# Homework part 3

Francesco MONTI

2022-11-17

```
library(stringr)
library(prettyR)
library(ggplot2)
library(ggthemes)
library(colorRamps)

knitr::opts_chunk$set(echo = F, warning = F) # setting chunk options globally
```

### 1) Import the file "BDD_VICAN.csv"

### 2) Display the first lines of the dataset. Display the lines 1; 4; 18; 103 of the dataset

```
##   fc_caisse ms_codcancer fc_agediag_r0 q5_sd4_r1 q5_sd5 q5_sd10_r2 q5_pcs12_r1
## 1         1            1            51         1      1          3    36.17192
## 2         1            1            50         2      1          2    44.73504
## 3         1            1            49         2      2          2    58.32207
## 4         1            7            50         1      1          3    56.38174
## 5         1           10            26         1      1          3    61.12935
## 6         1           81            35         1      1          3    20.59613
##   q5_mcs12_r1 q5_eortc_fatigue_r1 q5_anxiete q5_depression q5_jobv5.36_r1
## 1    32.31987           66,666667          1             0              3
## 2    33.16873           66,666667          2             0              1
## 3    58.63898           22,222222          0             0              2
## 4    51.82826           11,111111          0             0              2
## 5    44.05385           55,555556          1             0              2
## 6    52.33494                 100          0             0              1
##   ms_csp_enq_3c_r1 q5_med23.1 id q5_pain
## 1                3          5  2       0
## 2                1          5  3       0
## 3                1          5  4       0
## 4                2          5  5       0
## 5                1          5  6       1
## 6                2          2  7       1
```

```
##     fc_caisse ms_codcancer fc_agediag_r0 q5_sd4_r1 q5_sd5 q5_sd10_r2
## 1           1            1            51         1      1          3
## 4           1            7            50         1      1          3
## 18          1            1            50         2      1          3
## 103         1            3            51         1      1          3
```

```
##     q5_pcs12_r1 q5_mcs12_r1 q5_eortc_fatigue_r1 q5_anxiete q5_depression
## 1      36.17192    32.31987           66,666667          1             0
## 4      56.38174    51.82826           11,111111          0             0
## 18     40.27254    28.48350           66,666667          2             1
## 103    62.53169    38.25137           22,222222          1             0
##     q5_jobv5.36_r1 ms_csp_enq_3c_r1 q5_med23.1  id q5_pain
## 1                3                3          5   2       0
## 4                2                2          5   5       0
## 18               1                2          3  19       1
## 103              2                2          4 106       0
```

### 3. How many variables and observations are there?

```
## [1] 16
```

```
## [1] 3962
```

### 4. Does this file contain any missing values?

```
##           fc_caisse        ms_codcancer       fc_agediag_r0           q5_sd4_r1
##                   0                   0                   0                   0
##              q5_sd5          q5_sd10_r2         q5_pcs12_r1         q5_mcs12_r1
##                   0                   0                   0                   0
## q5_eortc_fatigue_r1          q5_anxiete       q5_depression      q5_jobv5.36_r1
##                   0                   0                   0                   0
##    ms_csp_enq_3c_r1          q5_med23.1                  id             q5_pain
##                   0                   0                   0                   0
```

```
## [1] 0
```

On a first impression, the dataframe looks to be free of any NAs. We'll see later that this is not true: the variable "q5_eortc_fatigue_r1" as been incorrectly identified as "character" as missing values have been tagged as "!NULL" rather than leaving the cells empty.

As a sidenote, there is no description of "q5_eortc_fatigue_r1" in the statement of the homework

### 5. What is the nature of the variables studied?

```
## 'data.frame':    3962 obs. of  16 variables:
##  $ fc_caisse          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ms_codcancer       : int  1 1 1 7 10 81 10 1 3 1 ...
##  $ fc_agediag_r0      : int  51 50 49 50 26 35 37 47 50 43 ...
##  $ q5_sd4_r1          : int  1 2 2 1 1 1 1 1 1 2 ...
##  $ q5_sd5             : int  1 1 2 1 1 1 1 2 1 2 ...
##  $ q5_sd10_r2         : int  3 2 2 3 3 3 3 3 2 2 ...
##  $ q5_pcs12_r1        : num  36.2 44.7 58.3 56.4 61.1 ...
##  $ q5_mcs12_r1        : num  32.3 33.2 58.6 51.8 44.1 ...
##  $ q5_eortc_fatigue_r1: chr  "66,666667" "66,666667" "22,222222" "11,111111" ...
##  $ q5_anxiete         : int  1 2 0 0 1 0 0 2 0 0 ...
##  $ q5_depression      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ q5_jobv5.36_r1     : int  3 1 2 2 2 1 1 2 2 1 ...
```

```
##  $ ms_csp_enq_3c_r1   : int  3 1 1 2 1 2 2 2 2 1 ...
##  $ q5_med23.1         : int  5 5 5 5 5 2 5 5 1 4 ...
##  $ id                 : int  2 3 4 5 6 7 8 9 10 11 ...
##  $ q5_pain            : int  0 0 0 0 1 1 0 0 1 0 ...
```

**6. Some of the variables are in the wrong format, for example, a qualitative variable in "numeric" format. Based on the description of each variable (found at the beginning of this exercise), recode the variable(s) into the correct format**

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

**7. Definition of clinically significant fatigue score: score >= 40 on the fatigue scale included in the survey, the threshold at which a fatigue condition was shown to be clinically significant. Create a categorical variable based on this definition.**

**8. Group the modalities of the variable sequelae into 3 modalities.**

This new variable, named "Q5_med23.1_rec" will be considered in the following analyses instead of "Q5_med23.1".

**9. Display the frequency table for this new variable.**

```
##   Important sequelae Moderate sequelae No sequelae
## a "907"              "1652"            "1403"
## b "22.89 %"          "41.7 %"          "35.41 %"
```

**10. Concerning age: What is the average age of our study population, then that of breast cancer.**

```
## [1] 54.71858
```

```
## [1] 50.34656
```

**Determine the 95% confidence intervals (CI) for each of the calculated means.**

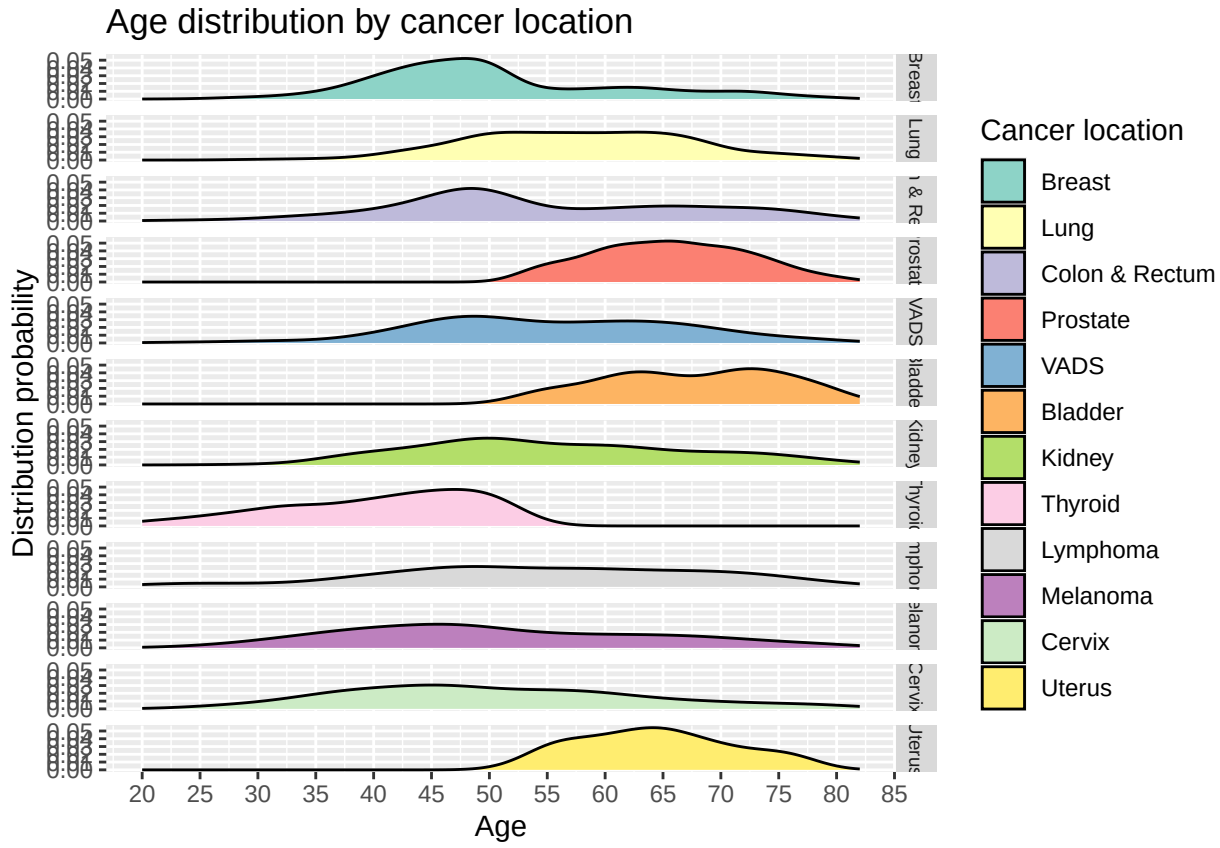
```
## [1] 54.33283 55.10433
## attr(,"conf.level")
## [1] 0.95
```

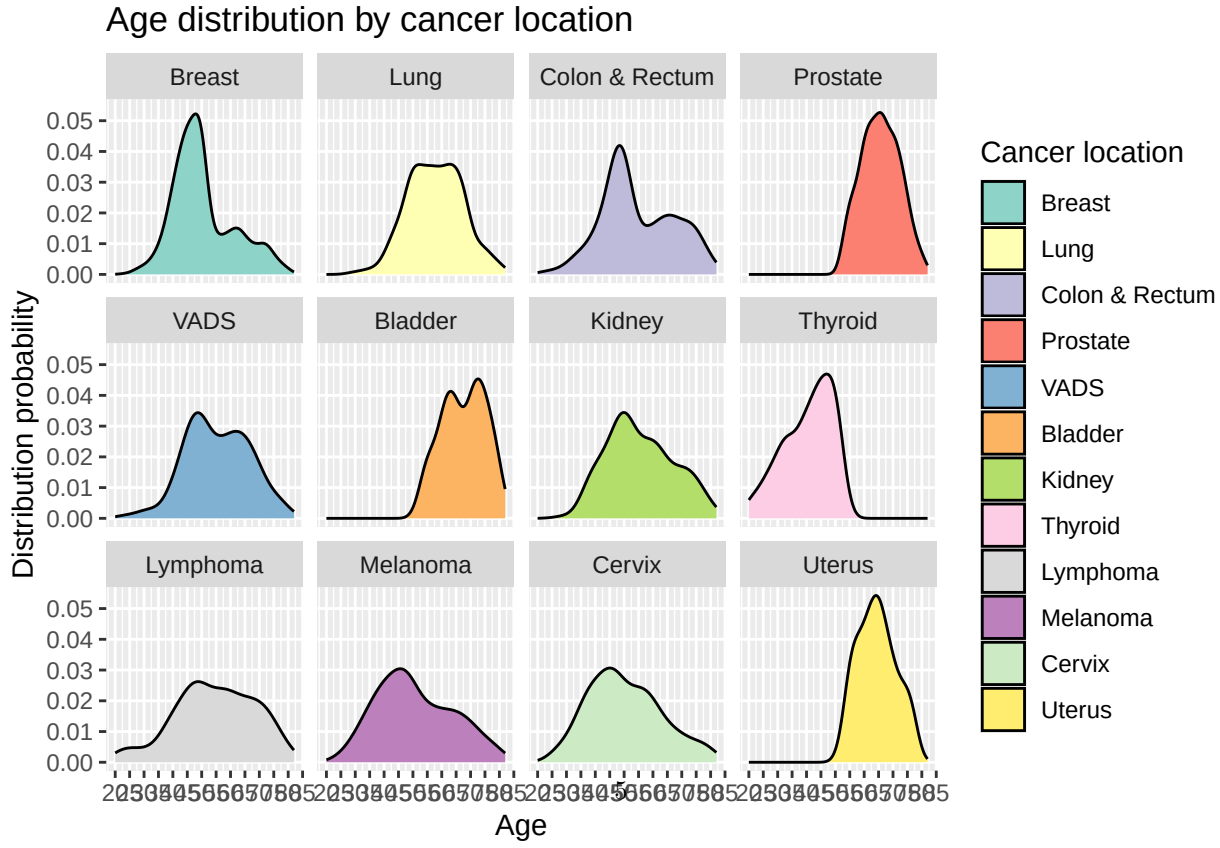
```
## [1] 49.72758 50.96554
## attr(,"conf.level")
## [1] 0.95
```

**Calculate the variance, standard deviation of the sample, then that of breast cancer.**

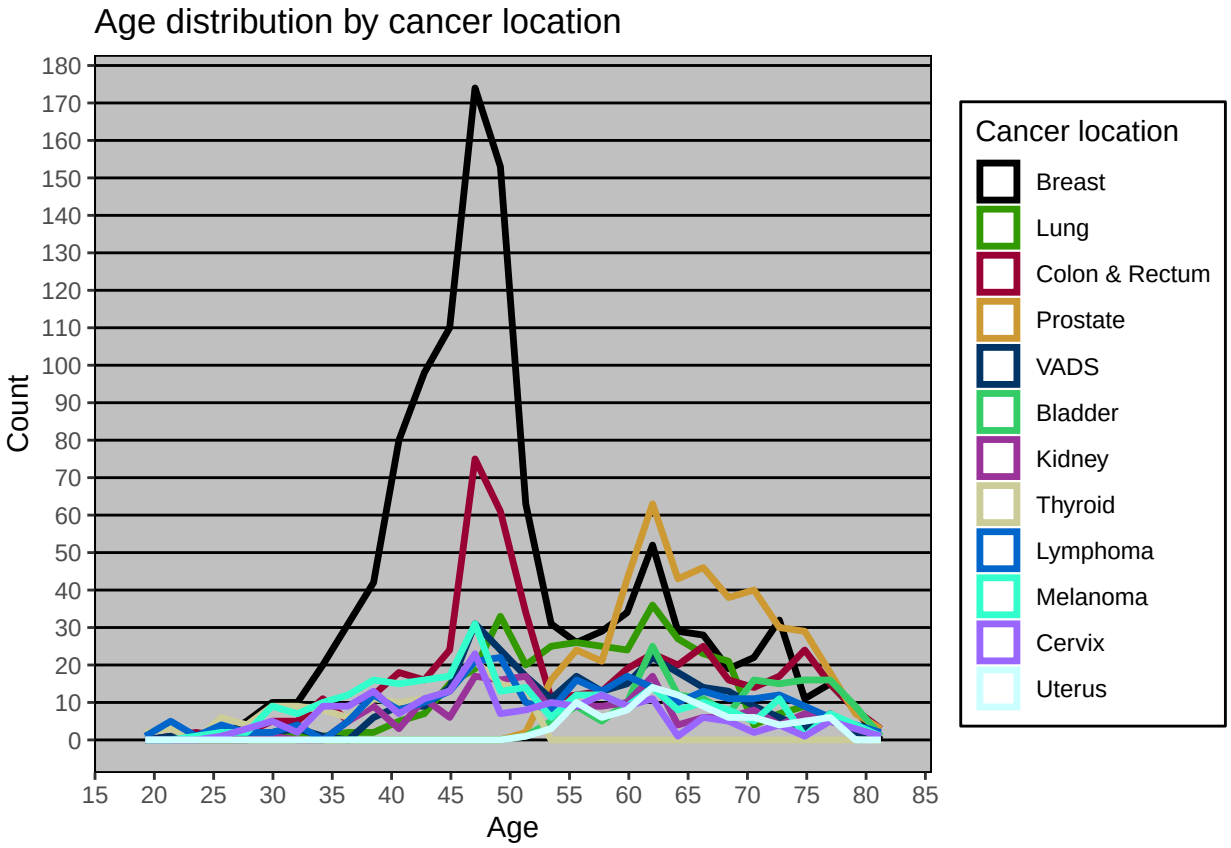
```
## Description of the age variable for the whole population

##
##  Numeric
##      var    sd
## x 153.38 12.38

## Description of the age variable for the breast cancer population

##
##  Numeric
##      var    sd
## x 112.86 10.62
```

**11. Draw a graph that will represent the distribution of age by location of the pathology. Choose the most appropriate graph. Export the graph in a pdf format.**



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

**12. Determine the factors associated with physical and then mental quality of life, including variables with a p-value $< 0.2$. Which model will you use? How will you proceed? Interpret the final result.**

**13. Export the new database in ".csv" format.**