Assignment 1 - Solutions

Isabella Gollini

Task 1: Manipulation

The dataset `EurostatCrime2015.csv` records offences (values per hundred thousand inhabitants) by offence category in 41 European Countries in 2015.

1.1:

Load the dataset. Use `row.names = 1` as an argument to set country names as row names instead of a column of data.

```
crime2015 <- read.csv("EurostatCrime2015.csv", header = TRUE, row.names = 1)
```

1.2:

The size and structure of the dataset is given by:

```
dim(crime2015)
```

```
## [1] 41  7
```

```
str(crime2015)
```

```
## 'data.frame':    41 obs. of  7 variables:
##  $ Assault             : num  NA 40.4 603.3 NA 35 ...
##  $ Intentional.homicide: num  NA 0.49 1.96 NA 1.79 0.88 1.42 0.8 0.81 NA ...
##  $ Rape                : num  NA 13.18 25.5 NA 1.65 ...
##  $ Robbery             : num  NA 39.8 196.7 NA 27 ...
##  $ Sexual.assault      : num  NA 27.39 65.92 NA 6.72 ...
##  $ Sexual.violence     : num  NA 40.57 91.42 NA 8.37 ...
##  $ Theft               : num  NA 1587 1660 NA 532 ...
```

From this output, we can see that there are 41 observations (41 countries) with 7 variables. All variables are numeric.

1.3:

1.3.1:

Add a new column called `Sex.crime` which contains the sum of all the crimes that have a sexual component: `Rape`, `Sexual.assault` and `Sexual.violence`.

This is easily done using the following code:

```
crime2015$Sex.crime <- crime2015$Rape +
                       crime2015$Sexual.assault + crime2015$Sexual.violence
```

1.3.2:

Remove the columns `Rape`, `Sexual.assault` and `Sexual.violence`. There are different ways to do this. The easiest is:

```
crime2015$Rape <- NULL
crime2015$Sexual.assault <- NULL
crime2015$Sexual.violence <- NULL
```

Let's look at the structure again to confirm that this worked:

```
str(crime2015)
```

```
## 'data.frame':    41 obs. of  5 variables:
##  $ Assault             : num  NA 40.4 603.3 NA 35 ...
##  $ Intentional.homicide: num  NA 0.49 1.96 NA 1.79 0.88 1.42 0.8 0.81 NA ...
##  $ Robbery             : num  NA 39.8 196.7 NA 27 ...
##  $ Theft               : num  NA 1587 1660 NA 532 ...
##  $ Sex.crime           : num  NA 81.1 182.8 NA 16.7 ...
```

Great! The dataframe now has 5 variables.

1.4:

Work with the dataset you created in question (3ii), and list the countries that contain any missing data.

I'll use the complete.cases() function for this. First, find the rows which *don't* have missing values:

```
vec1 <- complete.cases(crime2015)
```

Then use this logcial vector (or rather its opposite) to print the row names of rows which *have* missing data:

```
rownames(crime2015)[!vec1]
```

```
##  [1] "Albania"
##  [2] "Bosnia and Herzegovina"
##  [3] "England and Wales"
##  [4] "Former Yugoslav Republic of Macedonia, the"
##  [5] "Iceland"
##  [6] "Italy"
##  [7] "Kosovo (under United Nations Security Council Resolution 1244/99)"
##  [8] "Netherlands"
##  [9] "Norway"
## [10] "Scotland"
## [11] "Turkey"
```

1.5:

Remove the countries with missing data from the dataframe.

Using complete.cases() again:

```
crime2015b <- crime2015[complete.cases(crime2015), ]
```

1.6:

What is the size of this new dataframe?

The size is given by:

```
dim(crime2015b)
```

```
## [1] 30  5
```

We can see that there are 30 rows and 5 columns in this reduced dataframe.

Task 2: Analysis

2.1:

According to these data what was the most common crime in Ireland in 2015?

This can be found by:

```
names(which.max(crime2015b['Ireland',]))
```

```
## [1] "Theft"
```

Theft was the most common crime in Ireland in 2015.

2.2:

And the 3 least common crimes in Ireland in 2015 are found by:

```
names(sort(crime2015b['Ireland',])[1:3])
```

```
## [1] "Intentional.homicide" "Robbery"              "Sex.crime"
```

Intentional.homicide, Robbery, and Sex.crime are the least common crimes in Ireland.

2.3:

Which country has the highest record of offences (per hundred thousand inhabitants)?

This can be found by:

```
# Summing across rows:
crime2015b$Total <- apply(crime2015b, 1, sum)

# Picking out the row name which has the maximum Total:
row.names(crime2015b)[which.max(crime2015b$Total)]
```

```
## [1] "Sweden"
```

Sweden has the highest record of offences.

Task 2 - with the original dataframe:

If the original dataframe is used here (before removing missing values and removing columns), then:

```
# Reading in the dataset again:
crime2015 <- read.csv("EurostatCrime2015.csv", header = TRUE, row.names = 1)

# 2.1:
names(which.max(crime2015['Ireland',]))
```

```
## [1] "Theft"
```

```
# 2.2:
names(sort(crime2015['Ireland',])[1:3])
```

```
## [1] "Intentional.homicide" "Rape"                 "Sexual.assault"
```

```r
# 2.3:
# Summing across rows:
crime2015$Total <- apply(crime2015, 1, sum)

# Picking out the row name which has the maximum Total:
row.names(crime2015)[which.max(crime2015$Total)]
```

```
## [1] "Sweden"
```

Only the answer to 2.2 changes, and in this case: Intentional.homicide, Rape, and Sexual.assault are the three least common crimes in Ireland.

Task 3: Creativity

This task is up to you!