EDA

Joshua Tolhuis

9/14/2021

Assignment Introduction:

In this assignment I was tasked to research an existing data file given to me and apply data mining & machine learning to predict the outcome. I first was tasked creating a research question around the data. My question went as follows: "What's the minimal amount of data combinations to predict a patient has pancreatic ductal adenocarcinoma?"

Data

The data gained was from John Davis, uploaded to kaggle John Davis. He had gotten his data from a paper with the title "A combination of urinary biomarker panel and PancRisk score for earlier detection of pancreatic caner: A case-control study", this is the link to the paper.

setup and load code book:

Loading up libraries to be used in the EDA and log.

```
if(!require(devtools)) install.packages("devtools")
devtools::install_github("sinhrks/ggfortify")
devtools::install_github("AckerDWM/gg3D")
library("gg3D")
library(ggplot2)
library(tidyr)
library(gridExtra)
library(dplyr)
library(plotly)
library(cluster)
library(ggfortify)
```

Let's begin with loading in the required data and the code book describing it.

```
setwd("./Data")

data <- read.csv("Debernardi et al 2020 data.csv", header = T, sep = ",")
str(data[])
```

```
## '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 ...
```

```
codebook <- read.csv("Debernardi et al 2020 documentation.csv", sep = ",", header = T)

knitr::kable(codebook[2:3])
```

Original.column.name	Details
Sample ID	Unique string identifying each subject
Patient's Cohort	Cohort 1, previously used samples; Cohort 2, newly added samples
Sample Origin	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	Age in years
Sex	M = male, F = female
Diagnosis (1=Control, 2=Benign, 3=PDAC)	1 = control (no pancreatic disease), 2 = benign hepatobiliary disease (119 of which are chronic pancreatitis); 3 = Pancreatic ductal adenocarcinoma, i.e. pancreatic cancer
Stage	For those with pancratic cancer, what stage was it? One of IA, IB, IIA, IIIB, III, IV
Benign Samples Diagnosis	For those with a benign, non-cancerous diagnosis, what was the diagnosis?
Plasma CA19-9 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 mg/ml	Urinary biomarker of kidney function
LYVE1 ng/ml	Urinary levels of Lymphatic vessel endothelial hyaluronan receptor 1, a protein that may play a role in tumor metastasis
REG1B ng/ml	Urinary levels of a protein that may be associated with pancreas regeneration.
TFF1 ng/ml	Urinary levels of Trefoil Factor 1, which may be related to regeneration and repair of the urinary tract
REG1A 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)

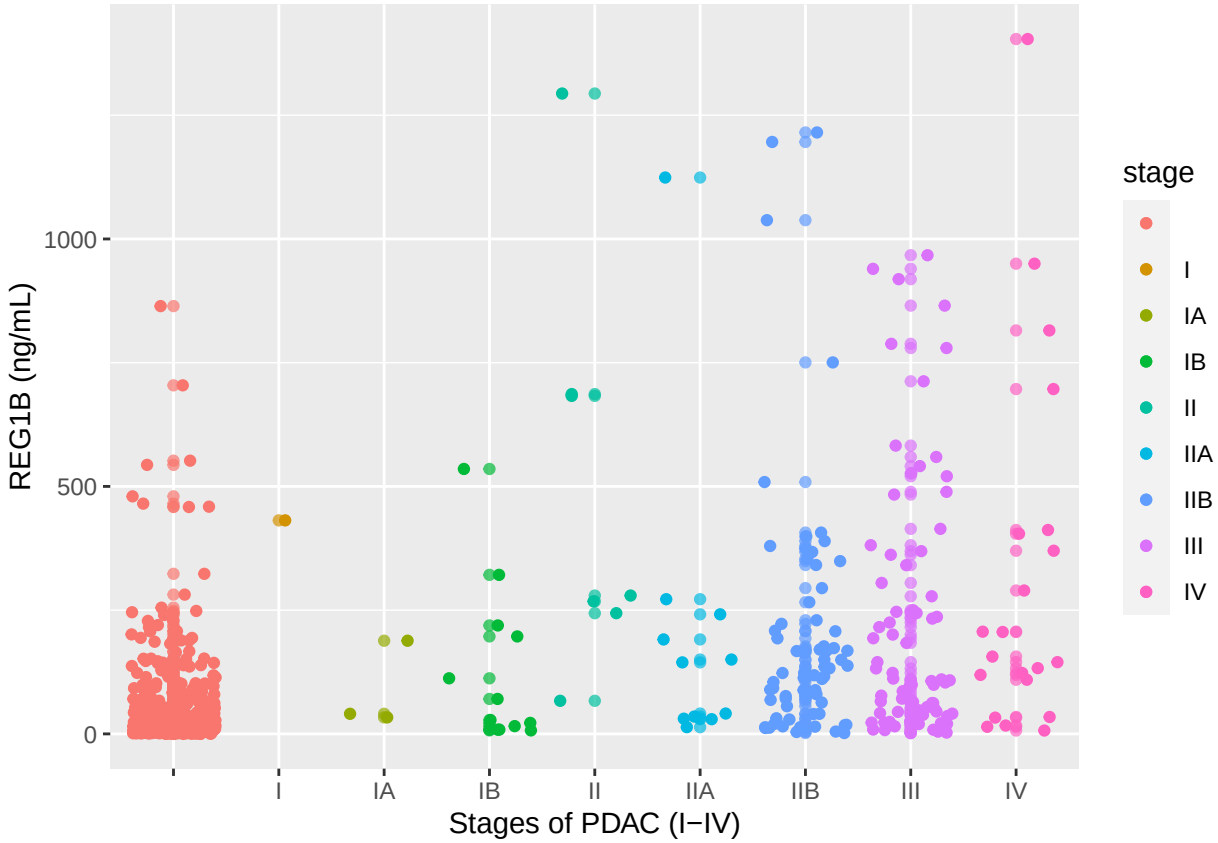
Intro

At first, I wanted to know if REG1B could be used only to predict PDAC, but after reconsideration I want to find out what the minimum required data is to find out if a patient has PDAC. Even so i first wanted to run some tests on the newly improved biomarker REG1B to find out it's impact, and it's change relative to REG1A.

Results

Let's take a first look at the effect of REG1B stand alone, what conclusions can be made by looking at the markers found at different stages.

```
ggplot(data = data, mapping = aes(x = stage, y = REG1B, col = stage)) +
        geom_point(alpha = 0.7) +
        geom_jitter() +
        xlab("Stages of PDAC (I-IV)") +
        ylab("REG1B (ng/mL)")
```



In this plot becomes clear, that REG1B has the same values at the stages "0, IIB, III and IV" relative to the others. This immediately shuts down the proposition to predict stages of cancer with only 1 variable.

In order to find out more I decided to view the other bio markers and this time with a box plot to see if there was a recognizable trend.I tried several box plots on the bio markers, first with normal data, second with scaled data, third with outliers removed and the last plot has been scaled and the outliers removed from it. I also had to create functions for the latter 3 plots.

scaling function:

```
scale_min_max <- function(x) {
     (x - min(x)) / (max(x) - min(x))
}
```

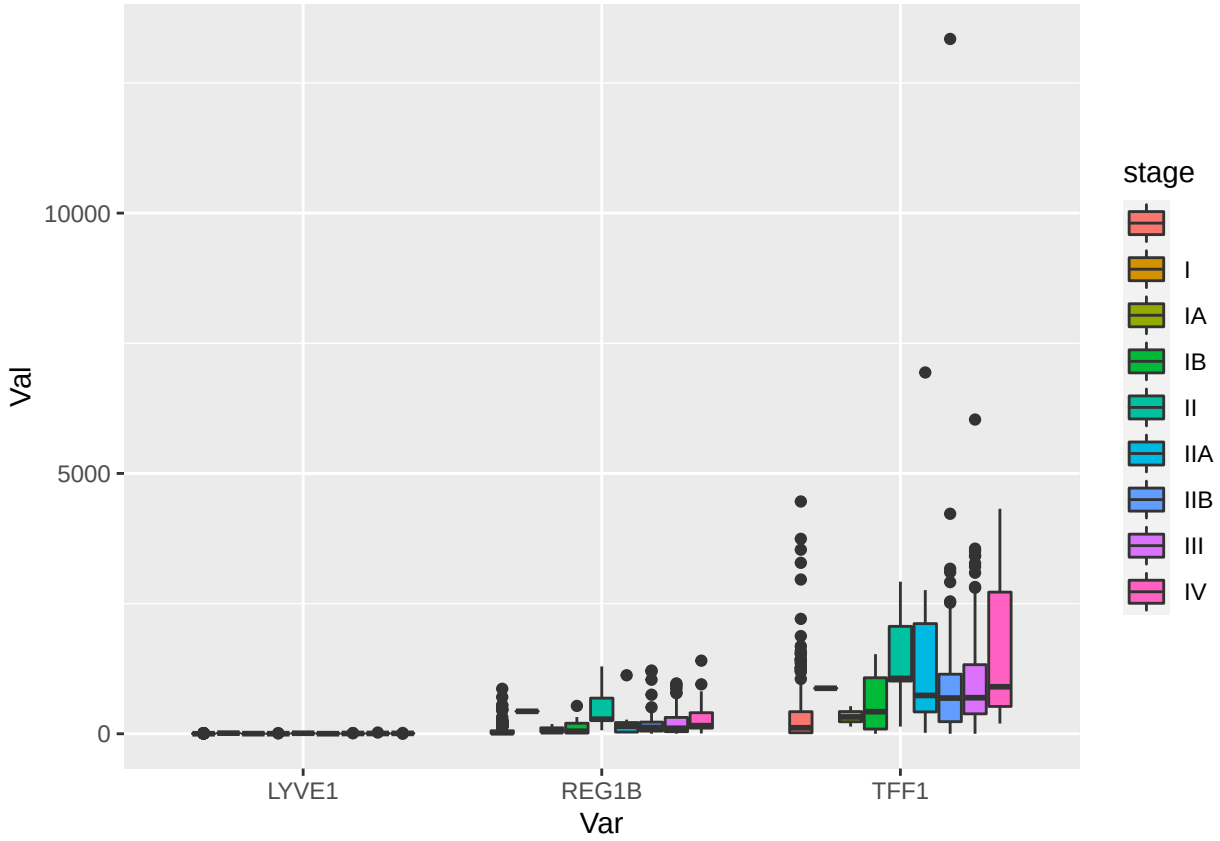
outlier removal function:

```
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
```

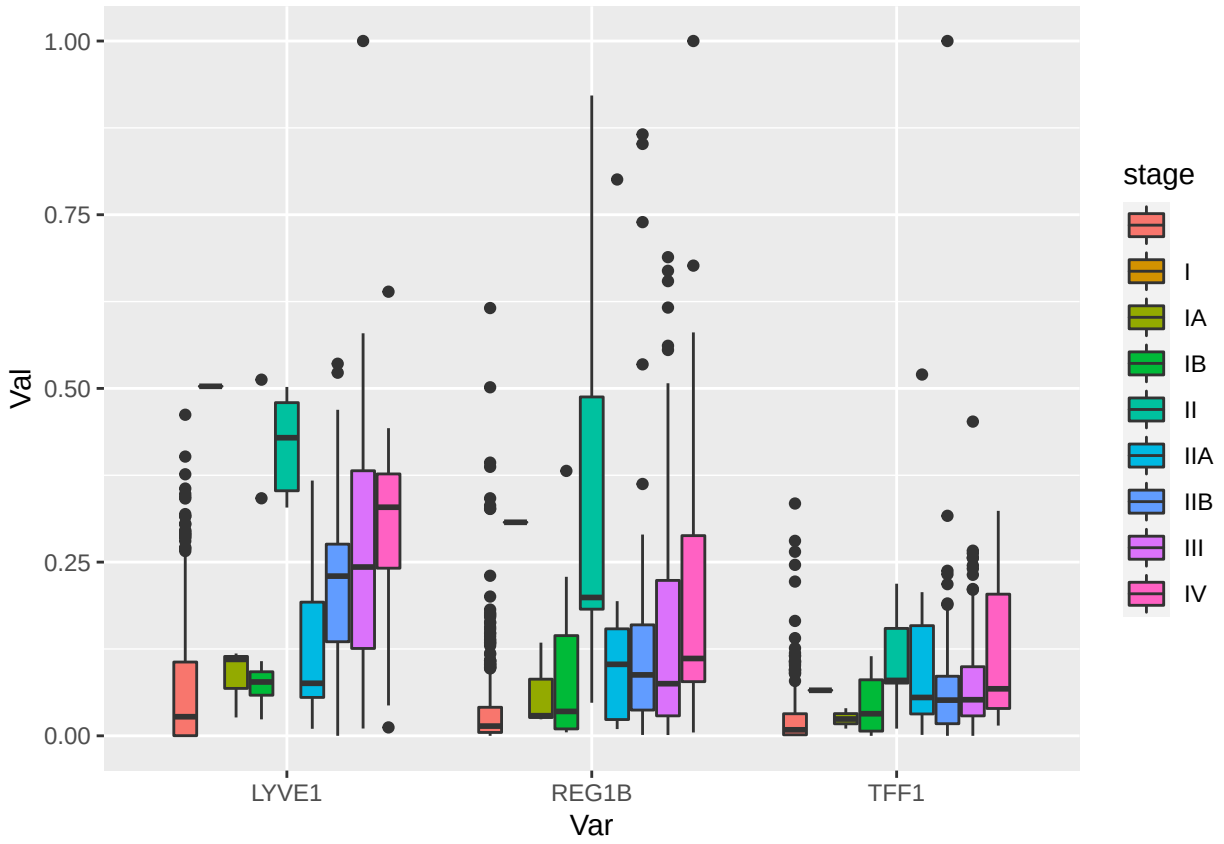
```
}
```

I printed them all separately to get a better view.

```
scaled_data <- data

scaled_data$REG1B <- scale_min_max(scaled_data$REG1B)
scaled_data$TFF1 <- scale_min_max(scaled_data$TFF1)
scaled_data$LYVE1 <- scale_min_max(scaled_data$LYVE1)

p1 <- pivot_longer(data = data, cols = c(REG1B,TFF1,LYVE1), names_to = "Var", values_to = "Val") %>%
        ggplot(aes(x = Var, y = Val, fill = stage)) +
        geom_boxplot()

p2 <- pivot_longer(data = scaled_data, cols = c(REG1B,TFF1,LYVE1), names_to = "Var", values_to = "Val")
        ggplot(aes(x = Var, y = Val, fill = stage)) +
        geom_boxplot()
p3<- pivot_longer(data = data, cols = c(REG1B,TFF1,LYVE1), names_to = "Var", values_to = "Val") %>%
        ggplot(aes(x = Var, y = remove_outliers(Val), fill = stage)) +
        geom_boxplot()
p4 <- pivot_longer(data = scaled_data, cols = c(REG1B,TFF1,LYVE1), names_to = "Var", values_to = "Val")
        ggplot(aes(x = Var, y = remove_outliers(Val), fill = stage)) +
        geom_boxplot()

p1 #normal data
```

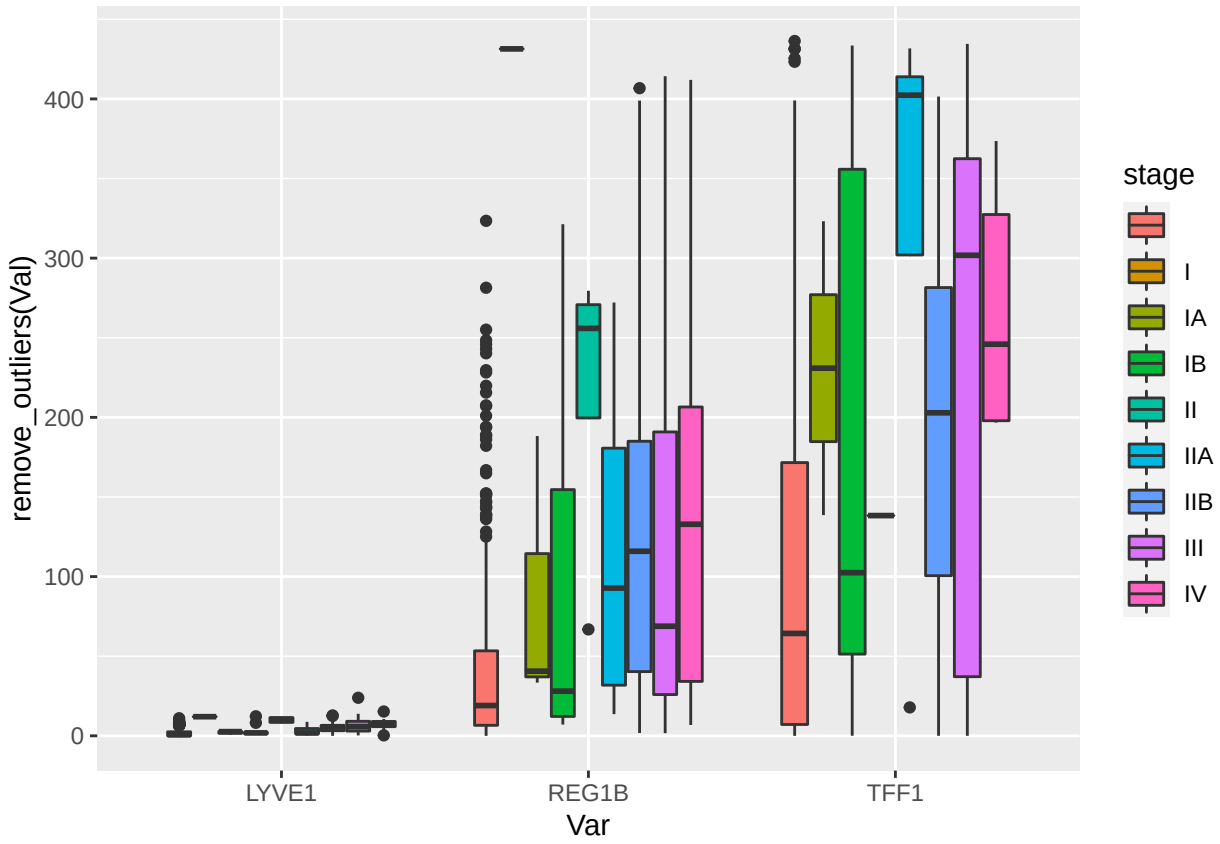



```
p2 #scaled data
```



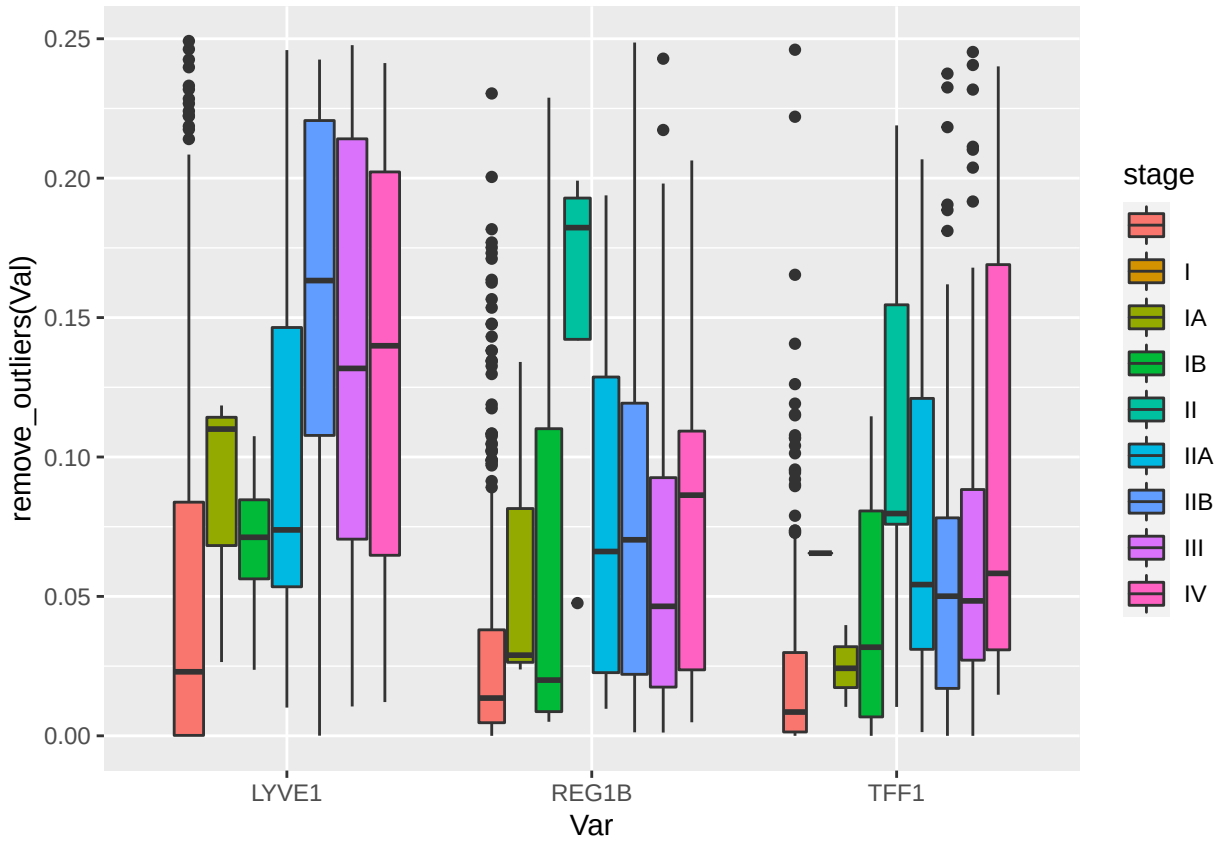

```
p3 #outliers removed
```

```
## Warning: Removed 257 rows containing non-finite values (stat_boxplot).
```




```
p4 #scaled data and outliers removed
```

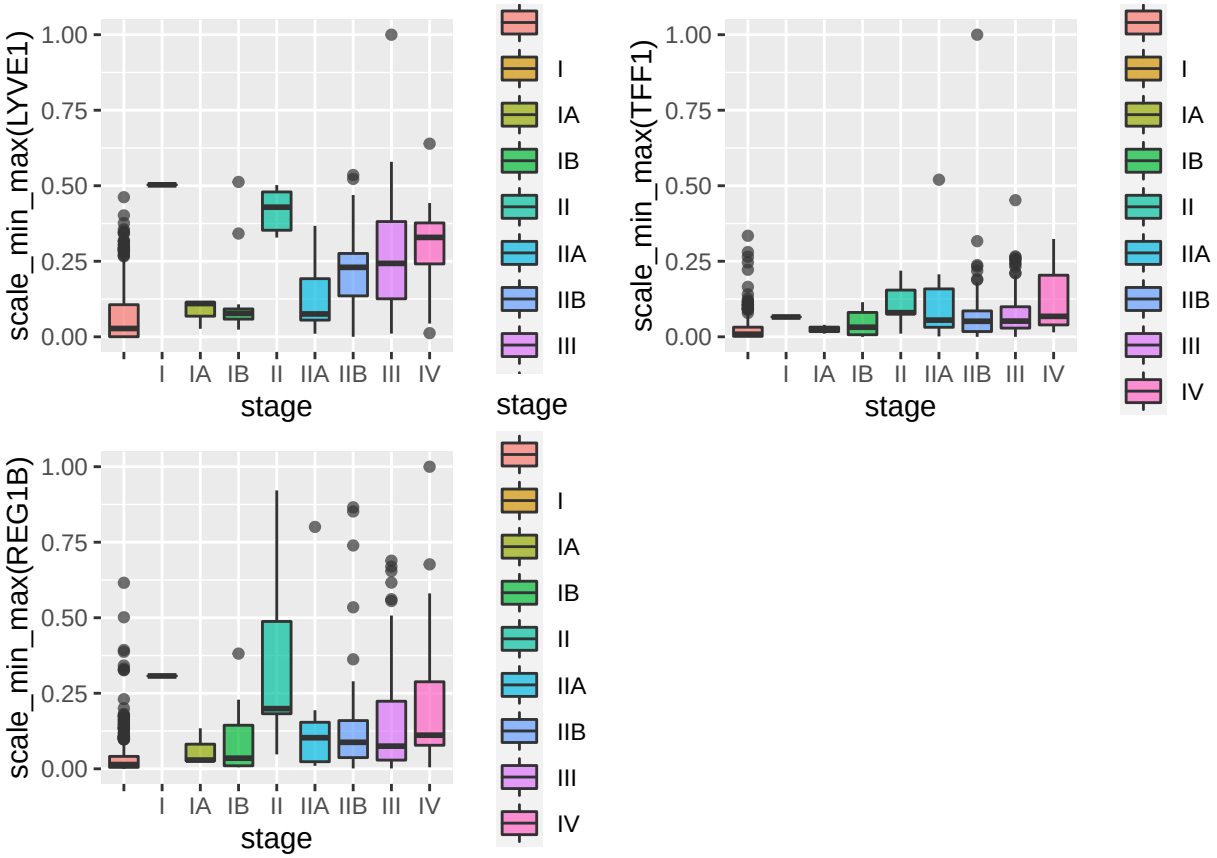
```
## Warning: Removed 179 rows containing non-finite values (stat_boxplot).
```



From these plots a few things are noticeable. First of all the without outliers the difference between stage 0 and other stages becomes compelling. This is good to know to be able to make predictions, but seeing as this is cancer and false positives or false negatives are dangerous. outliers will be necessary to make valid predictions. the second noticeable thing is that there is indeed a trend within the bio markers and stages. This is precious information to be able to predict the stages of cancer.

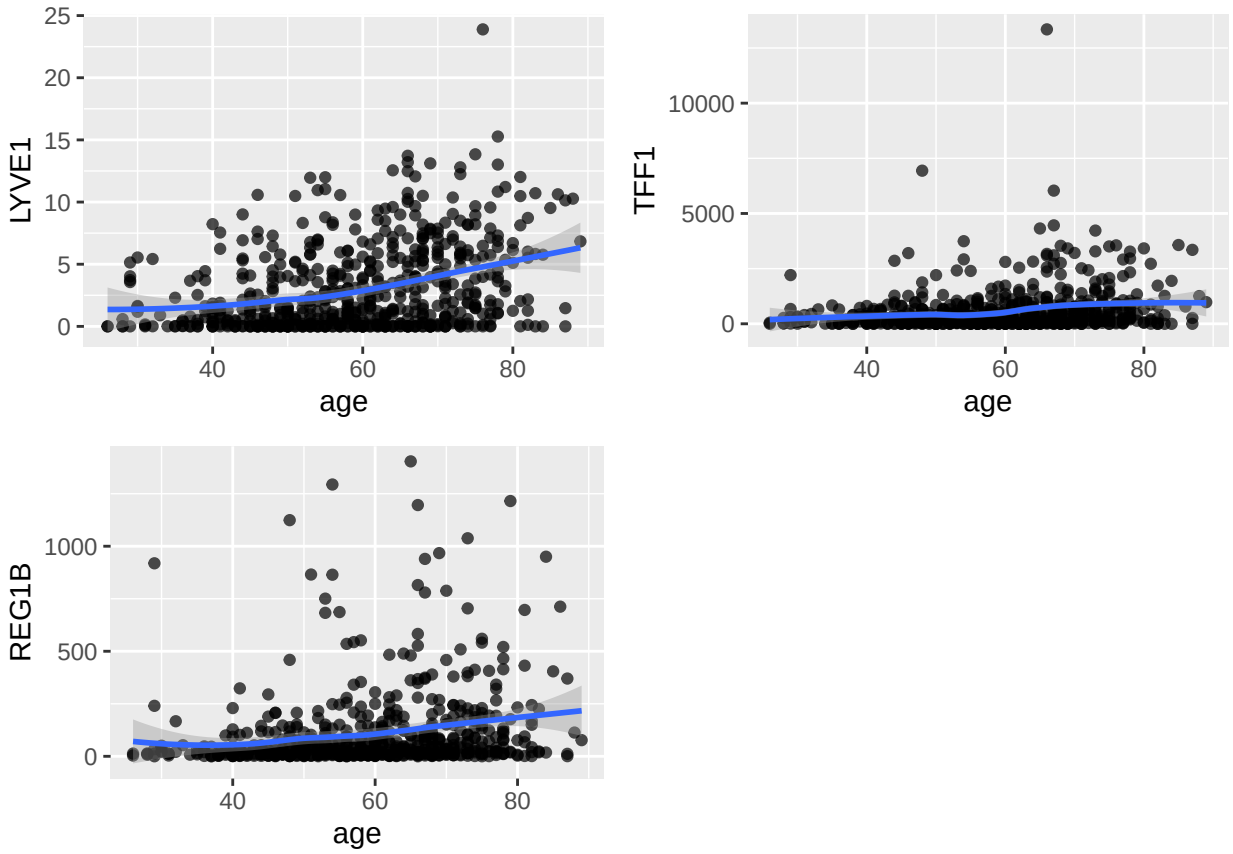
To find out the most prominent factor within the bio markers i decided to plot them again individually

```
p1 <- ggplot(data = data, mapping = aes(x = stage, y = scale_min_max(LYVE1), fill = stage)) +
        geom_boxplot(alpha = 0.7)

p2 <- ggplot(data = data, mapping = aes(x = stage, y = scale_min_max(TFF1), fill = stage)) +
        geom_boxplot(alpha = 0.7)

p3 <- ggplot(data = data, mapping = aes(x = stage, y = scale_min_max(REG1B), fill = stage)) +
        geom_boxplot(alpha = 0.7)

grid.arrange(p1, p2, p3, nrow = 2)
```



After this I wanted to know if age might be an important factor to the output of the bio markers so i tried a plot and added a trend line.

```
p1 <- ggplot(data = data, mapping = aes(x = age, y = LYVE1)) +
        geom_point(alpha = 0.7) +
        geom_smooth(method = "loess", formula = "y ~ x")
p2 <- ggplot(data = data, mapping = aes(x = age, y = TFF1)) +
        geom_point(alpha = 0.7) +
        geom_smooth(method = "loess", formula = "y ~ x")
p3 <- ggplot(data = data, mapping = aes(x = age, y = REG1B)) +
        geom_point(alpha = 0.7) +
        geom_smooth(method = "loess", formula = "y ~ x")

grid.arrange(p1, p2, p3, nrow = 2)
```



In these plots I concluded that age indeed has an effect on the outcome of the 'LYVE1' bio marker. The other 2 also had a ascending trend line but they show less of an increase that LYVE1.

There were also creatinine, and blood plasma factors within the data.
In order to find more correlations I decided to create a heat map to find more correlation within these variables.

```
selection <- c("age", "plasma_CA19_9", "creatinine", "LYVE1", "TFF1", "REG1B", "REG1A")

tmp <- data %>% select(selection) %>% drop_na()
```

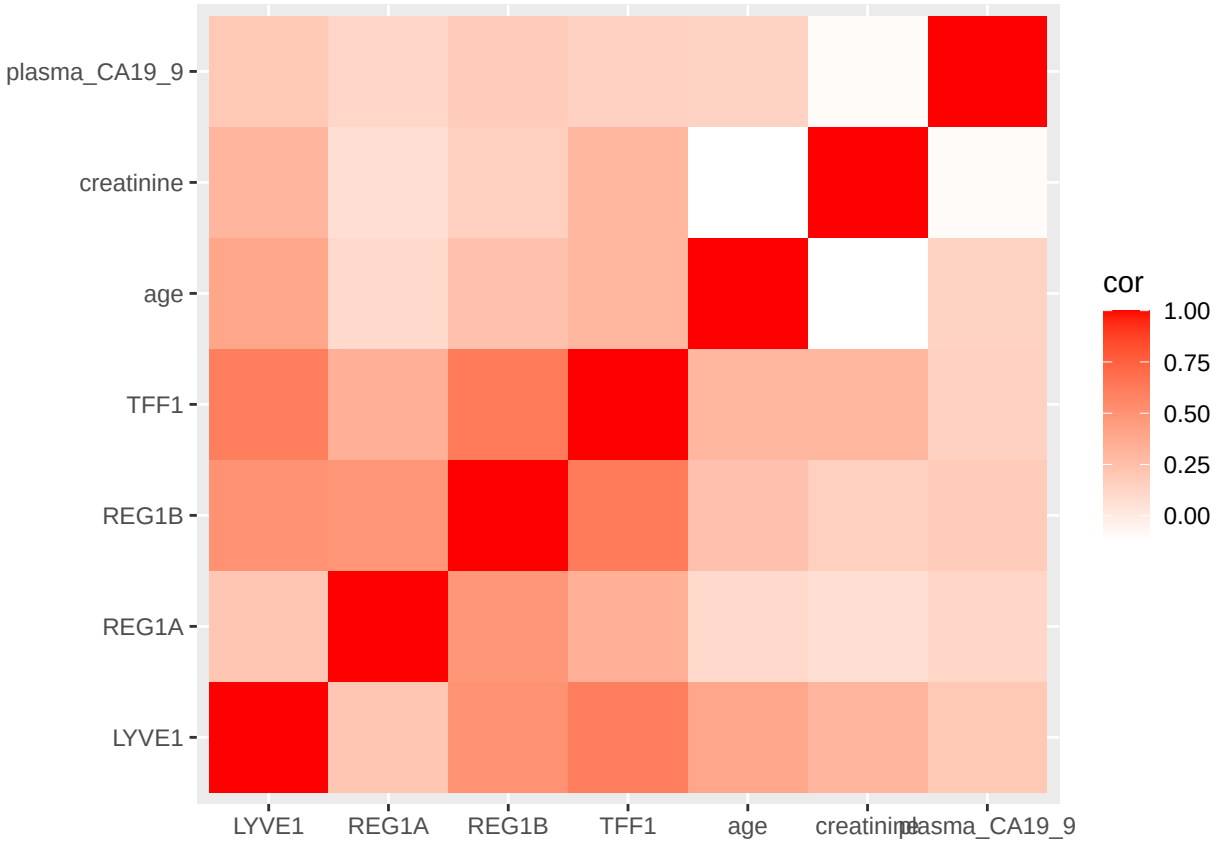
```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(selection)` instead of `selection` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
cor_matrix <- cor(tmp)

cor_matrix <- as_tibble(cor_matrix)
(cor_matrix <- cor_matrix %>% mutate(var1=selection) %>% select(8,1:7))
```

```
## # A tibble: 7 x 8
##   var1             age plasma_CA19_9 creatinine LYVE1  TFF1 REG1B  REG1A
##   <chr>          <dbl>         <dbl>      <dbl> <dbl> <dbl> <dbl>  <dbl>
## 1 age            1            0.139     -0.119  0.393 0.302 0.241 0.104
## 2 plasma_CA19_9  0.139        1         -0.0952 0.198 0.148 0.182 0.123
## 3 creatinine    -0.119       -0.0952     1      0.308 0.299 0.159 0.0689
## 4 LYVE1          0.393        0.198      0.308  1     0.611 0.506 0.214
## 5 TFF1           0.302        0.148      0.299  0.611 1     0.627 0.340
```

```
## 6 REG1B          0.241        0.182      0.159  0.506 0.627 1     0.484
## 7 REG1A          0.104        0.123      0.0689 0.214 0.340 0.484 1
```

```
cor_matrix_long <- pivot_longer(data = cor_matrix, cols = selection, names_to = "variable", values_to =


ggplot(data = cor_matrix_long, aes(x=var1, y=variable, fill=cor)) +
    geom_tile() +
    labs(x=NULL, y=NULL) +
    scale_fill_gradient(high = "red", low = "white" )
```

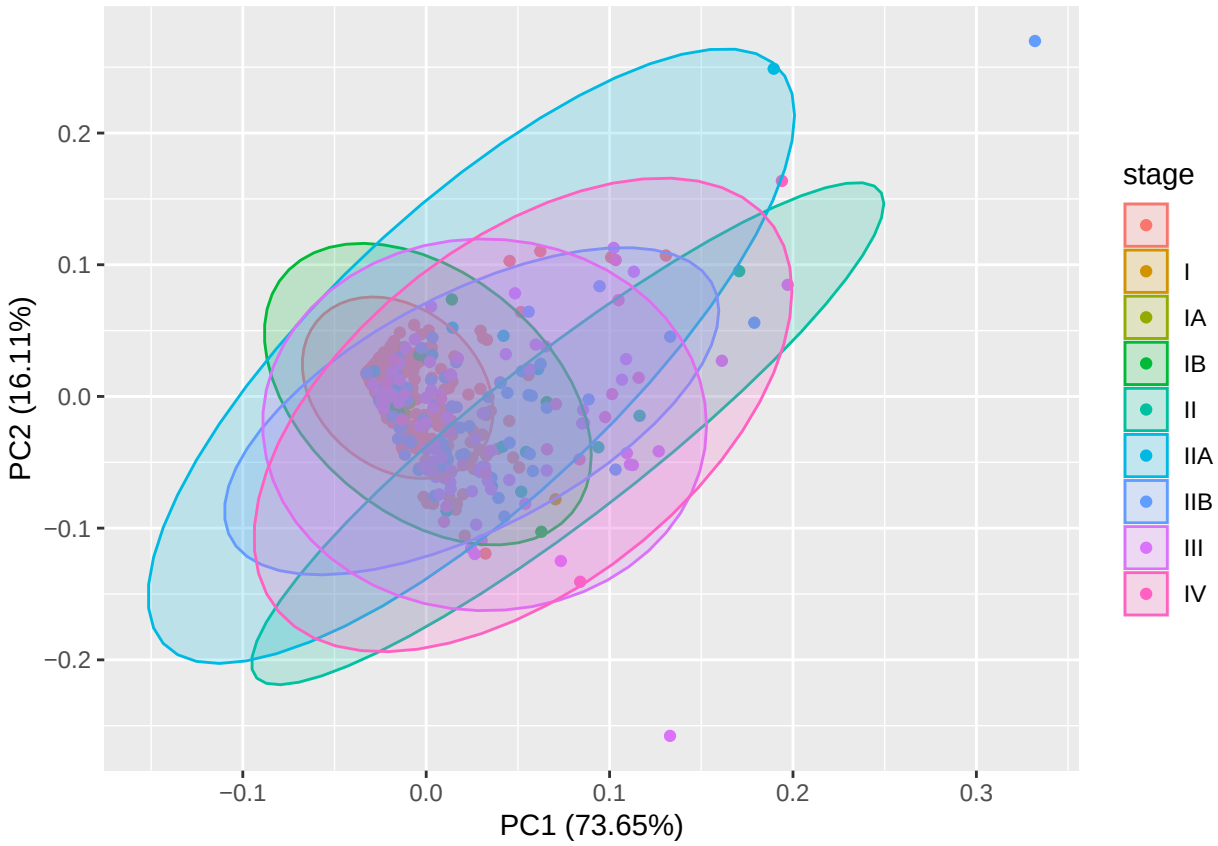


This heat map shows that bio markers have a high correlation, TFF1 and LYVE1 look like they have the highest correlation. LYVE1 is indeed as predicted very correlated to age and TFF1 and LYVE1 to creatinine. It's also clear that REG1A has been improved considerably and REG1B has a much higher correlation to other factors.

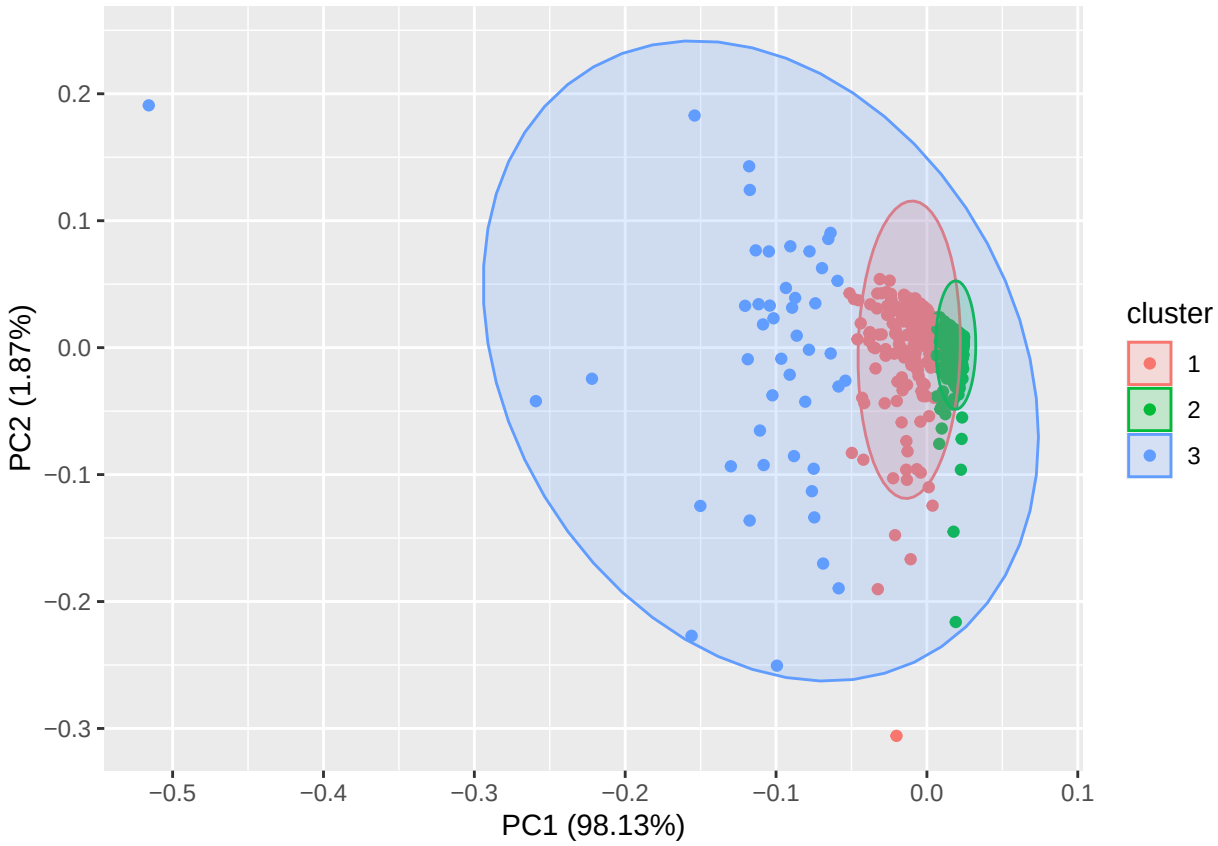
In order to further look at the correlation between the bio markers I decided to create a PCA.

```
df <- data[c(11:13)]
  #clustered by stage
pca_res <- prcomp(df, scale = TRUE)
  autoplot(pca_res, data = data, colour = 'stage', frane = TRUE, frame.type = 'norm')
```

```
## Too few points to calculate an ellipse
## Too few points to calculate an ellipse
```




```
#clustered by bio marker
autoplot(pam(data[11:13], 3), frame = TRUE, frame.type = 'norm')
```



Here I visualized the clusters decided by stage, and the second plot I clustered by bio marker.

Conclusion

I concluded that it should be possible to decide the stage of cancer by less data than given. The 3 bio markers should give a clear prediction and this can be helped with the age factor.