Homework 1 - part 3

Francesco MONTI

2022-11-17

```
library(stringr)
library(prettyR)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v tibble  3.1.8      v purrr   0.3.5
## v tidyr   1.2.1      v dplyr   1.0.10
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggthemes)
library(colorRamps)
library(psych)
```

```
##
## Caricamento pacchetto: 'psych'
##
## I seguenti oggetti sono mascherati da 'package:ggplot2':
##
##     %+%, alpha
##
## I seguenti oggetti sono mascherati da 'package:prettyR':
##
##     describe, skew
```

```
library(Hmisc)
```

```
## Caricamento del pacchetto richiesto: lattice
## Caricamento del pacchetto richiesto: survival
## Caricamento del pacchetto richiesto: Formula
##
## Caricamento pacchetto: 'Hmisc'
##
## Il seguente oggetto è mascherato da 'package:psych':
##
##     describe
##
```

```
## I seguenti oggetti sono mascherati da 'package:dplyr':
##
##     src, summarize
##
## Il seguente oggetto è mascherato da 'package:prettyR':
##
##     describe
##
## I seguenti oggetti sono mascherati da 'package:base':
##
##     format.pval, units
```

```
library(knitr)
```

Nota bene: output will go to the console, as defined in the global options.

1. Import the file "BDD_VICAN.csv"

```
data <- read.csv("BDD_VICAN.csv", sep = ";", dec = ",", encoding = "UTF-8")

names(data) = tolower(names(data)) # removing capital letters as working with them can be annoying
```

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?

```
sapply(data, function(x) sum(is.na(x))) # NAs by variable
```

```
##           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
```

```
sum(is.na(data)) # global NAs
```

```
## [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?

```
str(data)
```

```
## '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), re-code the variable(s) into the correct format

```r
# replacing "," with "." is necessary for as.numeric() to work correctly
data$q5_eortc_fatigue_r1 =
  as.numeric(str_replace(data$q5_eortc_fatigue_r1, ",", "."))
```

```
## Warning: NA introdotti per coercizione
```

```r
sum(is.na(data$q5_eortc_fatigue_r1))  # 6 NAs introduced where cells were "!NULL"
```

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

```r
# ---------------------------------- #
# For several variables it could be appropriate to convert them to factors but, at this stage of the an

# Health insurance
data$fc_caisse = factor(data$fc_caisse, labels = c("CNAMTS", "MSA", "RSI"))

# Pathology location
data$ms_codcancer = factor(data$ms_codcancer,
                           labels = c("Breast", "Lung", "Colon & Rectum",
                                      "Prostate", "VADS", "Bladder", "Kidney",
                                      "Thyroid", "Lymphoma", "Melanoma", "Cervix",
                                      "Uterus"))

# Marital status
data$q5_sd4_r1 = factor(data$q5_sd4_r1,
                        labels = c("Married/Partnered/Concubine",
                                   "Single/Divorced/Separated/Widowed"))
```

```r
# Children yes/non
data$q5_sd5 = factor(data$q5_sd5, labels = c("Yes", "Non"))

# Level of study
data$q5_sd10_r2 <- factor(data$q5_sd10_r2,
                          labels = c("No degree", "Less than Bachelor's degree",
                                     "High school diploma or more"))

# Pain
data$q5_pain <- factor(data$q5_pain, labels = c("Yes", "Non"))

# Anxiety
data$q5_anxiete <- factor(data$q5_anxiete,
                          labels = c("No anxiety", "Questionable anxiety state",
                                     "Certain anxiety state"))

# Depression
data$q5_depression <- factor(data$q5_depression,
                             labels = c("No depression",
                                        "Questionable depression state",
                                        "Certain depression state"))

# Net salary category
data$q5_jobv5.36_r1 <- factor(data$q5_jobv5.36_r1,
                              labels = c("<1500€", ">=1500€","Not employed"))

# Social category
data$ms_csp_enq_3c_r1 <- factor(data$ms_csp_enq_3c_r1,
                                labels = c("Executives", "Managerial occupations", "Not employed"))

# Sequels
data$q5_med23.1 <- factor(data$q5_med23.1,
                          labels = c("YES and they are very important",
                                     "YES and they are important",
                                     "YES but moderate", "YES but very moderate",
                                     "NO, i have no after-effects"))
```

```r
# Associationg a label with each variable, purely for QoL
label(data$fc_caisse) <- "Health insurance"
label(data$ms_codcancer) <- "Pathology's location"
label(data$fc_agediag_r0) <- "Age"
label(data$q5_sd4_r1) <- "Marital status"
label(data$q5_sd5) <- "Children yes/non"
label(data$q5_sd10_r2) <- "Level of study"
label(data$q5_pcs12_r1) <- "Physical QoL"
label(data$q5_mcs12_r1) <- "Mental QoL"
label(data$q5_pain) <- "Pain"
label(data$q5_anxiete) <- "Anxiety"
label(data$q5_depression) <- "Depression"
label(data$q5_jobv5.36_r1) <- "Net salary"
label(data$ms_csp_enq_3c_r1) <- "Socio-professional category"
label(data$q5_med23.1) <- "Sequels"
label(data$q5_eortc_fatigue_r1) <- "EORTC fatigue scale"
```

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.

```
data$q5_eortc_fatigue_r1_fac =
  cut(data$q5_eortc_fatigue_r1,
      breaks = c(0,40,100),
      labels = c("Not Clinically significant", "Clinically significant"),
      include.lowest = T)

label(data$q5_eortc_fatigue_r1_fac) <- "EORTC fatigue scale"
```

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”.

```
data$q5_med23.1_rec = factor(data$q5_med23.1,
                             labels = c("Important sequelae","Important sequelae",
                                        "Moderate sequelae", "Moderate sequelae",
                                        "No sequelae"))

label(data$q5_med23.1_rec) = "Sequels"
```

9. Display the frequency table for this new variable.

```
a = table(data$q5_med23.1_rec)
b = paste(round(prop.table(table(data$q5_med23.1_rec))*100,2),"%")

print(rbind(a,b))
```

```
##   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.

```
whole_pop = data$fc_agediag_r0
breast_pop = data$fc_agediag_r0[which(data$ms_codcancer=="Breast")]

mean(whole_pop) # Mean age of our population
```

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

```
mean(breast_pop, na.rm=T) # mean age for breast cancer subpopulation.
```

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

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

```
t.test(whole_pop)$"conf.int" # T confidence intervals for the whole population
```

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

```
t.test(breast_pop)$"conf.int" # T confidence interval for the breast cancer subpopulation
```

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

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

```
prettyR::describe(whole_pop,num.desc=c("var","sd"), xname="the age variable for the whole population", 
```

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

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

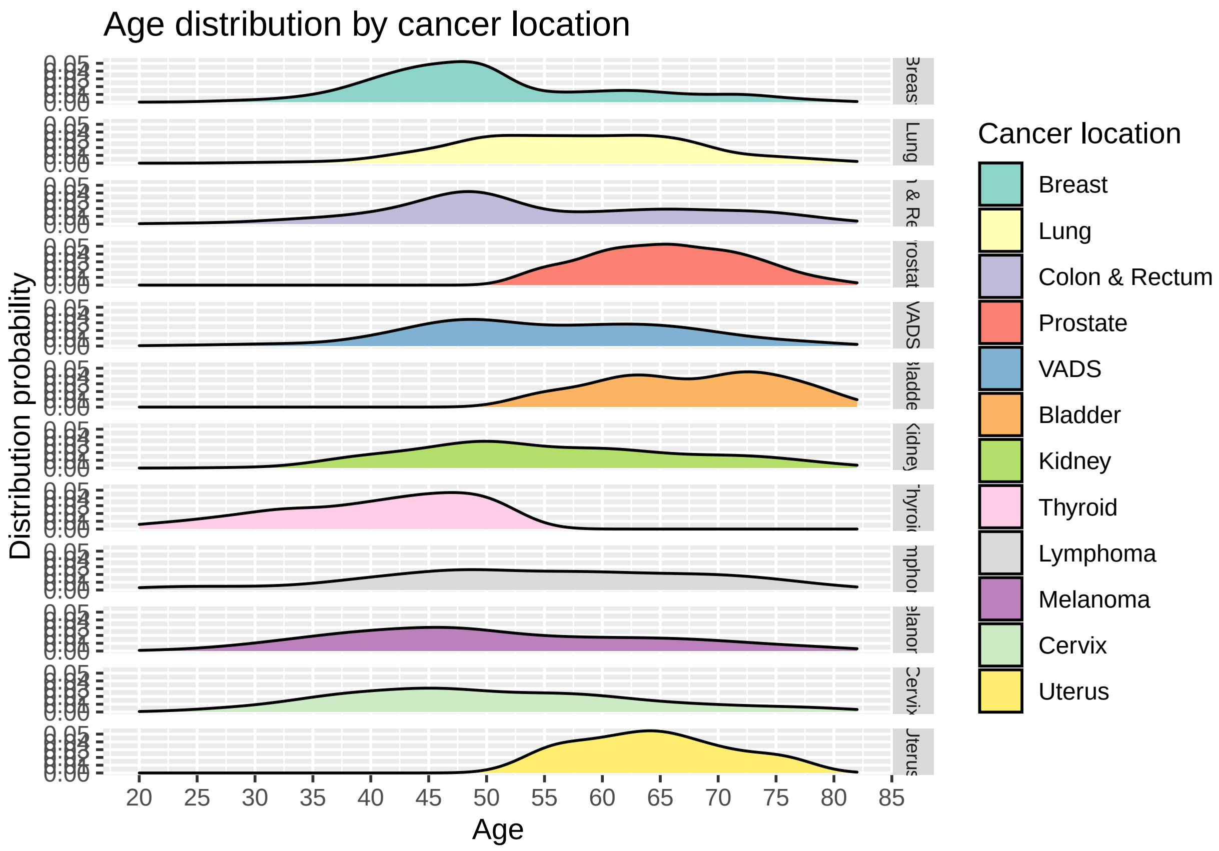
```
prettyR::describe(breast_pop,num.desc=c("var","sd"), xname="the age variable for the breast cancer popul
```

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

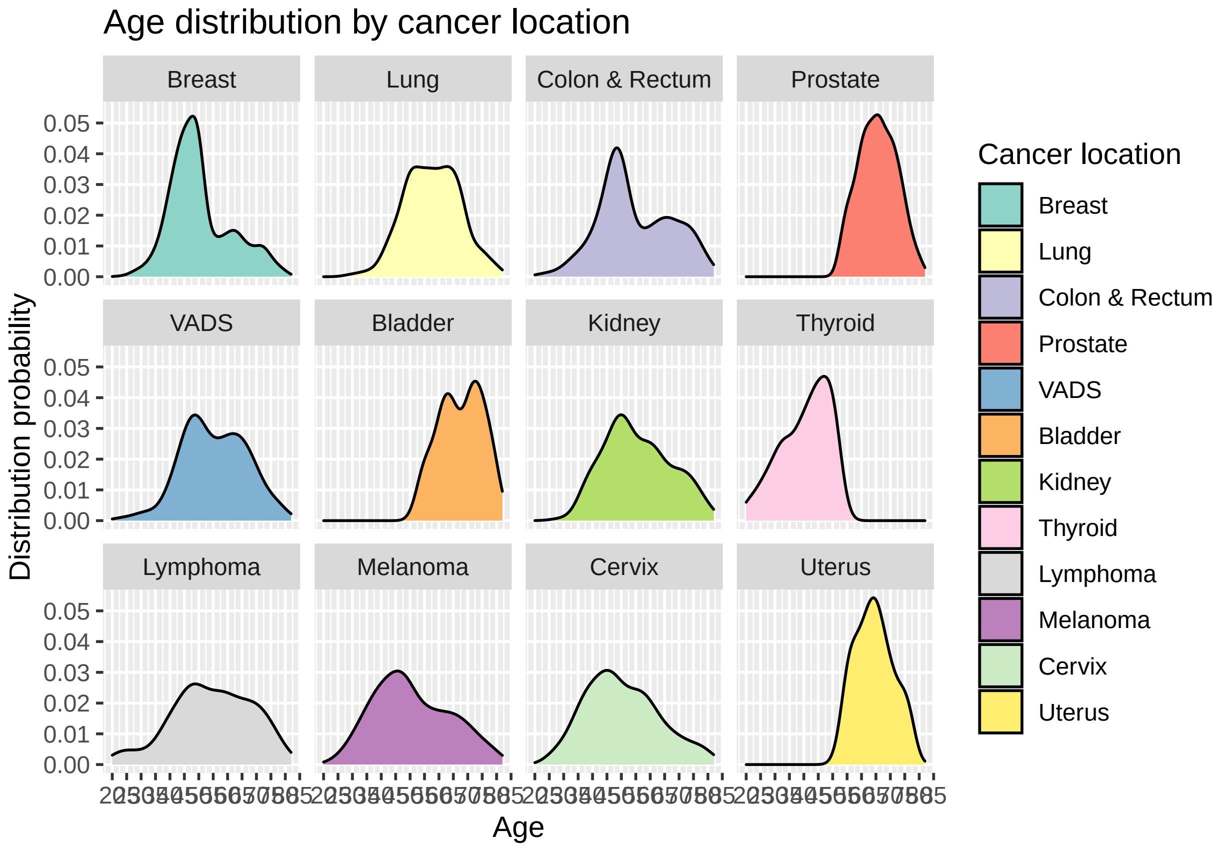
```
##
##  Numeric
##      var    sd
## x 112.86 10.62
```

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.

```
# Solution 1
ggplot(data = data,
       aes(x=fc_agediag_r0, group = ms_codcancer, fill = ms_codcancer)) +
    geom_density()+
    scale_fill_brewer(palette="Set3")+
    scale_y_continuous(breaks = seq(0,0.1, 0.01), minor_breaks = seq(0, 0.005, 0.01))+
    scale_x_continuous(breaks = seq(0,100, 5))+
    labs(x = "Age",
         y = "Distribution probability",
         title = "Age distribution by cancer location",
         fill = "Cancer location")+
    facet_grid(vars(ms_codcancer))+
    theme(strip.text.y = element_text(size = 7))
```

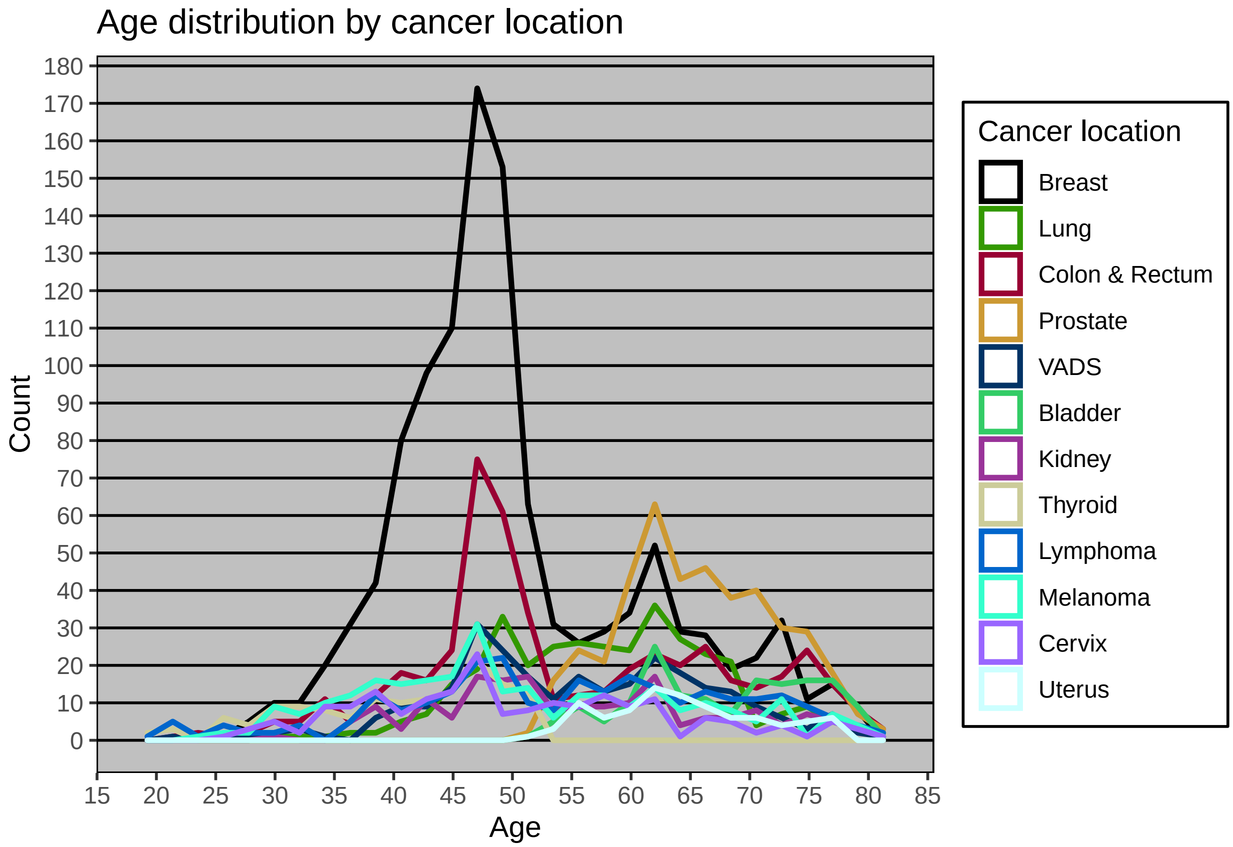


```
ggsave("Solution 1.pdf", plot = last_plot(), device = "pdf", dpi = 300, width = 20, height = 35, units =

# Solution 2
ggplot(data = data,
       aes(x=fc_agediag_r0, group = ms_codcancer, fill = ms_codcancer)) +
  geom_density()+
  scale_fill_brewer(palette="Set3")+
  scale_y_continuous(breaks = seq(0,0.1, 0.01), minor_breaks = seq(0, 0.005, 0.01))+
  scale_x_continuous(breaks = seq(0,100, 5))+
  labs(x = "Age",
      y = "Distribution probability",
       title = "Age distribution by cancer location",
       fill = "Cancer location")+
    facet_wrap(vars(ms_codcancer))
```



```
ggsave("Solution 2.pdf", plot = last_plot(), device = "pdf", dpi = 300, width = 40, height = 30, units

# Solution 3
ggplot(data = data,
       aes(x=fc_agediag_r0, group = ms_codcancer, colour = ms_codcancer)) +
    geom_density(stat = "bin", size = 1)+
    scale_color_manual(values = primary.colors(n=12, step = 6))+
    scale_x_continuous(breaks = seq(0,100, 5))+
    scale_y_continuous(breaks = seq(0,1000, 10))+
    labs(x = "Age",
         y = "Count",
         title = "Age distribution by cancer location",
         colour = "Cancer location")+ theme_excel()
```

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



```
ggsave("Solution 3.pdf", plot = last_plot(), device = "pdf", dpi = 300, width = 25, height = 20, units =
```

```
## ‘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.

```
# Listing the explicative variables
explicative_variables = c("fc_caisse", "ms_codcancer","fc_agediag_r0","q5_sd4_r1","q5_sd5","q5_sd10_r2"

# Printing the list of explicative variables, with their labels, that are gonna be used in the model
kable(label(data[explicative_variables]))
```

	x
fc_caisse	Health insurance
ms_codcancer	Pathology’s location
fc_agediag_r0	Age
q5_sd4_r1	Marital status
q5_sd5	Children yes/non
q5_sd10_r2	Level of study

	x
q5_eortc_fatigue_r1_fac	EORTC fatigue scale
q5_anxiete	Anxiety
q5_depression	Depression
q5_jobv5.36_r1	Net salary
ms_csp_enq_3c_r1	Socio-professional category
q5_med23.1_rec	Sequels
q5_pain	Pain

```
# Recoding the reference for qualitative variables
# List of factors to be recoded to make sure we correctly interpret the results
subset(sapply(data,is.factor), sapply(data,is.factor)==1)
```

```
##               fc_caisse            ms_codcancer               q5_sd4_r1
##                    TRUE                    TRUE                    TRUE
##                  q5_sd5              q5_sd10_r2              q5_anxiete
##                    TRUE                    TRUE                    TRUE
##           q5_depression          q5_jobv5.36_r1        ms_csp_enq_3c_r1
##                    TRUE                    TRUE                    TRUE
##              q5_med23.1                 q5_pain q5_eortc_fatigue_r1_fac
##                    TRUE                    TRUE                    TRUE
##          q5_med23.1_rec
##                    TRUE
```

```
data$fc_caisse <- relevel(data$fc_caisse, ref = "CNAMTS")
data$ms_codcancer <- relevel(data$ms_codcancer, ref = "Breast")
data$q5_sd4_r1 <- relevel(data$q5_sd4_r1, ref = "Married/Partnered/Concubine")
data$q5_sd5 <- relevel(data$q5_sd5, ref = "Non")
data$q5_sd10_r2 <- relevel(data$q5_sd10_r2, ref = "No degree")
data$q5_anxiete <- relevel(data$q5_anxiete, ref = "No anxiety")
data$q5_jobv5.36_r1 <- relevel(data$q5_jobv5.36_r1, ref = "Not employed")
data$ms_csp_enq_3c_r1 <- relevel(data$ms_csp_enq_3c_r1, ref = "Not employed")
data$q5_med23.1_rec <- relevel(data$q5_med23.1_rec, ref = "No sequelae")
data$q5_pain <- relevel(data$q5_pain, ref = "Non")
data$q5_depression <- relevel(data$q5_depression, ref = "No depression")
data$q5_eortc_fatigue_r1_fac <- relevel(data$q5_eortc_fatigue_r1_fac, ref = "Not Clinically significant

# Checking if any of the factor levels has a low effective
data %>% select(where(is.factor)) %>% apply(2,table)
```

```
## $fc_caisse
##
## CNAMTS    MSA    RSI
##   3005    531    426
##
## $ms_codcancer
##
##        Bladder         Breast         Cervix Colon & Rectum         Kidney
##            158           1134            190            492            182
##           Lung       Lymphoma       Melanoma       Prostate        Thyroid
##            343            266            273            423            141
##         Uterus           VADS
##             90            270
## 
## $q5_sd4_r1
## 
##       Married/Partnered/Concubine Single/Divorced/Separated/Widowed
##                              2870                              1092
## 
## $q5_sd5
## 
##  Non  Yes
##  435 3527
## 
## $q5_sd10_r2
## 
## High school diploma or more Less than Bachelor's degree
##                        1910                        1793
##                   No degree
##                         259
## 
## $q5_anxiete
## 
##      Certain anxiety state                 No anxiety
##                        901                       2128
## Questionable anxiety state
##                        933
## 
## $q5_depression
## 
##      Certain depression state                 No depression
##                           281                          3261
## Questionable depression state
##                           420
## 
## $q5_jobv5.36_r1
## 
##       <1500\200      >=1500\200 Not employed
##          740          901         2321
## 
## $ms_csp_enq_3c_r1
## 
##             Executives Managerial occupations           Not employed
##                    912                    727                   2323
## 
## $q5_med23.1
## 
##     NO, i have no after-effects      YES and they are important
##                            1403                             645
## YES and they are very important                YES but moderate
##                             262                            1062
##           YES but very moderate
##                             590
## 
## $q5_pain
```

```
##
##  Non  Yes
## 1019 2943
##
## $q5_eortc_fatigue_r1_fac
##
##     Clinically significant Not Clinically significant
##                       1938                       2018
##
## $q5_med23.1_rec
##
## Important sequelae  Moderate sequelae        No sequelae
##                907               1652               1403
```

```
# Physical QoL
mod1 <- lm(data = data, q5_pcs12_r1 ~ fc_caisse + ms_codcancer + fc_agediag_r0 + q5_sd4_r1 + q5_sd5 + q5

# Mental QoL
mod2 <- lm(data = data, q5_mcs12_r1 ~ fc_caisse + ms_codcancer + fc_agediag_r0 + q5_sd4_r1 + q5_sd5 + q5

# Factors associated to physical QoL score
subset(summary(mod1)$coefficients,
       summary(mod1)$coefficients[, 4] < 0.2) %>% # filter for p-value<0.2
  data.frame() %>%
  arrange(Estimate) %>% # ordering according to the column Estimate
  kable(caption = "Factors associated to physical QoL score",
        col.names = c("Estimate", "Std error", "t-value", "P-value"))
```

Table 2: Factors associated to physical QoL score

	Estimate	Std error	t-value	P-value
q5_eortc_fatigue_r1_facClinically significant	-7.9531291	0.2898231	-27.441319	0.0000000
q5_med23.1_recImportant sequelae	-5.8034204	0.3659216	-15.859738	0.0000000
q5_depressionCertain depression state	-5.6313473	0.5287921	-10.649454	0.0000000
q5_depressionQuestionable depression state	-3.0918624	0.4234795	-7.301091	0.0000000
ms_codcancerLung	-1.9713987	0.4908194	-4.016546	0.0000602
q5_med23.1_recModerate sequelae	-1.8759536	0.2915953	-6.433416	0.0000000
ms_codcancerKidney	-0.9102051	0.6173612	-1.474348	0.1404681
q5_mcs12_r1	-0.1461533	0.0154223	-9.476765	0.0000000
fc_agediag_r0	-0.0841834	0.0144784	-5.814419	0.0000000
ms_codcancerMelanoma	0.7944655	0.5243222	1.515224	0.1297962
ms_codcancerColon & Rectum	0.8968529	0.4200487	2.135117	0.0328132
q5_jobv5.36_r1<1500€	0.9992067	0.3976964	2.512486	0.0120282
ms_codcancerLymphoma	1.0189050	0.5261088	1.936681	0.0528560
ms_codcancerVADS	1.1165635	0.5317606	2.099749	0.0358146
ms_codcancerBladder	1.2017636	0.6813797	1.763721	0.0778568
q5_sd10_r2Less than Bachelor's degree	1.5646119	0.5083774	3.077658	0.0021007
ms_codcancerProstate	2.6837647	0.4796683	5.595042	0.0000000
q5_jobv5.36_r1>=1500€	2.7811554	0.3908075	7.116432	0.0000000
q5_sd10_r2High school diploma or more	3.1894703	0.5231439	6.096736	0.0000000
q5_painYes	3.6691743	0.3014531	12.171625	0.0000000
(Intercept)	56.5543576	1.3691412	41.306448	0.0000000

```r
# Factors associated to mental QoL score
subset(
  summary(mod2)$coefficients,
  summary(mod2)$coefficients[, 4] < 0.2) %>% # filter for p-value<0.2
  data.frame() %>%
  arrange(Estimate) %>% # ordering according to the column Estimate
  kable(caption = "Factors associated to mental QoL score",
        col.names = c("Estimate", "Std error", "t-value", "P-value"))
```

Table 3: Factors associated to mental QoL score

	Estimate	Std error	t-value	P-value
q5_depressionCertain depression state	-8.9875323	0.5297493	-16.965632	0.0000000
q5_anxieteCertain anxiety state	-7.8352006	0.3484953	-22.482947	0.0000000
q5_eortc_fatigue_r1_facClinically significant	-6.6832389	0.3056678	-21.864388	0.0000000
q5_depressionQuestionable depression state	-5.3710770	0.4277356	-12.557004	0.0000000
q5_anxieteQuestionable anxiety state	-4.0793367	0.3158164	-12.916797	0.0000000
q5_med23.1_recImportant sequelae	-2.5384246	0.3840921	-6.608896	0.0000000
q5_med23.1_recModerate sequelae	-1.0668308	0.2994483	-3.562654	0.0003715
q5_sd4_r1Single/Divorced/Separated/Widowed	-0.9770976	0.2884332	-3.387604	0.0007120
q5_sd5Yes	-0.7123245	0.4097851	-1.738288	0.0822385
q5_pcs12_r1	-0.1530166	0.0161465	-9.476765	0.0000000
fc_agediag_r0	0.0268096	0.0148719	1.802697	0.0715124
q5_jobv5.36_r1>=1500€	0.6593942	0.4023117	1.639013	0.1012906
ms_codcancerProstate	0.6675365	0.4926395	1.355020	0.1754890
ms_codcancerVADS	0.9938414	0.5441775	1.826319	0.0678782
ms_codcancerLung	1.4327711	0.5027227	2.850023	0.0043944
q5_painYes	1.5061492	0.3132951	4.807446	0.0000016
(Intercept)	58.3405934	1.3959080	41.794010	0.0000000

How to interpret the linear regression model Printed tables have already been filter to exclude results with a p-value >= 0.2 criteria.

Nota bene: these are *associations*, not implying causality.

Categorical variables: under the column "Estimate", the table shows the average difference between a given modality and the reference modality for the same factor variable, as defined in the code chunk "*recoding reference mod for factor variables*".
For example, patients in a "*Certain depression state*" score, on average, 5.63 points lower for Physical QoL compared to "*non-depressed*" patients.
According to the same logic, someone with a monthly salary over 1500€ scores on average 2.78 points higher compaired to un unemployed patient.
The same logic applies all others categorical variables and modalities.

Quantitative variables: two are the numeric variables taken into account by the model, patient's age and "the other" QoL score.
In a plot where x="Age" and Y=QoL score, *Estimate* is the coefficient that ties the two variables.
For example, for every additional year of age at diagnosis, physical QoL score lowers, on average, by 0.084 .

Key messages:

Nota bene: the huge population allow us to have great statistical power, lots of associations come out as statistically relevant but the magnitude of the effect is very small and frankly irrelevant on a scale that goes from 0 to 100 (or so it seems, we've no additional information on scale boundaries).

1) *Depressed* and *anxious* patients score lower on both scales with an impact proportional to the severity of the psychiatric pathology.

Patients with *sequels* and *clinically significant fatigue* are also associated with scores notably lower.

These 4 variables seems to be the only truly impact-full ones.

2) Educational degree and monthly salary seems to have a modest impact only on the physical QoL score.
3) Mental and physical QoL scores look to be inversely proportional but the coefficient is rather small, requiring huge variations on a scale to impact the other one. This is unexpected nevetheless, if both scales go in the same direction (0 --> 100).
4) Age at diagnosis seems to have only a very small impact on both score, especially the mental one.
5) A few cancer locations come out as statistically significant. As :

- we have no additional information on the kind of therapy the patients underwent to or on the stade of their disease at diagnosis
- coefficients are very small (the most important one being lung cancer patients scoring 2 points lower for physical QoL, on average)
- results are inconsistent between the two score (lung cancer scoring the lowest for physical QoL and the highest for mental QoL score)

  no meaningful hypothesis/explanation can be formulated.

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