Trabalho 1 de Estatística e Análise de Dados

Sónia Rocha up201704679

Simão Cardoso up201604595

17/04/2021

Este trabalho foi focado na database heart failure clinical records no kaggle, esta contém 299 valores com 13 variáveis, sendo que 6 destas são categóricas.

Inicialmente fazemos a mudança destas variáveis categóricas para fatores na nossa database que anteriormente se encontravam como numéricas (através de str() verificamos os tipos das variáveis), no entanto guardamos a database inicial numa outra variável pois vai ser necessária mais tarde.

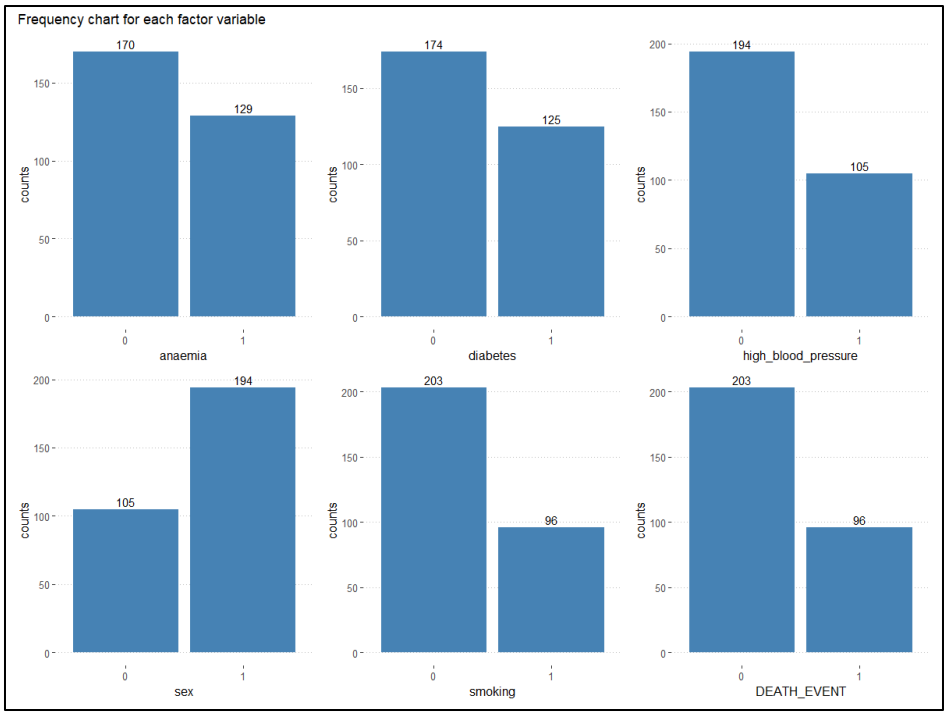
Decidimos após o referido, fazer uma tabela descritiva desta database, para uma melhor compreensão da mesma como podemos mostrar a seguir:

```
str(dataset)
dataset_t <- dataset
dataset <- dataset %>%mutate_at(c("anaemia", "diabetes", "high_blood_pres
sure", "sex", "smoking", "DEATH_EVENT"), as.factor)
```

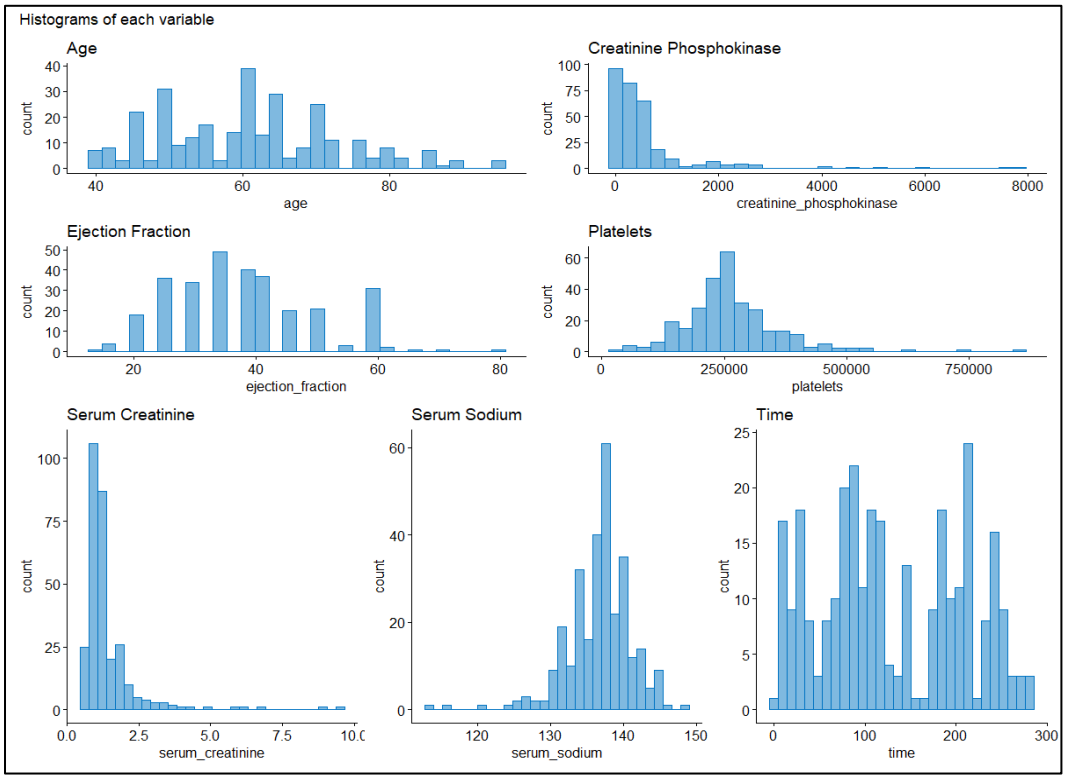
attributes	description
age	Describes the age of the subjects in the dataset
anaemia	Decrease of red blood cells or hemoglobin (boolean); 0 if absent, 1 if present
creatinine_phosphokinase	Level of the CPK enzyme in the blood (mcg/L)
diabetes	If the patient has diabetes (boolean); 0 if absent, 1 if present
ejection_fraction	Percentage of blood leaving the heart at each contraction (percentage)
high_blood_pressure	If the patient has hypertension (boolean); 0 if absent, 1 if present
platelets	Platelets in the blood (kiloplatelets/mL)
serum_creatinine	Level of serum creatinine in the blood (mg/dL)
serum_sodium	Level of serum sodium in the blood (mEq/L)
sex	Woman or man (binary); 0 if woman, 1 if man
smoking	If the patient smokes or not (boolean); 0 if absent, 1 if present
time	Follow up peroid (days)
DEATH_EVENT	If the patient deceased during the follow-up period (boolean); 0 if no, 1 if yes.

Univariate analysis:

De seguida, realizamos bar plots para todas as variáveis categóricas para saber a frequência de cada uma na dataset.



Realizamos também histogramas das variáveis contínuas:



Observação para: creatinine_phosphokinase

Podemos observar no plot desta variável que contém muitos pacientes em zero. Também podemos concluir que a maior parte destes pacientes têm um nível normal de creatinine phosphokinase , pois a normalidade varia entre 10-120. Observamos alguns outliers que serão analisados mais tarde.

Observação para: Platelets

Esta variável parece ter uma distribuição aproximadamente normal com uma média aproximada dos 250000. The Platelets distribution looks fairly like a normal distribution with a mean of around 250000 and a std of 100000. Observamos alguns outliers que serão analisados mais tarde.

Observação para: Serum Creatinine

Esta variável tem uma distribuição bastante enviesada para a esquerda e tem também alguns outliers.

Observação para: Serum Sodium

Sabendo que a normalidade de serum sodio na população é entre 135-145, concluimos que a maior parte dos pacientes analisados têm um nível normal.

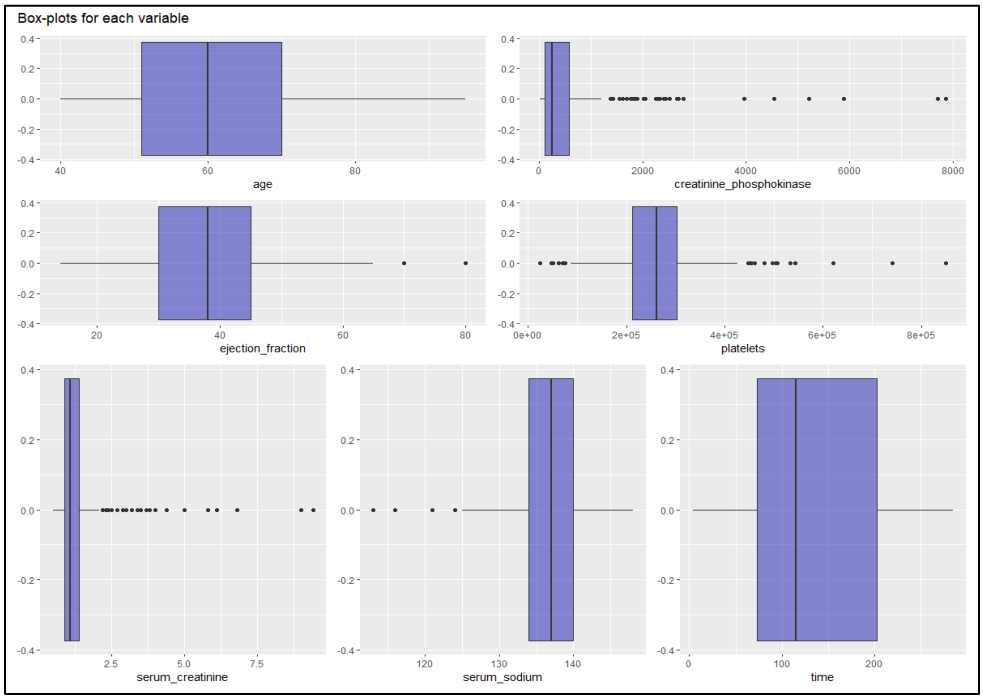
Observação para: Ejection Fraction

Sabendo que a normalidade de ejection fraction está entre 50-70%, podemos observar que a maior parte dos pacientes analisados têm valores inferiores a 50%. Também é possível ver alguns valores pequenos o que poderão ser outliers que serão confirmados mais tarde.

Observação para: Time

Observa-se um gráfico bastante inconsistente, com muita variedade de valores.

De seguida realizamos box-plots para estas mesmas variáveis para ter uma mais fácil perspetiva da sua distribuição e quais os teus outliers.



Através dos mesmos temos uma maior visualização os outliers que se encontram em muitas das variáveis e por isso foram analisadas com maior pormenor:

```
boxplot.stats(dataset$age)$out
```

```
## numeric(0)
```

```
boxplot.stats(dataset$creatinine_phosphokinase)$out
```

```
##  [1] 7861 2656 1380 3964 7702 5882 5209 1876 1808 4540 1548 1610 2261
1846 2334
## [16] 2442 3966 1419 1896 1767 2281 2794 2017 2522 2695 1688 1820 2060
2413
```

```
boxplot.stats(dataset$ejection_fraction)$out
```

```
## [1] 80 70
```

```
boxplot.stats(dataset$platelets)$out
```

```
##  [1] 454000  47000 451000 461000 497000 621000 850000 507000 448000  7
5000
## [11]  70000  73000 481000 504000  62000 533000  25100 451000  51000 54
3000
## [21] 742000
```

```
boxplot.stats(dataset$serum_creatinine)$out
```

```
##  [1] 2.7 9.4 4.0 5.8 3.0 3.5 2.3 3.0 4.4 6.8 2.2 2.7 2.3 2.9 2.5 2.3 3
.2 3.7 3.4
## [20] 6.1 2.5 2.4 2.5 3.5 9.0 5.0 2.4 2.7 3.8
```

Assim, concluímos que a variável age não tem outliers, confirmando que creatinine phosphokinase,platelets e sérum creatinine têm muitos outliers, e que a ejection fraction tem apenas 2.

Após as representações gráficas e análise de outliers, fizemos a análise da posição, dispersão, e forma das variáveis continuas guardadas na variável cont_dataset.

```
cont_dataset <- dataset[,-c(2,4,6,10,11,13)]

table_describe <- describe(cont_dataset, IQR=TRUE,quant=c(.1,.25,.75,.90)
)
coef <- cont_dataset %>%
            gather("variable", "value") %>%
            group_by(variable) %>%
            summarize(
            coef_var = cv(value))

table_describe <- mutate(table_describe ,coef)
table_describe <- table_describe[,-c(1,19)]
```

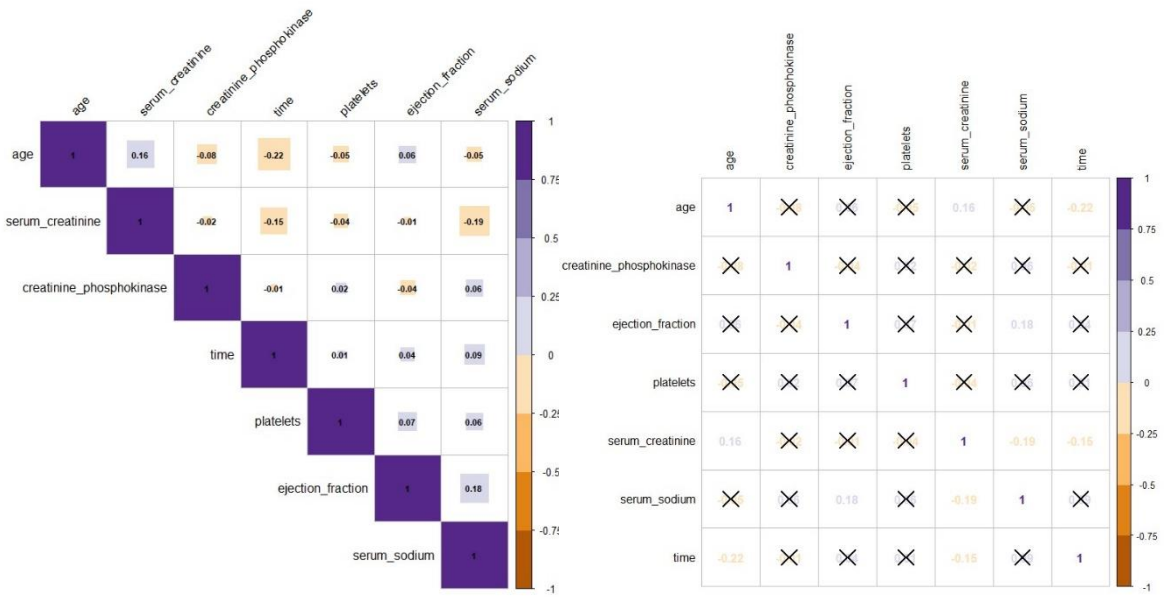
	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	IQR	Q0.1	Q0.25	Q0.75	Q0.9	coef_var
age	299	60.83389	11.894809	60.0	6.021715e+01	14.82600	40.0	95.0	55.0	0.4188266	-0.2204793	0.6878946	19.0	45.0	51.0	70.0	75.4	0.1955293
creatinine_phosphokinase	299	581.83946	970.287881	250.0	3.654938e+02	269.83320	23.0	7861.0	7838.0	4.4184296	24.5254138	56.1131970	465.5	67.6	116.5	582.0	1203.8	1.6676213
ejection_fraction	299	38.08361	11.834841	38.0	3.742739e+01	11.86080	14.0	80.0	66.0	0.5498228	0.0005484	0.6844265	15.0	25.0	30.0	45.0	60.0	0.3107594
platelets	299	263358.02926	97804.236869	262000.0	2.567301e+05	65234.40000	25100.0	850000.0	824900.0	1.4476814	6.0252322	5656.1650591	91000.0	153000.0	212500.0	303500.0	374600.0	0.3713737
serum_creatinine	299	1.39388	1.034510	1.1	1.189295e+00	0.29652	0.5	9.4	8.9	4.4113866	25.1888415	0.0598273	0.5	0.8	0.9	1.4	2.1	0.7421804
serum_sodium	299	136.62542	4.412477	137.0	1.368216e+02	4.44780	113.0	148.0	35.0	-1.0376430	3.9841899	0.2551802	6.0	132.0	134.0	140.0	141.2	0.0322962
time	299	130.26087	77.614208	115.0	1.292780e+02	105.26460	4.0	285.0	281.0	0.1265232	-1.2238150	4.4885455	130.0	26.8	73.0	203.0	244.0	0.5958367

Olhando para a Skewness e Kurtosis das nossas variáveis, podemos concluir que tanto a 'creatine_phosphokinase', 'platelets', 'serum_creatine' e 'serum_sodium' são bastante enviesados e contêm uma distribuição não normal. As restantes variáveis analisadas contêm valores que mais se adequam a uma distribuição normal.

Bivariate analysis:

Após a univariate analysis, seguimos para a bivariate analysis, começando por analisar as correlações entre as variáveis contínuas. No plot que vai ser apresentado de seguida, é de notar que correlações positivas estão a roxo e as negativas a amarelo. A intensidade da cor e o tamanho do quadrado são proporcionais ao coeficiente de correlação.

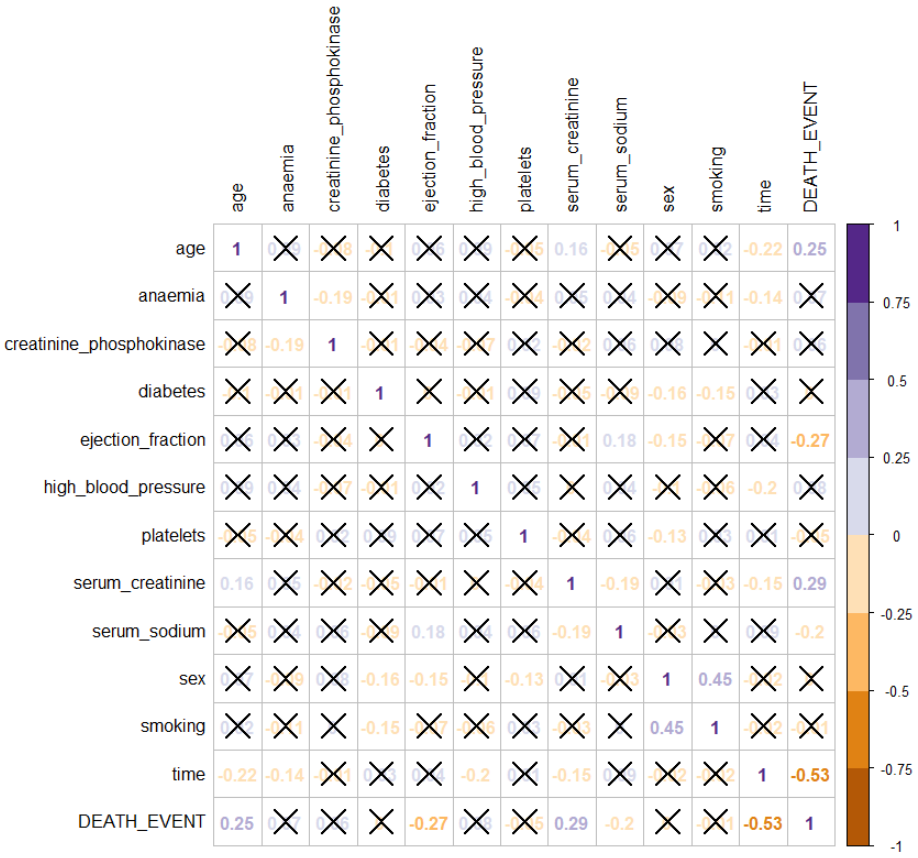
```
corrplot(cor(cont_dataset), type = "upper",method = "square", order = "hc
lust",tl.col = "black", tl.srt = 45,addCoef.col = TRUE,number.cex = .7,
         col = brewer.pal(n = 8, name = "PuOr"))

corrplot(cor(cont_dataset), method = "number", tl.col = "black",p.mat = c
or.mtest(cont_dataset)$p,sig.level = 0.05,col = brewer.pal(n = 8, name =
"PuOr"))
```



Assim, concluímos que as variáveis não têm uma forte correlação entre si, no entanto, a um nível de significância de 0.05, podemos concluir que a variável age tem uma correlação positiva com a serum creatinine e uma correlação negativa com o time. O ejection fraction tem uma correlação positiva com serum sodio. O serum creatinine uma correlação positiva com a age e negativas com serum sodio e time. Serum sodio tem uma correlação positiva com ejection fraction e negativa com serum creatinine. Por fim, tem correlações negativas com age e serum creatinine.

Para uma compreensão mais completa da correlação entre as diferentes variáveis, foi feito outro plot de correlação, mas agora entre as variáveis contínuas e categóricas. É de notar que foi usado a database inicial pois não é permitido variáveis do tipo factor neste tipo de plot.

```
corrplot(cor(dataset_t), method = "number", tl.col = "black",p.mat = cor.
mtest(dataset_t)$p,sig.level = 0.05,col = brewer.pal(n = 8, name = "PuOr"
))
```



Verificamos então, que a um nível de significância de 0.05, age tem uma maior correlação com o fator death (sendo esta positiva), creatinine phosphokinase uma correlação negativa com o fator anaemia, ejection fraction duas correlações negativas com os fatores sex e death, platelets tem correlações pequenas com todos os fatores, serum creatinine tem uma maior correlação com death (positiva), serum sodio correlação negativa com o fator death, e por fim, time tem uma correlações negativas com os fatores high blood pressure, anaemia e death.

Estas correlações vão ser informações importantes e serão utilizadas mais adiante.

Fizemos algumas contingency tables para alguns pares de variáveis categóricas que achamos relevantes:

```
sex_diab <- as.data.frame.matrix(table(dataset$sex,dataset$diabetes))
colnames(sex_diab) <- c("absent diabetes","present diabetes")
rownames(sex_diab) <- c("woman","man")

death_smk <- as.data.frame.matrix(table(dataset$DEATH_EVENT,dataset$smoki
ng))
colnames(death_smk) <- c("survived","dead")
rownames(death_smk) <- c("absent smoking","present smoking")
ana_hbp <- as.data.frame.matrix(table(dataset$anaemia,dataset$high_blood_
pressure))
colnames(ana_hbp) <- c("absent anaemia","present anaemia")
rownames(ana_hbp) <- c("absent high blood pressure","present high blood p
ressure")

sex_diab %>% kable(caption = "Number of man and woman that have or not di
abetes") %>% kable_styling()
death_smk %>% kable(caption = "Death and smoking ratio") %>% kable_stylin
g()
ana_hbp %>% kable(caption = "Anaemia and high blood pressure ratio") %>%
kable_styling()
```

Number of man and woman that have or not diabetes

	absent diabetes	present diabetes
woman	50	55
man	124	70

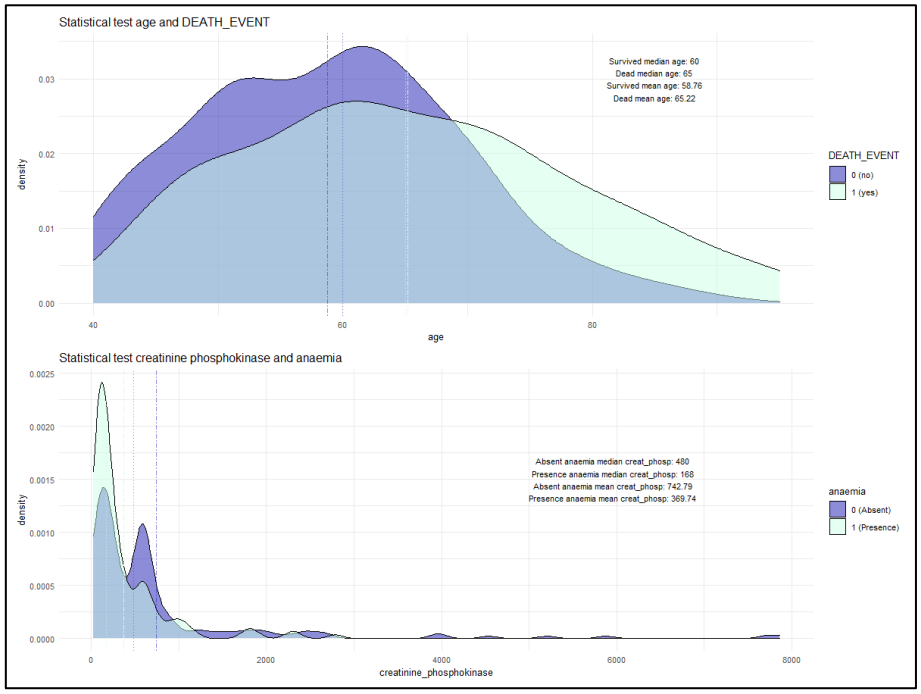
Death and smoking ratio

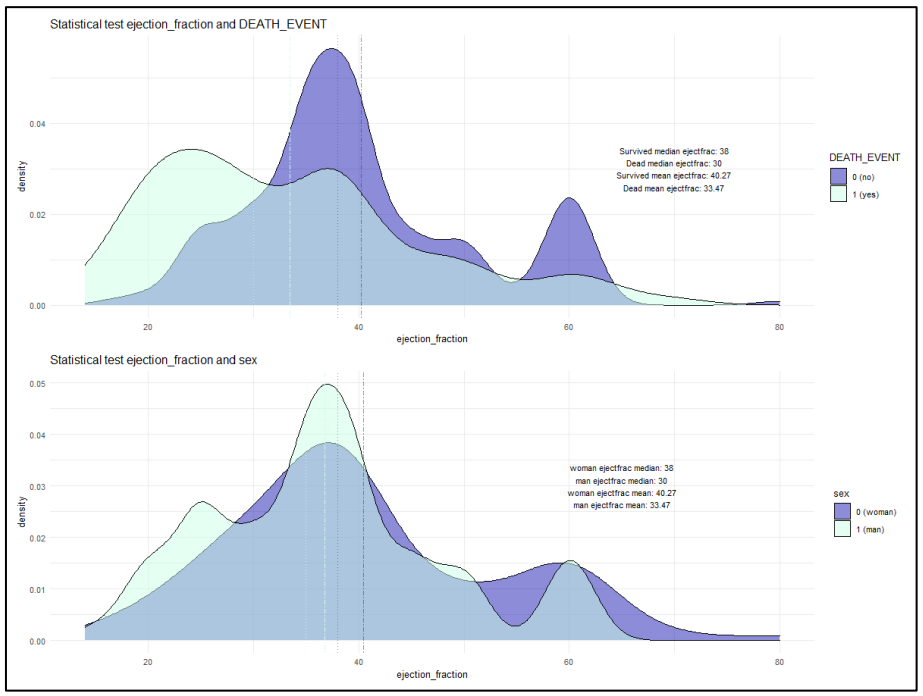
	survived	dead
absent smoking	137	66
present smoking	66	30

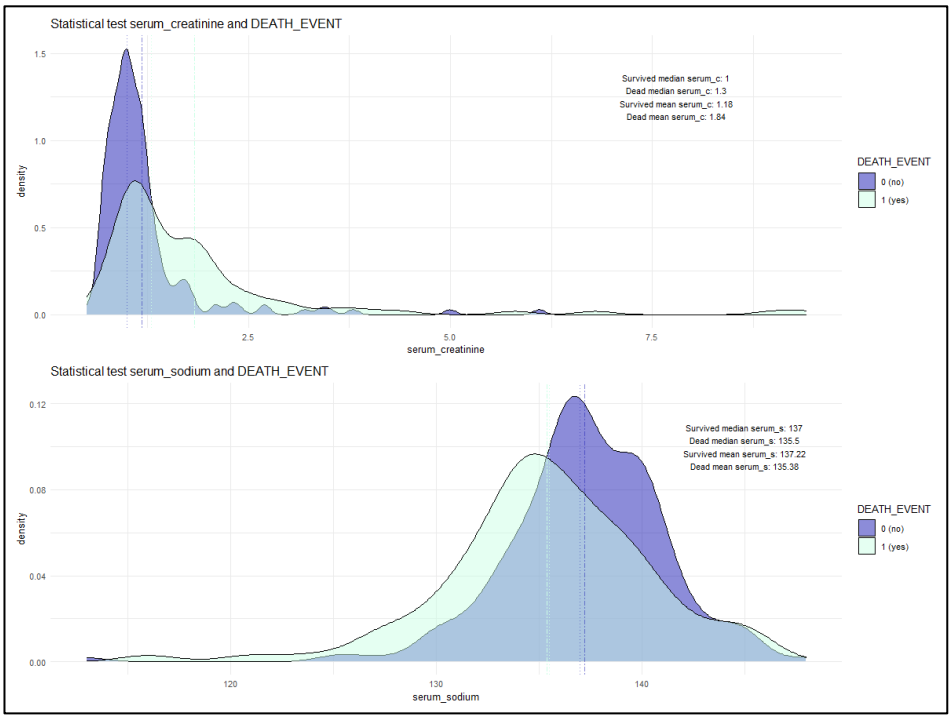
Anaemia and high blood pressure ratio

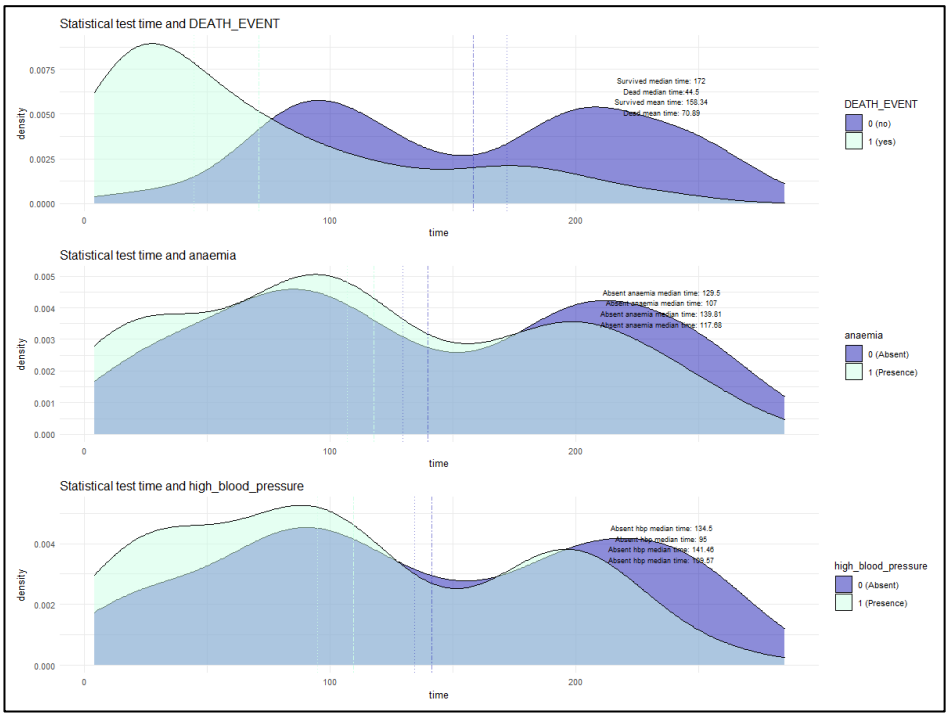
	absent anaemia	present anaemia
absent high blood pressure	113	57
present high blood pressure	81	48

De seguida, realizamos alguns testes estatísticos do comportamento de algumas variáveis contínuas com os fatores que mais têm correlação (que já tinham sido indicamos acima), nestes gráficos calculamos a média e a mediana.






Statistical test serum_creatinine and DEATH_EVENT
Survived median serum_c: 1
Dead median serum_c: 1.3
Survived mean serum_c: 1.18
Dead mean serum_c: 1.84
density
serum_creatinine
DEATH_EVENT
0 (no)
1 (yes)
Statistical test serum_sodium and DEATH_EVENT
Survived median serum_s: 137
Dead median serum_s: 135.5
Survived mean serum_s: 137.22
Dead mean serum_s: 135.38
density
serum_sodium
DEATH_EVENT
0 (no)
1 (yes)


Statistical test time and DEATH_EVENT
Survived median time: 172
Dead median time:44.5
Survived mean time: 158.34
Dead mean time: 70.89
density
time
DEATH_EVENT
0 (no)
1 (yes)
Statistical test time and anaemia
Absent anaemia median time: 129.5
Absent anaemia median time: 107
Absent anaemia median time: 139.81
Absent anaemia median time: 117.68
density
time
anaemia
0 (Absent)
1 (Presence)
Statistical test time and high_blood_pressure
Absent hbp median time: 134.5
Absent hbp median time: 95
Absent hbp median time: 141.46
Absent hbp median time: 103.57
density
time
high_blood_pressure
0 (Absent)
1 (Presence)

Multivariate analysis:

Agora que temos um bom conhecimento da distribuição da data e relações, tentamos aplicar o método PCA nas variáveis contínuas para melhores visualizações. Usamos PCA para reduzir a dimensionalidade da nossa database, esta dimensão é feita capturando a variância da mesma.

Começando por aplicar o método, calculamos os eigen values e os respetivos eigen_vectors:

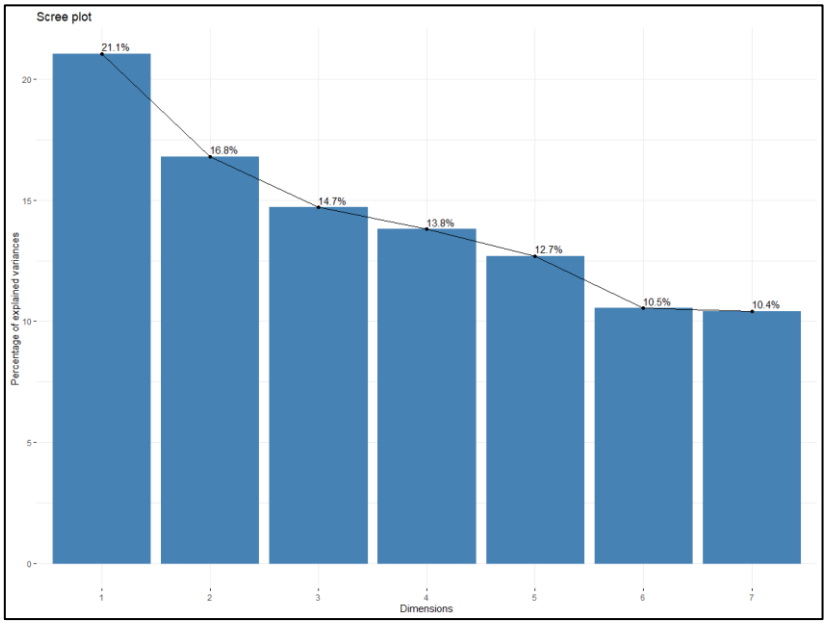
```
PCA_data <-PCA(cont_dataset,graph = FALSE)
summary(PCA_data)

get_eig(PCA_data)#Eigen values

##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  1.4745726         21.06532                    21.06532
## Dim.2  1.1755914         16.79416                    37.85949
## Dim.3  1.0294792         14.70685                    52.56633
## Dim.4  0.9662257         13.80322                    66.36956
## Dim.5  0.8877341         12.68192                    79.05147
## Dim.6  0.7374427         10.53490                    89.58637
## Dim.7  0.7289544         10.41363                   100.00000

PCA_data$svd$V # Eigen vectors

##            [,1]        [,2]        [,3]        [,4]       [,5]
## [1,] -0.4649617  0.45213222 -0.00779977 -0.19809211 -0.1912135
## [2,]  0.1379593 -0.19389349  0.81505355 -0.33440577  0.2948224
## [3,]  0.1788924  0.68147830 -0.10671326 -0.01299509  0.4694857
## [4,]  0.1992576  0.24678636  0.40331735  0.82095373 -0.1807563
## [5,] -0.5117770  0.04569638  0.10167226  0.18226520  0.6335802
## [6,]  0.4474108  0.42971962  0.11797610 -0.36260682 -0.1513990
## [7,]  0.4806034 -0.21428597 -0.37056533  0.10046937  0.4461860
```

Obetemos também as coordenadas, contribuições e qualidade de representação tanto de indíviduos como variáveis, não iremos mostrar nesta análise o print que proporciona na consola pois é bastante extenso.

```
res.var <- get_pca_var(PCA_data)
res.var$coord          # Coordinates
res.var$contrib        # Contributions to the PCs
res.var$cos2           # Quality of representation
# Results for individuals
res.ind <- get_pca_ind(PCA_data)
res.ind$coord          # Coordinates
res.ind$contrib        # Contributions to the PCs
res.ind$cos2           # Quality of representation
#########################
```

Após estes resultados, para uma mais fácil demonstração, fizemos os seguintes gráficos:

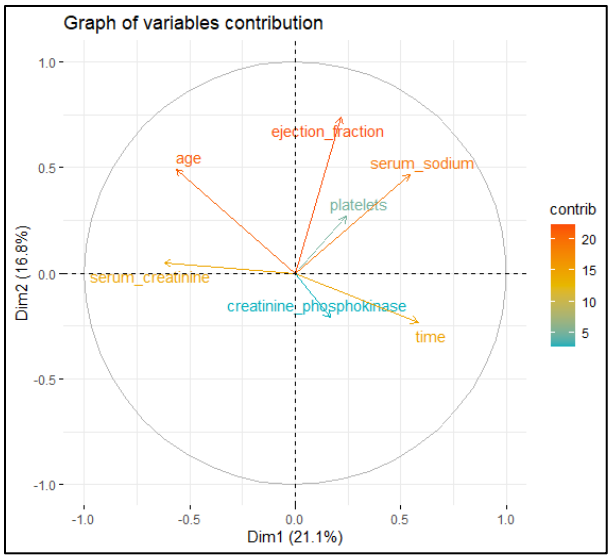
```
fviz_eig(PCA_data, addlabels=TRUE) #Plot the eigenvalues/variances against the number of dimensions
```

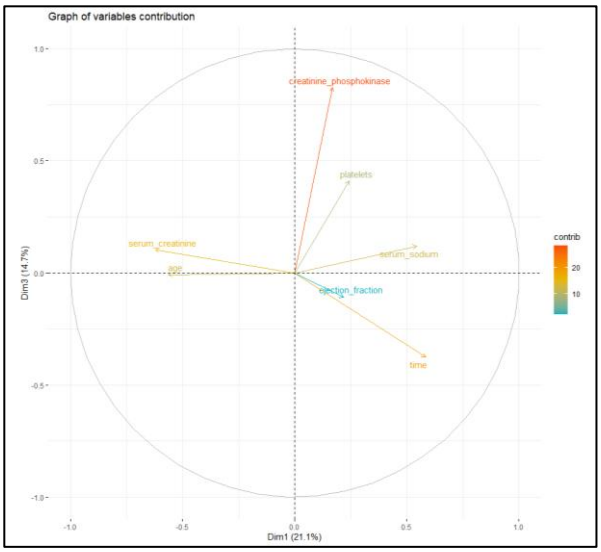


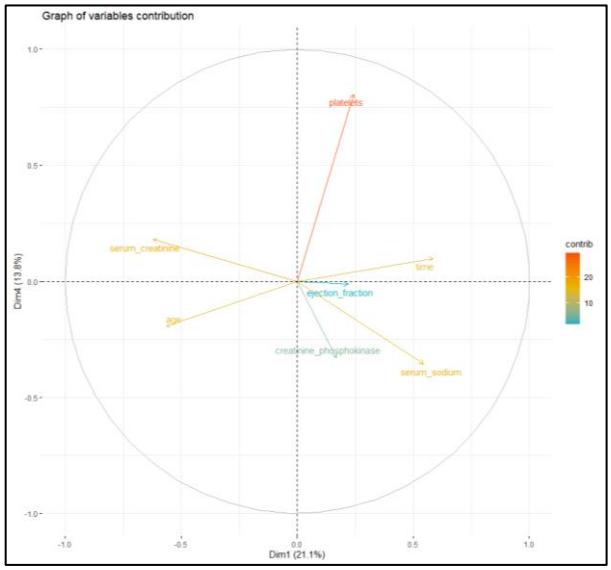
Assim, como podemos ver no plot acima iremos parar a análise na 5º dimensão pois, 76% da informação (variâncias) da data, estão retidas nos primeiros 5 principal components.

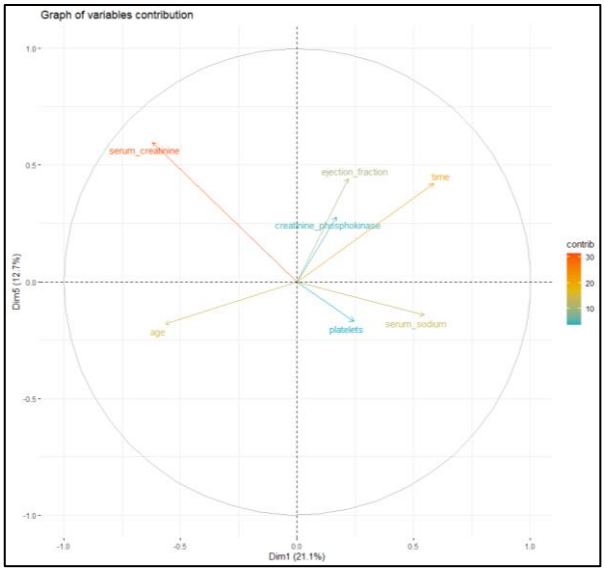
De seguida fazemos um gráfico com as variáveis nas dimensões 1-2 , 1-3, 1-4 e 1-5 relacionando o nível de contribuição das mesmas, abaixo está um excerto do código utilizado para um dos plots.

```
#Graph of variables DIM 1 e 2
fviz_pca_var(PCA_data,
             col.var = "contrib", # Color by contributions to the PC
             title = "Graph of variables contribution",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```







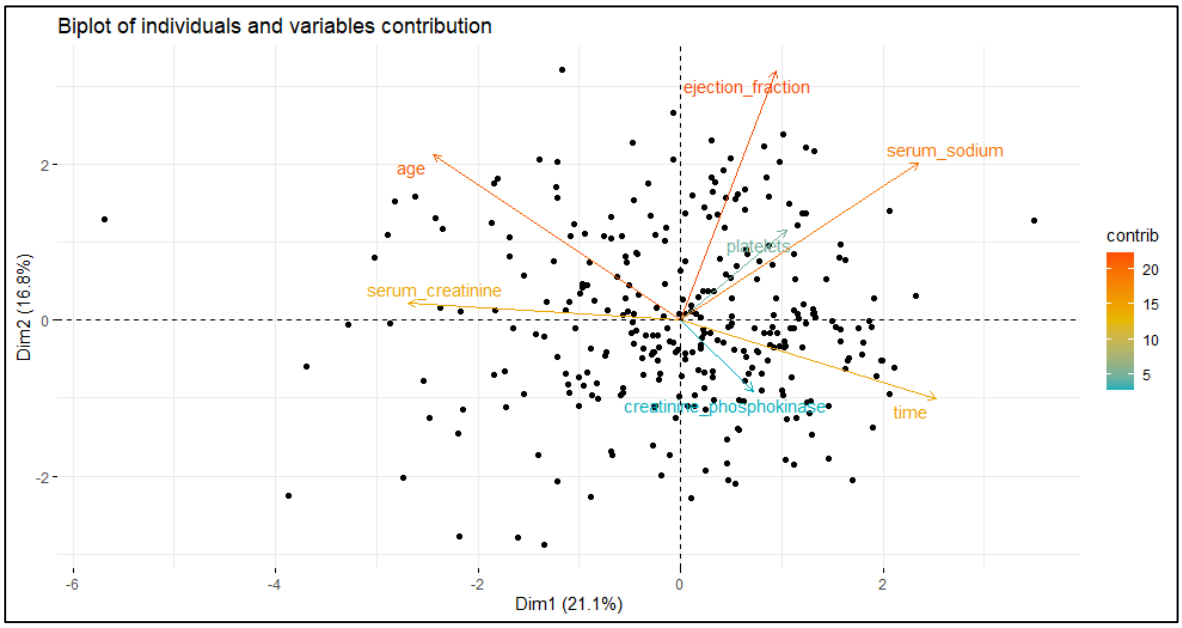


Pelos diferentes plots, podemos concluir distribuições de contribuição muito distintas, na dimensão 1-2 ejection faction é o que contêm mais contribuição, e creatine phosphokinase menor. Na dimensão 1-3 creatine phosphokinase é a que tem maior contribuição e ejection fraction menor. Na dimensão 1-4 platelets é a que tem maior contribuição e ejection fraction menor. E por fim, na dimensão 1-5 serum creatinine é a que tem maior contribuição e platelets menor.

Concluímos então que nas 5 dimensões a variável time mantém-se com um nível de contribuição médio/baixo.

Fizemos também um biplot para ver os valores e variáveis em conjunto, mas apenas da dimensão 1-2 que é a mais importante.

```
#Biplot of individuals and variables 1 2
fviz_pca_biplot(PCA_data,
             axes= c(1,2),
             label="var",
             title = "Biplot of individuals and variables contribution",
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)
```



Por fim, fizemos um plot para ver a qualidade de representação das variáveis na dimensão 1-2: `fviz_cos2(PCA_data, choice = "var", axes = 1:2)`