Exploratory_Analysis

Alex Wei

10/8/2021

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Complete an initial round of exploratory analyses on your data that would be relevant to your plan and responses above, and include any plots, summaries, code and output. Please include exploratory analysis for outcome(s) of continuous form however/wherever possible even if your ultimate goals/questions involve a different form of outcome data such as binary, polytomous, etc. (You may consider this initial analysis as a potential sub-analysis later on.)

Basic Information of the Data

Data set

```
dat <- read.csv(file = 'heart_failure_clinical_records_dataset.csv')
head(dat)
```

```
##   age anaemia creatinine_phosphokinase diabetes ejection_fraction
## 1  75       0                      582        0                20
## 2  55       0                     7861        0                38
## 3  65       0                      146        0                20
## 4  50       1                      111        0                20
## 5  65       1                      160        1                20
## 6  90       1                       47        0                40
##   high_blood_pressure platelets serum_creatinine serum_sodium sex smoking time
## 1                   1    265000              1.9          130   1       0    4
## 2                   0    263358              1.1          136   1       0    6
## 3                   0    162000              1.3          129   1       1    7
## 4                   0    210000              1.9          137   1       0    7
## 5                   0    327000              2.7          116   0       0    8
## 6                   1    204000              2.1          132   1       1    8
##   DEATH_EVENT
## 1           1
## 2           1
## 3           1
## 4           1
## 5           1
## 6           1
```

Number of Variables

```
ncol(dat)
```

```
## [1] 13
```

```
nrow(dat)
```

```
## [1] 299
```

There are 13 variables and 299 samples in the data set.

Mean and Standard Deviation of continuous variable

Mean and deviation of variable age

```
mean(dat$age)
```

```
## [1] 60.83389
```

```
sd(dat$age)
```

```
## [1] 11.89481
```

Mean and deviation of variable creatinine phosphokinase concentration

```
mean(dat$creatinine_phosphokinase)
```

```
## [1] 581.8395
```

```
sd(dat$creatinine_phosphokinase)
```

```
## [1] 970.2879
```

Mean and deviation of variable ejection fraction

```
mean(dat$ejection_fraction)
```

```
## [1] 38.08361
```

```
sd(dat$ejection_fraction)
```

```
## [1] 11.83484
```

Mean and deviation of variable platelets concentration

```
mean(dat$platelets)
```

```
## [1] 263358
```

```
sd(dat$platelets)
```

```
## [1] 97804.24
```

Mean and deviation of variable serum creatine concentration

```
mean(dat$serum_creatinine)
```

```
## [1] 1.39388
```

```
sd(dat$serum_creatinine)
```

```
## [1] 1.03451
```

Mean and deviation of variable serum sodium concentration

```
mean(dat$serum_sodium)
```

```
## [1] 136.6254
```

```
sd(dat$serum_sodium)
```

```
## [1] 4.412477
```

Frequency of discrete variable

Anaemia frequency

```
nrow(dat)
```

```
## [1] 299
```

```
nrow(filter(dat,anaemia == 0))
```

```
## [1] 170
```

```
nrow(filter(dat,anaemia == 1))
```

```
## [1] 129
```

There are 299 samples in the data set, 170 of them are clear with anaemia, while 129 of them have anaemia.

Diabete frequency

```
nrow(dat)
```

```
## [1] 299
```

```
nrow(filter(dat,diabetes == 0))
```

```
## [1] 174
```

```
nrow(filter(dat,diabetes == 1))
```

```
## [1] 125
```

There are 299 samples in the data set, 174 of them are clear with diabete, while 125 of them have diabete.

High Blood Pressure frequency

```
nrow(dat)
```

```
## [1] 299
```

```
nrow(filter(dat,high_blood_pressure == 0))
```

```
## [1] 194
```

```
nrow(filter(dat,high_blood_pressure == 1))
```

```
## [1] 105
```

There are 299 samples in the data set, 194 of them are clear with high blood pressure, while 105 of them have high blood pressure.

Gender frequency

```
nrow(dat)
```

```
## [1] 299
```

```
nrow(filter(dat,sex == 0))
```

```
## [1] 105
```

```
nrow(filter(dat,sex == 1))
```

```
## [1] 194
```

There are 299 samples in the data set, 105 of them are female, while 194 of them are male.

Smoking condition frequency

```
nrow(dat)
```

```
## [1] 299
```

```
nrow(filter(dat,smoking == 0))
```

```
## [1] 203
```

```
nrow(filter(dat,smoking == 1))
```

```
## [1] 96
```

There are 299 samples in the data set, 203 of them are non-smoker, while 96 of them are smoker.

Death frequency

```
nrow(dat)
```

```
## [1] 299
```

```
nrow(filter(dat,DEATH_EVENT == 0))
```

```
## [1] 203
```

```
nrow(filter(dat,DEATH_EVENT == 1))
```

```
## [1] 96
```

There are 299 samples in the data set, 203 of them are alive during the following up period, while 96 of them died during the following up period.

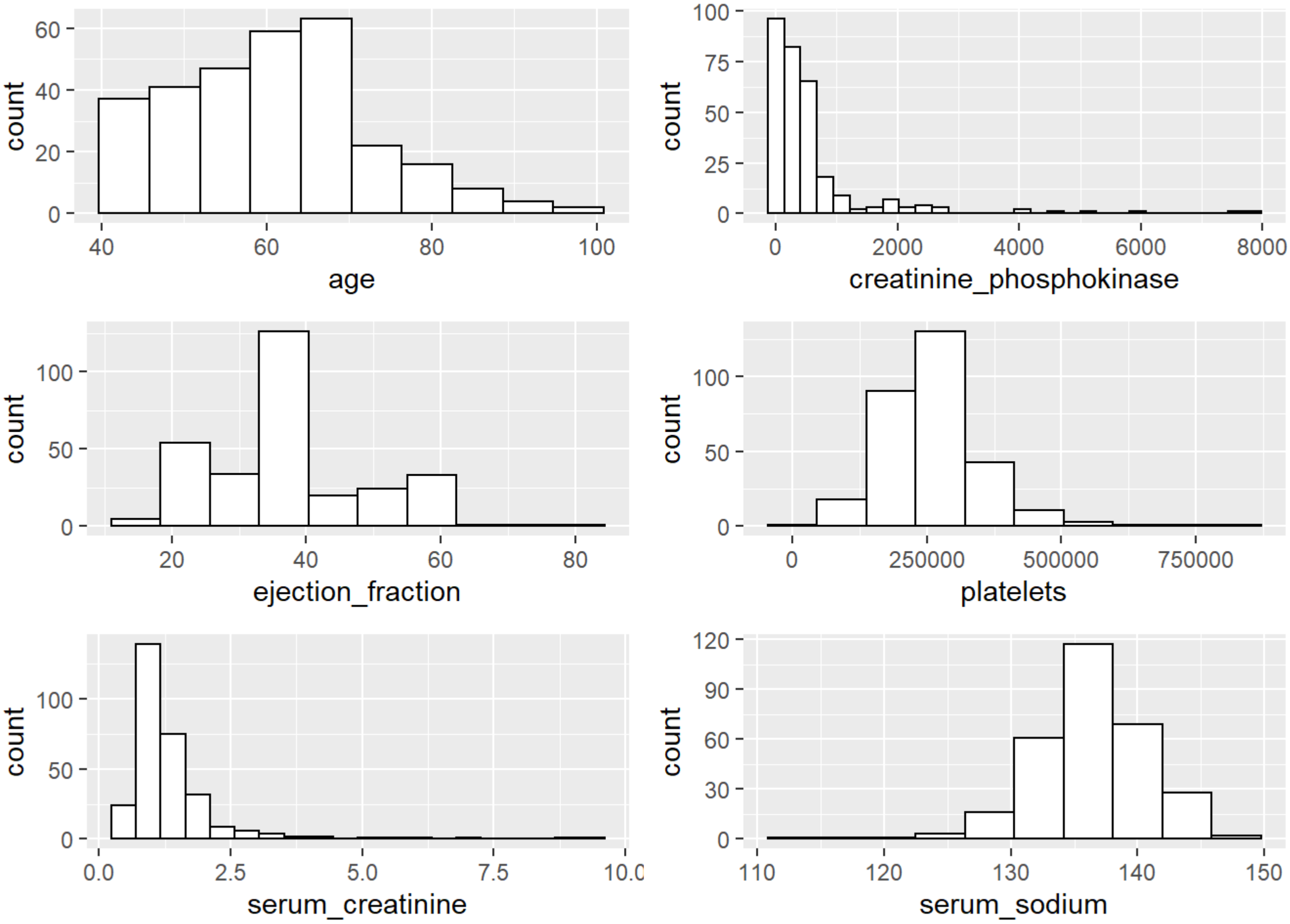
Normality Test

```
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

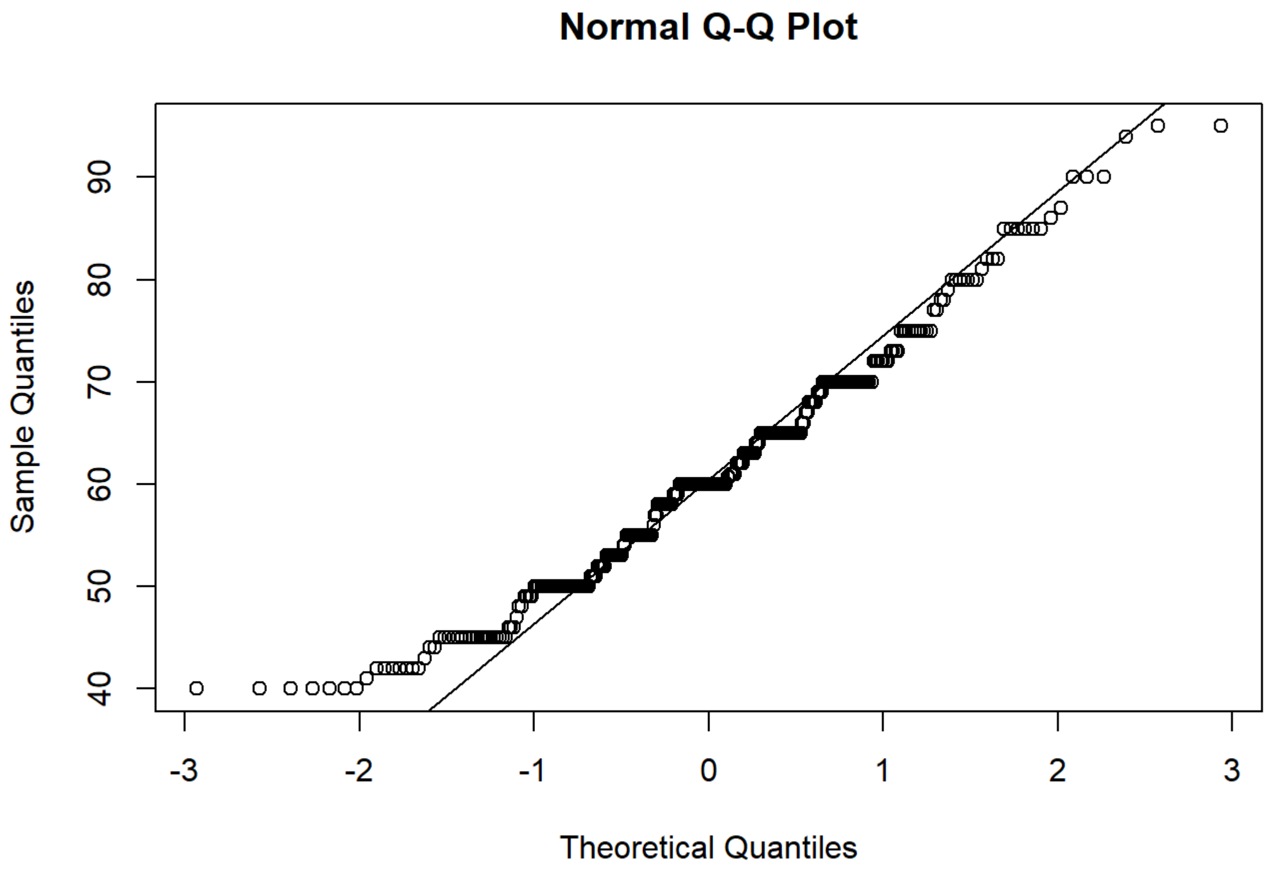
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
age_dis <- ggplot(dat, aes(x=age)) +
              geom_histogram(bins=10,colour="black", fill="white")

crea_pho_dis <- ggplot(dat, aes(x=creatinine_phosphokinase)) +
                  geom_histogram(bins=30,colour="black", fill="white")

ef_dis <- ggplot(dat, aes(x=ejection_fraction)) +
            geom_histogram(bins=10,colour="black", fill="white")

plate_dis <- ggplot(dat, aes(x=platelets)) +
                geom_histogram(bins=10,colour="black", fill="white")

s_crea_dis <- ggplot(dat, aes(x=serum_creatinine)) +
                geom_histogram(bins=20,colour="black", fill="white")

s_sodium_dis <- ggplot(dat, aes(x=serum_sodium)) +
                   geom_histogram(bins=10,colour="black", fill="white")

grid.arrange(age_dis,crea_pho_dis,ef_dis,plate_dis,s_crea_dis,s_sodium_dis)
```



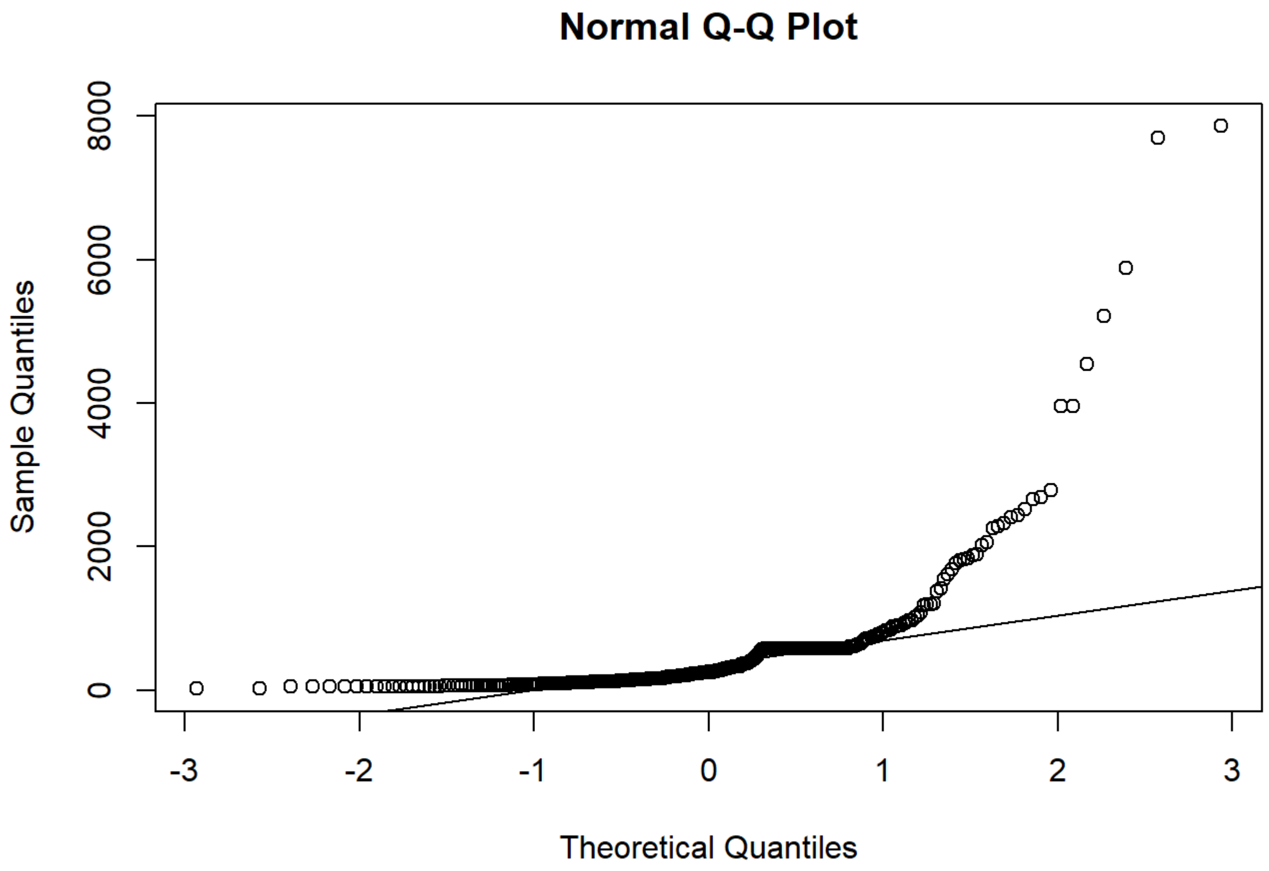
Age QQplot

```
qqnorm(dat$age)
qqline(dat$age)
```



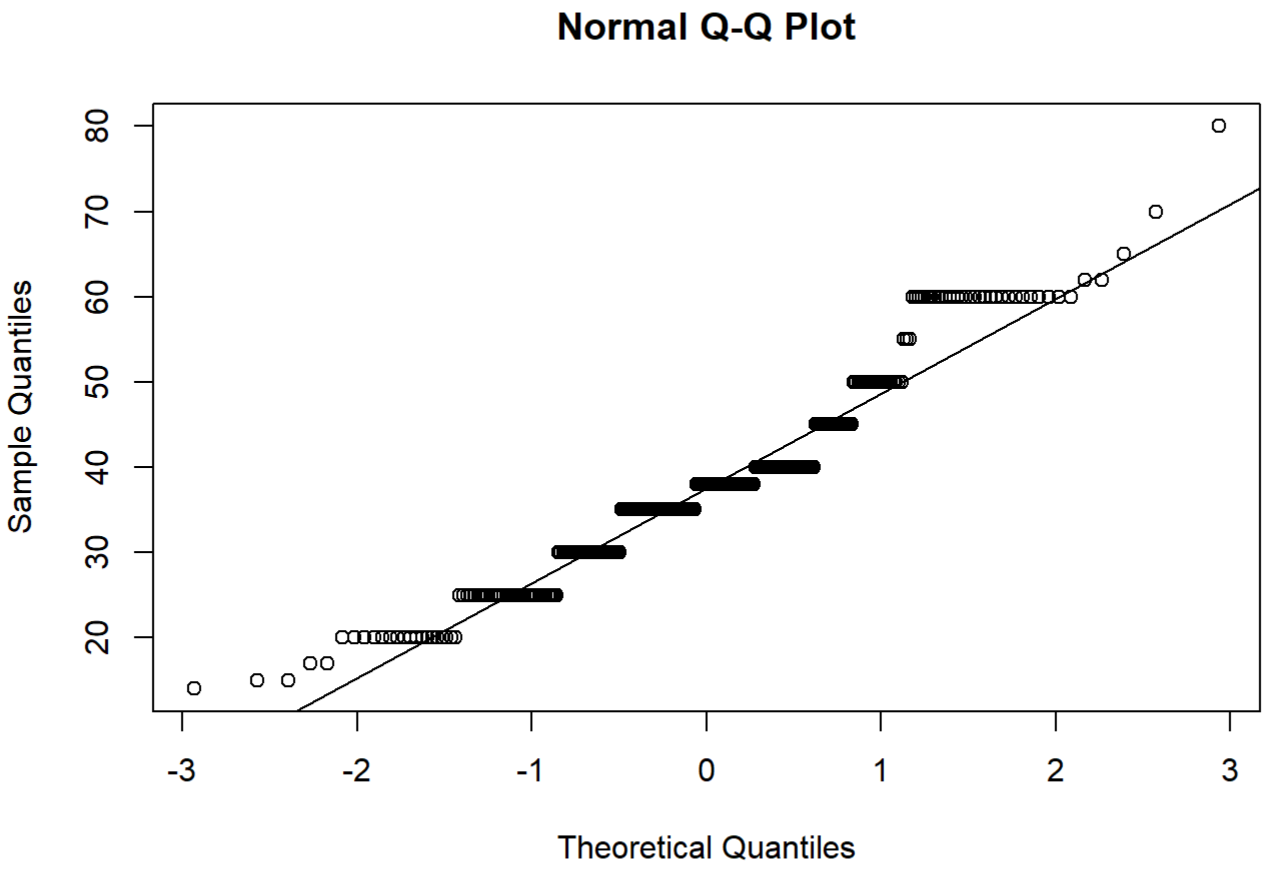
Creatinine Phosphokinase concentration QQplot

```
qqnorm(dat$creatinine_phosphokinase)
qqline(dat$creatinine_phosphokinase)
```



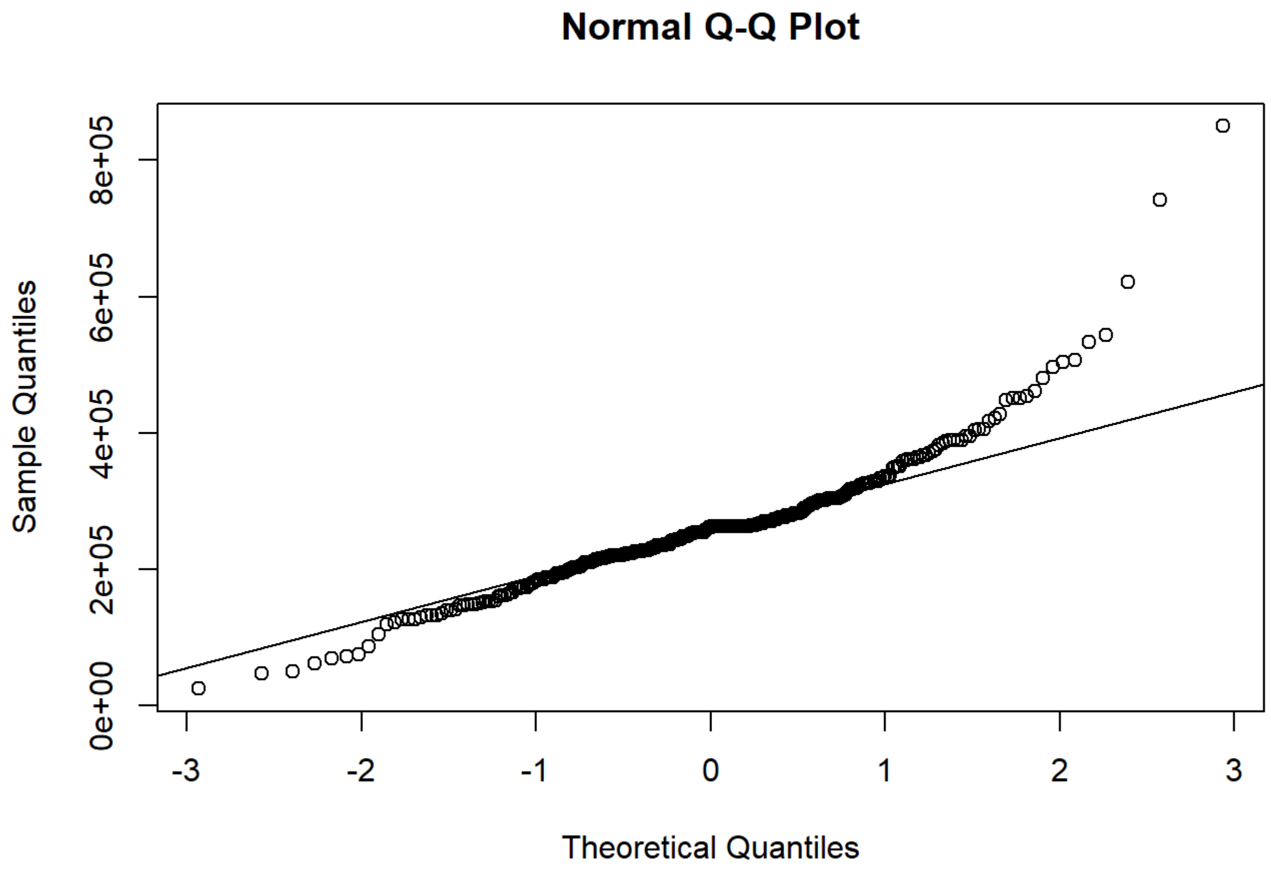
Ejection fraction QQplot

```
qqnorm(dat$ejection_fraction)
qqline(dat$ejection_fraction)
```



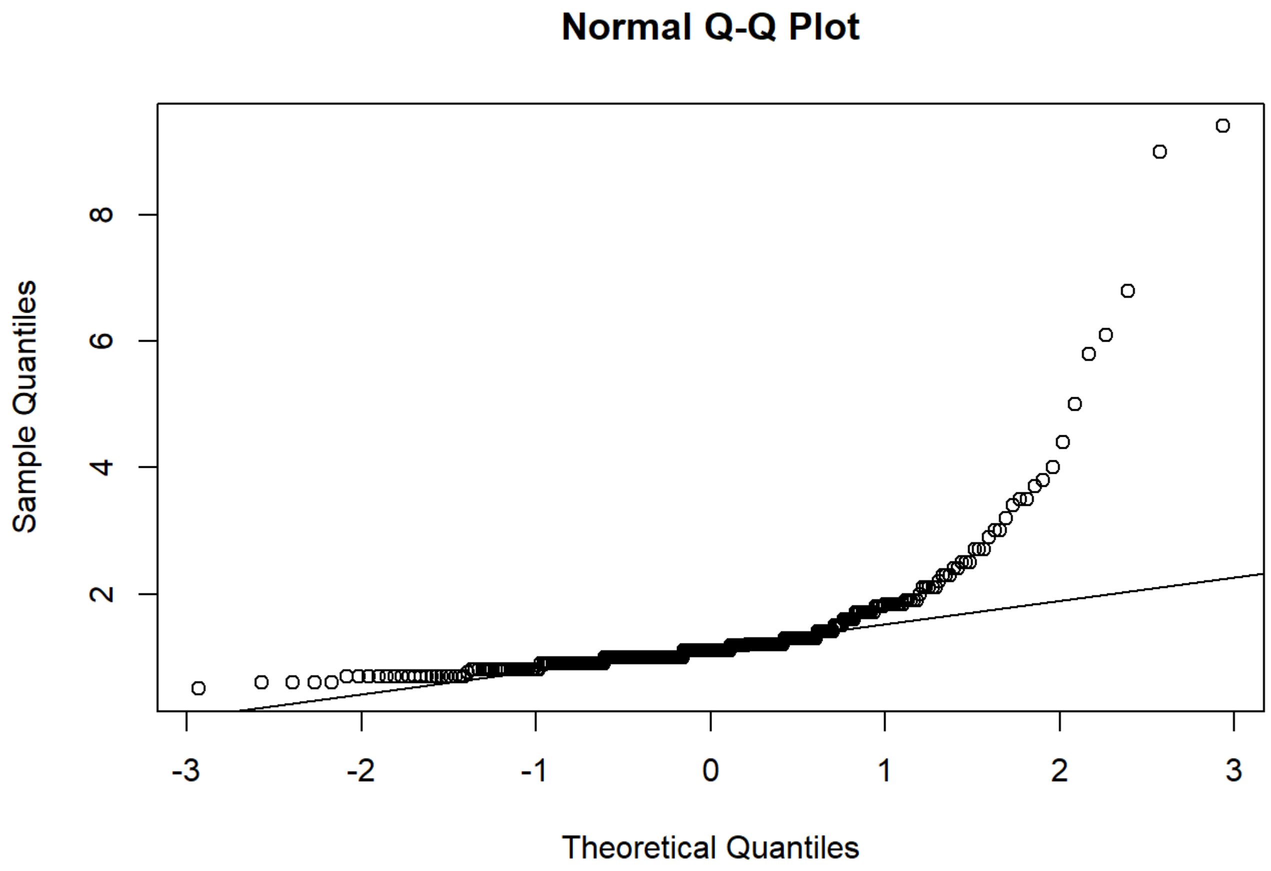
Platelets QQplot

```
qqnorm(dat$platelets)
qqline(dat$platelets)
```



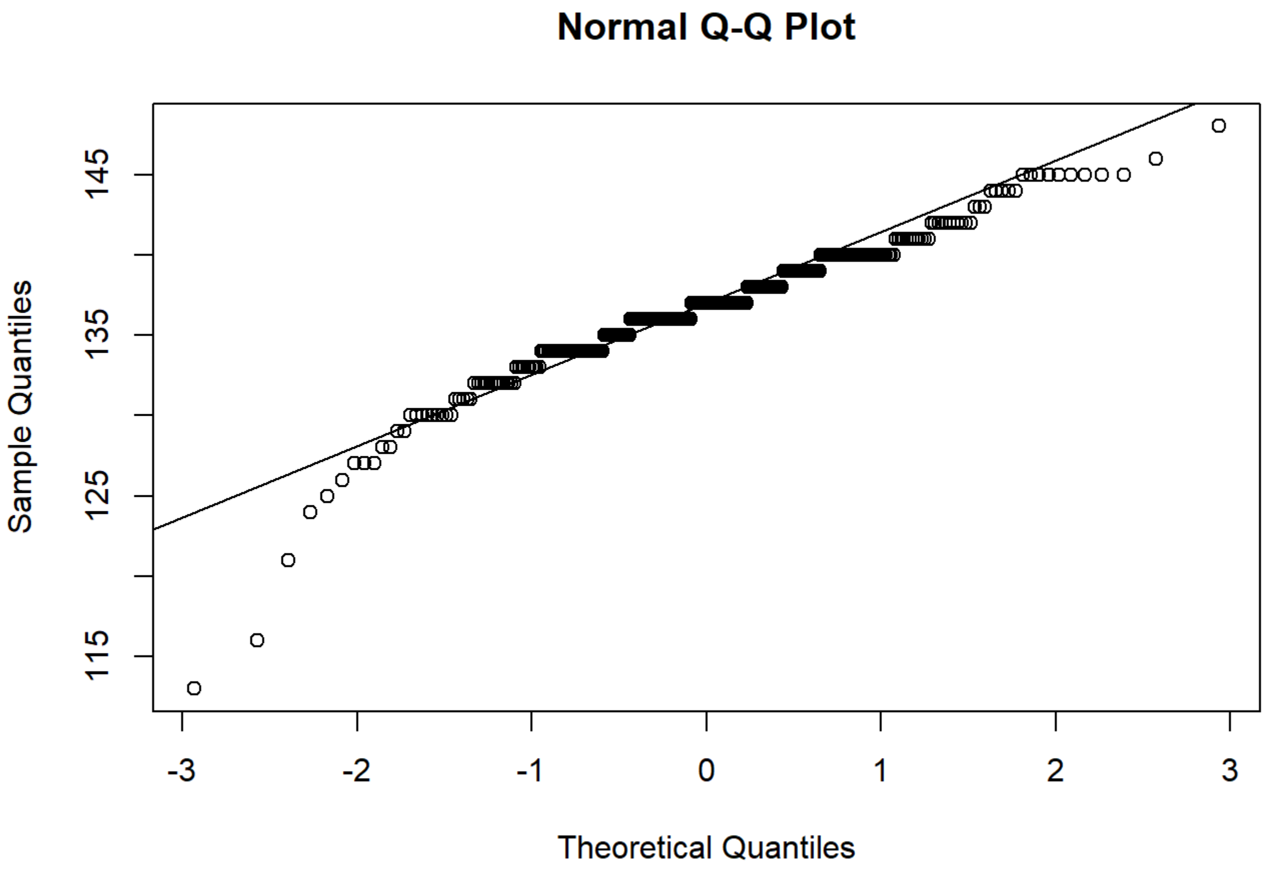
Serum creatinine QQplot

```
qqnorm(dat$serum_creatinine)
qqline(dat$serum_creatinine)
```



Serum Sodium QQplot

```
qqnorm(dat$serum_sodium)
qqline(dat$serum_sodium)
```

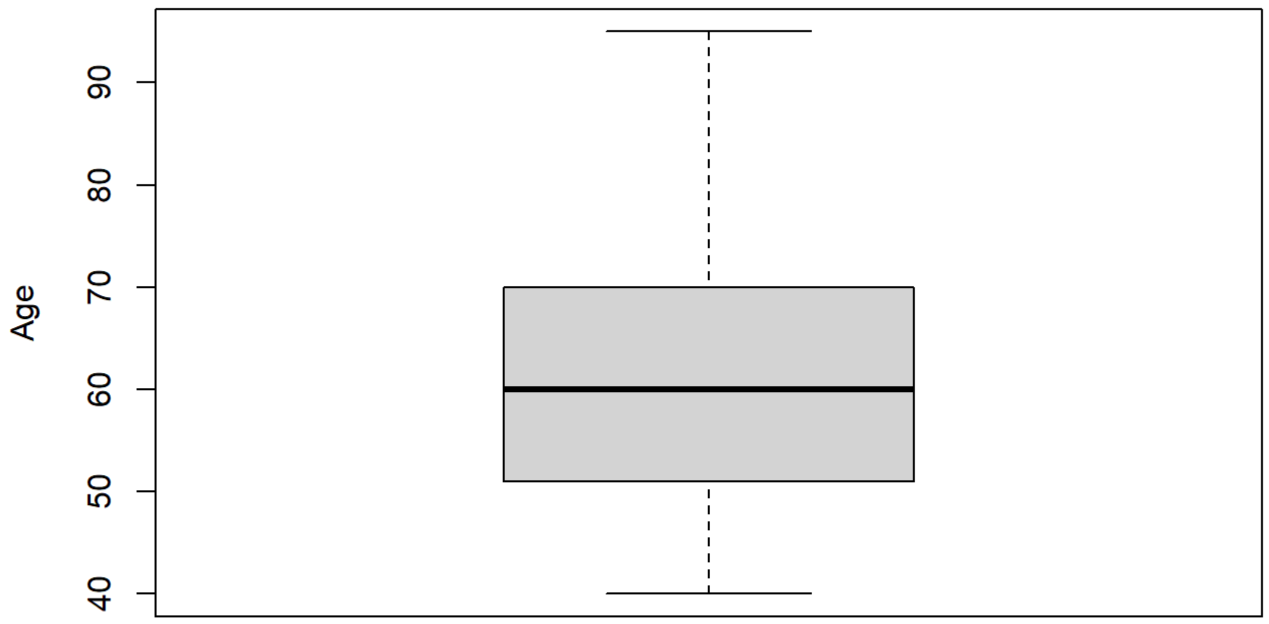


Combine the result of the histogram and qqplot, we could see that age, ejection_fraction, platelets, and serum_sodium is relatively normally distributed, while creatinine_phosphokinase and serum_creatinine are quite not normally distributed in the data set.

Boxplot

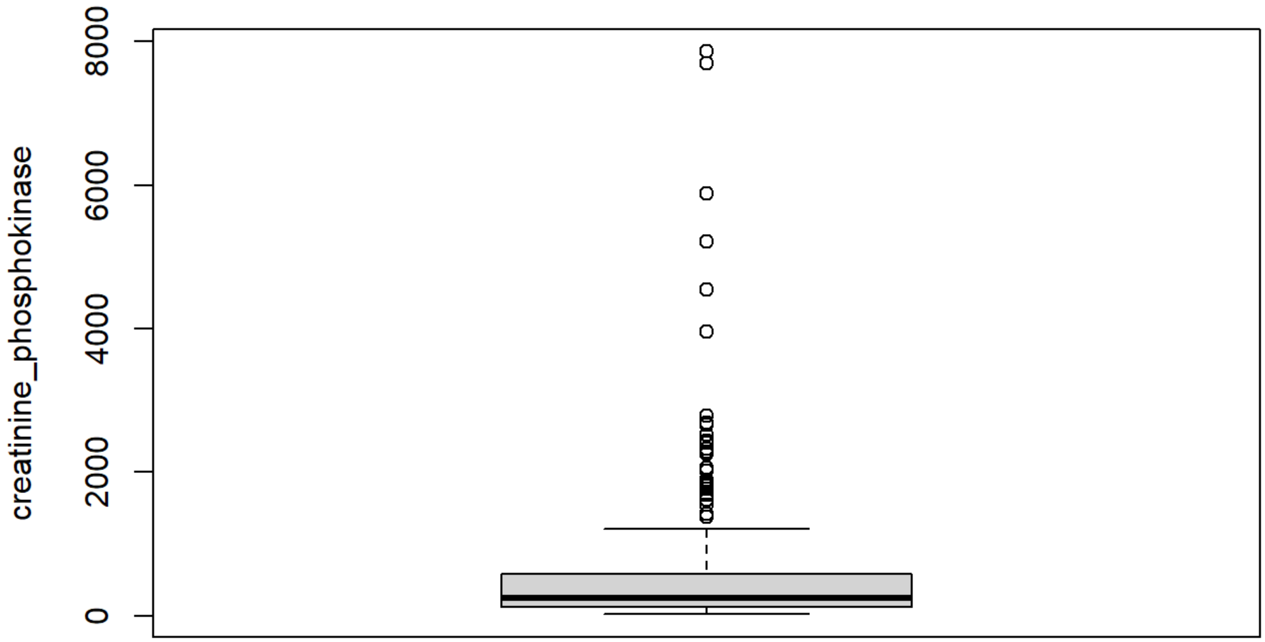
Age

```
age_box <- boxplot(dat$age, ylab = "Age")
```



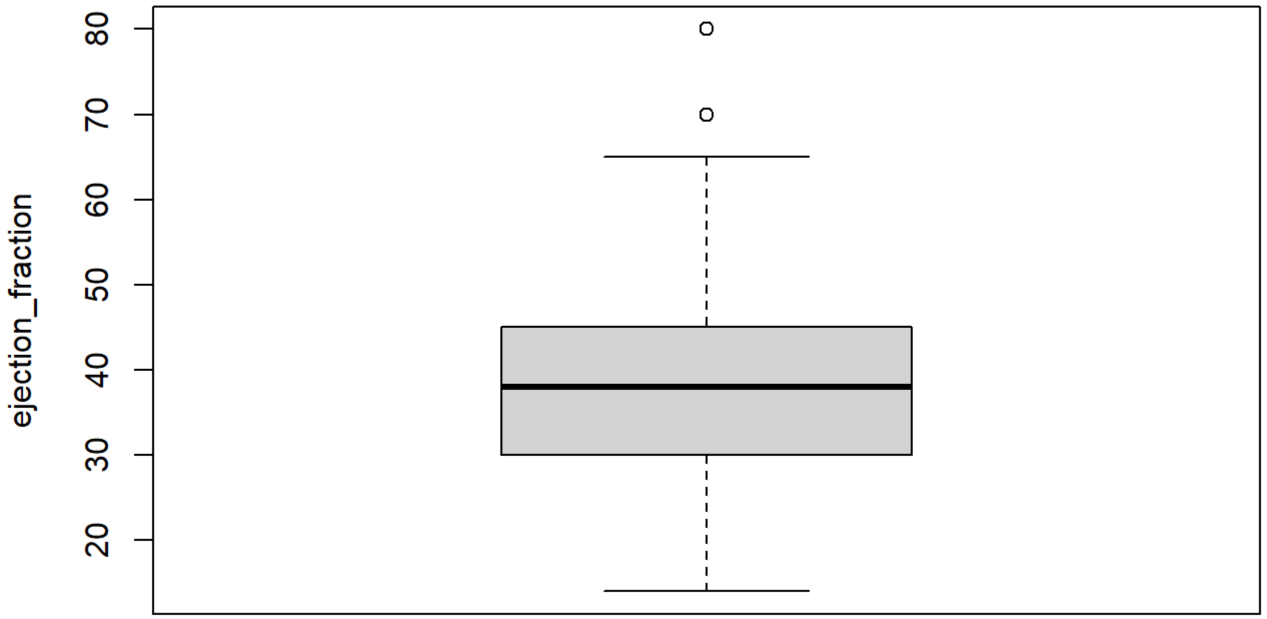
Creatinine Phosphokinease Concentration

```
cp_box <- boxplot(dat$creatinine_phosphokinase, ylab = "creatinine_phosphokinase")
```



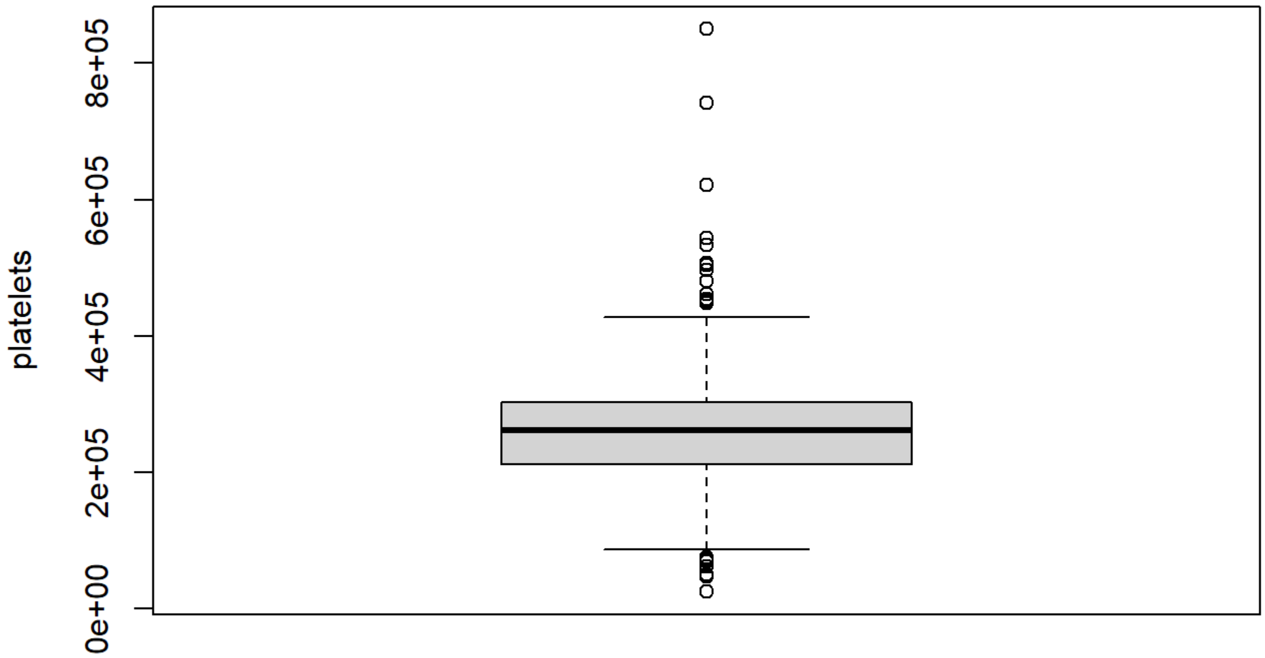
Ejection Fraction

```
ef_box <- boxplot(dat$ejection_fraction, ylab = "ejection_fraction")
```



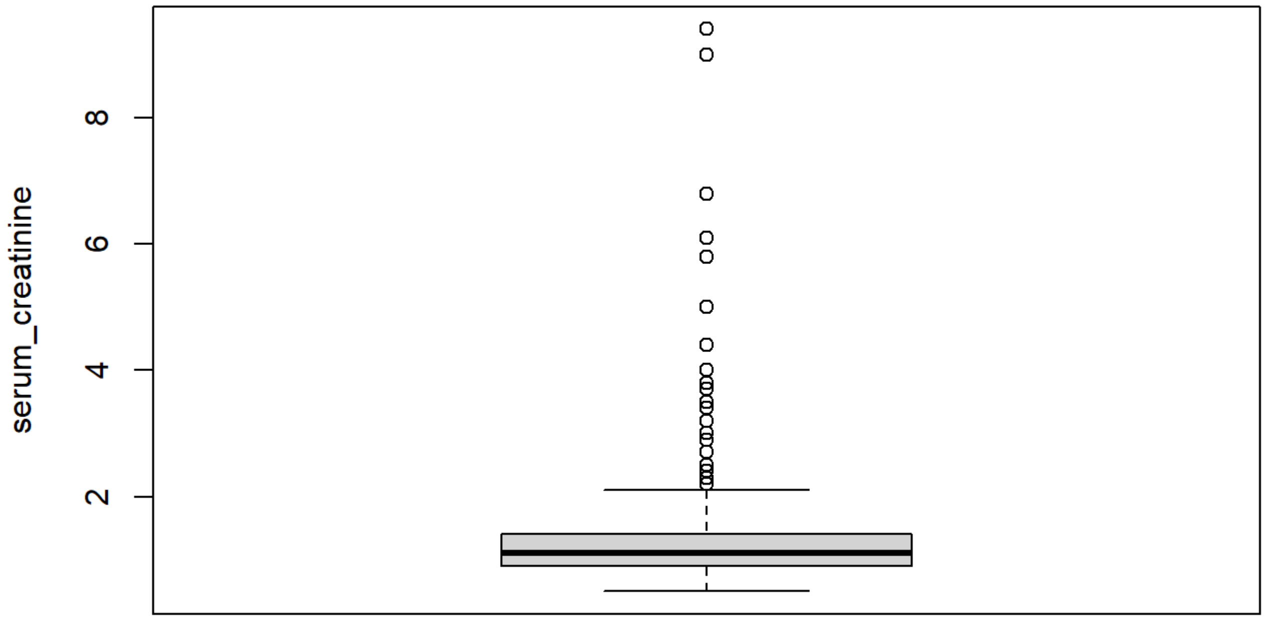
Plateletse Concentration

```
platelets_box <- boxplot(dat$platelets, ylab = "platelets")
```



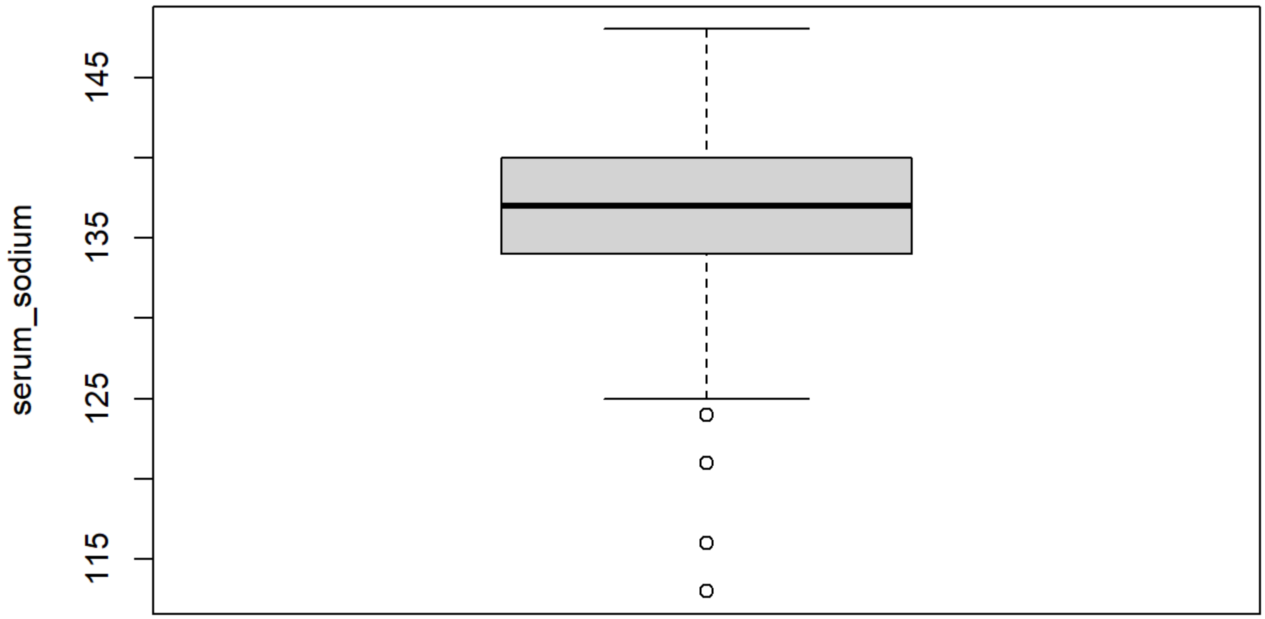
Serum Creatinine Concentration

```
sc_box <- boxplot(dat$serum_creatinine, ylab = "serum_creatinine")
```



Serum Sodium Concentration

```
ss_box <- boxplot(dat$serum_sodium, ylab = "serum_sodium")
```


serum_sodium
145
135
125
115

Correlation Matrix

The critical pearson correlation value for degree of freedom of 11 and p value 0.05 is 0.553.

```
pearson_table <- cor(dat,method = "pearson")

pearson_table
```

```
##                                  age     anaemia creatinine_phosphokinase
## age                       1.00000000  0.08800644             -0.081583900
## anaemia                   0.08800644  1.00000000             -0.190741030
## creatinine_phosphokinase -0.08158390 -0.19074103              1.000000000
## diabetes                 -0.10101239 -0.01272905             -0.009638514
## ejection_fraction         0.06009836  0.03155697             -0.044079554
## high_blood_pressure       0.09328868  0.03818200             -0.070589980
## platelets                -0.05235437 -0.04378555              0.024463389
## serum_creatinine          0.15918713  0.05217360             -0.016408480
## serum_sodium             -0.04596584  0.04188161              0.059550156
## sex                       0.06542952 -0.09476896              0.079790629
## smoking                   0.01866787 -0.10728984              0.002421235
## time                     -0.22406842 -0.14141398             -0.009345653
## DEATH_EVENT               0.25372854  0.06627010              0.062728160
##                              diabetes ejection_fraction high_blood_pressure
## age                      -0.101012385        0.06009836         0.093288685
## anaemia                  -0.012729046        0.03155697         0.038182003
## creatinine_phosphokinase -0.009638514       -0.04407955        -0.070589980
## diabetes                  1.000000000       -0.00485031        -0.012732382
## ejection_fraction        -0.004850310        1.00000000         0.024444731
## high_blood_pressure      -0.012732382        0.02444473         1.000000000
## platelets                 0.092192828        0.07217747         0.049963481
## serum_creatinine         -0.046975315       -0.01130247        -0.004934525
## serum_sodium             -0.089550619        0.17590228         0.037109470
## sex                      -0.157729504       -0.14838597        -0.104614629
## smoking                  -0.147173413       -0.06731457        -0.055711369
## time                      0.033725509        0.04172924        -0.196439479
## DEATH_EVENT              -0.001942883       -0.26860331         0.079351058
##                            platelets serum_creatinine serum_sodium          sex
## age                      -0.05235437      0.159187133 -0.045965841  0.065429524
## anaemia                  -0.04378555      0.052173604  0.041881610 -0.094768961
## creatinine_phosphokinase  0.02446339     -0.016408480  0.059550156  0.079790629
## diabetes                  0.09219283     -0.046975315 -0.089550619 -0.157729504
## ejection_fraction         0.07217747     -0.011302475  0.175902282 -0.148385965
## high_blood_pressure       0.04996348     -0.004934525  0.037109470 -0.104614629
## platelets                 1.00000000     -0.041198077  0.062124619 -0.125120483
## serum_creatinine         -0.04119808      1.000000000 -0.189095210  0.006969778
## serum_sodium              0.06212462     -0.189095210  1.000000000 -0.027566123
## sex                      -0.12512048      0.006969778 -0.027566123  1.000000000
## smoking                   0.02823445     -0.027414135  0.004813195  0.445891712
## time                      0.01051391     -0.149315418  0.087640000 -0.015608220
## DEATH_EVENT              -0.04913887      0.294277561 -0.195203596 -0.004316376
##                               smoking         time  DEATH_EVENT
## age                       0.018667868 -0.224068420  0.253728543
## anaemia                  -0.107289838 -0.141413982  0.066270098
## creatinine_phosphokinase  0.002421235 -0.009345653  0.062728160
## diabetes                 -0.147173413  0.033725509 -0.001942883
## ejection_fraction        -0.067314567  0.041729235 -0.268603312
## high_blood_pressure      -0.055711369 -0.196439479  0.079351058
## platelets                 0.028234448  0.010513909 -0.049138868
## serum_creatinine         -0.027414135 -0.149315418  0.294277561
## serum_sodium              0.004813195  0.087640000 -0.195203596
## sex                       0.445891712 -0.015608220 -0.004316376
## smoking                   1.000000000 -0.022838942 -0.012623153
## time                     -0.022838942  1.000000000 -0.526963779
## DEATH_EVENT              -0.012623153 -0.526963779  1.000000000
```

```
which(pearson_table > 0.553 | pearson_table < -0.553)
```

```
##  [1]   1  15  29  43  57  71  85  99 113 127 141 155 169
```

All the correlation in the table, except the diagnosis, are all smaller than the critical value, so there’s no multicollinearity among the variables. There are several possible risk factors realted to death (with higher correlation with death in the matrix):age, ejection fraction and serum creatinine that worth research on.