

## Introduction to R and RStudio (Session 1, Practice 1)

Read in my COMMA DELIMITED file called example\_data and save it as a data frame called practice1

```
practice1 <- as.data.frame(read.table("practice1.csv", sep=",", header=TRUE))
```

Where is your working directory?

```
getwd()
```

```
## [1] "C:/Users/mfazzari/OneDrive - Montefiore Medicine/R TRAINING WORKSHOPS/R WORKSHOP 1 MATERIALS"
```

Examine how practice1 is stored in R

```
class(practice1)
```

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

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

```
head(practice1)
```

```
##   ID age weight_kg height_m male_sex
## 1  1  45    98.8    1.80      1
## 2  2  67    75.3    1.71      1
## 3  3  32   103.2    1.80      1
## 4  4  29    78.0    1.55      1
## 5  5  41    95.6    1.90      1
## 6  6  44    88.1    1.78      1
```

```
dim(practice1)
```

```
## [1] 20  5
```

```
str(practice1)
```

```
## 'data.frame':  20 obs. of  5 variables:
## $ ID      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ age     : int  45 67 32 29 41 44 66 78 75 61 ...
## $ weight_kg: num  98.8 75.3 103.2 78 95.6 ...
## $ height_m : num  1.8 1.71 1.8 1.55 1.9 1.78 1.88 1.77 1.9 1.74 ...
## $ male_sex : int  1 1 1 1 1 1 1 1 1 1 ...
```

Create a new variable age.cat as part of the practice1 data frame that is equal to 0 if patient is younger than 30 or 1 if patient is 30+

```
practice1$age.cat <- ifelse(practice1$age >= 30,1,0)
practice1
```

##	ID	age	weight_kg	height_m	male_sex	age.cat
## 1	1	45	98.8	1.80	1	1
## 2	2	67	75.3	1.71	1	1
## 3	3	32	103.2	1.80	1	1
## 4	4	29	78.0	1.55	1	0
## 5	5	41	95.6	1.90	1	1
## 6	6	44	88.1	1.78	1	1
## 7	7	66	84.0	1.88	1	1
## 8	8	78	NA	1.77	1	1
## 9	9	75	120.0	1.90	1	1
## 10	10	61	94.6	1.74	1	1
## 11	11	60	68.9	1.60	0	1
## 12	12	58	67.0	1.67	0	1
## 13	13	33	NA	1.70	0	1
## 14	14	32	NA	NA	0	1
## 15	15	38	62.9	1.63	0	1
## 16	16	46	75.6	1.64	0	1
## 17	17	60	101.1	1.81	0	1
## 18	18	59	79.9	1.75	0	1
## 19	19	71	83.3	1.73	0	1
## 20	20	74	96.8	1.67	0	1

Create a new variable age.cat.out (not part of practice1) that is equal to 0 if patient is younger than 50 or 1 if patient is 50+

```
age.cat.out <- ifelse(practice1$age >= 50,1,0)
practice1
```

##	ID	age	weight_kg	height_m	male_sex	age.cat
## 1	1	45	98.8	1.80	1	1
## 2	2	67	75.3	1.71	1	1
## 3	3	32	103.2	1.80	1	1
## 4	4	29	78.0	1.55	1	0
## 5	5	41	95.6	1.90	1	1
## 6	6	44	88.1	1.78	1	1
## 7	7	66	84.0	1.88	1	1
## 8	8	78	NA	1.77	1	1
## 9	9	75	120.0	1.90	1	1
## 10	10	61	94.6	1.74	1	1
## 11	11	60	68.9	1.60	0	1
## 12	12	58	67.0	1.67	0	1
## 13	13	33	NA	1.70	0	1
## 14	14	32	NA	NA	0	1
## 15	15	38	62.9	1.63	0	1
## 16	16	46	75.6	1.64	0	1
## 17	17	60	101.1	1.81	0	1
## 18	18	59	79.9	1.75	0	1
## 19	19	71	83.3	1.73	0	1
## 20	20	74	96.8	1.67	0	1

Create a new data frame called subset1 that consists of only those patients who do not have missing values for height (in meters) or weight (in kg). How many people are in this data set?

```
subset1 <- practice1[complete.cases(practice1[,c('height_m','weight_kg'))],]  
dim(subset1)
```

```
## [1] 17 6
```

```
subset1
```

```
##      ID age weight_kg height_m male_sex age.cat  
## 1    1  45    98.8    1.80        1        1  
## 2    2  67    75.3    1.71        1        1  
## 3    3  32   103.2    1.80        1        1  
## 4    4  29    78.0    1.55        1        0  
## 5    5  41    95.6    1.90        1        1  
## 6    6  44    88.1    1.78        1        1  
## 7    7  66    84.0    1.88        1        1  
## 9    9  75   120.0    1.90        1        1  
## 10  10  61    94.6    1.74        1        1  
## 11  11  60    68.9    1.60        0        1  
## 12  12  58    67.0    1.67        0        1  
## 15  15  38    62.9    1.63        0        1  
## 16  16  46    75.6    1.64        0        1  
## 17  17  60   101.1    1.81        0        1  
## 18  18  59    79.9    1.75        0        1  
## 19  19  71    83.3    1.73        0        1  
## 20  20  74    96.8    1.67        0        1
```

Create a new variable in subset1 called BMI by computing BMI (weight/ height<sup>2</sup>)

```
subset1$BMI <- subset1$weight_kg/(subset1$height_m*subset1$height_m)  
practice1
```

```
##      ID age weight_kg height_m male_sex age.cat  
## 1    1  45    98.8    1.80        1        1  
## 2    2  67    75.3    1.71        1        1  
## 3    3  32   103.2    1.80        1        1  
## 4    4  29    78.0    1.55        1        0  
## 5    5  41    95.6    1.90        1        1  
## 6    6  44    88.1    1.78        1        1  
## 7    7  66    84.0    1.88        1        1  
## 8    8  78     NA    1.77        1        1  
## 9    9  75   120.0    1.90        1        1  
## 10  10  61    94.6    1.74        1        1  
## 11  11  60    68.9    1.60        0        1  
## 12  12  58    67.0    1.67        0        1  
## 13  13  33     NA    1.70        0        1  
## 14  14  32     NA     NA        0        1  
## 15  15  38    62.9    1.63        0        1  
## 16  16  46    75.6    1.64        0        1  
## 17  17  60   101.1    1.81        0        1
```

```
## 18 18 59      79.9      1.75      0      1
## 19 19 71      83.3      1.73      0      1
## 20 20 74      96.8      1.67      0      1
```

Create a factor variable for biological sex with values 'Male' and 'Female'

Create a FACTOR for treated with labels

```
subset1$sexf <- factor(subset1$male_sex,levels=c(0,1), labels=c('Female','Male'))
subset1
```

```
##      ID age weight_kg height_m male_sex age.cat      BMI      sexf
## 1  1  45      98.8      1.80      1      1 30.49383    Male
## 2  2  67      75.3      1.71      1      1 25.75151    Male
## 3  3  32     103.2      1.80      1      1 31.85185    Male
## 4  4  29      78.0      1.55      1      0 32.46618    Male
## 5  5  41      95.6      1.90      1      1 26.48199    Male
## 6  6  44      88.1      1.78      1      1 27.80583    Male
## 7  7  66      84.0      1.88      1      1 23.76641    Male
## 9  9  75     120.0      1.90      1      1 33.24100    Male
## 10 10 61      94.6      1.74      1      1 31.24587    Male
## 11 11 60      68.9      1.60      0      1 26.91406 Female
## 12 12 58      67.0      1.67      0      1 24.02381 Female
## 15 15 38      62.9      1.63      0      1 23.67421 Female
## 16 16 46      75.6      1.64      0      1 28.10827 Female
## 17 17 60     101.1      1.81      0      1 30.85986 Female
## 18 18 59      79.9      1.75      0      1 26.08980 Female
## 19 19 71      83.3      1.73      0      1 27.83254 Female
## 20 20 74      96.8      1.67      0      1 34.70903 Female
```

```
str(subset1)
```

```
## 'data.frame':  17 obs. of  8 variables:
## $ ID      : int  1 2 3 4 5 6 7 9 10 11 ...
## $ age      : int  45 67 32 29 41 44 66 75 61 60 ...
## $ weight_kg: num  98.8 75.3 103.2 78 95.6 ...
## $ height_m : num  1.8 1.71 1.8 1.55 1.9 1.78 1.88 1.9 1.74 1.6 ...
## $ male_sex : int  1 1 1 1 1 1 1 1 1 0 ...
## $ age.cat  : num  1 1 1 0 1 1 1 1 1 1 ...
## $ BMI      : num  30.5 25.8 31.9 32.5 26.5 ...
## $ sexf     : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
```

Create a categorical variable called BMI\_cat that assigns a value of 1 if BMI < 25, 2 if 25 <= BMI < 30, and 3 is BMI >= 30.

Then create a factor variable with values "Normal", "Overweight" and "Obese"

```
subset1$BMI_cat <- ifelse(subset1$BMI > 30,2,ifelse(subset1$BMI > 25,1,0))
subset1$BMI_cat <- factor(subset1$BMI_cat,levels=c(0,1,2),
                          labels=c('Normal BMI','Overweight','Obese'))
subset1
```

```
##      ID age weight_kg height_m male_sex age.cat      BMI      sexf      BMI_cat
```

```
## 1 1 45 98.8 1.80 1 1 30.49383 Male Obese
## 2 2 67 75.3 1.71 1 1 25.75151 Male Overweight
## 3 3 32 103.2 1.80 1 1 31.85185 Male Obese
## 4 4 29 78.0 1.55 1 0 32.46618 Male Obese
## 5 5 41 95.6 1.90 1 1 26.48199 Male Overweight
## 6 6 44 88.1 1.78 1 1 27.80583 Male Overweight
## 7 7 66 84.0 1.88 1 1 23.76641 Male Normal BMI
## 9 9 75 120.0 1.90 1 1 33.24100 Male Obese
## 10 10 61 94.6 1.74 1 1 31.24587 Male Obese
## 11 11 60 68.9 1.60 0 1 26.91406 Female Overweight
## 12 12 58 67.0 1.67 0 1 24.02381 Female Normal BMI
## 15 15 38 62.9 1.63 0 1 23.67421 Female Normal BMI
## 16 16 46 75.6 1.64 0 1 28.10827 Female Overweight
## 17 17 60 101.1 1.81 0 1 30.85986 Female Obese
## 18 18 59 79.9 1.75 0 1 26.08980 Female Overweight
## 19 19 71 83.3 1.73 0 1 27.83254 Female Overweight
## 20 20 74 96.8 1.67 0 1 34.70903 Female Obese
```

```
str(subset1)
```

```
## 'data.frame': 17 obs. of 9 variables:
## $ ID : int 1 2 3 4 5 6 7 9 10 11 ...
## $ age : int 45 67 32 29 41 44 66 75 61 60 ...
## $ weight_kg: num 98.8 75.3 103.2 78 95.6 ...
## $ height_m : num 1.8 1.71 1.8 1.55 1.9 1.78 1.88 1.9 1.74 1.6 ...
## $ male_sex : int 1 1 1 1 1 1 1 1 1 0 ...
## $ age.cat : num 1 1 1 0 1 1 1 1 1 1 ...
## $ BMI : num 30.5 25.8 31.9 32.5 26.5 ...
## $ sexf : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...
## $ BMI_cat : Factor w/ 3 levels "Normal BMI","Overweight",...: 3 2 3 3 2 2 1 3 3 2 ...
```

Summarize subset1

```
summary(subset1)
```

```
##      ID      age      weight_kg      height_m
## Min.   : 1.00   Min.   :29.00   Min.   : 62.90   Min.   :1.550
## 1st Qu.: 5.00   1st Qu.:44.00   1st Qu.: 75.60   1st Qu.:1.670
## Median :10.00   Median :59.00   Median : 84.00   Median :1.740
## Mean   :10.29   Mean   :54.47   Mean   : 86.65   Mean   :1.739
## 3rd Qu.:16.00   3rd Qu.:66.00   3rd Qu.: 96.80   3rd Qu.:1.800
## Max.   :20.00   Max.   :75.00   Max.   :120.00   Max.   :1.900
##  male_sex      age.cat      BMI      sexf      BMI_cat
## Min.   :0.0000   Min.   :0.0000   Min.   :23.67   Female:8   Normal BMI:3
## 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:26.09   Male  :9   Overweight:7
## Median :1.0000   Median :1.0000   Median :27.83           Obese    :7
## Mean   :0.5294   Mean   :0.9412   Mean   :28.55
## 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:31.25
## Max.   :1.0000   Max.   :1.0000   Max.   :34.71
```