1:41] NA 1.21 0.02 NA 5.88 0 1.8 0.94
   0.18 NA ...
$ Sexual violence : num [1:41] 92.04 8.29 17.32 107.19 59.7
   ...
$ Rape : num [1:41] 37 1.4 16.3 48.2 15.1 ...
$ Sexual assault : num [1:41] 55.01 6.89 0.86 58.99 44.61
   ...
$ Sexual exploitation : num [1:41] 18.18 0.95 9.01 57.29
   56.84 ...
$ Child pornography : logi [1:41] NA NA NA NA NA NA ...
$ Robbery : num [1:41] 104.4 13.2 13.5 23.9 45.9 ...
$ Burglary : num [1:41] 445 63.5 320.9 537.8 313.1 ...
$ Burglary of private residential premises : num [1:41]
   336.8 63.6 56 390.4 79.2 ...
$ Theft : num [1:41] 1686 413 331 2496 1254 ...
$ Theft of a motorized vehicle or parts thereof : num
   [1:41] 92.2 22 33 20.4 NA ...
$ Unlawful acts involving controlled drugs or precursors:
   num [1:41] 484 78.3 39.9 532.4 409.3 ...
$ Fraud : num [1:41] 977.2 33.6 153.1 973.6 962.8 ...
$ Corruption : num [1:41] 43.21 7.81 15.31 59.27 5.92 ...
$ Bribery : num [1:41] 0.59 1.54 1.71 0.17 5.92 8.11 NA
   1.25 0.37 0.35 ...
$ Money laundering : num [1:41] 35.2 0.6 6.57 100.5 27.17
   ...
$ Acts against computer systems : num [1:41] 82.15 0.88
   27.08 NA 19.9 ...
$ Participation in an organized criminal group : num [1:41]
   2.94 0.72 0.21 NA 0.77 0.15 NA 3.27 4.29 3.09 ...
```

Ans:-

Number of rows , columns - 41, 22 respectively.

3. Change the column name of first column to "Country"

I am using **colnames** function to rename the column name at index 1 ~ first column name

```
colnames(ds_a1)[1] = "Country"
```

4. Remove certain columns from the dataset.

I am using **subset** function with the list of columns that needs to be removed from dataset and putting it in a modified data set for further use.

```
ds_a_modified = subset(ds_a1,
select=-c(`Child pornography`,`Rape`, `Sexual assault`,`Theft`
          ,`Theft of a motorized vehicle or parts thereof`
          ,`Burglary`, `Burglary of private residential premises` ))
```

5. Work with the dataset you just created, and write some code to list the countries that contain any missing data.

Here I have used **complete.case** function which gets all complete value ~ no missing values and used **not operator** to get all rows having atleast 1 missing value.

```
rows_with_na <- ds_a_modified[!complete.cases(ds_a_modified), ]
rows_with_na$Country
```

```
 [1] "Belgium"                    "Denmark"
 [3] "Estonia"                    "Ireland"
 [5] "France"                     "Cyprus"
 [7] "Latvia"                     "Luxembourg"
 [9] "Hungary"                    "Netherlands"
[11] "Poland"                     "Portugal"
[13] "Slovakia"                   "Sweden"
[15] "Iceland"                    "Liechtenstein"
[17] "Norway"                     "Switzerland"
[19] "England and Wales"          "Scotland (NUTS 2021)"
```

```
[21] "Northern Ireland (UK) (NUTS 2021)" "Bosnia and Herzegovina"
[23] "Montenegro"                        "North Macedonia"
[25] "Serbia"                            "Türkiye"
[27] "Kosovo*"
```

6. Remove the countries with missing data (i.e. countries with at least one NA)

I am using **na.exclude** function to exclude all country with any missing data .

```
ds_a_modified= na.exclude(ds_a_modified)
```

7. Add a column containing the overall record of offences for each country (per hundred thousand inhabitants)

Adding a column named **overall_ofnc_record** using function **rowSums**

```
ds_a_modified$overall_ofnc_record = rowSums(ds_a_modified[,2:length(ds_a_modified)])
```

8. How many observations and variables are in this new dataset?

I will use **str** function to see the modified dataframe structure

```
str(ds_a_modified, width = 60, strict.width = "wrap")
```

```
tibble [14 x 16] (S3: tbl_df/tbl/data.frame)
$ Country : chr [1:14] "Bulgaria" "Czechia" "Germany"
   "Greece" ...
$ Intentional homicide : num [1:14] 1.11 0.75 0.74 0.76
   0.69 0.8 0.55 2.21 1.54 0.72 ...
$ Attempted intentional homicide : num [1:14] 0.67 0.6 2.07
   1.6 2.56 2.8 1.72 0.46 0.96 1.49 ...
$ Serious assault : num [1:14] 45.8 39.7 173.8 12.7 55.8
   ...
$ Kidnapping : num [1:14] 1.21 0.02 5.88 0.94 0.18 0 0.24 0
   0 0.1 ...
$ Sexual violence : num [1:14] 8.29 17.32 59.7 3.93 35.4
   ...
$ Sexual exploitation : num [1:14] 0.95 9.01 56.84 1.02
   4.23 ...
```

```
$ Robbery : num [1:14] 13.2 13.5 45.9 27.1 133.7 ...
$ Unlawful acts involving controlled drugs or precursors:
   num [1:14] 78.3 39.9 409.3 115 40.2 ...
$ Fraud : num [1:14] 33.6 153.1 962.8 98.3 923.9 ...
$ Corruption : num [1:14] 7.81 15.31 5.92 6.23 2.94 ...
$ Bribery : num [1:14] 1.54 1.71 5.92 1.25 0.37 ...
$ Money laundering : num [1:14] 0.6 6.57 27.17 1.18 0.87
   ...
$ Acts against computer systems : num [1:14] 0.88 27.08
   19.9 7.66 15.26 ...
$ Participation in an organized criminal group : num [1:14]
   0.72 0.21 0.77 3.27 4.29 0.93 0.75 0.21 0 0.98 ...
$ overall_ofnc_record : num [1:14] 195 325 1777 281 1220
   ...
- attr(*, "na.action")= 'exclude' Named int [1:27] 1 4 6 7
   10 13 14 16 17 19 ...
..- attr(*, "names")= chr [1:27] "1" "4" "6" "7" ...
```

Ans:- The modified dataframe have 14 rows /Observations and 16 columns /variables

Task 2 : Analysis

Work with the dataset produced at the end of Task 1.

```
ds_a2= ds_a_modified
```

1. Produce a table showing the country names and their record of participation in an organized #criminal group in 2022 sorted by highest to lowest value, and display one decimal digit.

I am loading **dplyr** package to use function **select** to take few columns from the dataframe and using **knitr::kable** to print table and attribute **d=1** is for decimal point 1.

```
library(dplyr)
ds_a2_q1 = select(ds_a2[
  with(ds_a2, order(`Participation in an organized criminal group`, decreasing = T)),
],Country,`Participation in an organized criminal group`)

knitr::kable(ds_a2_q1, align = "ccc",d=1)
```

Country	Participation in an organized criminal group
Spain	4.3
Greece	3.3
Romania	2.7
Albania	1.4
Austria	1.0
Croatia	0.9
Germany	0.8
Italy	0.8
Bulgaria	0.7
Czechia	0.2
Lithuania	0.2
Finland	0.0
Malta	0.0
Slovenia	0.0

2 Which country has the highest record of participation in an organized criminal group in 2022 #(per hundred thousand inhabitants)?

I am using above dataframe and printing the first row first column as the above is sorted in descending order and 1st column and 1st row will give the country with highest record.

```
knitr::kable(ds_a2_q1[1,1],align='ccc')
```

Country
Spain

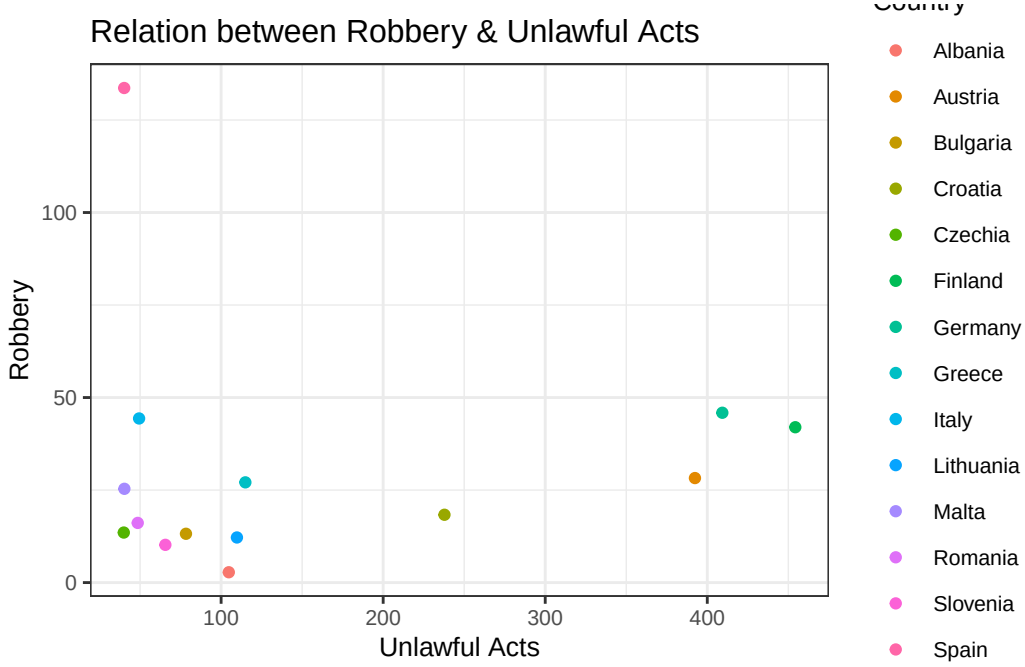
3. Create a plot displaying the relationship between robbery and unlawful acts involving controlled drugs or precursors. Make the plot "nice" i.e., show country names, change size of the plot, change axis labels, etc.

I am loading **ggplot2** package for plotting

```
library(ggplot2)
#|ndefined echo: false
p <- ggplot(ds_a2,
          aes(y = Robbery,
              x = `Unlawful acts involving controlled drugs or precursors`
```

```
  geom_point() +
  labs(x = "Unlawful Acts"
       , y = "Robbery"
       , color = "Country"
       ,title = "Relation between Robbery & Unlawful Acts") +
  theme_bw() + theme(text=element_text(size=10))


p
```

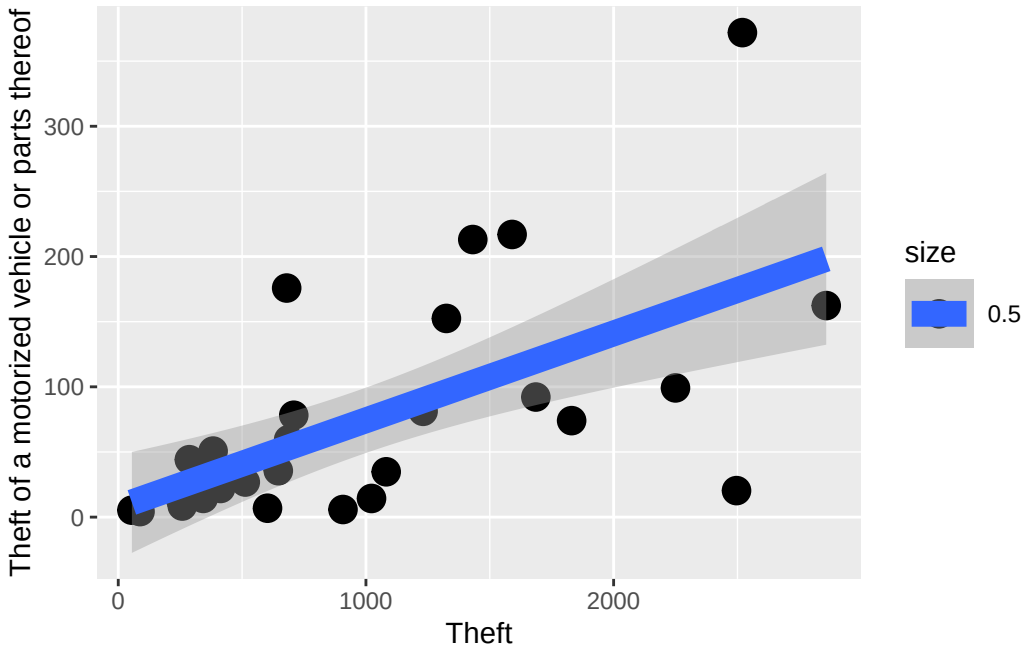


Task 3 Creativity

I see a correlation in theft and theft of a motorized vehicle both are linearly dependent.

```
#|ndefined echo: false

c <- ggplot(ds_a1,aes(x=Theft
                      ,y=`Theft of a motorized vehicle or parts thereof`
                      ,size=0.5)) + geom_point() + stat_smooth(method=lm)
c
```

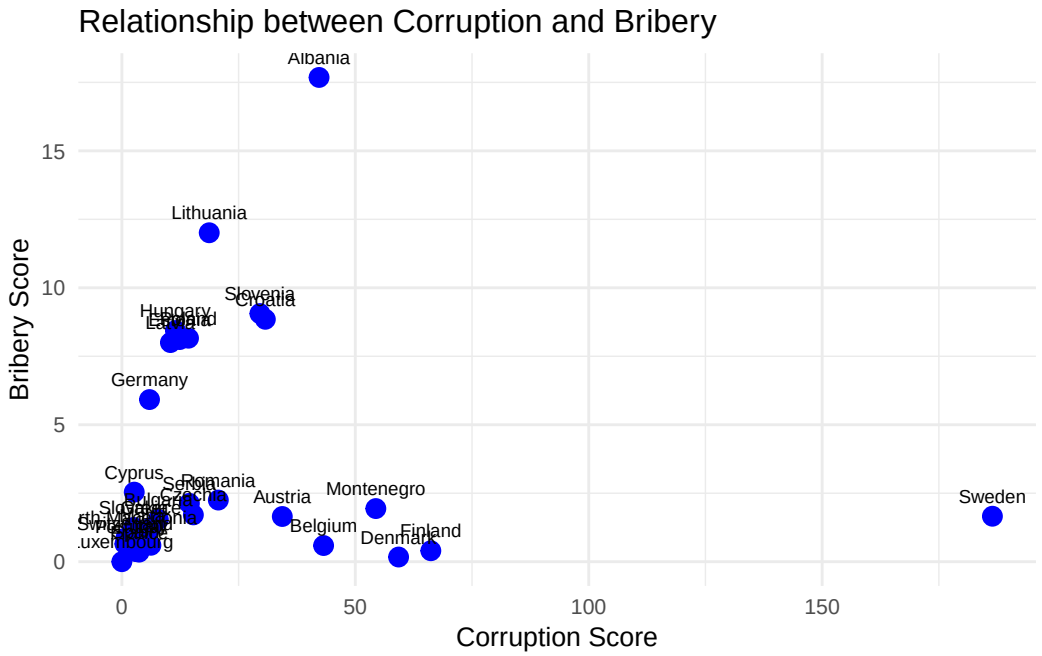


Country - Albania and Sweden stands out in terms of Bribery and corruption scores respectively with highest score .

```
#|ndefined echo: false

d <- ggplot(ds_a1, aes(x = Corruption, y = Bribery, label = Country)) +
  geom_point(color = "blue", size = 3) +  # Points for each country
  geom_text(vjust = -1,size = 2.5) +                 # Add country labels
  labs(title = "Relationship between Corruption and Bribery",
       x = "Corruption Score",
       y = "Bribery Score") +
  theme_minimal() +
  theme(text = element_text(size = 10))

d
```


Relationship between Corruption and Bribery
Albania
Lithuania
Slovenia
Croatia
Hungary
Germany
Cyprus
Romania
Austria
Montenegro
Belgium
Finland
Denmark
Sweden
Bribery Score
Corruption Score
0
5
10
15
0
50
100
150