Assignment 1

Vaishnavi Mohan (22200292)

2022-10-09

Task 1 : Data manipulation

1. Load the dataset EurostatCrime2019.csv.

```
data <- read.csv("C:/Users/HP/Desktop/UCD/assignments/R/EurostatCrime2019.csv",
    row.names = 1)
```

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

The dataset has 41 rows, each representing a country and 13 columns. Structure of the dataset is as below.

```
# size of dataset
nrow(data)
```

```
## [1] 41
```

```
ncol(data)
```

```
## [1] 13
```

```
# structure of dataset
str(data)
```

```
## 'data.frame':    41 obs. of  13 variables:
##  $ Intentional.homicide                                : num  2.03 0.84 1.27 NA 1.14 0.81 1.48 0.7
##  $ Attempted.intentional.homicide                      : num  3.25 1.93 8.87 NA 0.54 2.4 1.71 0.58
##  $ Assault                                             : num  5.52 43.29 556.36 NA 39.54 ...
##  $ Kidnapping                                          : num  0.14 0.07 NA NA 1.03 0.02 0.91 0.11 I
##  $ Sexual.violence                                     : num  5.38 50.9 77.45 NA 8.64 ...
##  $ Rape                                                : num  2.69 18.92 33.33 NA 1.87 ...
##  $ Sexual.assault                                      : num  2.69 26.64 44.12 NA NA ...
##  $ Robbery                                             : num  3.42 29.67 140.14 NA 16.9 ...
##  $ Burglary                                            : num  NA 613.2 565.9 NA 79.8 ...
##  $ Burglary.of.private.residential.premises            : num  40.4 99.3 410.1 NA NA ...
##  $ Theft                                               : num  169 1303 1952 NA 474 ...
##  $ Theft.of.a.motorized.land.vehicle                   : num  11.1 44.2 109.8 NA 18.9 ...
##  $ Unlawful.acts.involving.controlled.drugs.or.precursors: num  70.3 494.1 547.7 NA 78.1 ...
```

3. Produce appropriate commands to do the following actions:

(i) For most countries sexual violence figures are the sum of rape and sexual assault. Remove the columns Rape and Sexual.assault.

Since the columns are in position 6 and 7, the below code can be used to remove the columns and store the result back into original variable.

```
data <- data[, c(-6, -7)]
```

The dataset structure now looks like:

```
print(str(data))
```

```
## 'data.frame':    41 obs. of  11 variables:
##  $ Intentional.homicide                                 : num  2.03 0.84 1.27 NA 1.14 0.81 1.48 0.7
##  $ Attempted.intentional.homicide                       : num  3.25 1.93 8.87 NA 0.54 2.4 1.71 0.58
##  $ Assault                                              : num  5.52 43.29 556.36 NA 39.54 ...
##  $ Kidnapping                                           : num  0.14 0.07 NA NA 1.03 0.02 0.91 0.11 I
##  $ Sexual.violence                                      : num  5.38 50.9 77.45 NA 8.64 ...
##  $ Robbery                                              : num  3.42 29.67 140.14 NA 16.9 ...
##  $ Burglary                                             : num  NA 613.2 565.9 NA 79.8 ...
##  $ Burglary.of.private.residential.premises             : num  40.4 99.3 410.1 NA NA ...
##  $ Theft                                                : num  169 1303 1952 NA 474 ...
##  $ Theft.of.a.motorized.land.vehicle                    : num  11.1 44.2 109.8 NA 18.9 ...
##  $ Unlawful.acts.involving.controlled.drugs.or.precursors: num  70.3 494.1 547.7 NA 78.1 ...
## NULL
```

(ii) For some countries Theft includes also burglary, and theft of motorised land vehicle, in others they are recorded separately. In order to compare the different countries, remove the columns involving theft and burglary.

The subset function has been used to de-select the unwanted columns.

```
data <- subset(data, select = -c(Theft, Theft.of.a.motorized.land.vehicle,
    Burglary, Burglary.of.private.residential.premises))
```

The dataset structure now looks like:

```
print(str(data))
```

```
## 'data.frame':    41 obs. of  7 variables:
##  $ Intentional.homicide                                 : num  2.03 0.84 1.27 NA 1.14 0.81 1.48 0.7
##  $ Attempted.intentional.homicide                       : num  3.25 1.93 8.87 NA 0.54 2.4 1.71 0.58
##  $ Assault                                              : num  5.52 43.29 556.36 NA 39.54 ...
##  $ Kidnapping                                           : num  0.14 0.07 NA NA 1.03 0.02 0.91 0.11 I
##  $ Sexual.violence                                      : num  5.38 50.9 77.45 NA 8.64 ...
##  $ Robbery                                              : num  3.42 29.67 140.14 NA 16.9 ...
##  $ Unlawful.acts.involving.controlled.drugs.or.precursors: num  70.3 494.1 547.7 NA 78.1 ...
## NULL
```

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

```
data["Overall record of offences"] <- apply(data, 1, sum)
```

We can see that the new column has been added now.

```
print(str(data))
```

```
## 'data.frame':    41 obs. of  8 variables:
##  $ Intentional.homicide                                 : num  2.03 0.84 1.27 NA 1.14 0.81 1.48 0.7
##  $ Attempted.intentional.homicide                       : num  3.25 1.93 8.87 NA 0.54 2.4 1.71 0.58
##  $ Assault                                              : num  5.52 43.29 556.36 NA 39.54 ...
##  $ Kidnapping                                           : num  0.14 0.07 NA NA 1.03 0.02 0.91 0.11 I
##  $ Sexual.violence                                      : num  5.38 50.9 77.45 NA 8.64 ...
##  $ Robbery                                              : num  3.42 29.67 140.14 NA 16.9 ...
##  $ Unlawful.acts.involving.controlled.drugs.or.precursors: num  70.3 494.1 547.7 NA 78.1 ...
##  $ Overall record of offences                           : num  90 621 NA NA 146 ...
## NULL
```

4. Work with the dataset you just created, and list the countries that contain any missing data.

Since, if a country contains any missing data for the given variables it will be reflected in the Overall record of offences (as NA has not been removed while taking the sum), to analyse countries with missing (NA) values, it would suffice to look at this column alone.

The is.na() function can be used to select rows that have NA values.

```
countries_with_missing_data <- rownames(data[is.na(data$`Overall record of offences`),
    ])
print(countries_with_missing_data)
```

```
##  [1] "Belgium"                "Bosnia and Herzegovina" "Denmark"
##  [4] "England and Wales"      "Estonia"                "France"
##  [7] "Hungary"                "Iceland"                "Liechtenstein"
## [10] "Netherlands"            "North Macedonia"        "Northern Ireland (UK)"
## [13] "Norway"                 "Poland"                 "Portugal"
## [16] "Scotland"               "Slovakia"               "Sweden"
## [19] "Turkey"
```

5. Remove the countries with missing data from the dataframe.

The subset function can be used as below to filter out countries which do not have any missing data. The is.na() function can be used in a logical condition to leave out rows that have missing values and only select ones that don't.

```
sub_data <- subset(data, !is.na(`Overall record of offences`))
colnames(sub_data)[7] <- "Drugs involvement"
colnames(sub_data)[5] <- "Sexual violence"
```

Column names have been renamed where necessary for convenience and easy readability. The new modified dataframe now looks as shown below, and has been stored in a new variable called sub_data.

```
str(sub_data)
```

```
## 'data.frame':    22 obs. of  8 variables:
##  $ Intentional.homicide          : num  2.03 0.84 1.14 0.81 1.48 0.76 1.59 0.71 0.71 0.71 ...
##  $ Attempted.intentional.homicide: num  3.25 1.93 0.54 2.4 1.71 0.58 5.96 2.18 1.09 0.55 ...
##  $ Assault                       : num  5.52 43.29 39.54 18.06 20.09 ...
##  $ Kidnapping                    : num  0.14 0.07 1.03 0.02 0.91 0.11 0.02 5.44 0.66 1.71 ...
##  $ Sexual violence               : num  5.38 50.9 8.64 21.05 1.94 ...
##  $ Robbery                       : num  3.42 29.67 16.9 20.56 6.28 ...
##  $ Drugs involvement             : num  70.3 494.1 78.1 272.2 117.8 ...
##  $ Overall record of offences    : num  90 621 146 335 150 ...
```

```
head(sub_data, 4)
```

```
##          Intentional.homicide Attempted.intentional.homicide Assault Kidnapping
## Albania                  2.03                           3.25    5.52       0.14
## Austria                  0.84                           1.93   43.29       0.07
## Bulgaria                 1.14                           0.54   39.54       1.03
## Croatia                  0.81                           2.40   18.06       0.02
##          Sexual violence Robbery Drugs involvement Overall record of offences
## Albania             5.38    3.42             70.26                      90.00
## Austria            50.90   29.67            494.05                     620.75
## Bulgaria            8.64   16.90             78.14                     145.93
## Croatia            21.05   20.56            272.16                     335.06
```

6. How many observations and variables are in this new dataframe?

There are 22 observations(rows) and 8 variables(columns) in the new dataframe, as found using the below code.

```
print(str(sub_data))
```

```
## 'data.frame':    22 obs. of  8 variables:
##  $ Intentional.homicide          : num  2.03 0.84 1.14 0.81 1.48 0.76 1.59 0.71 0.71 0.71 ...
##  $ Attempted.intentional.homicide: num  3.25 1.93 0.54 2.4 1.71 0.58 5.96 2.18 1.09 0.55 ...
##  $ Assault                       : num  5.52 43.29 39.54 18.06 20.09 ...
##  $ Kidnapping                    : num  0.14 0.07 1.03 0.02 0.91 0.11 0.02 5.44 0.66 1.71 ...
##  $ Sexual violence               : num  5.38 50.9 8.64 21.05 1.94 ...
##  $ Robbery                       : num  3.42 29.67 16.9 20.56 6.28 ...
##  $ Drugs involvement             : num  70.3 494.1 78.1 272.2 117.8 ...
##  $ Overall record of offences    : num  90 621 146 335 150 ...
## NULL
```

```
nrow(sub_data)
```

```
## [1] 22
```

```
ncol(sub_data)
```

```
## [1] 8
```

Task 2 : Analysis

1. According to these data what were the 3 most common crimes in Ireland in 2019?

```
sub_data["Ireland", order(sub_data["Ireland", 1:7], decreasing = T)][,
    1:3]
```

```
##         Drugs involvement Assault Sexual violence
## Ireland            421.84  102.18           67.86
```

To find this, the data of Ireland is filtered, and columns, each of which contains a crime, are ordered from highest to lowest. Then, the first three are further filtered to get the 3 most common crimes.

We can see that the 3 most common crimes in Ireland in 2019 are 'Unlawful acts involving controlled drugs or precursors', 'Assault' and 'Sexual violence'.

2. What proportion of the overall crimes was due to Assault in Ireland in 2019?

```
ireland_stats <- sub_data["Ireland", ]
ireland_stats$Assault/ireland_stats$`Overall record of offences`
```

```
## [1] 0.1605316
```

Therefore, about 16% of crimes in Ireland in 2019 was due to Assault.

3. Which country had the highest record of kidnapping in 2019 (per hundred thousand inhabitants)?

```
sub_data[sub_data$Kidnapping == max(sub_data$Kidnapping), ]
```

```
##            Intentional.homicide Attempted.intentional.homicide Assault
## Luxembourg                 0.65                            9.61  103.76
##            Kidnapping Sexual violence Robbery Drugs involvement
## Luxembourg       7.17           48.22   74.44            690.35
##            Overall record of offences
## Luxembourg                      934.2
```

Luxembourg had the highest record of kidnapping at 7.17 per thousand inhabitants.

4. Which country had the lowest overall record of offences in 2019 (per hundred thousand inhabitants)?

```
sub_data[sub_data$`Overall record of offences` == min(sub_data$`Overall record of offences`),
    ]
```

```
##         Intentional.homicide Attempted.intentional.homicide Assault Kidnapping
## Romania                 1.31                           1.86    1.46       2.01
##         Sexual violence Robbery Drugs involvement Overall record of offences
## Romania            9.85   17.85             35.72                      70.06
```

Romania had the lowest overall record of offences in 2019 at a value of 140.12 per thousand inhabitants.

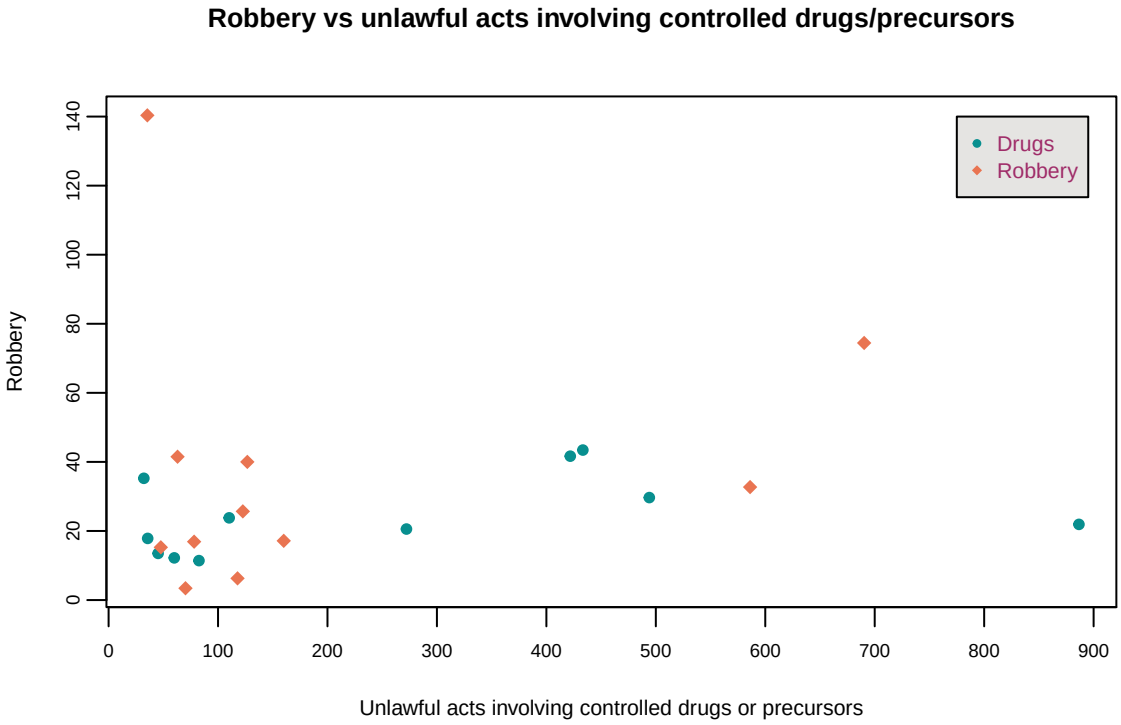
5. Create a plot displaying the relationship between robbery and unlawful acts involving controlled drugs or precursors. Make the plot look "nice" i.e. change axis labels etc.

```
Robbery <- sub_data$Robbery
Drugs <- sub_data$`Drugs involvement`

par(mgp = c(2, 0.5, 0), cex.main = 0.8)
plot(Drugs, Robbery, main = "Robbery vs unlawful acts involving controlled drugs/precursors",
    xlab = "Unlawful acts involving controlled drugs or precursors",
    col = c("#E97451", "#088F8F"), pch = c(18, 20), xaxt = "n",
    cex.axis = 0.6, cex.lab = 0.7)

axis(1, at = seq(0, 1000, 100), cex.axis = 0.6)  # customize ticks of x axis

legend(775, 140, pch = c(20, 18), legend = c("Drugs", "Robbery"),
    col = c("#088F8F", "#E97451"), cex = 0.7, text.col = "#9F2B68",
    bg = "#E5E4E2")
```



Saving for a few outliers, there is a small positive correlation evident between the above two variables.

Task 3 : Creativity

Do something interesting with these data (either the original dataset or the modified one)! Create a nice plot which shows something we have not discovered above already and outline your findings.

Let us take a look at the overall record of offences statistics for 2019.

```
summary(sub_data$`Overall record of offences`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.06  136.91  195.69  328.43  549.33  950.32
```

I would like to filter out countries in which the overall record statistics falls in the top 75%, that is, greater than the thirs quantile value 549.33 (approx 550). I will call this chunk of data as countries_with_top_overallCrime.

```
countries_with_top_overallCrime <- sub_data[sub_data$`Overall record of offences` >
    550, ]
print(countries_with_top_overallCrime)
```

```
##                                                 Intentional.homicide
## Austria                                                         0.84
## Finland                                                         1.59
## Germany (until 1990 former territory of the FRG)                0.71
## Ireland                                                         0.71
## Luxembourg                                                      0.65
## Switzerland                                                     0.54
##                                                 Attempted.intentional.homicide
## Austria                                                                   1.93
## Finland                                                                   5.96
## Germany (until 1990 former territory of the FRG)                          2.18
## Ireland                                                                   0.55
## Luxembourg                                                                9.61
## Switzerland                                                               1.88
##                                                 Assault Kidnapping
## Austria                                           43.29       0.07
## Finland                                           28.71       0.02
## Germany (until 1990 former territory of the FRG) 160.31       5.44
## Ireland                                          102.18       1.71
## Luxembourg                                       103.76       7.17
## Switzerland                                        7.46       0.06
##                                                 Sexual violence Robbery
## Austria                                                   50.90   29.67
## Finland                                                   72.65   32.71
## Germany (until 1990 former territory of the FRG)          49.05   43.43
## Ireland                                                   67.86   41.66
## Luxembourg                                                48.22   74.44
## Switzerland                                               31.86   21.91
##                                                 Drugs involvement
## Austria                                                    494.05
## Finland                                                    586.14
## Germany (until 1990 former territory of the FRG)           433.33
## Ireland                                                    421.84
## Luxembourg                                                 690.35
## Switzerland                                                886.61
##                                                 Overall record of offences
## Austria                                                             620.75
## Finland                                                             727.78
## Germany (until 1990 former territory of the FRG)                    694.45
## Ireland                                                             636.51
## Luxembourg                                                          934.20
## Switzerland                                                         950.32
```

To increase readability and convenience, the rowname for Germany can be shortened.

```
row.names(countries_with_top_overallCrime)[3] <- "Germany"
```

Now let us order the countries in increasing order of the number of overall crimes.

```
countries_with_top_overallCrime <- countries_with_top_overallCrime[order(countries_with_top_overallCrime
    ]
print(countries_with_top_overallCrime[8])
```

```
##             Overall record of offences
```

```
## Austria                    620.75
## Ireland                    636.51
## Germany                    694.45
## Finland                    727.78
## Luxembourg                 934.20
## Switzerland                950.32
```

Country wise analysis

We can now take a look at this data through visualization.

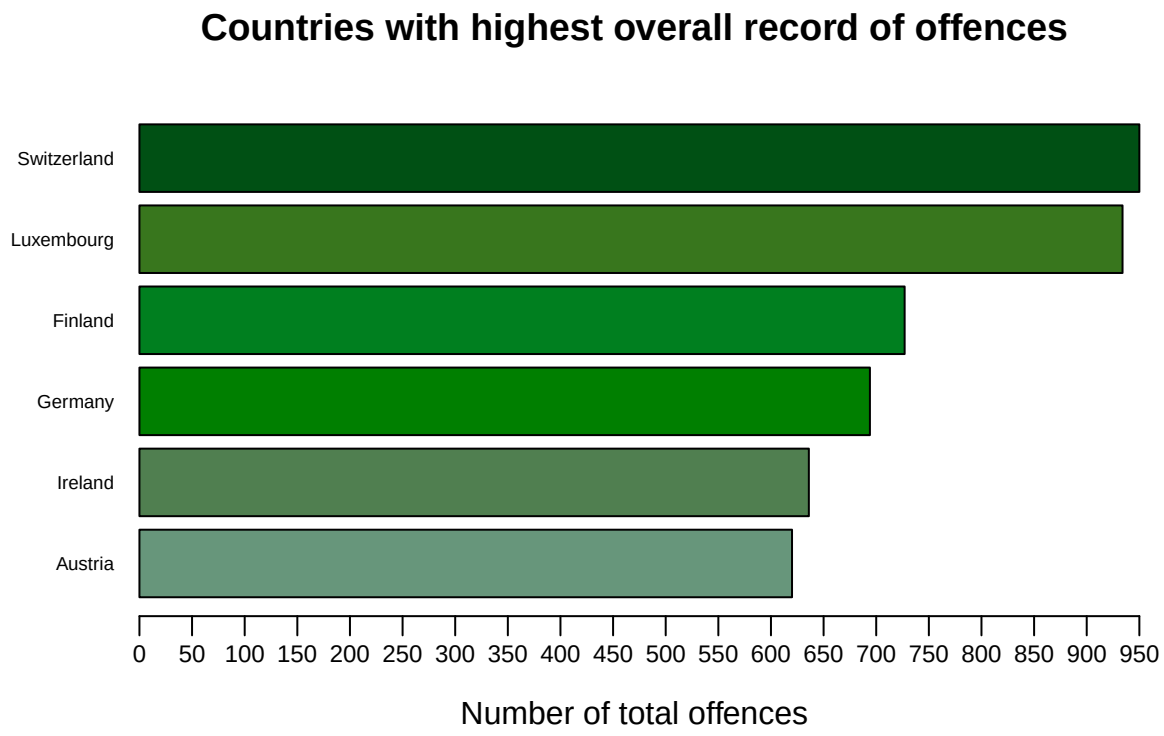
```
# rownames, i.e., country names form labels
labels <- rownames(countries_with_top_overallCrime)

top_countries <- factor(rep(labels, countries_with_top_overallCrime[labels,
    ]$`Overall record of offences`))

par(mgp = c(2, 0.4, 0))

barplot(height = sort(table(top_countries)), horiz = T, las = 1,
    main = "Countries with highest overall record of offences",
    cex.names = 0.6, cex.axis = 0.75, xaxt = "n", col = c("#67967b",
        "#507f50", "#007f00", "#007f1f", "#38761d", "#005014"),
    xlab = "Number of total offences")

# customisation of x-axis ticks and limits
axis(1, at = seq(0, 1000, 50), gap.axis = 0.1, cex.axis = 0.75)
```

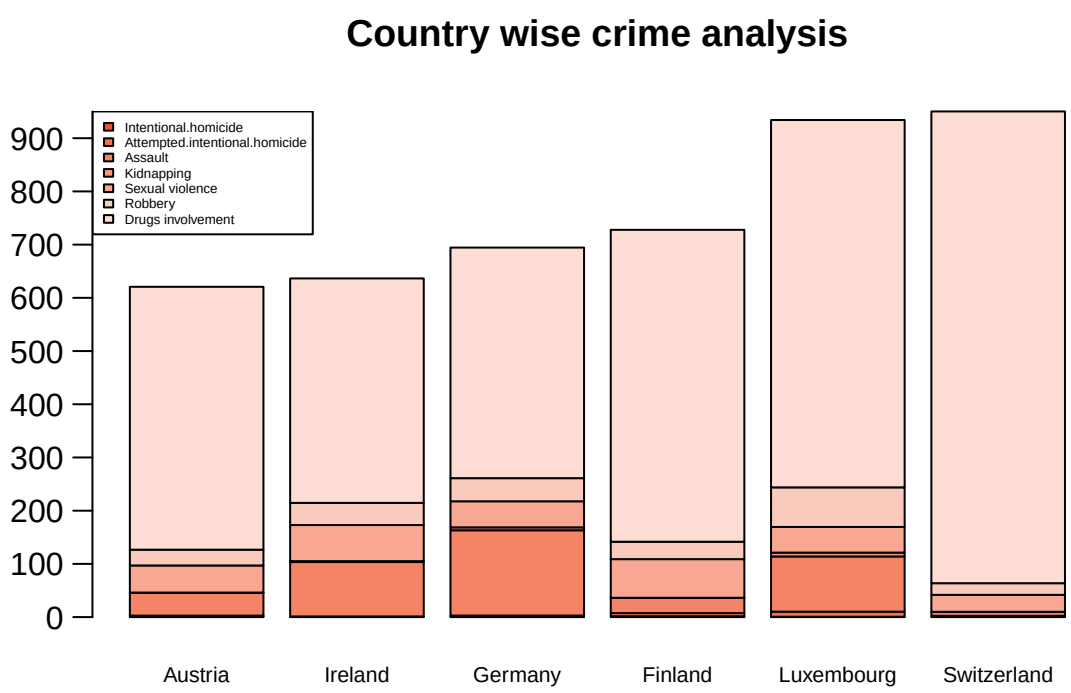


It is evident that Switzerland has the highest record, nearly 950. Austria is the last of top 6 with about 650 crimes. Luxembourg is a close competition to Switzerland. In comparison, Finland, Germany, Ireland have lesser crimes.

Let us further analyse the type of crimes that have taken place in these top countries.

```
# excluding overall count and converting df to matrix
countries_matrix <- as.matrix(countries_with_top_overallCrime[,
    1:7])
par(mgp = c(2, 0.75, 0))

# Create grouped bar-chart
barplot(t(countries_matrix)[c(-8), ], col = c("#f15025", "#f37250",
    "#f58466", "#f6967c", "#f8a792", "#facabd", "#fcdcd3"), yaxt = "n",
    main = "Country wise crime analysis", cex.names = 0.65)
# customization of y axis
axis(2, at = seq(0, 1000, 100), las = 1, cex = 0.5)
# Add legend to bar-plot
legend("topleft", legend = colnames(countries_with_top_overallCrime[,
    1:7]), fill = c("#f15025", "#f37250", "#f58466", "#f6967c",
    "#f8a792", "#facabd", "#fcdcd3"), cex = 0.4)
```



Inferences- 1) The lion's share of the total record is taken by drug offences in each of the top 6 countries. The occurrences of drug offences even overpower the sum of all other crimes put together. 2) Robbery and assault are the next major crimes. The other crimes seem negligible in numbers compared to these two. 3) Personal inference : It could be possible, in my point of view, that robbery and assault are behaviours that result from the influence of drugs.

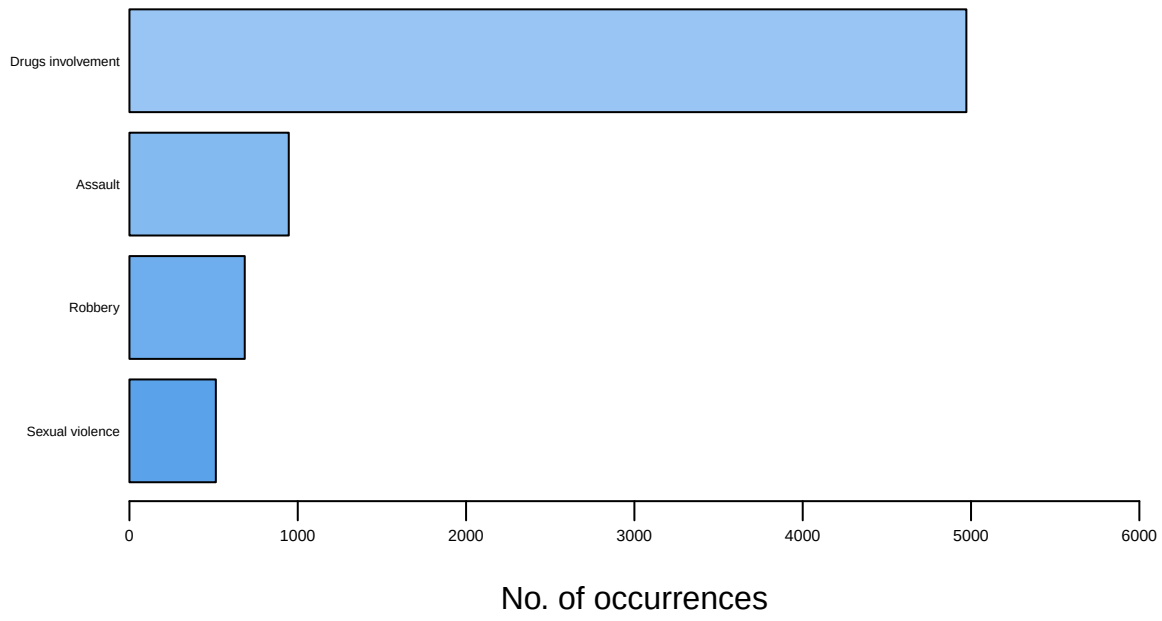
Crime wise analysis

Let us take a look at the top crimes in Europe in 2019.

```
crime_wise_sum <- apply(sub_data[-8], 2, sum)

crime_wise_sum <- crime_wise_sum[order(crime_wise_sum, decreasing = T)]

par(mgp = c(2, 0.25, 0))
barplot(crime_wise_sum[4:1], horiz = T, las = 1, cex.names = 0.45,
    cex.axis = 0.5, col = c("#5aa3eb", "#6eaeee", "#83baf0",
        "#98c5f3"), xlim = c(0, 6000), xlab = "No. of occurrences")
```



Inference : Drugs seems to be a primary shareholder of the crime stats. The others of the top 4 crimes, don't even come close to it. Their number of occurrences are significantly smaller than drug cases. Again, my personal inference is that assault, sexual violence and robbery maybe consequences of drugged stupor.