Thema09_Log

install kableExtra install knitr

Urinary biomarkers for pancreatic cancer

https://www.kaggle.com/johnjdavisiv/urinary-biomarkers-for-pancreatic- cancer

Research question: Is it possible to detect pancreatic cancer using values of the urinary biomarkers?

Introduction

Pancreatic ductal adenocarcinoma (PDAC) is one of the deadiest cancers. The chances of survival are increased when diagnosed in an early stage. However, PDAC shows symptoms when it already spread throughout the body. Most of the time, it's too late by then. There may be a way to detect PCAD in an early stage with a simple urine test, with the use of the following biomarkers: creatinine (Urinary biomarker of kidney function) LYVE1 (Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis), REG1A and REG1B (Urinary levels of a protein that may be associated with pancreas regeneration.), and TFF1 (Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract)

the attributes in the data interesting for this research are the biomarker values mentioned before. There is also an attribute called diagnosis, in which the diagnosis of the sample is stated, where 1 means no PDAC, 2 means benign hepatobiliary disease (non cancerous, non harmful pancreatic condition), and 3 means that the sample has PDAC.

EDA

Codebook

```
myData <- read.csv("Data/Debernardi et al 2020 data.csv")

columns <- colnames(myData)
type <- c("character", "character", "character", "double", "character", "double", "logical", "logical",
unit <- c(NA, NA, NA, "years", "F/M", NA, NA, NA, "U/ml", "mg/ml", "ng/ml", "ng/ml", "ng/ml", "ng/ml")
descriptions = c("Unique string identifying each subject", "Cohort 1, previously used samples; Cohort 2
codebook <- data.frame(columns, type, unit, descriptions)
write.csv(codebook, "Codebook.csv", row.names = FALSE)
knitr::kable(codebook, format = 'latex') %>%
  kable_styling(full_width = F) %>%
  column_spec(1, bold = T) %>%
  column_spec(4, width = "22em")
```

columns	type	unit	descriptions
sample_id	character	NA	Unique string identifying each subject
patient_cohort	character	NA	Cohort 1, previously used samples; Cohort 2, newly added samples
sample_origin	character	NA	BPTB: Barts Pancreas Tissue Bank, London, UK; ESP: Spanish National Cancer Research Centre, Madrid, Spain; LIV: Liverpool University, UK; UCL: University College London, UK
age	double	years	Age in years
sex	character	F/M	M = male, F = female
diagnosis	double	NA	1 = control (no pancreatic disease), 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis); 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer
stage	logical	NA	For those with pancratic cancer, what stage was it? One of IA, IB, IIA, IIIB, III, IV
benign_sample_diagnosis	logical	NA	For those with a benign, non-cancerous diagnosis, what was the diagnosis?
plasma_CA19_9	double	U/ml	Blood plasma levels of CA 19–9 monoclonal antibody that is often elevated in patients with pancreatic cancer. Only assessed in 350 patients (one goal of the study was to compare various CA 19-9 cutpoints from a blood sample to the model developed using urinary samples).
creatinine	double	mg/ml	Urinary biomarker of kidney function
LYVE1	double	ng/ml	Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
REG1B	double	ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration.
TFF1	double	ng/ml	Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract
REG1A	double	ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration. Only assessed in 306 patients (one goal of the study was to assess REG1B vs REG1A)

Data exploration

Visualization

```
str(myData)
```

```
## 'data.frame':    590 obs. of  14 variables:
##  $ sample_id              : chr  "S1" "S10" "S100" "S101" ...
##  $ patient_cohort         : chr  "Cohort1" "Cohort1" "Cohort2" "Cohort2" ...
##  $ sample_origin          : chr  "BPTB" "BPTB" "BPTB" "BPTB" ...
##  $ age                    : int  33 81 51 61 62 53 70 58 59 56 ...
##  $ sex                    : chr  "F" "F" "M" "M" ...
##  $ diagnosis              : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ stage                  : chr  "" "" "" "" ...
##  $ benign_sample_diagnosis: chr  "" "" "" "" ...
##  $ plasma_CA19_9          : num  11.7 NA 7 8 9 NA NA 11 NA 24 ...
##  $ creatinine             : num  1.832 0.973 0.78 0.701 0.215 ...
##  $ LYVE1                  : num  0.89322 2.03758 0.14559 0.0028 0.00086 ...
##  $ REG1B                  : num  52.9 94.5 102.4 60.6 65.5 ...
##  $ TFF1                   : num  654.3 209.5 461.1 142.9 41.1 ...
##  $ REG1A                  : num  1262 228 NA NA NA ...
```

For every column, you can find information below

```
summary(myData[c(4, 9:14)])
```

```
##       age         plasma_CA19_9       creatinine          LYVE1
##  Min.   :26.00   Min.   :    0.0   Min.   :0.05655   Min.   : 0.000129
##  1st Qu.:50.00   1st Qu.:    8.0   1st Qu.:0.37323   1st Qu.: 0.167179
##  Median :60.00   Median :   26.5   Median :0.72384   Median : 1.649862
##  Mean   :59.08   Mean   :  654.0   Mean   :0.85538   Mean   : 3.063530
##  3rd Qu.:69.00   3rd Qu.:  294.0   3rd Qu.:1.13948   3rd Qu.: 5.205037
##  Max.   :89.00   Max.   :31000.0   Max.   :4.11684   Max.   :23.890323
##                  NA's   :240
##      REG1B               TFF1              REG1A
##  Min.   :   0.0011   Min.   :    0.005   Min.   :    0.00
##  1st Qu.:  10.7572   1st Qu.:   43.961   1st Qu.:   80.69
##  Median :  34.3034   Median :  259.874   Median :  208.54
##  Mean   : 111.7741   Mean   :  597.869   Mean   :  735.28
##  3rd Qu.: 122.7410   3rd Qu.:  742.736   3rd Qu.:  649.00
##  Max.   :1403.8976   Max.   :13344.300   Max.   :13200.00
##                                          NA's   :284
```

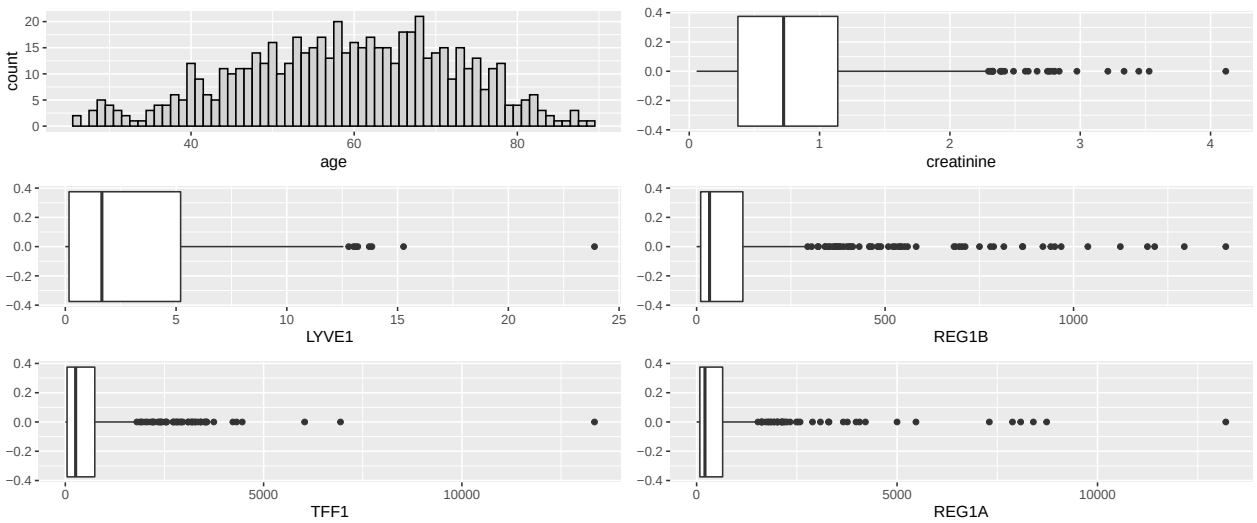
Let's have a look at the important data with plots.

```
p1 <- ggplot(myData, aes(x=age)) +
  geom_histogram(fill = "lightgrey", col = "black", binwidth = 1)

p2 <- ggplot(myData, aes(x=creatinine)) +
  geom_boxplot()
p3 <- ggplot(myData, aes(x=LYVE1)) +
  geom_boxplot()
p4 <- ggplot(myData, aes(x=REG1B)) +
  geom_boxplot()
p5 <- ggplot(myData, aes(x=TFF1)) +
  geom_boxplot()
p6 <- ggplot(myData, aes(x=REG1A)) +
```

```
  geom_boxplot()

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)
```



as you can see, there are a lot of outliers which we need to check. There are also a lot of missing values in the REG1A column, We should remove those.

```
Q <- quantile(myData$creatinine, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(myData$creatinine)
up <- Q[2]+1.5*iqr
low <- Q[1]-1.5*iqr

myData <- subset(myData, myData$creatinine > low & myData$creatinine < up)

Q <- quantile(myData$LYVE1, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(myData$LYVE1)
up <- Q[2]+1.5*iqr
low <- Q[1]-1.5*iqr

myData <- subset(myData, myData$LYVE1 > low & myData$LYVE1 < up)

Q <- quantile(myData$REG1B, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(myData$REG1B)
up <- Q[2]+1.5*iqr
low <- Q[1]-1.5*iqr

myData <- subset(myData, myData$REG1B > low & myData$REG1B < up)

Q <- quantile(myData$TFF1, probs=c(.25, .75), na.rm = FALSE)
iqr <- IQR(myData$TFF1)
up <- Q[2]+1.5*iqr
low <- Q[1]-1.5*iqr

myData <- subset(myData, myData$TFF1 > low & myData$TFF1 < up)

Q <- quantile(myData$REG1A, probs=c(.25, .75), na.rm = TRUE)
iqr <- IQR(myData$REG1A, na.rm = TRUE)
```

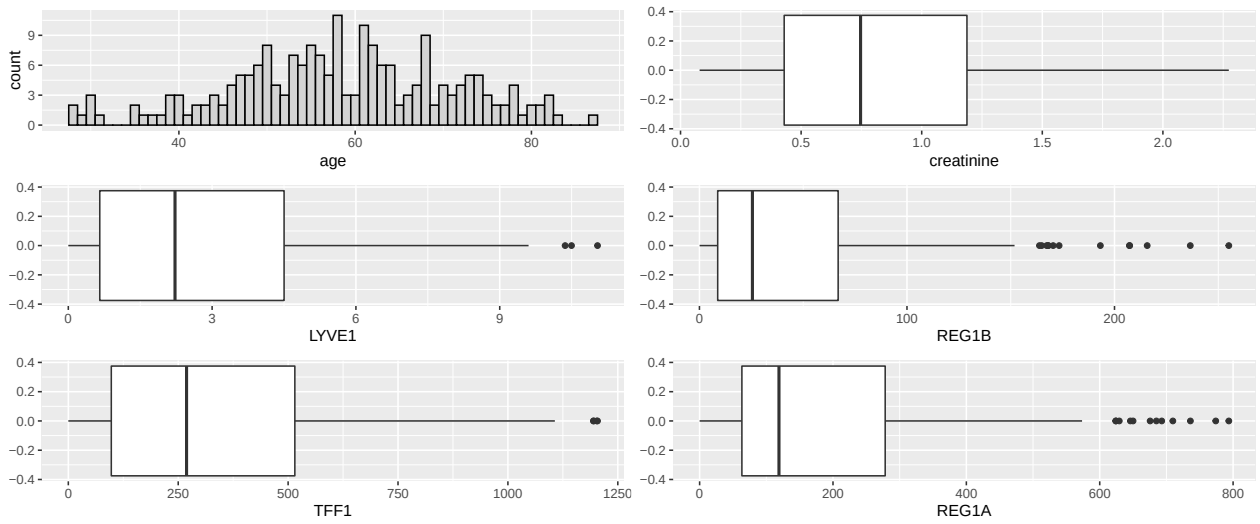
```
up <- Q[2]+1.5*iqr
low <- Q[1]-1.5*iqr

myData <- subset(myData, myData$REG1A > low & myData$REG1A < up)

p1 <- ggplot(myData, aes(x=age)) +
  geom_histogram(fill = "lightgrey", col = "black", binwidth = 1)

p2 <- ggplot(myData, aes(x=creatinine)) +
  geom_boxplot()
p3 <- ggplot(myData, aes(x=LYVE1)) +
  geom_boxplot()
p4 <- ggplot(myData, aes(x=REG1B)) +
  geom_boxplot()
p5 <- ggplot(myData, aes(x=TFF1)) +
  geom_boxplot()
p6 <- ggplot(myData, aes(x=REG1A)) +
  geom_boxplot()

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)
```



This looks a lot cleaner and clearer already.
What we should do now, is look at the values of those urinary levels with each diagnosis and see if there are any obvious differences.
Let's start with the samples that do not have PDAC

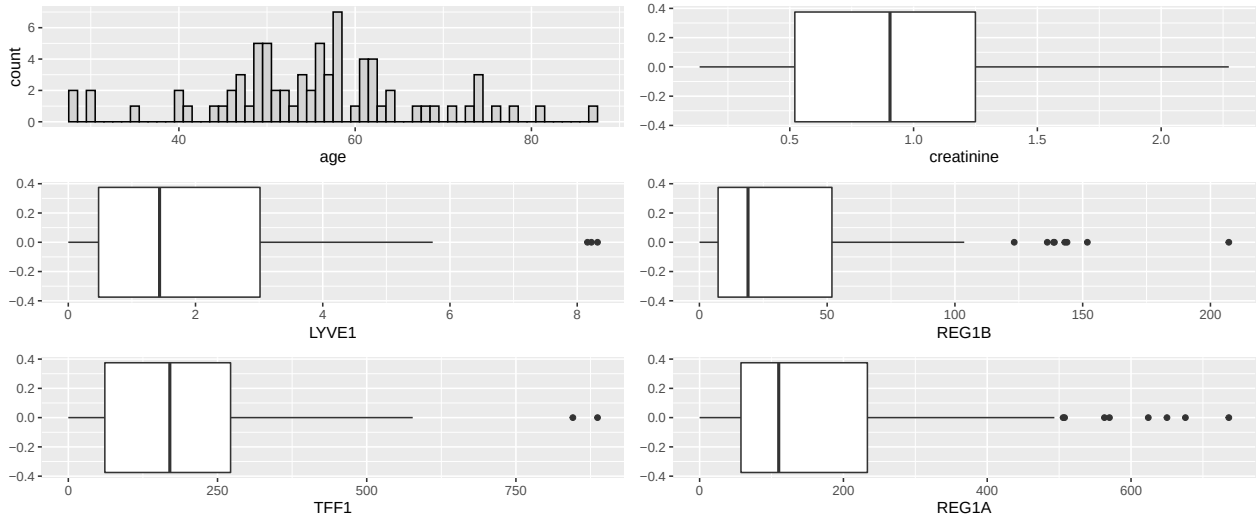
```
myData1 <- subset(myData, myData$diagnosis == 1)

p1 <- ggplot(myData1, aes(x=age)) +
  geom_histogram(fill = "lightgrey", col = "black", binwidth = 1)

p2 <- ggplot(myData1, aes(x=creatinine)) +
  geom_boxplot()
p3 <- ggplot(myData1, aes(x=LYVE1)) +
  geom_boxplot()
p4 <- ggplot(myData1, aes(x=REG1B)) +
  geom_boxplot()
```

```
p5 <- ggplot(myData1, aes(x=TFF1)) +
  geom_boxplot()
p6 <- ggplot(myData1, aes(x=REG1A)) +
  geom_boxplot()

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)
```

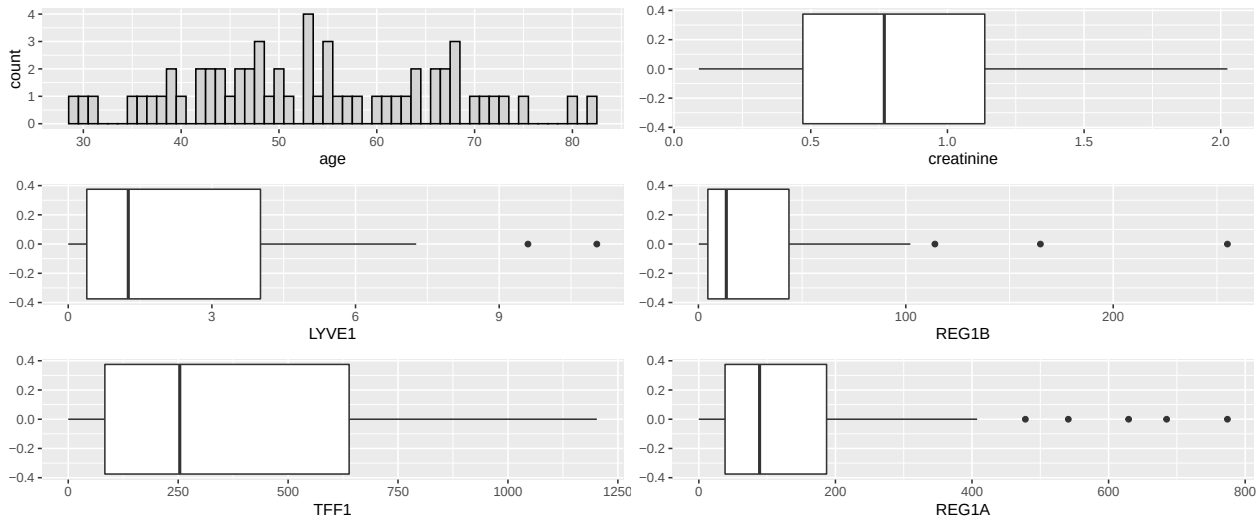



```
myData2 <- subset(myData, myData$diagnosis == 2)

p1 <- ggplot(myData2, aes(x=age)) +
  geom_histogram(fill = "lightgrey", col = "black", binwidth = 1)

p2 <- ggplot(myData2, aes(x=creatinine)) +
  geom_boxplot()
p3 <- ggplot(myData2, aes(x=LYVE1)) +
  geom_boxplot()
p4 <- ggplot(myData2, aes(x=REG1B)) +
  geom_boxplot()
p5 <- ggplot(myData2, aes(x=TFF1)) +
  geom_boxplot()
p6 <- ggplot(myData2, aes(x=REG1A)) +
  geom_boxplot()

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)
```




```
myData3 <- subset(myData, myData$diagnosis == 3)

p1 <- ggplot(myData3, aes(x=age)) +
  geom_histogram(fill = "lightgrey", col = "black", binwidth = 1)

p2 <- ggplot(myData3, aes(x=creatinine)) +
  geom_boxplot()
p3 <- ggplot(myData3, aes(x=LYVE1)) +
  geom_boxplot()
p4 <- ggplot(myData3, aes(x=REG1B)) +
  geom_boxplot()
p5 <- ggplot(myData3, aes(x=TFF1)) +
  geom_boxplot()
p6 <- ggplot(myData3, aes(x=REG1A)) +
  geom_boxplot()

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)
```