

Introduction to R and RStudio (Session 1)

Call all R packages (libraries) you will need

Note: You will need to do this every R session

```
library('readxl')
```

Set up your working directory where you will put files

```
setwd('C:/Users/mfazz/OneDrive/Desktop/R intensive course/Workshop 1')
```

Now read in my EXCEL file called example_data and save it as a data frame called ex11

This data frame will be stored in my working directory I specified

Note: <- is the R assignment operator

```
ex11 <- as.data.frame(  
  read_excel("C:/Users/mfazz/OneDrive/Desktop/R intensive course/Workshop 1/example_data.xlsx"))
```

Look at the data sets available in your working directory

```
ls()
```

```
## [1] "ex11"
```

Examine how ex11 is stored in R

```
class(ex11)
```

```
## [1] "data.frame"
```

Print out the first few rows, the dimension (rows x columns), and detail about the columns (variables)

```
head(ex11)
```

```
##   ID age  BMI HR      race      date treated  
## 1  1  30 22.5 70   white 2024-03-14        1  
## 2  2  45 30.1 NA    white 2022-05-01        1  
## 3  3  21 20.8 56   black 2023-07-10        0  
## 4  4   NA 26.2 61 hispanic 2022-07-16        0  
## 5  5  33 24.9 73    other 2020-02-02        1
```

```
dim(ex11)
```

```
## [1] 5 7
```

```
str(ex11)
```

```
## 'data.frame': 5 obs. of 7 variables:
## $ ID : num 1 2 3 4 5
## $ age : num 30 45 21 NA 33
## $ BMI : num 22.5 30.1 20.8 26.2 24.9
## $ HR : num 70 NA 56 61 73
## $ race : chr "white" "white" "black" "hispanic" ...
## $ date : POSIXct, format: "2024-03-14" "2022-05-01" ...
## $ treated: num 1 1 0 0 1
```

Extract only the age column

```
ex11['age']
```

```
## age
## 1 30
## 2 45
## 3 21
## 4 NA
## 5 33
```

```
ex11[2]
```

```
## age
## 1 30
## 2 45
## 3 21
## 4 NA
## 5 33
```

```
ex11$age
```

```
## [1] 30 45 21 NA 33
```

Create a FACTOR for treated with labels

```
ex11$treatedf <- factor(ex11$treated, levels=c(0,1), labels=c('yes','no'))
ex11
```

```
## ID age BMI HR race date treated treatedf
## 1 1 30 22.5 70 white 2024-03-14 1 no
## 2 2 45 30.1 NA white 2022-05-01 1 no
## 3 3 21 20.8 56 black 2023-07-10 0 yes
## 4 4 NA 26.2 61 hispanic 2022-07-16 0 yes
## 5 5 33 24.9 73 other 2020-02-02 1 no
```

```
str(ex11)
```

```
## 'data.frame': 5 obs. of 8 variables:
## $ ID : num 1 2 3 4 5
## $ age : num 30 45 21 NA 33
## $ BMI : num 22.5 30.1 20.8 26.2 24.9
## $ HR : num 70 NA 56 61 73
## $ race : chr "white" "white" "black" "hispanic" ...
## $ date : POSIXct, format: "2024-03-14" "2022-05-01" ...
## $ treated : num 1 1 0 0 1
## $ treatedf: Factor w/ 2 levels "yes","no": 2 2 1 1 2
```

Create a new column in ex11 that indicates whether the patient is older than 50

```
ex11$age_cat_50 <- ifelse(ex11$age >= 50,1,0)
ex11
```

```
## ID age BMI HR race date treated treatedf age_cat_50
## 1 1 30 22.5 70 white 2024-03-14 1 no 0
## 2 2 45 30.1 NA white 2022-05-01 1 no 0
## 3 3 21 20.8 56 black 2023-07-10 0 yes 0
## 4 4 NA 26.2 61 hispanic 2022-07-16 0 yes NA
## 5 5 33 24.9 73 other 2020-02-02 1 no 0
```

Create a new column in ex11 that indicates whether the patient reported race as Black.

Note: == is the logical question of whether the value equals “black”, we do not use = because that is an assignment operator

```
ex11$black<- ifelse(ex11$race == "black",1,0)
ex11
```

```
## ID age BMI HR race date treated treatedf age_cat_50 black
## 1 1 30 22.5 70 white 2024-03-14 1 no 0 0
## 2 2 45 30.1 NA white 2022-05-01 1 no 0 0
## 3 3 21 20.8 56 black 2023-07-10 0 yes 0 1
## 4 4 NA 26.2 61 hispanic 2022-07-16 0 yes NA 0
## 5 5 33 24.9 73 other 2020-02-02 1 no 0 0
```

```
ex11$notblack<- ifelse(ex11$race != "black",1,0)
ex11
```

```
## ID age BMI HR race date treated treatedf age_cat_50 black notblack
## 1 1 30 22.5 70 white 2024-03-14 1 no 0 0 1
## 2 2 45 30.1 NA white 2022-05-01 1 no 0 0 1
## 3 3 21 20.8 56 black 2023-07-10 0 yes 0 1 0
## 4 4 NA 26.2 61 hispanic 2022-07-16 0 yes NA 0 1
## 5 5 33 24.9 73 other 2020-02-02 1 no 0 0 1
```

Create a new column in ex11 that indicates whether the patient reported race as Black.

```
ex11$black<- ifelse(ex11$race == "black",1,0)
ex11
```

```
##   ID age  BMI HR    race      date treated treatedf age_cat_50 black notblack
## 1  1  30 22.5 70   white 2024-03-14      1      no          0      0        1
## 2  2  45 30.1 NA   white 2022-05-01      1      no          0      0        1
## 3  3  21 20.8 56   black 2023-07-10      0     yes          0      1        0
## 4  4   NA 26.2 61 hispanic 2022-07-16      0     yes         NA      0        1
## 5  5  33 24.9 73   other 2020-02-02      1      no          0      0        1
```

Create a new column called age_centered in ex11 that subtracts the mean age

```
ex11$age_centered <- ex11$age - mean(ex11$age,na.rm=TRUE)
ex11
```

```
##   ID age  BMI HR    race      date treated treatedf age_cat_50 black notblack
## 1  1  30 22.5 70   white 2024-03-14      1      no          0      0        1
## 2  2  45 30.1 NA   white 2022-05-01      1      no          0      0        1
## 3  3  21 20.8 56   black 2023-07-10      0     yes          0      1        0
## 4  4   NA 26.2 61 hispanic 2022-07-16      0     yes         NA      0        1
## 5  5  33 24.9 73   other 2020-02-02      1      no          0      0        1
##   age_centered
## 1         -2.25
## 2         12.75
## 3        -11.25
## 4             NA
## 5          0.75
```

Subsetting data from the ex11 data frame Creates a new data frame of people younger than 50 using the subset function

```
younger <- subset(ex11, age<50)
younger
```

```
##   ID age  BMI HR    race      date treated treatedf age_cat_50 black notblack
## 1  1  30 22.5 70   white 2024-03-14      1      no          0      0        1
## 2  2  45 30.1 NA   white 2022-05-01      1      no          0      0        1
## 3  3  21 20.8 56   black 2023-07-10      0     yes          0      1        0
## 5  5  33 24.9 73   other 2020-02-02      1      no          0      0        1
##   age_centered
## 1         -2.25
## 2         12.75
## 3        -11.25
## 5          0.75
```

Creates a new data frame of people younger than 50 retaining columns 1,2 and 3

```
younger <- subset(ex11, age<50, select=c(1,2,3))
younger
```

```
##   ID age  BMI
## 1  1  30 22.5
## 2  2  45 30.1
## 3  3  21 20.8
## 5  5  33 24.9
```

Create a new data frame with only people with BMI > 30, keeping all columns

```
high_BMI <- ex11[ex11$BMI > 30,]
high_BMI2 <- subset(ex11, BMI > 30)
high_BMI
```

```
##   ID age  BMI HR  race      date treated treatedf age_cat_50 black notblack
## 2  2  45 30.1 NA  white 2022-05-01      1      no          0      0          1
##   age_centered
## 2             12.75
```

Sort data by age

```
sorted.byage=ex11[order(ex11$age),]
sorted.byage
```

```
##   ID age  BMI HR  race      date treated treatedf age_cat_50 black notblack
## 3  3  21 20.8 56  black 2023-07-10      0      yes          0      1          0
## 1  1  30 22.5 70  white 2024-03-14      1      no          0      0          1
## 5  5  33 24.9 73  other 2020-02-02      1      no          0      0          1
## 2  2  45 30.1 NA  white 2022-05-01      1      no          0      0          1
## 4  4  NA 26.2 61 hispanic 2022-07-16      0      yes         NA      0          1
##   age_centered
## 3             -11.25
## 1              -2.25
## 5               0.75
## 2              12.75
## 4               NA
```

Descending order

```
sorted.byage=ex11[order(-ex11$age),]
sorted.byage
```

```
##   ID age  BMI HR  race      date treated treatedf age_cat_50 black notblack
## 2  2  45 30.1 NA  white 2022-05-01      1      no          0      0          1
## 5  5  33 24.9 73  other 2020-02-02      1      no          0      0          1
## 1  1  30 22.5 70  white 2024-03-14      1      no          0      0          1
## 3  3  21 20.8 56  black 2023-07-10      0      yes          0      1          0
## 4  4  NA 26.2 61 hispanic 2022-07-16      0      yes         NA      0          1
##   age_centered
## 2              12.75
## 5               0.75
## 1              -2.25
## 3             -11.25
## 4               NA
```

Examine missingness in the ex11 data frame

```
ex11.complete <- ex11[complete.cases(ex11),]  
ex11.complete
```

```
##   ID age  BMI HR  race      date treated treatedf age_cat_50 black notblack  
## 1  1  30 22.5 70 white 2024-03-14      1      no          0      0          1  
## 3  3  21 20.8 56 black 2023-07-10      0     yes          0      1          0  
## 5  5  33 24.9 73 other 2020-02-02      1      no          0      0          1  
##   age_centered  
## 1           -2.25  
## 3           -11.25  
## 5            0.75
```

Alternative approach

```
ex11.complete <- na.exclude(ex11)  
ex11.complete
```

```
##   ID age  BMI HR  race      date treated treatedf age_cat_50 black notblack  
## 1  1  30 22.5 70 white 2024-03-14      1      no          0      0          1  
## 3  3  21 20.8 56 black 2023-07-10      0     yes          0      1          0  
## 5  5  33 24.9 73 other 2020-02-02      1      no          0      0          1  
##   age_centered  
## 1           -2.25  
## 3           -11.25  
## 5            0.75
```

Complete data, but only include only specific columns

```
ex11.complete2 <- na.exclude(ex11[,c(2:4)])  
ex11.complete2
```

```
##   age  BMI HR  
## 1  30 22.5 70  
## 3  21 20.8 56  
## 5  33 24.9 73
```

All patients with non-missing age

```
ex11.complete <- ex11[complete.cases(ex11[, 'age']),]  
ex11.complete
```

```
##   ID age  BMI HR  race      date treated treatedf age_cat_50 black notblack  
## 1  1  30 22.5 70 white 2024-03-14      1      no          0      0          1  
## 2  2  45 30.1 NA white 2022-05-01      1      no          0      0          1  
## 3  3  21 20.8 56 black 2023-07-10      0     yes          0      1          0  
## 5  5  33 24.9 73 other 2020-02-02      1      no          0      0          1  
##   age_centered  
## 1           -2.25  
## 2           12.75  
## 3           -11.25  
## 5            0.75
```

All patients with missing age

```
ex11.missing_age <- ex11[!complete.cases(ex11[, 'age']),]  
ex11.missing_age
```

```
##   ID age  BMI HR      race      date treated treatedf age_cat_50 black notblack  
## 4   4  NA 26.2 61 hispanic 2022-07-16      0      yes      NA      0      1  
##   age_centered  
## 4              NA
```