M2 – FA1 Data Cleaning

For this assessment, the researchers were tasked to download a dataset involving children and run its accompanying R script. They are then asked to observe the data cleaning process.

I. Running the R Script

Unlike the previous R script from M1-FA2, the R script for this assessment ran successfully without any errors, as seen in the code snippet below.

```
> library(data.table)
> setwd('C:/Users/User/Documents/Mapua/Third Year - 3rd Term/CS174 BM2 DATA SCIENCE 4/Submissions/M2-FA1')
>
> children.dt <- fread('children.csv')
> children.dt
    children Room
 1:        1    2
 2:        3    3
 3:        2   na
 4:        0    .
 5:        0    3
 6:     <NA>    4
 7:  missing    4
 8:      N/A    4
 9:        m    3
10:        M    2
11:             4
12:      -99    2
13:        4    3
14:        1    3
> sum(is.na(children.dt))    ## only 1 NA in dataset
[1] 1
> which(is.na(children.dt))  ## That one NA is in row 6 [i.e. value coded: NA]
[1] 6
> ## Source data has 9 different codes for missing value but only one code is auto-recognized by R.
>
>
> # Use na.strings to define all human codes for missing values to be NA.
> children2.dt <- fread('children.csv', na.strings = c("NA", "missing", "N/A", -99, "", "m", "M", "na", "."))
> children2.dt  ## All the 9 ways to code missing value are now recoded as NA.
    children Room
 1:        1    2
 2:        3    3
 3:        2   NA
 4:        0   NA
 5:        0    3
 6:       NA    4
 7:       NA    4
 8:       NA    4
 9:       NA    3
10:       NA    2
11:       NA    4
12:       NA    2
13:        4    3
14:        1    3
> sum(is.na(children2.dt))    ## 9 NAs in dataset
[1] 9
> which(is.na(children2.dt))
[1]  6  7  8  9 10 11 12 17 18
> which(is.na(children2.dt$children))  # where are the NAs in children column.
[1]  6  7  8  9 10 11 12
> which(is.na(children2.dt$Room))      # where are the NAs in Room column.
[1] 3 4
```

II. Determining the number of "NA" on the output of is.na(children.dt))

In the R script, the *sum(is.na(children.dt))* line determines the number of NA in the dataset by R. It uses two functions from R, *sum()* and *is.na()*. The *sum()* function in R determines the sum of elements (Prajwal, 2022), and the *is.na()* function handles missing values in the dataset or data frame (Bajwa, n.d.). The number of "NA" that was determined in the dataset by R is only **one (1)**.

```
> sum(is.na(children.dt))     ## only 1 NA in dataset
[1] 1
```

III. Determining the rows that have "NA" on the output

The *which(is.na(children.dt))* line determines which row the "NA" was located. The line used the *which()* and is.na functions. The precise indexes of NA values were extracted using the *which()* function (Bajwa, n.d.). In this case, the only "NA" was found to be in row 6, as seen in the data table below.

```
> which(is.na(children.dt))  ## That one NA is in row 6 [i.e. value coded: NA]
[1] 6
```

```
> children.dt
    children Room
 1:        1    2
 2:        3    3
 3:        2   na
 4:        0    .
 5:        0    3
 6:     <NA>    4
 7:  missing    4
 8:      N/A    4
 9:        m    3
10:        M    2
11:             4
12:      -99    2
13:        4    3
14:        1    3
```

IV. Printing the newly cleansed output with "NA"

The code snippet below was able to clean the data table using *na.strings()*, which was used to match strings that should be replaced with "NA". The following strings were replaced with NA: “NA”, “missing”, “N/A”, “-99”, “”. “m”, “M”, “na”, and “.”. Overall, there were **nine (9)** values that were replaced with NA.

```
> # Use na.strings to define all human codes for missing values to be NA.
> children2.dt <- fread('children.csv', na.strings = c("NA", "missing", "N/A", -99, "", "m", "M", "na", "."))
> children2.dt  ## All the 9 ways to code missing value are now recoded as NA.
    children Room
 1:        1    2
 2:        3    3
 3:        2   NA
 4:        0   NA
 5:        0    3
 6:       NA    4
 7:       NA    4
 8:       NA    4
 9:       NA    3
10:       NA    2
11:       NA    4
12:       NA    2
13:        4    3
14:        1    3
```

V. Determining which rows that "NAs" appear in the "Children" column

The number of "NAs" in the "Children" column was determined from **rows 6, 7, 8, 9, 10, 11, and 12**. Therefore, **seven (7) "NAs"** are in the "Children" column.

```
> which(is.na(children2.dt$children))  # where are the NAs in children column.
[1]  6  7  8  9 10 11 12
```

VI. Determining which rows that "NAs" appear in the "Room" column

The number of "NAs" in the "Room" column was determined from **rows 3 and 4**. Therefore, **two (2) "NAs"** are in the "Room" column.

```
> which(is.na(children2.dt$Room))      # where are the NAs in Room column.
[1] 3 4
```

References

Bajwa, A. (n.d.). *What is is.na() function in R?* Educative: Interactive Courses for Software Developers. https://www.educative.io/answers/what-is-isna-function-in-r

na.strings = c() in R. (n.d.). Stack Overflow. https://stackoverflow.com/questions/45765944/na-strings-c-in-r

Prajwal, C. (2022, August 3). *How to use sum() in R - Find the sum of elements in R*. DigitalOcean. https://www.digitalocean.com/community/tutorials/sum-in-r