Assignment1

Beyza Kordan

Crime Offences in European Countries (2022): Data Manipulation and Analysis

Introduction

This report analyzes the 2022 crime offences dataset for 41 European countries, focusing on data manipulation and analysis using R. The dataset records various crime categories, with values representing offences per hundred thousand inhabitants.

The report is divided into three tasks: cleaning and manipulating the dataset, conducting specific analyses on organized crime, and creating visualizations to uncover additional insights. Key steps include handling missing data, removing irrelevant columns, and calculating overall crime rates. The final goal is to gain a clearer understanding of crime patterns across Europe through data-driven insights.

Task 1 : Data Manipulation

Question 1

In this question, the dataset `crim_off_cat_2022.xlsx` was loaded using the `read_excel` function. Since the relevant data begins at row 9, the first eight rows were skipped. Missing values, represented by colons (`":"`), were replaced with `NA` to handle them appropriately in the analysis. Additionally, rows with invalid data, such as "Special value," were removed. After these steps, the dataset was verified to include 41 rows and 22 variables, confirming that the data was loaded correctly and is ready for further manipulation.

```r
library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.3.3
```

```r
file_path <- "C:/Users/pc/Downloads/crim_off_cat_2022.xlsx"

# Load the dataset starting from row 9, with ":" replaced by NA
dataset <- read_excel(file_path, skip = 8, na = ":")

# Remove rows that don't have valid country data and filter out the "Special value" row
dataset <- dataset[!is.na(dataset[[1]]) & dataset[[1]] != "Special value", ]

# Display the first few rows of the dataset
head(dataset, 10)
```

```
# A tibble: 10 x 22
   `ICCS (Labels)` `Intentional homicide` `Attempted intentional homicide`
   <chr>           <chr>                                             <dbl>
 1 Belgium         1.54                                              10.3 
 2 Bulgaria        1.1100000000000001                                 0.67
 3 Czechia         0.75                                               0.6 
 4 Denmark         1                                                  2.38
 5 Germany         0.74                                               2.07
 6 Estonia         1.35                                               0.68
 7 Ireland         0.87                                               0.24
 8 Greece          0.76                                               1.6 
 9 Spain           0.69                                               2.56
10 France          1.21                                               5.28
# i 19 more variables: `Serious assault` <dbl>, Kidnapping <dbl>,
#   `Sexual violence` <dbl>, Rape <dbl>, `Sexual assault` <dbl>,
#   `Sexual exploitation` <dbl>, `Child pornography` <lgl>, Robbery <dbl>,
#   Burglary <dbl>, `Burglary of private residential premises` <dbl>,
#   Theft <dbl>, `Theft of a motorized vehicle or parts thereof` <dbl>,
#   `Unlawful acts involving controlled drugs or precursors` <dbl>,
#   Fraud <dbl>, Corruption <dbl>, Bribery <dbl>, `Money laundering` <dbl>, ...
```

```r
# Check the dimensions
dim(dataset)
```

```
[1] 41 22
```

Question 2

In this question, I examined the size and structure of the dataset. The dataset contains 41 rows and 22 columns, representing crime data for 41 European countries across various offence categories. The `str()` function was used to provide a summary of the structure, displaying the data types of the columns and a sample of the values for each variable.

```r
size <- dim(dataset)
print(paste("Number of rows:", size[1]))
```

```
[1] "Number of rows: 41"
```

```r
print(paste("Number of columns:", size[2]))
```

```
[1] "Number of columns: 22"
```

```r
# Structure of the dataset
str(dataset, max.level = 1)
```

```
tibble [41 x 22] (S3: tbl_df/tbl/data.frame)
```

Question 3

In this step, the first column of the dataset was renamed to "Country" to make it clearer. The change was confirmed by checking the first few column names.

```
colnames(dataset)[1] <- "Country"

# Check the column names to confirm the change
colnames(dataset)
```

```
 [1] "Country"
 [2] "Intentional homicide"
 [3] "Attempted intentional homicide"
 [4] "Serious assault"
 [5] "Kidnapping"
 [6] "Sexual violence"
 [7] "Rape"
 [8] "Sexual assault"
 [9] "Sexual exploitation"
[10] "Child pornography"
[11] "Robbery"
[12] "Burglary"
[13] "Burglary of private residential premises"
[14] "Theft"
[15] "Theft of a motorized vehicle or parts thereof"
[16] "Unlawful acts involving controlled drugs or precursors"
[17] "Fraud"
[18] "Corruption"
[19] "Bribery"
[20] "Money laundering"
[21] "Acts against computer systems"
[22] "Participation in an organized criminal group"
```

Question 4

The columns "Child pornography," "Rape," "Sexual assault," "Theft," and related categories were removed to improve data consistency. After the removal, the first few column names were checked to confirm the changes. The dataset now includes only the relevant variables, such as "Sexual violence" and "Robbery," ensuring the data is ready for analysis.

```
dataset <- dataset[, !colnames(dataset) %in% c(
  "Child pornography",
  "Rape",
  "Sexual assault",
  "Theft",
  "Theft of a motorized vehicle or parts thereof",
  "Burglary",
  "Burglary of private residential premises")]

# Checking the remaining column names
colnames(dataset)
```

```
 [1] "Country"
 [2] "Intentional homicide"
 [3] "Attempted intentional homicide"
 [4] "Serious assault"
 [5] "Kidnapping"
 [6] "Sexual violence"
 [7] "Sexual exploitation"
 [8] "Robbery"
 [9] "Unlawful acts involving controlled drugs or precursors"
[10] "Fraud"
[11] "Corruption"
[12] "Bribery"
[13] "Money laundering"
[14] "Acts against computer systems"
[15] "Participation in an organized criminal group"
```

Question 5

In response to this question, I identified the countries with missing data by checking all relevant columns, excluding the "Country" column. The countries with missing data include France, Montenegro, Serbia, Türkiye, and Kosovo*. In total, 27 rows contained missing values, indicating these countries may need further data cleaning or adjustments in the analysis.

```
# Checking for missing values in the dataset
countries_with_missing_data <- dataset[rowSums(is.na(dataset[, -1])) > 0,"Country"]

# Display the list of countries with missing data
print(countries_with_missing_data)
```

```
# A tibble: 27 x 1
   Country
   <chr>
 1 Belgium
 2 Denmark
 3 Estonia
 4 Ireland
 5 France
 6 Cyprus
 7 Latvia
 8 Luxembourg
 9 Hungary
10 Netherlands
# i 17 more rows
```

Question 6

In this step, all countries with missing data were removed using the `complete.cases()` function. After removal, the dataset now includes only countries with complete data, such as Bulgaria, Czechia, and Germany.

```
dataset_clean <- dataset[complete.cases(dataset), ]
cat("Remaining countries after removing those with missing data:\n")
```

```
Remaining countries after removing those with missing data:
```

```r
print(dataset_clean$Country)
```

```
 [1] "Bulgaria"  "Czechia"   "Germany"   "Greece"    "Spain"     "Croatia"  
 [7] "Italy"     "Lithuania" "Malta"     "Austria"   "Romania"   "Slovenia" 
[13] "Finland"   "Albania"  
```

Question 7

A new column, "Overall_Offences," was added to the dataset to represent the total number of offences per country. The total was calculated using the `rowSums()` function across all offence-related columns. For instance, the total offences for Bulgaria, Czechia, and Germany were 194.69, 324.79, and 1776.69, respectively.

```r
offense_columns <- dataset_clean[, -1]

# Ensure all offense-related columns are numeric
offense_columns[] <- lapply(offense_columns, as.numeric)

# Add a new column with the sum of all offense categories
dataset_clean$Overall_Offences <- rowSums(offense_columns, na.rm = TRUE)

head(dataset_clean[, c("Country", "Overall_Offences")])
```

```
# A tibble: 6 x 2
  Country  Overall_Offences
  <chr>               <dbl>
1 Bulgaria             195.
2 Czechia              325.
3 Germany             1777.
4 Greece               281.
5 Spain               1220.
6 Croatia              487.
```

Question 8

In this step, the new dataset was checked for its dimensions. The cleaned dataset contains 14 observations (countries) and 16 variables, including the new "Overall_Offences" column. This confirms the size of the dataset after removing countries with missing data and adding the new column.

```r
# Dimensions of the new dataset
num_rows <- nrow(dataset_clean)
num_columns <- ncol(dataset_clean)

cat("Number of observations (rows):", num_rows, "\n")
```

```
Number of observations (rows): 14
```

```r
cat("Number of variables (columns):", num_columns, "\n")
```

```
Number of variables (columns): 16
```

Task 2 : Analysis

Question 1

A table was created to show the participation in organized criminal groups for each country in 2022, sorted from highest to lowest. Spain had the highest rate at 4.3, followed by Greece with 3.3, and Romania with 2.7. Several countries, including Finland, Malta, and Slovenia, reported no participation (0.0) in organized crime.

```
# Extract the relevant columns
organized_crime_data <-
  dataset_clean[, c("Country", "Participation in an organized criminal group")]

# Sort the data from highest to lowest
organized_crime_data <-
  organized_crime_data[order(-organized_crime_data$`Participation in an organized criminal group`), ]

# Round the participation values to one decimal place
organized_crime_data$`Participation in an organized criminal group` <-
  round(organized_crime_data$`Participation in an organized criminal group`, 1)

if(!require(knitr)) install.packages("knitr")
```

```
Zorunlu paket yükleniyor: knitr
```

```
Warning: package 'knitr' was built under R version 4.3.2
```

```
# Create a nicely formatted table for the PDF or Markdown output
kable(organized_crime_data, format = "markdown", col.names =
        c("Country", "Participation in Organized Crime (2022)"))
```

Country	Participation in Organized Crime (2022)
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

Question 2

To find the country with the highest participation in organized criminal groups, the `filter()` function was used to identify the maximum value in the “Participation in an organized criminal group” column. The result shows that Spain had the highest participation rate in 2022.

```
# Find the country with the highest participation in 2022
max_participation <-
  max(dataset_clean$`Participation in an organized criminal group`, na.rm = TRUE)

country_with_max_participation <-
  dataset_clean$Country[dataset_clean$`Participation in an organized criminal group` == max_participation]

print(country_with_max_participation)
```

```
[1] "Spain"
```

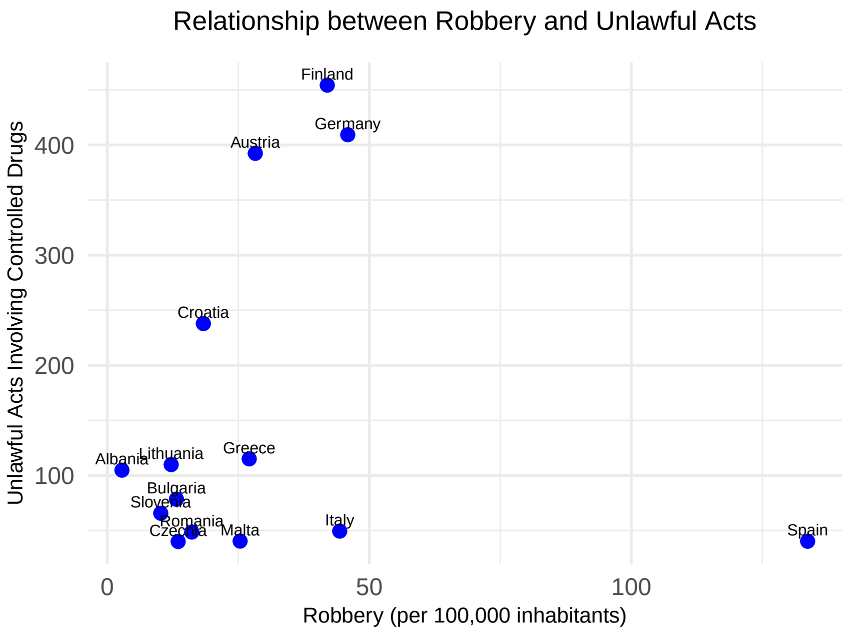
Question 3

A scatter plot was created to display the relationship between robbery and unlawful acts involving controlled drugs. Each point represents a country, with country names labeled for clarity. The plot includes clear axis labels and a centered title, making it easy to interpret the relationship between the two variables.

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
ggplot(dataset_clean,
  aes(x = `Robbery`, y = `Unlawful acts involving controlled drugs or precursors`, label = Country)) +
  geom_point(color = "blue", size = 2) +
  geom_text(aes(label = Country), hjust = 0.5, vjust = -0.5, size = 2) +
  labs(
    title = "Relationship between Robbery and Unlawful Acts",
    x = "Robbery (per 100,000 inhabitants)",
    y = "Unlawful Acts Involving Controlled Drugs"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, hjust = 0.5, margin = margin(t = 10, b = 10)),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8)
  )
```

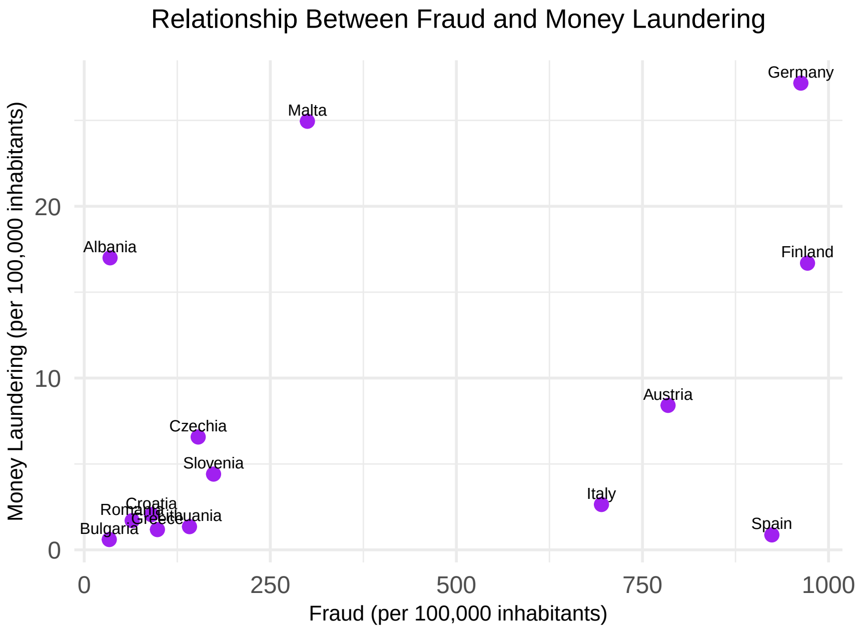


Task 3 : Creativity

Plot 1: Relationship Between Fraud and Money Laundering

This plot will show the correlation between fraud and money laundering. Both are financial crimes, so it could be interesting to see if countries with high levels of fraud also report high levels of money laundering.

```r
ggplot(dataset_clean, aes(x = Fraud, y = `Money laundering`, label = Country)) +
  geom_point(color = "purple", size = 2) +
  geom_text(aes(label = Country), hjust = 0.5, vjust = -0.5, size = 2) +
  labs(
    title = "Relationship Between Fraud and Money Laundering",
    x = "Fraud (per 100,000 inhabitants)",
    y = "Money Laundering (per 100,000 inhabitants)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, hjust = 0.5, margin = margin(t = 10, b = 10)),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8)
  )
```



The scatter plot reveals an interesting relationship between fraud and money laundering across different countries. Notably, Malta has a high rate of money laundering despite relatively low fraud rates. On the other hand, countries like Germany and Finland exhibit both high fraud and moderate money laundering rates. Countries such as Italy, Spain, and Austria display high fraud rates with lower levels of money laundering. The overall trend suggests that while there is some correlation between fraud and money laundering, the relationship varies significantly across countries, with some countries showing high rates of one crime but not the other.

Plot 2: Crime Composition Breakdown (Stacked Bar Plot)

This plot shows the breakdown of different types of crimes, such as fraud, money laundering, robbery, serious assault, and corruption, across several countries. By visualizing the composition of crimes in each country, it becomes easier to identify which types of crime are most prevalent in different regions and how the overall crime rates vary between countries.

```
library(ggplot2)
library(tidyr)
```

```
Warning: package 'tidyr' was built under R version 4.3.2
```

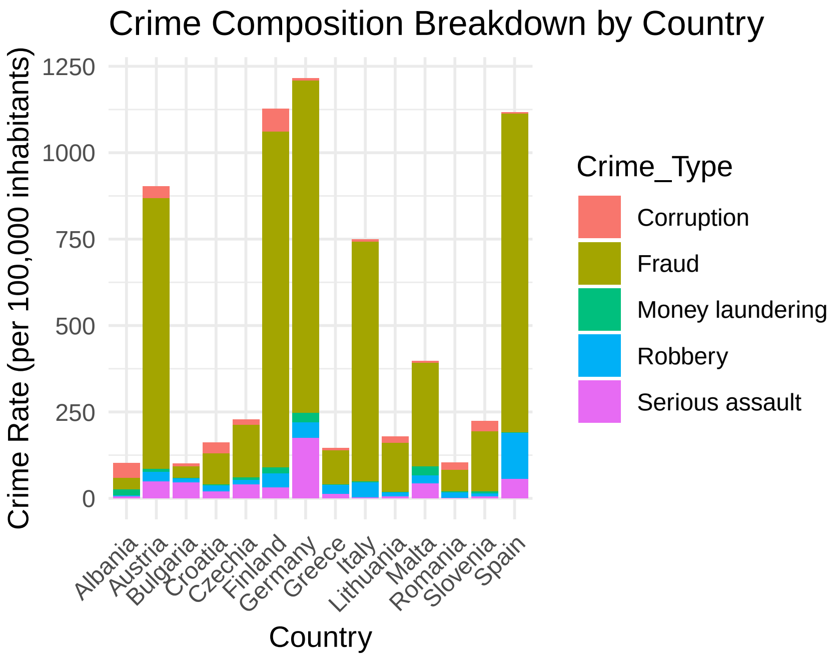
```
library(dplyr)
```

```
Warning: package 'dplyr' was built under R version 4.3.3
```

```
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
```

```
crime_data <- dataset_clean %>%
  select(Country, Robbery, Fraud, `Money laundering`, `Serious assault`, Corruption)


crime_data_long <- pivot_longer(crime_data,
                                cols = -Country,
                                names_to = "Crime_Type",
                                values_to = "Crime_Rate")

# Create a stacked bar plot
ggplot(crime_data_long, aes(x = Country, y = Crime_Rate, fill = Crime_Type)) +
  geom_bar(stat = "identity") +  # Create stacked bars
  labs(
    title = "Crime Composition Breakdown by Country",
    x = "Country",
    y = "Crime Rate (per 100,000 inhabitants)"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1))
```



This stacked bar plot shows the types of crimes across different countries. Germany and Spain have the highest overall crime rates, with fraud being the most common crime. In countries like Finland and Austria, fraud also makes up a large part of the total crime. Other countries, like Romania and Slovenia, have lower overall crime rates and a more balanced distribution of crimes. Corruption is a small part of the crime in most countries, except for Greece, where it's more noticeable.