Assignment 1

Conor Heffron 23211267

Task 1: Manipulation

1. Load dataset

- Load the dataset EurostatCrime2021.xlsx. Notice that the data starts in row 6, with row 7 containing the variable names, and that the missing values are represented by ":". [Note: do not modify the original file EurostatCrime2021.xlsx directly (by opening it with Excel, for example) as we need to be able to render your .Qmd file and reproduce your final document using the original dataset.

```
library(readxl)

EurostatCrime2021 <- read_excel("/Users/conorheffron/Library/CloudStorage/GoogleDrive-cono
```

2. Size / Dimensions of data set

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

```
dim(EurostatCrime2021)
```

```
[1] 41 18
```

- 41 records and 18 columns / 41 observations and 18 variables

3. Remove fraud columns X2

- Remove the columns Fraud and Money laundering (they contain no data).

```r
# Load dplyr for df manipulation
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# Remove Fraud and Money Laundering columns
df <- select(EurostatCrime2021, -Fraud, -"Money laundering")
```

4. Remove theft/burglary related columns X2

- For some countries Theft includes also Theft of a motorized vehicle or parts thereof, Burglary, and Burglary of private residential premises, in others they are recorded separately. To compare different countries, remove the columns involving theft and burglary:
- Theft
- Theft of a motorized vehicle or parts thereof
- Burglary
- Burglary of private residential premises

```r
# Remove x4 columns related to theft / burglary
df <- subset(select(df, -Theft, -"Theft of a motorized vehicle or parts thereof", -Burglar
```

5. Add a column

- Add a column containing the overall record of offences for each country (per hundred thousand inhabitants) [Hint: there is a function in base R that allow you to do this].

```r
# Create copy of df and remove non numeric column (Country)
df_numeric <- select(df, -Country)
```

```
# Convert all numeric columns to the numeric type for further computation via numeric colu
df[colnames(df_numeric)] <- sapply(df[colnames(df_numeric)], as.numeric)
```

```
Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion
```

```
# Create offences column (row sums)
df$offences <- rowSums(select(df, -Country))
```

6. List countries with missing data

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

```
countries<-list()
for(i in 1:nrow(df)) {
 if(any(is.na(df[i,]))) {
  countries <- append(countries, df[[i,"Country"]])
 }
}
str(countries)
```

```
List of 29
 $ : chr "Belgium"
 $ : chr "Bulgaria"
 $ : chr "Denmark"
 $ : chr "Germany"
 $ : chr "Estonia"
 $ : chr "France"
 $ : chr "Italy"
 $ : chr "Luxembourg"
 $ : chr "Hungary"
 $ : chr "Malta"
 $ : chr "Netherlands"
 $ : chr "Poland"
 $ : chr "Portugal"
 $ : chr "Slovenia"
 $ : chr "Slovakia"
 $ : chr "Sweden"
 $ : chr "Iceland"
 $ : chr "Liechtenstein"
 $ : chr "Norway"
 $ : chr "Switzerland"
 $ : chr "England and Wales"
 $ : chr "Scotland"
 $ : chr "Northern Ireland (UK)"
 $ : chr "Bosnia and Herzegovina"
 $ : chr "Montenegro"
 $ : chr "North Macedonia"
 $ : chr "Serbia"
 $ : chr "Türkiye"
 $ : chr "Kosovo (under United Nations Security Council Resolution 1244/99)"
```

7. Remove the countries with missing data.

```r
# Keep copy of original data for creativity task
df_original <- df

# Omit records with NA values
df <- na.omit(df)
```

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

```
dim(df)
```

```
[1] 12 13
```

- 12 observations and 13 variables

Task 2: Analysis

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

1. Which country has the highest overall record of offences in 2021 (per hundred thousand inhab- itants)? To get full marks you must also provide the R code that returns that country name only. [1]

```
# library(tidyverse)
# df[which.max(df$offences),"Country"] %>%
#   pluck("Country", 1)

df[which.max(df$offences),"Country"][[1,1]]
```

```
[1] "Finland"
```

2. Produce a table showing the countries and the proportion of the overall crimes due to acts against computer system, sort the rows in ascending order of proportions, and display only the first three decimal digits. [2]

```
# Create copy of df for plot in next task
df_copy <- df

# Create df with country, offences and acts against computer system columns
df <- select(df, Country, offences, "Acts against computer systems")

# Create proportion column
df$proportion <- (df$"Acts against computer systems" / df$offences) * 100
```

```r
# Round proportion column to 3 decimal places
df[,'proportion'] = round(df[, 'proportion'], 3)

# Sort by proportion in ascending order
df <- df[order(as.integer(df$proportion), decreasing = FALSE),]
df <- select(df, -"Acts against computer systems")

# Print table
library(knitr)
#| label: tbl-LABEL
#| tbl-cap: CAPTION
knitr::kable(df)
```

Country	offences	proportion
Ireland	577.69	0.370
Latvia	172.10	0.552
Greece	162.93	2.222
Finland	785.50	4.537
Albania	176.96	4.035
Spain	249.70	6.320
Romania	109.92	7.114
Croatia	346.16	9.282
Cyprus	168.97	9.641
Czechia	138.13	12.626
Lithuania	196.16	12.857
Austria	733.71	23.023

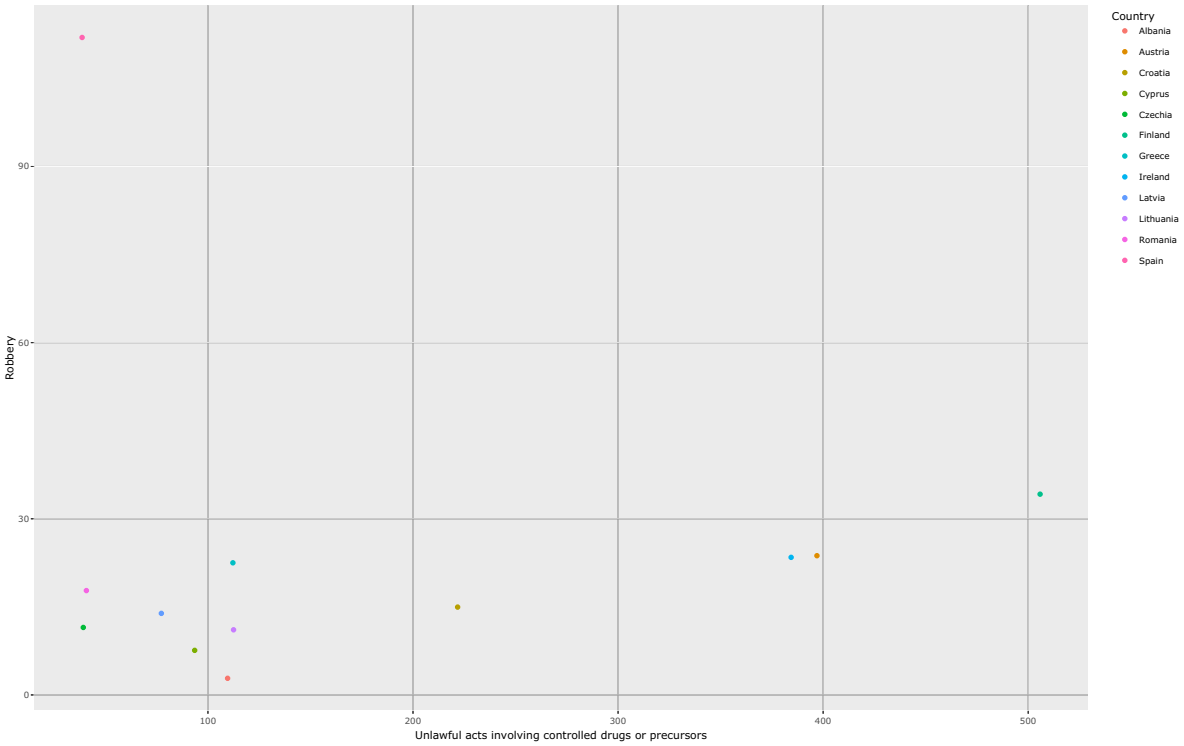
3. Create a plot displaying the relationship between robbery and unlawful acts involving controlled drugs or precursors. Make the plot "nice" i.e., show country names, change size of the plot, change axis labels, etc.

```r
# Load plot libraries
library(ggplot2)
library(plotly)
```

```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':
```

```
    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```r
# Replace spaces in column names with underscores for convenience
colnames(df_copy) <- gsub(" ", "_", colnames(df_copy))

# Change plot size
options(repr.plot.width=5, repr.plot.height=4)

# Plot and label graph
p <- df_copy |>
  ggplot(mapping = aes(x = Unlawful_acts_involving_controlled_drugs_or_precursors, y = Rob
  geom_point(aes(colour = Country)) +
  labs(x = "Unlawful acts involving controlled drugs or precursors")

ggplotly(p)
```



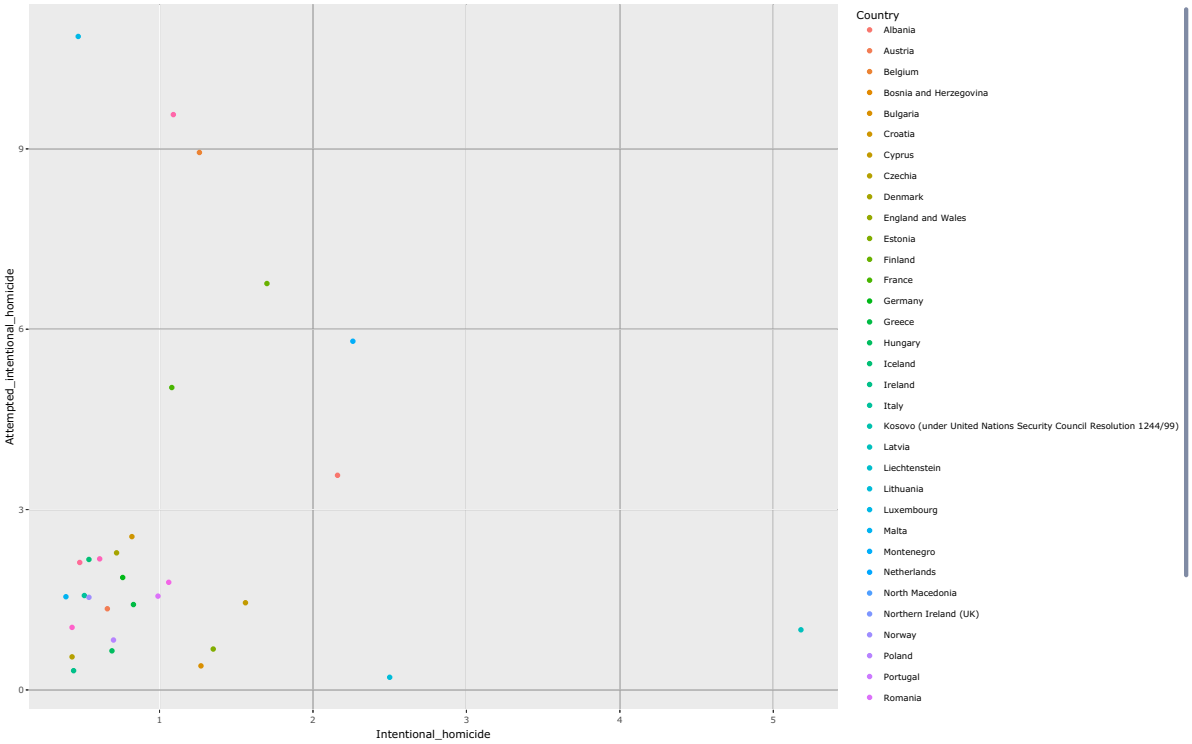
Task 3: Creativity

- Do something interesting with these data! Create two plots showing something we have not discovered above already and outline your findings. For this Task you can decide if you want to use the original dataset from Task 1 Question 1 or the modified one.

```
# Plot 1

# Replace spaces in column names with underscores for convenience
colnames(df_original) <- gsub(" ", "_", colnames(df_original))

# Check correlation of the 2 cols in original dataset before NA omit
cor(x = df_original$Intentional_homicide, y = df_original$Attempted_intentional_homicide,
```

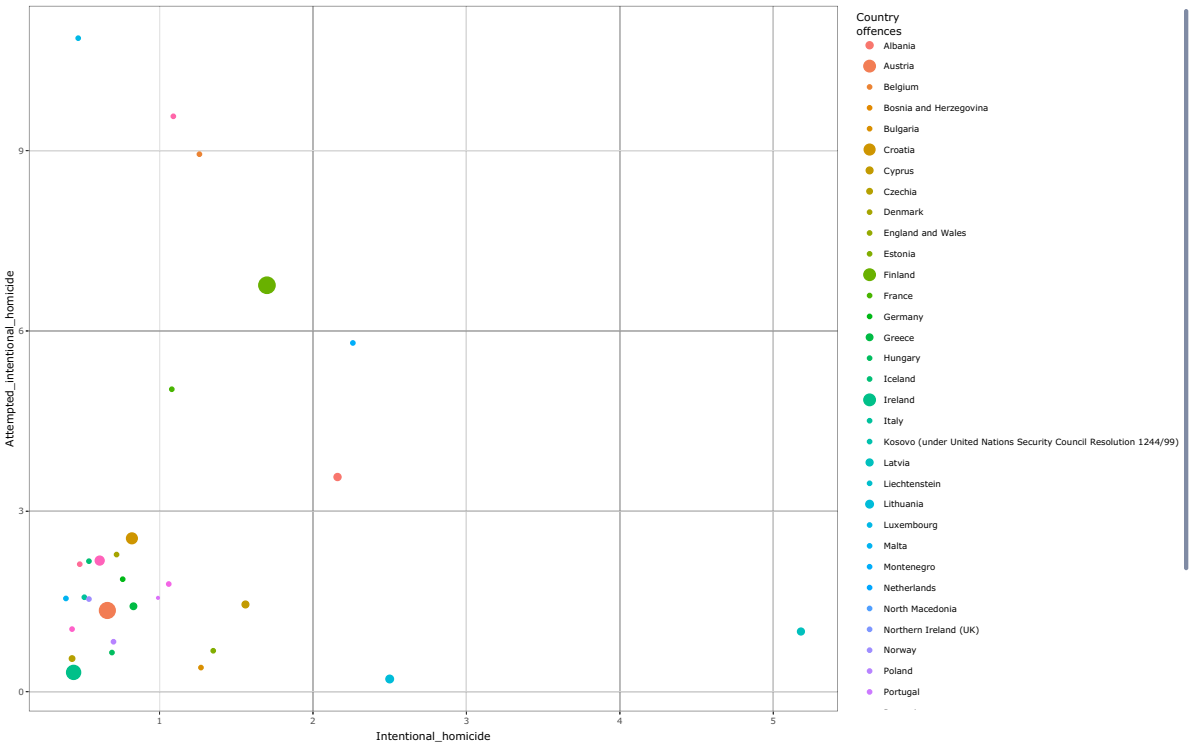
```
[1] 0.01422267
```

```r
# Plot and label graph
p1 <- df_original |>
  ggplot(mapping = aes(x = Intentional_homicide, y = Attempted_intentional_homicide)) +
  geom_point(aes(colour = Country))

ggplotly(p1)
```



- I created a plot to display relationship between attempted intentional homicide and intentional homicide.
- I think it is interesting that Latvia is an outlier with a lot more intentional than attempted intentional homicide (only right side point - everything else lower than intentional rate of 3
- Most countries fall into ([0:1], [0:3]) quadrant, bottom left corner of plot
- However, 0.01422267 is correlation coefficient between the two columns in original dataset so this implies there is only negligible correlation between the two variables.
- Note: For plotly pdf renedering support, run install.packages(webshot") & webshot::install_phantomjs()

```
# Plot 2

# Plot and label graph
p2 <- df_original |>
  ggplot(mapping = aes(x = Intentional_homicide, y = Attempted_intentional_homicide)) +
  geom_point(aes(size = offences, colour = Country)) +
  theme_bw()

ggplotly(p2)
```



- I created a second plot where the size of the plot point is weighted by offences value.
- It is much easier to see Finland, Ireland and Austria have a higher overall rate of offences now but Latvia do not so this may explain that less data or lower overall offences recorded may have contributed to its outlier position (bad or missing data).