

# MVA\_Assignment\_3

Aman

10/01/2020

## Assignment 3 - Data Cleaning, EDA, Tests

This document does a preliminary analysis on the Heart Failure Prediction dataset

We will start by loading libraries and data

```
# clear environment
rm(list = ls())

# defining libraries

library(ggplot2)
library(dplyr)
library(PerformanceAnalytics)
library(data.table)
library(sqldf)
library(nortest)
library(tidyverse)
library(MASS)
library(rpart)
library(class)
library(ISLR)
library(scales)
library(ClustOfVar)
library(GGally)
library(reticulate)
library(ggthemes)
library(RColorBrewer)
library(gridExtra)
library(kableExtra)
library(Hmisc)
library(corrplot)
library(energy)
library(nnet)
library(Hotelling)
library(car)

# reading data
data <- read.csv('/Users/mac/Downloads/heart_failure_clinical_records_dataset.csv')
```

```
# structure of data
str(data)
```

```
## 'data.frame':    299 obs. of  13 variables:
## $ age           : num  75 55 65 50 65 90 75 60 65 80 ...
## $ anaemia       : int   0 0 0 1 1 1 1 0 1 ...
## $ creatinine_phosphokinase: int 582 7861 146 111 160 47 246 315 157 123 ...
## $ diabetes      : int   0 0 0 0 1 0 0 1 0 0 ...
## $ ejection_fraction : int  20 38 20 20 20 40 15 60 65 35 ...
## $ high_blood_pressure : int   1 0 0 0 0 1 0 0 0 1 ...
## $ platelets     : num 265000 263358 162000 210000 327000 ...
## $ serum_creatinine : num  1.9 1.1 1.3 1.9 2.7 2.1 1.2 1.1 1.5 9.4 ...
## $ serum_sodium    : int  130 136 129 137 116 132 137 131 138 133 ...
## $ sex            : int   1 1 1 1 0 1 1 1 0 1 ...
## $ smoking        : int   0 0 1 0 0 1 0 1 0 1 ...
## $ time           : int   4 6 7 7 8 8 10 10 10 10 ...
## $ DEATH_EVENT    : int   1 1 1 1 1 1 1 1 1 1 ...
```

```
glimpse(data)
```

```
## Observations: 299
## Variables: 13
## $ age           <dbl> 75, 55, 65, 50, 65, 90, 75, 60, 65, 8...
## $ anaemia       <int> 0, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1...
## $ creatinine_phosphokinase <int> 582, 7861, 146, 111, 160, 47, 246, 31...
## $ diabetes      <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0...
## $ ejection_fraction <int> 20, 38, 20, 20, 20, 40, 15, 60, 65, 3...
## $ high_blood_pressure <int> 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0...
## $ platelets     <dbl> 265000, 263358, 162000, 210000, 32700...
## $ serum_creatinine <dbl> 1.90, 1.10, 1.30, 1.90, 2.70, 2.10, 1...
## $ serum_sodium   <int> 130, 136, 129, 137, 116, 132, 137, 13...
## $ sex           <int> 1, 1, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1...
## $ smoking       <int> 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 0...
## $ time          <int> 4, 6, 7, 7, 8, 8, 10, 10, 10, 10, 10,...
## $ DEATH_EVENT   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
```

Let us summarise the data and note observations

## summary of data

```
summary(data)
```

```
##      age      anaemia  creatinine_phosphokinase
## Min.   :40.00  Min.   :0.0000  Min.    : 23.0
## 1st Qu.:51.00  1st Qu.:0.0000  1st Qu.: 116.5
## Median :60.00  Median :0.0000  Median : 250.0
## Mean   :60.83  Mean   :0.4314  Mean    : 581.8
## 3rd Qu.:70.00  3rd Qu.:1.0000  3rd Qu.: 582.0
## Max.   :95.00  Max.   :1.0000  Max.    :7861.0
##      diabetes  ejection_fraction high_blood_pressure  platelets
## Min.   :0.0000  Min.   :14.00    Min.    :0.0000    Min.    : 25100
## 1st Qu.:0.0000  1st Qu.:30.00    1st Qu.:0.0000    1st Qu.:212500
## Median :0.0000  Median :38.00    Median :0.0000    Median :262000
```

```
## Mean :0.4181 Mean :38.08 Mean :0.3512 Mean :263358
## 3rd Qu.:1.0000 3rd Qu.:45.00 3rd Qu.:1.0000 3rd Qu.:303500
## Max. :1.0000 Max. :80.00 Max. :1.0000 Max. :850000
## serum_creatinine serum_sodium sex smoking
## Min. :0.500 Min. :113.0 Min. :0.0000 Min. :0.0000
## 1st Qu.:0.900 1st Qu.:134.0 1st Qu.:0.0000 1st Qu.:0.0000
## Median :1.100 Median :137.0 Median :1.0000 Median :0.0000
## Mean :1.394 Mean :136.6 Mean :0.6488 Mean :0.3211
## 3rd Qu.:1.400 3rd Qu.:140.0 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :9.400 Max. :148.0 Max. :1.0000 Max. :1.0000
## time DEATH_EVENT
## Min. : 4.0 Min. :0.0000
## 1st Qu.: 73.0 1st Qu.:0.0000
## Median :115.0 Median :0.0000
## Mean :130.3 Mean :0.3211
## 3rd Qu.:203.0 3rd Qu.:1.0000
## Max. :285.0 Max. :1.0000
```

Observations -

1. 299 observations for 13 variables
2. Age is between 40 and 95 so not much outliers by intuition
3. Death\_event should be converted to factor variable as they take only 2 values
4. Creatinine phosphokinase, platelets clearly has an outlier from max value which we will confirm later by univariate analysis

## Missing/ NAs check

```
data2 <- na.omit(data)
```

‘data2’ has same rows as ‘data’ so there are no missing values in data

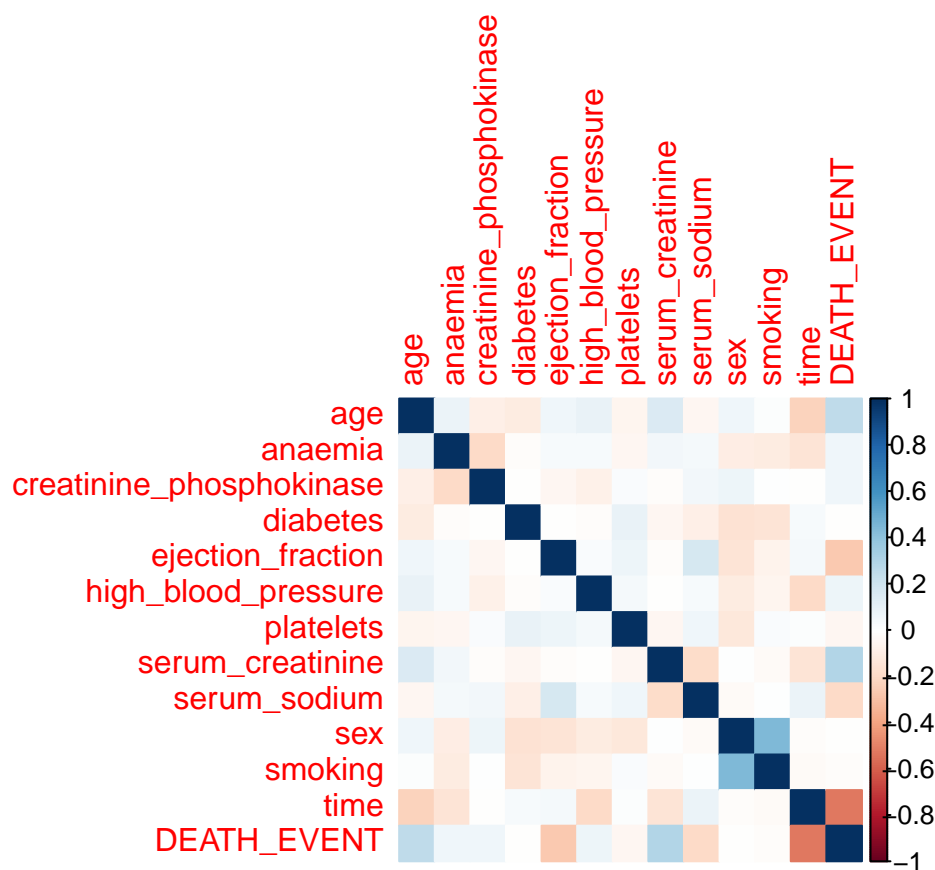
## Correlation Plot

```
M<-cor(data)
head(round(M,2))
```

```
##          age anaemia creatinine_phosphokinase diabetes
## age          1.00  0.09                -0.08  -0.10
## anaemia       0.09  1.00                -0.19  -0.01
## creatinine_phosphokinase -0.08 -0.19                1.00  -0.01
## diabetes      -0.10 -0.01                -0.01  1.00
## ejection_fraction  0.06  0.03                -0.04  0.00
## high_blood_pressure  0.09  0.04                -0.07  -0.01
##          ejection_fraction high_blood_pressure platelets
## age              0.06              0.09  -0.05
## anaemia           0.03              0.04  -0.04
## creatinine_phosphokinase -0.04          -0.07  0.02
## diabetes           0.00          -0.01  0.09
## ejection_fraction   1.00              0.02  0.07
## high_blood_pressure  0.02              1.00  0.05
```

```
##          serum_creatinine serum_sodium  sex smoking  time
## age                0.16      -0.05  0.07   0.02 -0.22
## anaemia             0.05       0.04 -0.09  -0.11 -0.14
## creatinine_phosphokinase -0.02    0.06  0.08   0.00 -0.01
## diabetes            -0.05    -0.09 -0.16  -0.15  0.03
## ejection_fraction   -0.01    0.18 -0.15  -0.07  0.04
## high_blood_pressure  0.00     0.04 -0.10  -0.06 -0.20
##          DEATH_EVENT
## age                0.25
## anaemia             0.07
## creatinine_phosphokinase 0.06
## diabetes            0.00
## ejection_fraction   -0.27
## high_blood_pressure  0.08
```

```
corrplot(M, method="color")
```



We see that age, anaemia, creatinine\_phosphokinase, high\_blood\_pressure, serum\_creatinine have +ve correlation with death\_event

We see that ejection\_fraction, platelets, serum\_sodium, and time have -ve correlation with death\_event

But we will need deeper analysis to confirm these relationships.

## Data Information

```
# Converting to factor (dependent variable)
data$DEATH_EVENT <- factor(data$DEATH_EVENT)

# Let's check how many zeros are in dataset
colSums(data==0)
```

```
##           age           anaemia creatinine_phosphokinase
##           0           170           0
##      diabetes ejection_fraction high_blood_pressure
##      174           0           194
##    platelets serum_creatinine serum_sodium
##           0           0           0
##           sex           smoking           time
##      105           203           0
##    DEATH_EVENT
##      203
```

```
# Let's check their proportion to dataset as well
round(colSums(data==0)/nrow(data)*100,2)
```

```
##           age           anaemia creatinine_phosphokinase
##           0.00           56.86           0.00
##      diabetes ejection_fraction high_blood_pressure
##      58.19           0.00           64.88
##    platelets serum_creatinine serum_sodium
##           0.00           0.00           0.00
##           sex           smoking           time
##      35.12           67.89           0.00
##    DEATH_EVENT
##      67.89
```

Smoking, High BP, Diabetes, Anaemia are over 50% while sex is below 35%  
The Event rate of survival is ~67.9%

Let's classify independent variables into -

1. Categorical -> Anaemia, Diabetes, High\_blood\_pressure, Sex, Smoking
2. Numeric -> Age, Creatinine\_phosphokinase, Ejection\_fraction, Platelets, serum\_creatinine, serum\_sodium, time

We also see that

Sex - Gender of patient Male = 1, Female =0

Diabetes - 0 = No, 1 = Yes

Anaemia - 0 = No, 1 = Yes

High\_blood\_pressure - 0 = No, 1 = Yes

Smoking - 0 = No, 1 = Yes

DEATH\_EVENT - 0 = No, 1 = Yes

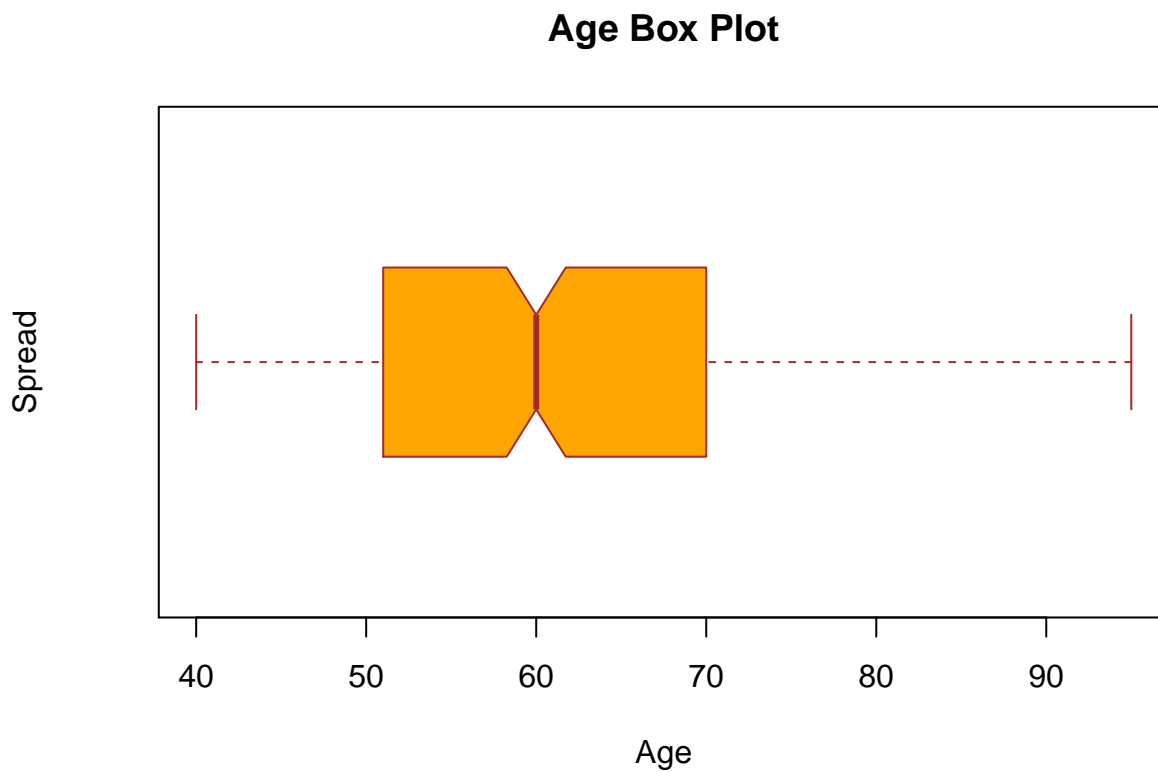
We note the scale of few variables like creatinine\_phosphokinase, platelets, ejection\_fraction and time. We can normalize the same before modeling but for now we will keep them as-is for the EDA.

## Analysis

Let's start with Outlier Analysis

Age

```
# Outlier Analysis
boxplot(data$age,
  main = "Age Box Plot",
  xlab = "Age",
  ylab = "Spread",
  col = "orange",
  border = "brown",
  horizontal = TRUE,
  notch = TRUE
)
```



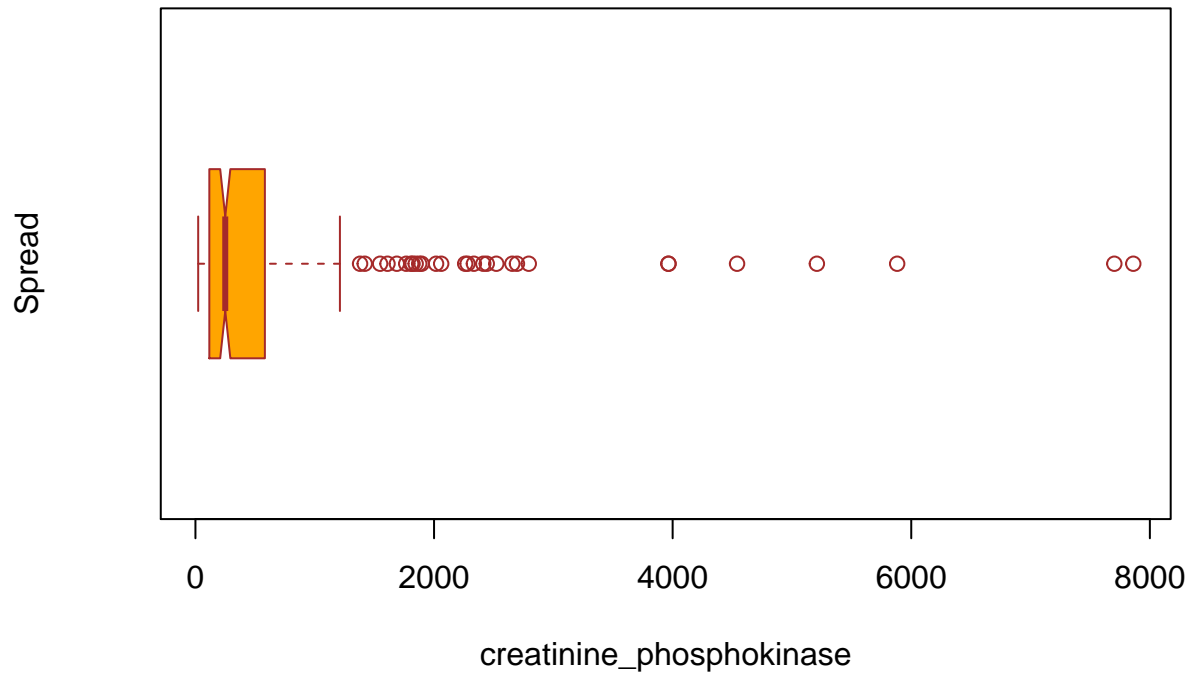
We note no observed outlier in age

creatinine\_phosphokinase

```
boxplot(data$creatinine_phosphokinase,
  main = "creatinine_phosphokinase Box Plot",
  xlab = "creatinine_phosphokinase",
  ylab = "Spread",
  col = "orange",
```

```
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```

### creatinine\_phosphokinase Box Plot

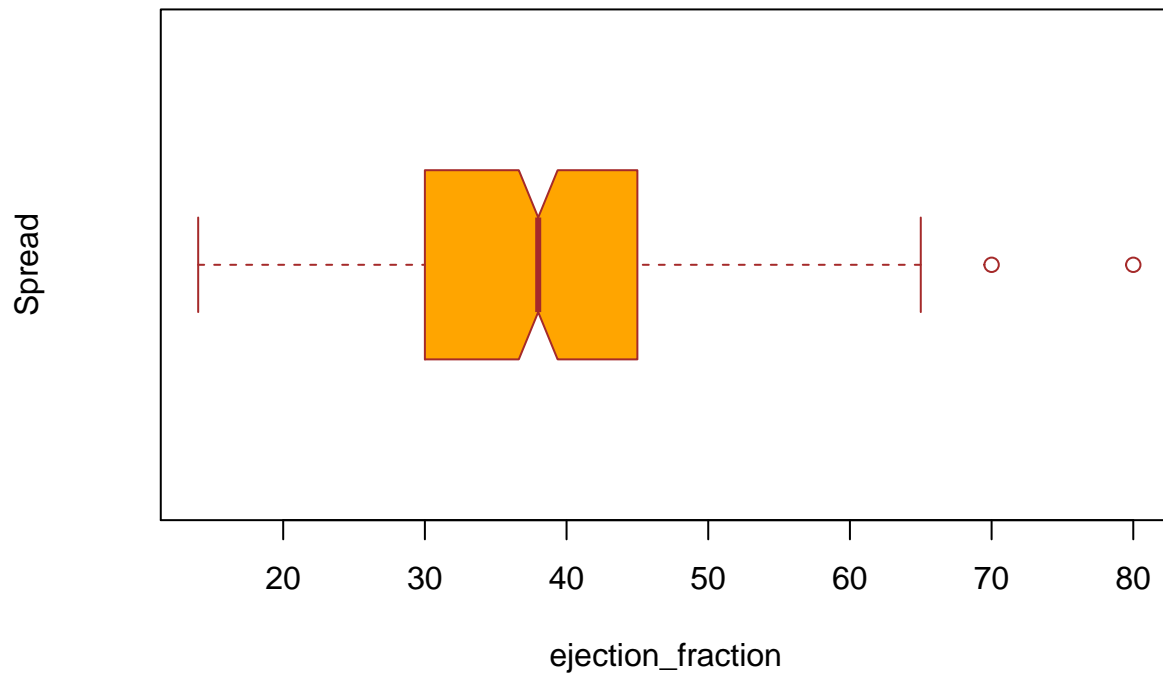


We notice some outliers at the positive side in creatinine\_phosphokinase with data above median more dispersed

### ejection\_fraction

```
boxplot(data$ejection_fraction,
main = "ejection_fraction Box Plot",
xlab = "ejection_fraction",
ylab = "Spread",
col = "orange",
border = "brown",
horizontal = TRUE,
notch = TRUE
)
```

## ejection\_fraction Box Plot

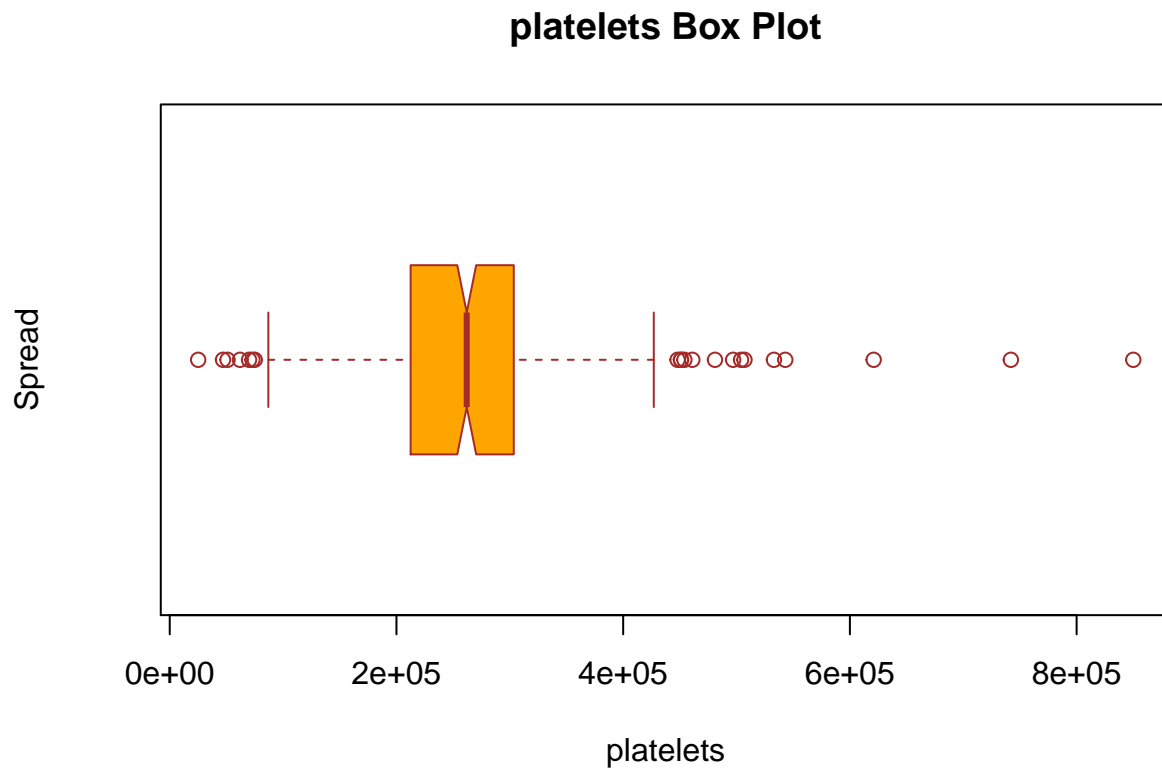


We notice 2 data points as outliers in ejection\_fraction

platelets

```
boxplot(data$platelets,  
  main = "platelets Box Plot",  
  xlab = "platelets",  
  ylab = "Spread",  
  col = "orange",  
  border = "brown",  
  horizontal = TRUE,  
  notch = TRUE  
)
```



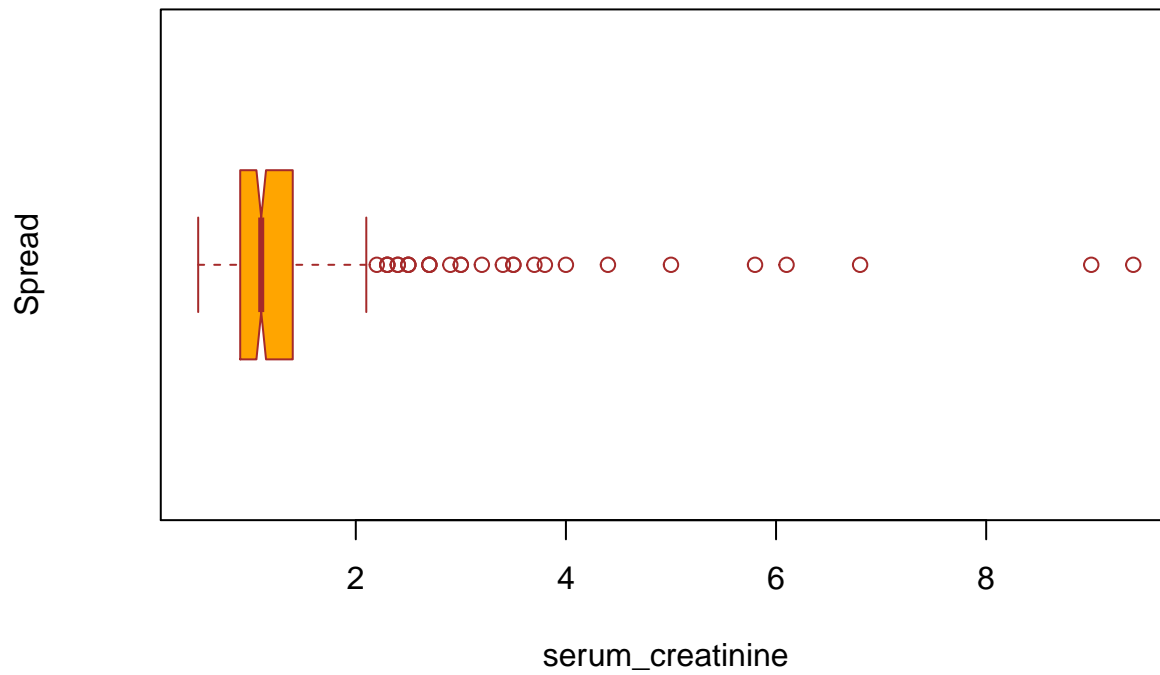


We notice outliers on both spectrum (high and low) in platelets

serum\_creatinine

```
boxplot(data$serum_creatinine,  
  main = "serum_creatinine Box Plot",  
  xlab = "serum_creatinine",  
  ylab = "Spread",  
  col = "orange",  
  border = "brown",  
  horizontal = TRUE,  
  notch = TRUE  
)
```

## serum\_creatinine Box Plot

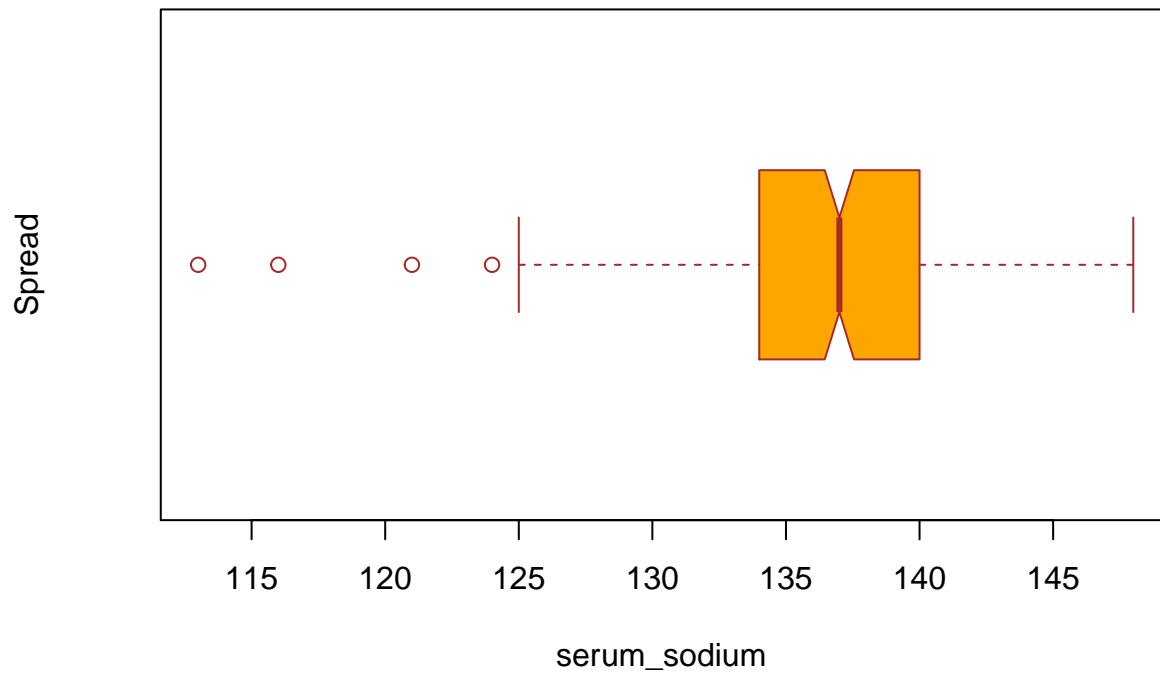


We notice some outliers in serum\_creatinine on higher end (similar to creatinine\_phosphokinase). However these are in possible ranges medically.

## serum\_sodium

```
boxplot(data$serum_sodium ,
  main = "serum_sodium Box Plot",
  xlab = "serum_sodium ",
  ylab = "Spread",
  col = "orange",
  border = "brown",
  horizontal = TRUE,
  notch = TRUE
)
```

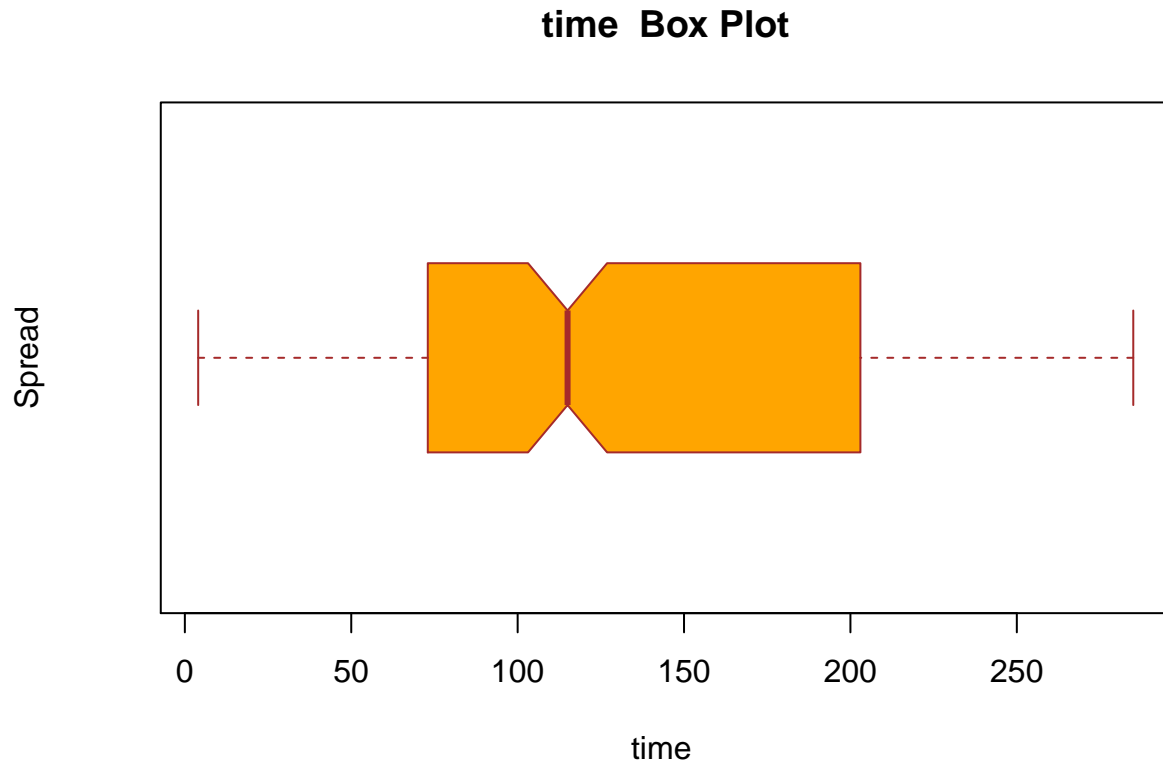
## serum\_sodium Box Plot



We notice some outliers in serum\_sodium on lower end

time

```
boxplot(data$time ,  
        main = "time Box Plot",  
        xlab = "time ",  
        ylab = "Spread",  
        col = "orange",  
        border = "brown",  
        horizontal = TRUE,  
        notch = TRUE  
)
```



We notice no outliers in time (follow up period) however data above median is more dispersed

**Note:** While some of these are clear outliers, others must be checked with possible medical range

## Data Cleaning - Let's remove these outliers

```
data <- data[data$ejection_fraction < 70,]  
data <- data[data$creatinine_phosphokinase < 7000,]
```

The new data now has only 295 observations - 4 observations were removed

## Univariate analysis

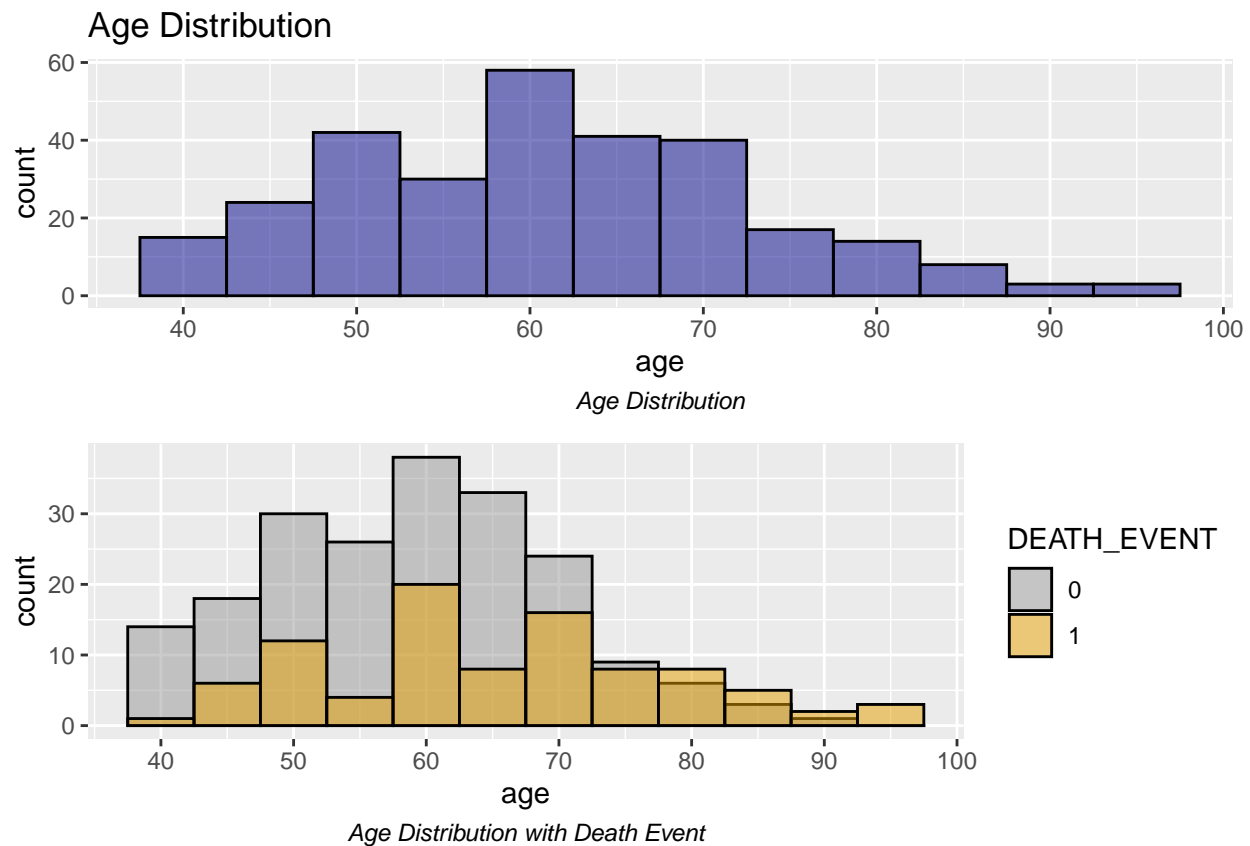
We will plot density distributions for numerical variables  
And check splits for categorical variables

### Age

```
# Let's check Age distribution and see it with death event
a<-ggplot(data,aes(x = age))+geom_histogram(binwidth = 5, color = "black",
                                             fill = "dark blue",alpha = 0.5)+
  labs(title = "Age Distribution",
       caption = "Age Distribution")+
  theme(plot.caption = element_text(hjust = 0.5,face = "italic"))+
  scale_x_continuous(breaks = seq(40,100,10))

b<-ggplot(data,aes(x = age, fill = DEATH_EVENT))+geom_histogram(binwidth = 5,
                                                                position = "identity",
                                                                alpha = 0.5,color = "black")+scale_fill_manual(values = c("#999999", "#E69F00"))+
  labs(caption = "Age Distribution with Death Event")+
  theme(plot.caption = element_text(hjust = 0.5,face = "italic"))+
  scale_x_continuous(breaks = seq(40,100,10))

gridExtra::grid.arrange(a,b)
```



We see that as age increases, chances of death event go up as well

We can also confirm this numerically ;

```
# Let's create age ranges
data$age_tr[data$age < 50 & data$age >= 40]="40-50"
data$age_tr[data$age < 60 & data$age >= 50]="50-60"
data$age_tr[data$age < 70 & data$age >= 60]="60-70"
data$age_tr[data$age < 80 & data$age >= 70]="70-80"
data$age_tr[data$age < 90 & data$age >= 80]="80-90"
data$age_tr[data$age < 100 & data$age >= 90]="90-100"
```

```
table(data$DEATH_EVENT, data$age_tr)
```

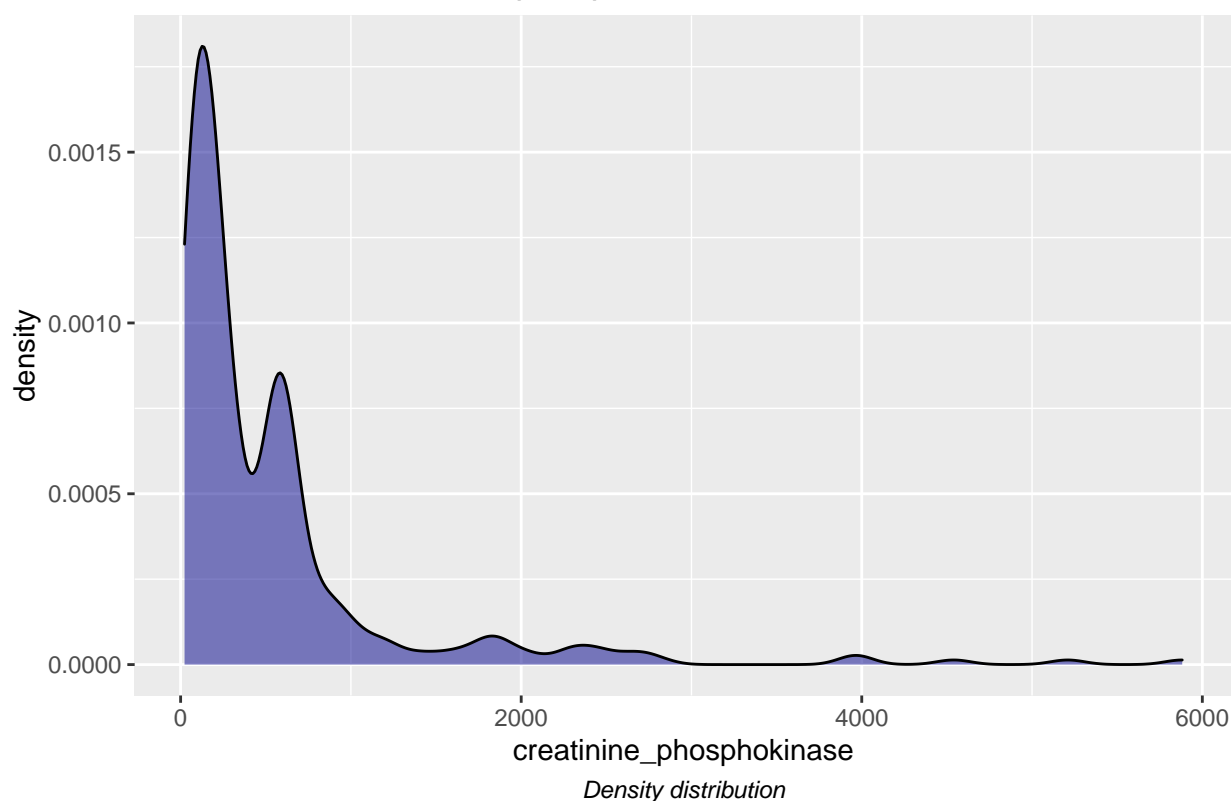
```
##
##      40-50 50-60 60-70 70-80 80-90 90-100
##    0      35    62    66    32     6     1
##    1      10    18    27    20    13     5
```

Numerically, we can confirm the same observation (Higher death rate in higher ages)

### Creatinine\_phosphokinase

```
# density plot of Creatinine_phosphokinase
ggplot(data,aes(x = creatinine_phosphokinase))+geom_density(fill = "dark blue",
                                                             alpha = 0.5)+
  labs(title = "Distribution of creatinine phosphokinase", caption =
        "Density distribution")+
  theme(plot.caption = element_text(hjust = 0.5, face = "italic"))
```

## Distribution of creatinine phosphokinase



The distribution looks skewed

```
#let's create 10 splits of this variable
data$creatinine_phosphokinase_tr <- cut(data$creatinine_phosphokinase, 10)
table(data$DEATH_EVENT, data$creatinine_phosphokinase_tr)
```

```
##
##      (17.1,609] (609,1.19e+03] (1.19e+03,1.78e+03] (1.78e+03,2.37e+03]
## 0          156           24              7              7
## 1           79            6              2              3
##
##      (2.37e+03,2.95e+03] (2.95e+03,3.54e+03] (3.54e+03,4.12e+03]
## 0                  5              0              1
## 1                  1              0              1
##
##      (4.12e+03,4.71e+03] (4.71e+03,5.3e+03] (5.3e+03,5.89e+03]
## 0                  1              1              0
## 1                  0              0              1
```

Numerically, we can see that for creatinine levels above 4000, death event seems to be higher but what about averages ?

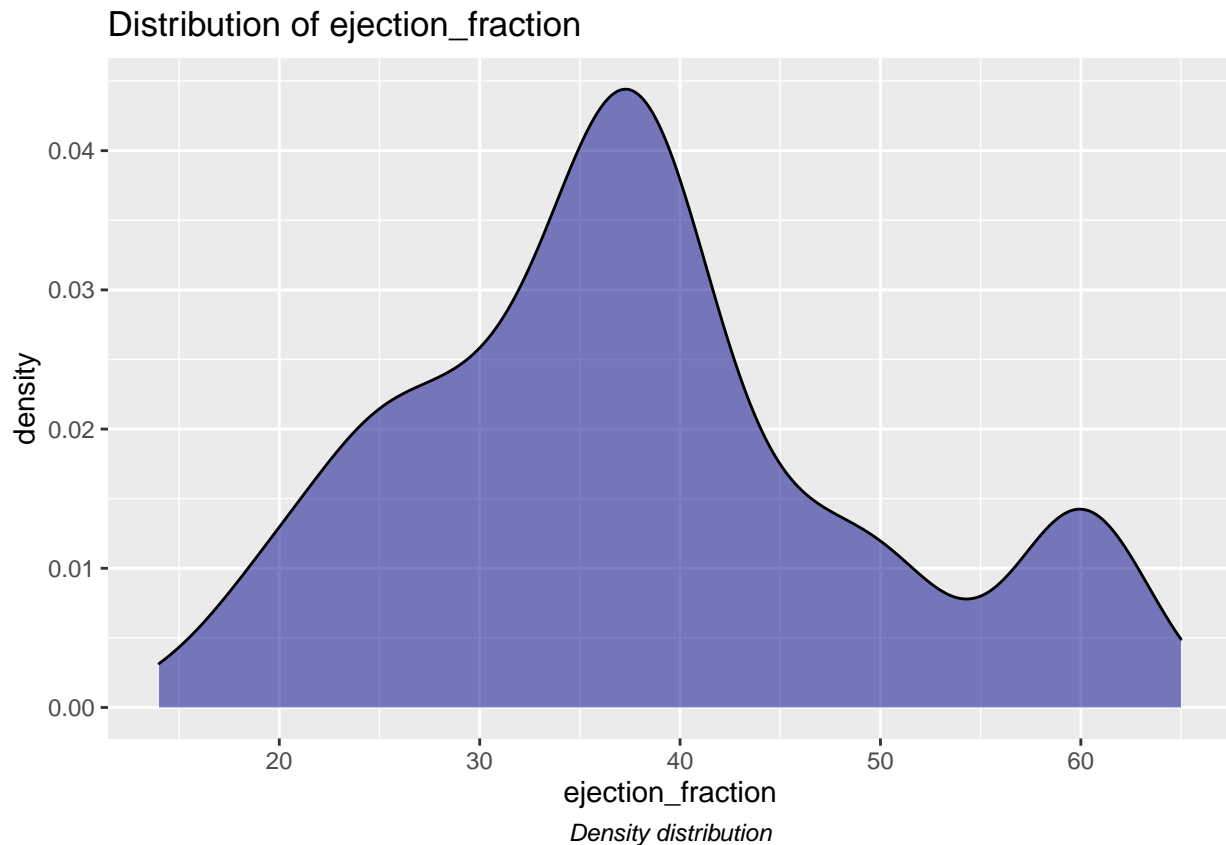
```
aggregate(data[, c('creatinine_phosphokinase')], list(data$DEATH_EVENT), mean)
```

```
##   Group.1      x
## 1      0 539.8465
## 2      1 519.8817
```

Numerically, we can see that average creatinine levels are lower in case of death event

## Ejection\_fraction

```
ggplot(data,aes(x = ejection_fraction))+geom_density(fill = "dark blue",
                                                       alpha = 0.5)+
  labs(title = "Distribution of ejection_fraction",
       caption = "Density distribution")+
  theme(plot.caption = element_text(hjust = 0.5, face = "italic"))
```



The distribution has a major and a minor peak

```
#let's create 10 splits of this variable
data$ejection_fraction_tr <- cut(data$ejection_fraction, 10)
table(data$DEATH_EVENT, data$ejection_fraction_tr)
```

```
##
##      (13.9,19.1] (19.1,24.2] (24.2,29.3] (29.3,34.4] (34.4,39.5]
## 0           1           2           18           21           67
## 1           4          16          17          13           21
##
##      (39.5,44.6] (44.6,49.7] (49.7,54.8] (54.8,59.9] (59.9,65.1]
## 0           33          15          15           2           28
## 1           4           5           6           1           6
```

Numerically, we can see that ejection fraction is low in case of death event



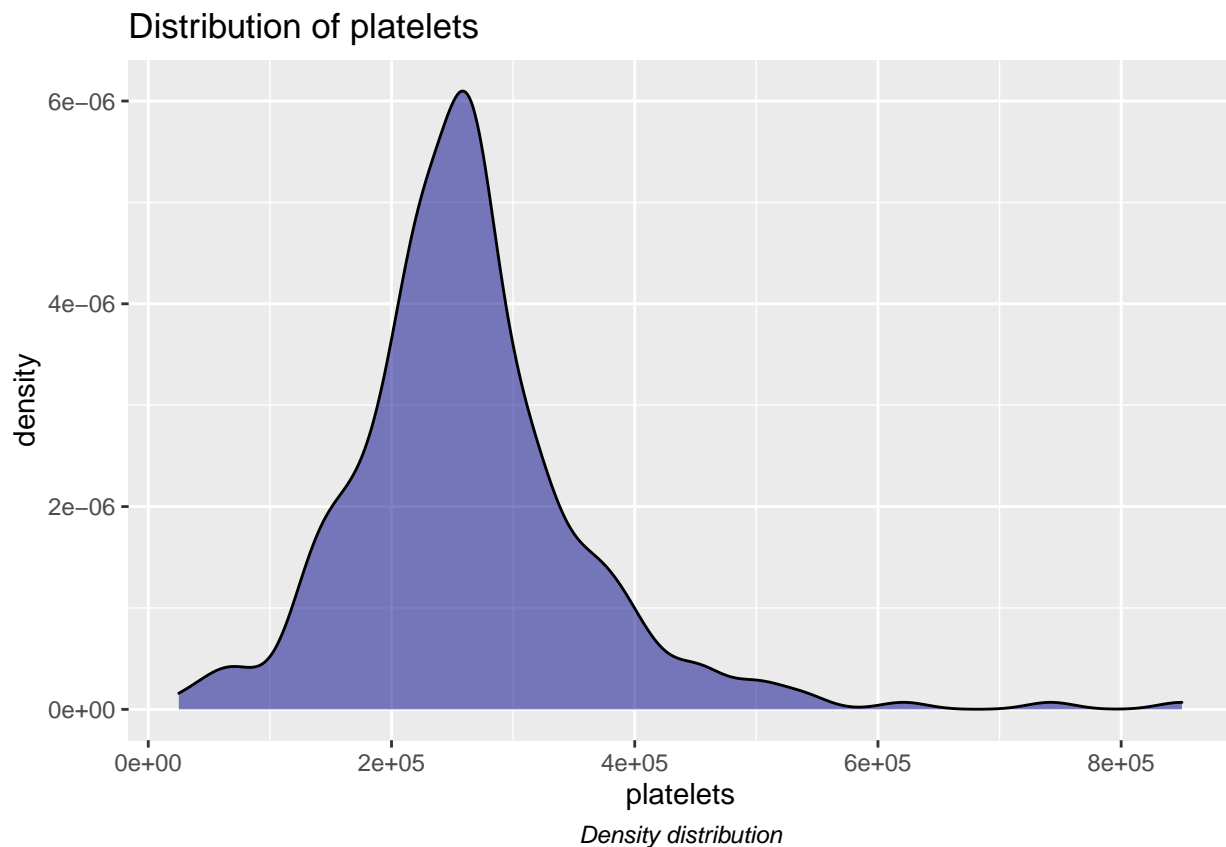
```
aggregate(data[, c('ejection_fraction')], list(data$DEATH_EVENT), mean)
```

```
##   Group.1      x
## 1      0 40.06931
## 2      1 33.11828
```

Numerically, we can see that average ejection fraction is also low in case of death event

## platelets

```
ggplot(data,aes(x = platelets ))+geom_density(fill = "dark blue", alpha = 0.5)+
  labs(title = "Distribution of platelets ", caption = "Density distribution")+
  theme(plot.caption = element_text(hjust = 0.5, face = "italic"))
```



The distribution looks pretty normal with some tail noticeable at the right extreme

```
#let's create 10 splits of this variable
data$platelets_tr <- cut(data$platelets , 10)
table(data$DEATH_EVENT, data$platelets_tr)
```

```
##
##   (2.43e+04,1.08e+05] (1.08e+05,1.9e+05] (1.9e+05,2.73e+05]
## 0      4              31              91
## 1      5              14              43
##
```

```
##      (2.73e+05,3.55e+05] (3.55e+05,4.38e+05] (4.38e+05,5.2e+05]
## 0              49              18              5
## 1              18              8              4
##
##      (5.2e+05,6.03e+05] (6.03e+05,6.85e+05] (6.85e+05,7.68e+05]
## 0              2              0              1
## 1              0              1              0
##
##      (7.68e+05,8.51e+05]
## 0              1
## 1              0
```

Numerically, we can see that platelets are low in case of death event

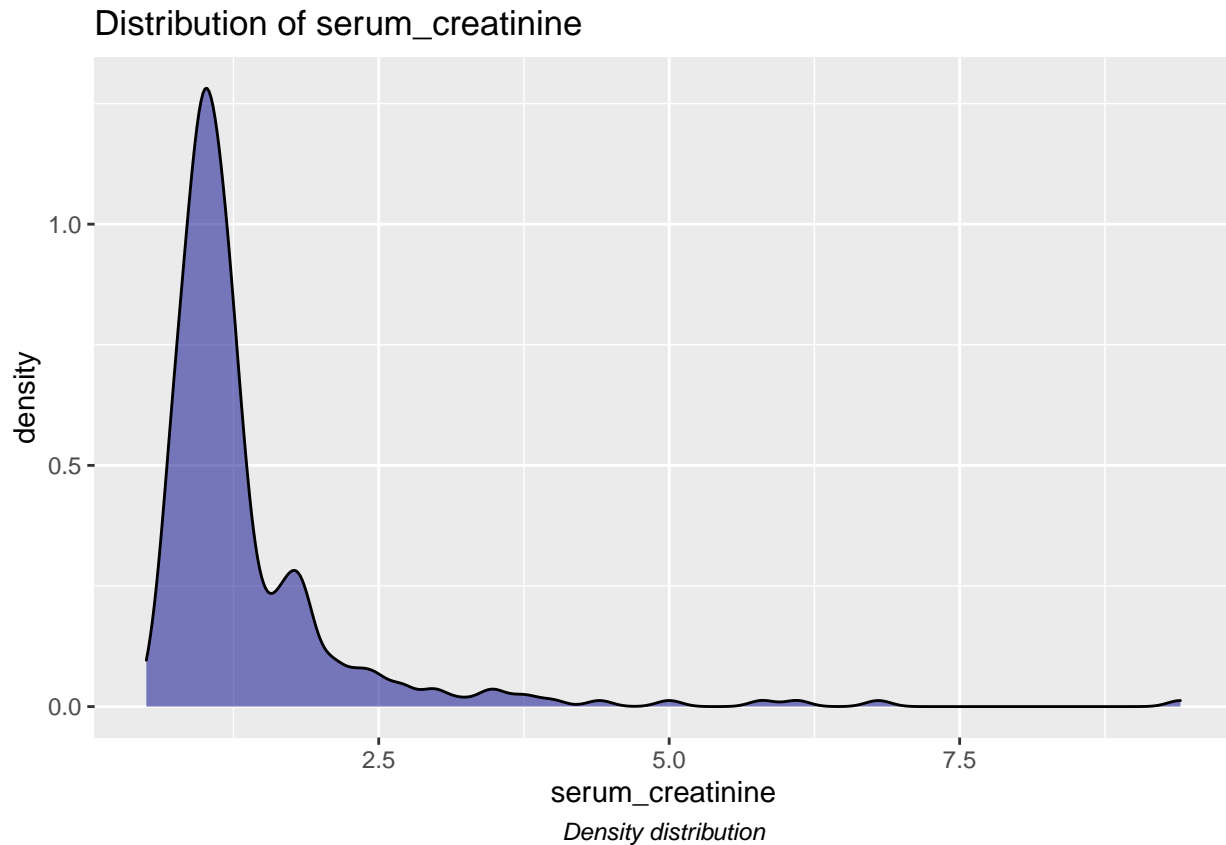
```
aggregate(data[, c('platelets')], list(data$DEATH_EVENT), mean)
```

```
##   Group.1      x
## 1      0 266673.8
## 2      1 256002.4
```

Numerically, we can see that average platelets are marginally lower in case of death event

### serum\_creatinine

```
ggplot(data,aes(x = serum_creatinine ))+geom_density(fill =
  "dark blue", alpha = 0.5)+
  labs(title = "Distribution of serum_creatinine ",
    caption = "Density distribution")+
  theme(plot.caption = element_text(hjust = 0.5, face = "italic"))
```



The distribution looks similar to creatinine\_phosphokinase

```
#let's create 10 splits of this variable
data$serum_creatinine_tr <- cut(data$serum_creatinine , 10)
table(data$DEATH_EVENT, data$serum_creatinine_tr)
```

```
##
##      (0.491,1.39] (1.39,2.28] (2.28,3.17] (3.17,4.06] (4.06,4.95]
## 0          168          23           5           4           0
## 1           47          30           9           3           1
##
##      (4.95,5.84] (5.84,6.73] (6.73,7.62] (7.62,8.51] (8.51,9.41]
## 0              1              1              0              0              0
## 1              1              0              1              0              1
```

Numerically, we can see that death event is high when serum\_creatinine levels are above 1.39 and very high above 2.28

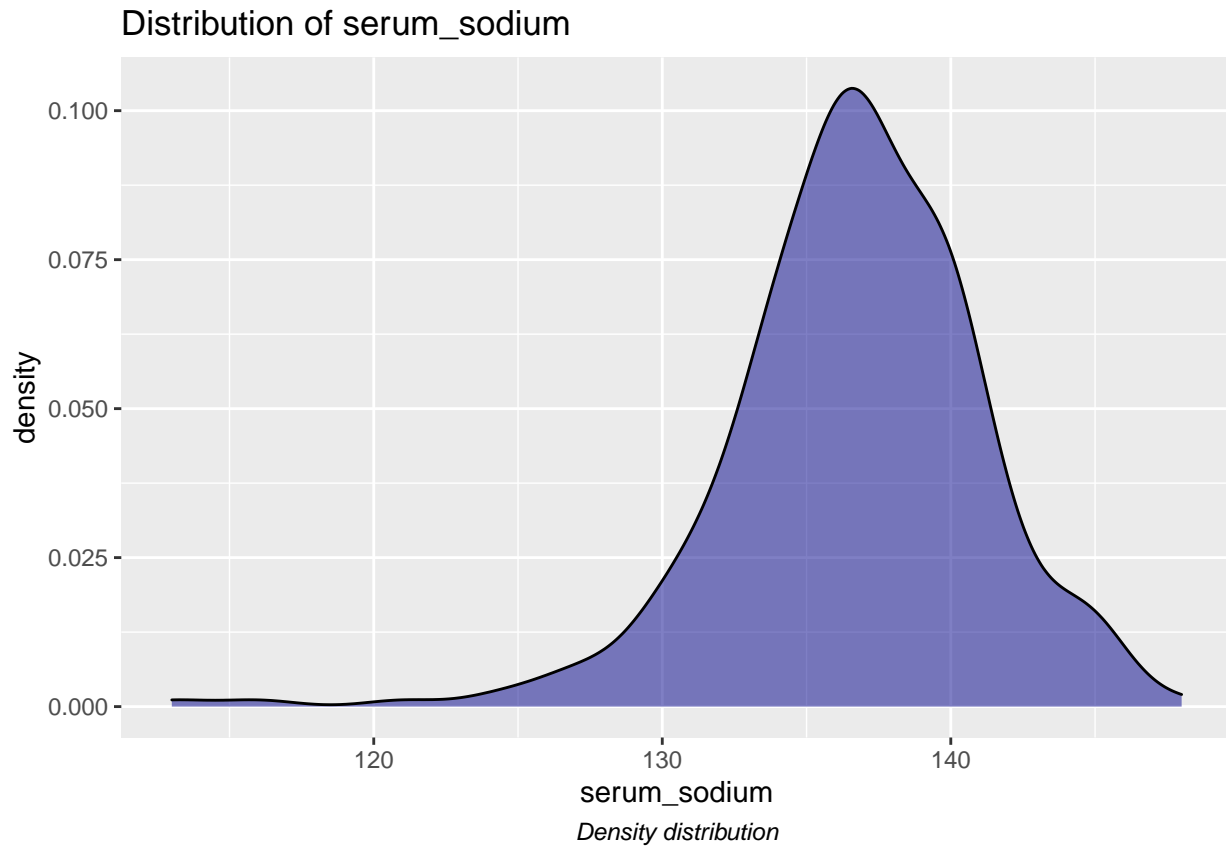
```
aggregate(data[, c('serum_creatinine')], list(data$DEATH_EVENT), mean)
```

```
##   Group.1      x
## 1      0 1.184901
## 2      1 1.775699
```

Numerically, we can see that average serum\_creatinine is high in case of death event

serum\_sodium

```
ggplot(data,aes(x = serum_sodium ))+geom_density(fill = "dark blue",
                                                    alpha = 0.5)+
  labs(title = "Distribution of serum_sodium ",
        caption = "Density distribution")+
  theme(plot.caption = element_text(hjust = 0.5, face = "italic"))
```



The distribution looks pretty normal with some tail noticeable at the left extreme

```
#let's create 10 splits of this variable
data$serum_sodium_tr <- cut(data$serum_sodium , 10)
table(data$DEATH_EVENT, data$serum_sodium_tr)
```

```
##
##      (113,116] (116,120] (120,124] (124,127] (127,130] (130,134] (134,138]
## 0           1           0           0           2           7           31           69
## 1           1           0           1           4           6           30           22
##
##      (138,141] (141,144] (144,148]
## 0           72          13           7
## 1           19           6           4
```

Numerically, we can see that serum\_sodium are low in case of death event

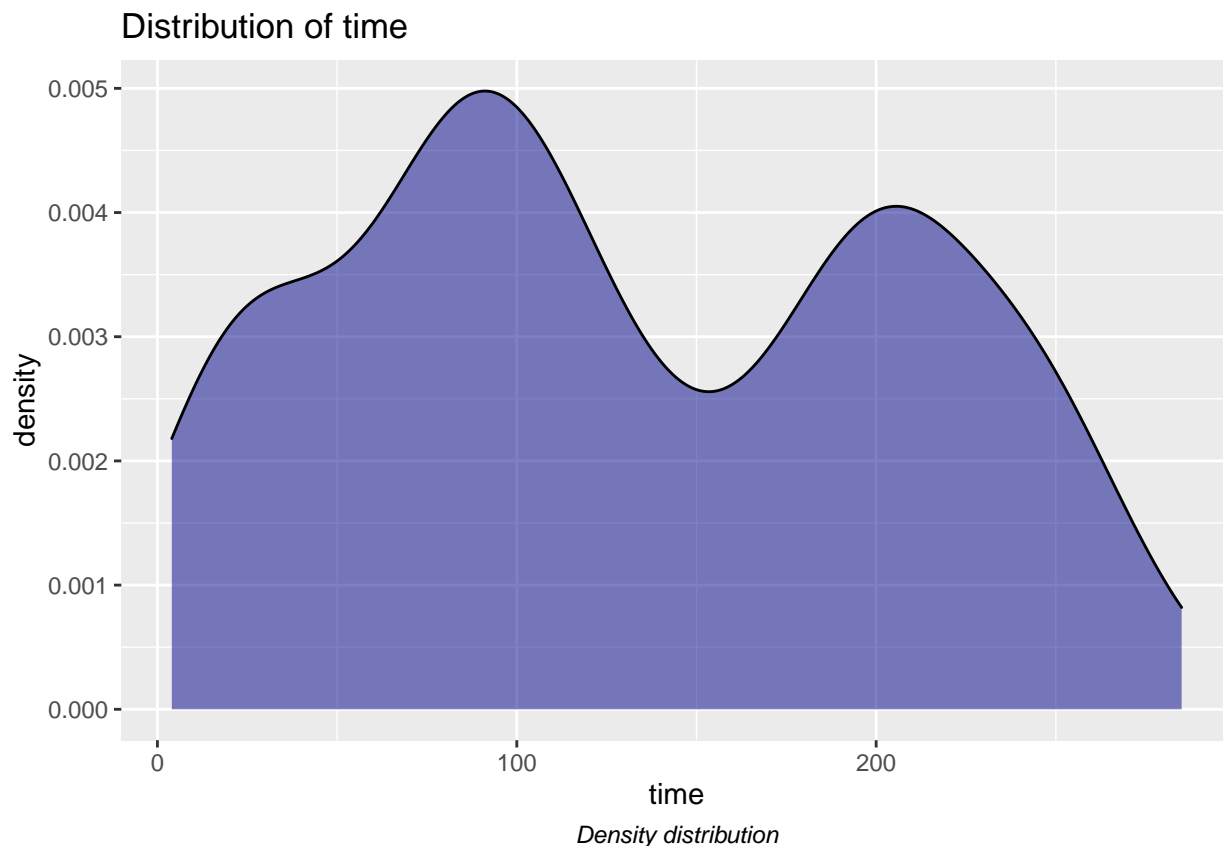
```
aggregate(data[, c('serum_sodium')], list(data$DEATH_EVENT), mean)
```

```
##   Group.1      x
## 1      0 137.2178
## 2      1 135.3118
```

Numerically, we can see that average serum sodium are marginally lower in case of death event

time

```
ggplot(data,aes(x = time ))+geom_density(fill = "dark blue", alpha = 0.5)+
  labs(title = "Distribution of time ", caption = "Density distribution")+
  theme(plot.caption = element_text(hjust = 0.5, face = "italic"))
```



We note a bi-modal peak

```
#let's create 10 splits of this variable
data$time_tr <- cut(data$time , 10)
table(data$DEATH_EVENT, data$time_tr)
```

```
##
##      (3.72,32.1] (32.1,60.2] (60.2,88.3] (88.3,116] (116,144] (144,173]
## 0           5           4           33           32           14           13
## 1          36          16          13           9           4           7
##
##      (173,201] (201,229] (229,257] (257,285]
## 0          28          33          31           9
## 1           4           2           2           0
```

Numerically, we can see that follow up period was small in case of death event

```
aggregate(data[, c('time')], list(data$DEATH_EVENT), mean)
```

```
##   Group.1      x
## 1      0 158.81188
## 2      1  70.35484
```

Numerically, we can see that average follow up period is low in case of death event. This simply may illustrate that once deemed healthy, the patients may have stopped following up whereas diseased patients would undergo more checkups

Let's do EDA for categorical variables now

```
# Anaemia, Diabetes, High_blood_pressure, Sex, Smoking

a <- ggplot(data, aes(x = DEATH_EVENT, fill = factor(anaemia)))+
  geom_bar(position = "fill")+
  scale_x_discrete(labels = c("Death Event:No", "Death Event:Yes"))+
  scale_fill_manual(values = c("#999999", "#E69F00"), name = "Anaemia",
    labels = c("No", "Yes"))+labs(subtitle = "Anaemia")

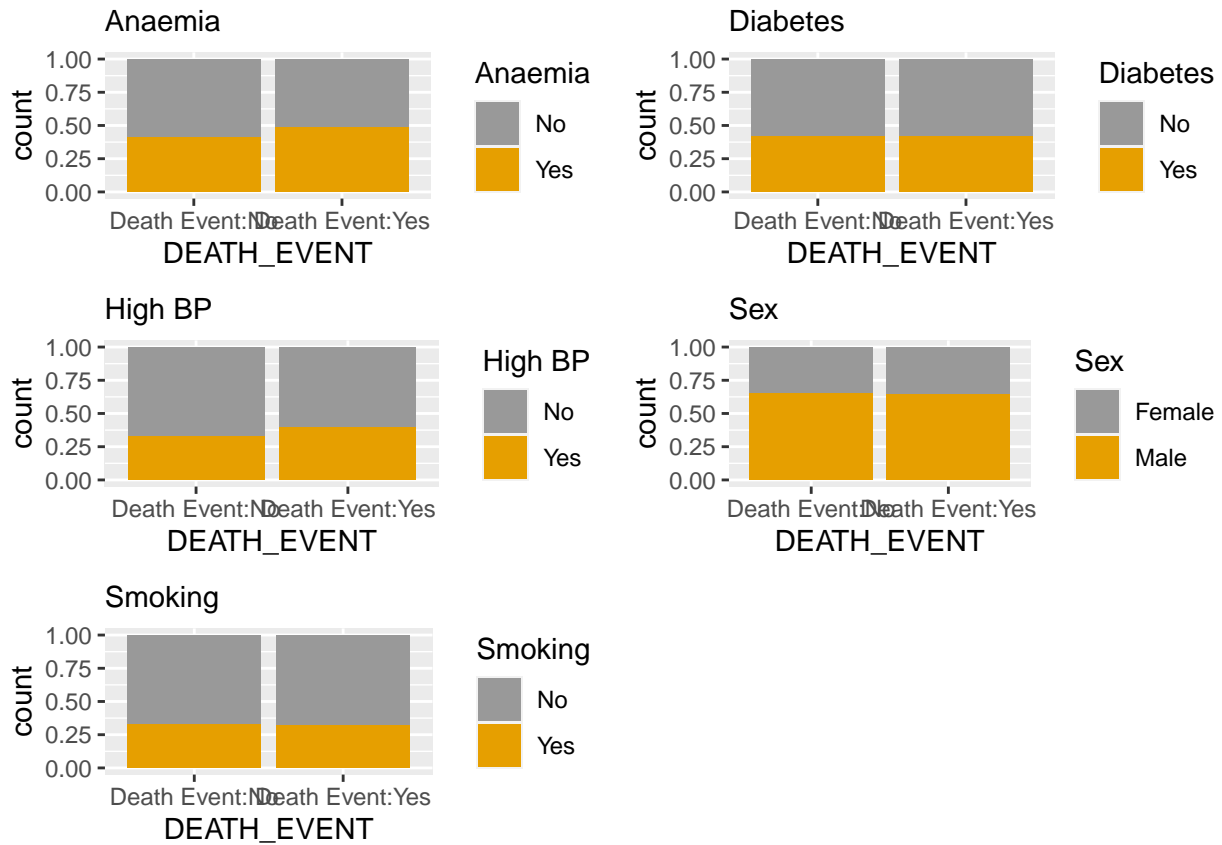
b<-ggplot(data, aes(x = DEATH_EVENT, fill = factor(diabetes)))+
  geom_bar(position = "fill")+
  scale_x_discrete(labels = c("Death Event:No", "Death Event:Yes"))+
  scale_fill_manual(values = c("#999999", "#E69F00"), name = "Diabetes",
    labels = c("No", "Yes"))+labs(subtitle = "Diabetes")

c<-ggplot(data, aes(x = DEATH_EVENT, fill = factor(high_blood_pressure)))+
  geom_bar(position = "fill")+
  scale_x_discrete(labels = c("Death Event:No", "Death Event:Yes"))+
  scale_fill_manual(values = c("#999999", "#E69F00"), name = "High BP",
    labels = c("No", "Yes"))+labs(subtitle = "High BP")

d<-ggplot(data, aes(x = DEATH_EVENT, fill = factor(sex)))+
  geom_bar(position = "fill")+
  scale_x_discrete(labels = c("Death Event:No", "Death Event:Yes"))+
  scale_fill_manual(values = c("#999999", "#E69F00"), name = "Sex",
    labels = c("Female", "Male"))+labs(subtitle = "Sex")

e<-ggplot(data, aes(x = DEATH_EVENT, fill = factor(smoking)))+
  geom_bar(position = "fill")+
  scale_x_discrete(labels = c("Death Event:No", "Death Event:Yes"))+
  scale_fill_manual(values = c("#999999", "#E69F00"), name = "Smoking",
    labels = c("No", "Yes"))+labs(subtitle = "Smoking")

grid.arrange(a,b,c,d,e)
```



We can see that Anaemia and High BP has significant difference for death event whereas others not so much

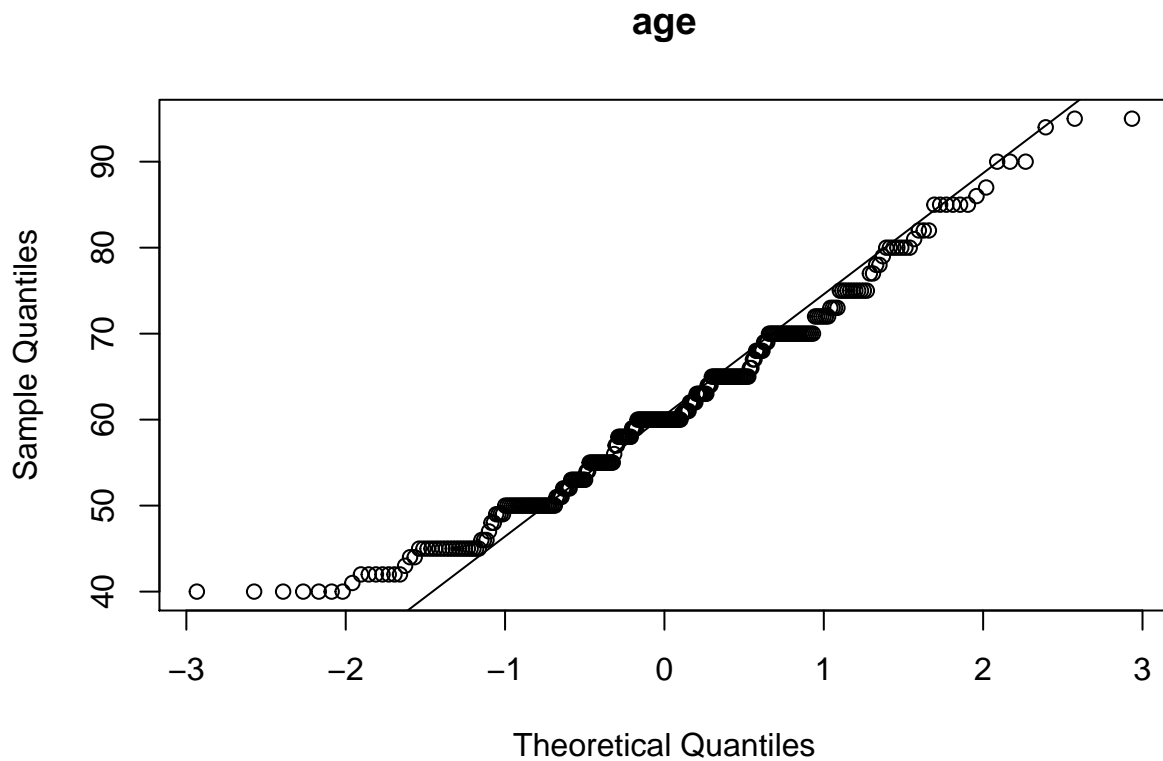
# Tests

## Normality Test

```
# univariate normality
data <- read.csv('/Users/mac/Downloads/heart_failure_clinical_records_dataset.csv')
cm <- colMeans(data)
S <- cov(data)
d <- apply(data, MARGIN = 1, function(data)t(data - cm) %*% solve(S) %*% (data - cm))
```

## Age

```
qqnorm(data[, "age"], main = "age")
qqline(data[, "age"])
```

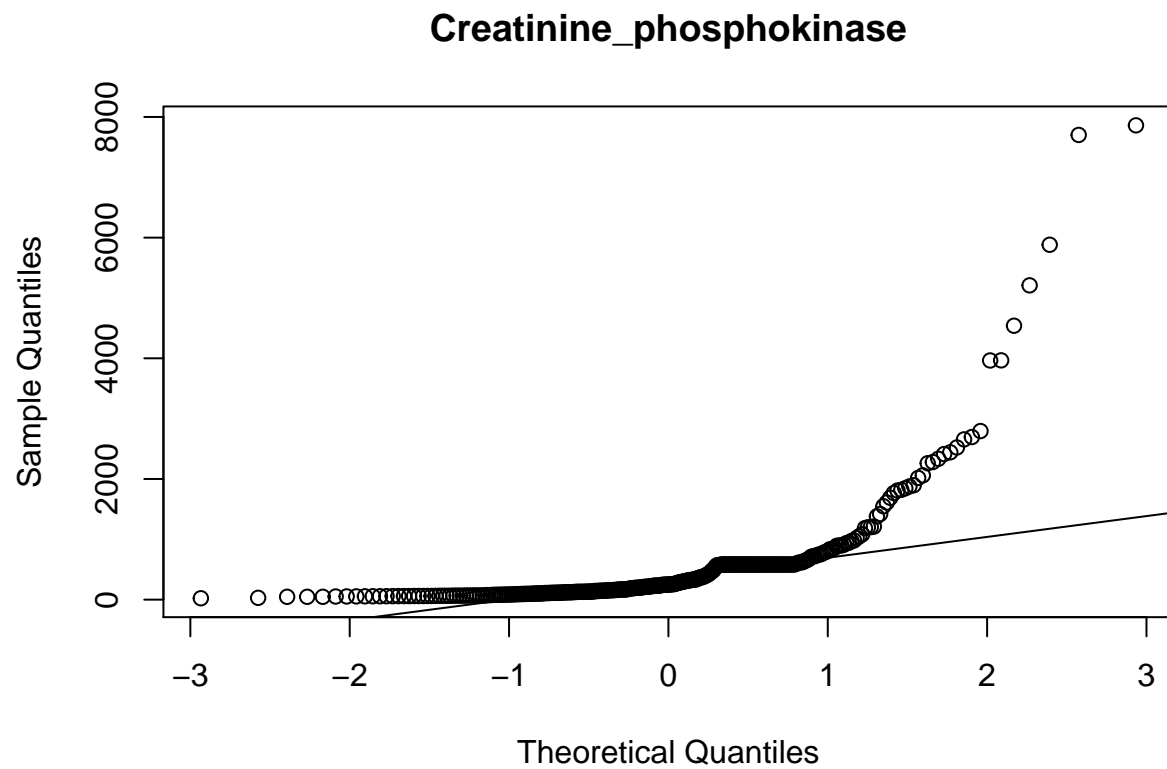


Age looks normally distributed

## Creatinine\_phosphokinase

```
qqnorm(data[, "creatinine_phosphokinase"], main = "Creatinine_phosphokinase")
qqline(data[, "creatinine_phosphokinase"])
```

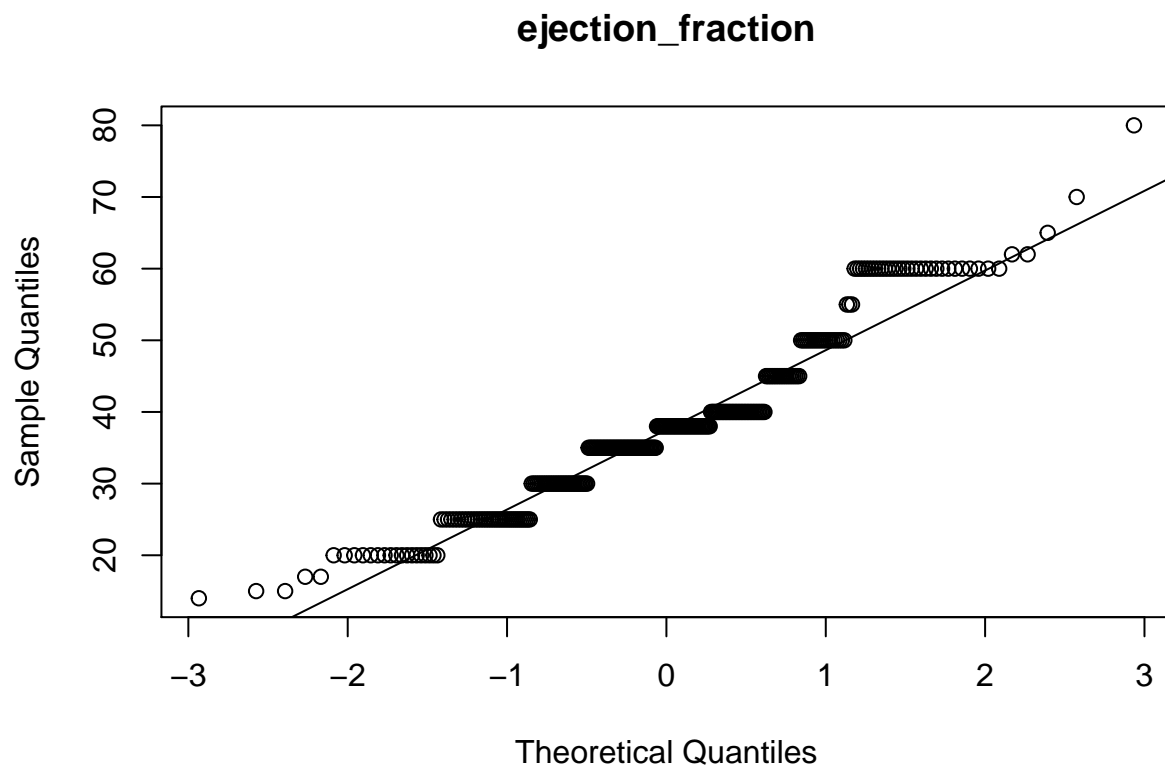




Creatinine\_phosphokinase doesn't look normal but skewed

ejection\_fraction

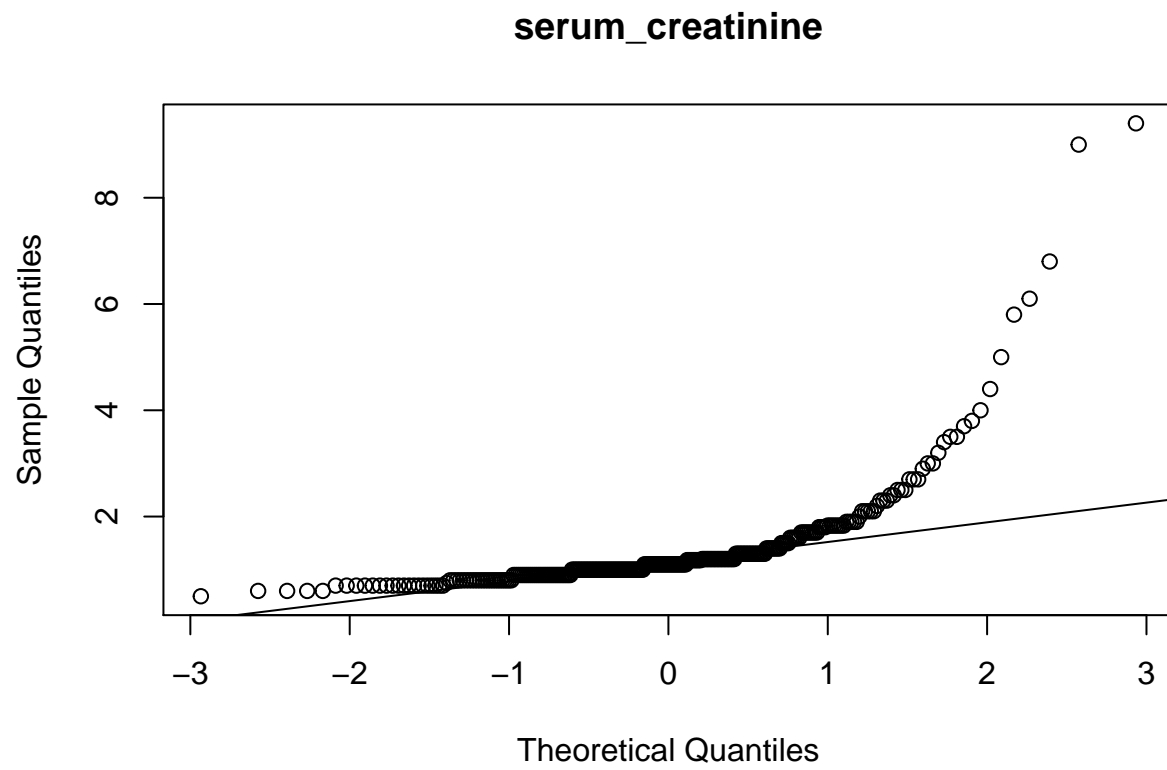
```
qqnorm(data[, "ejection_fraction"], main = "ejection_fraction")  
qqline(data[, "ejection_fraction"])
```



ejection\_fraction doesn't look normal as well

serum\_creatinine

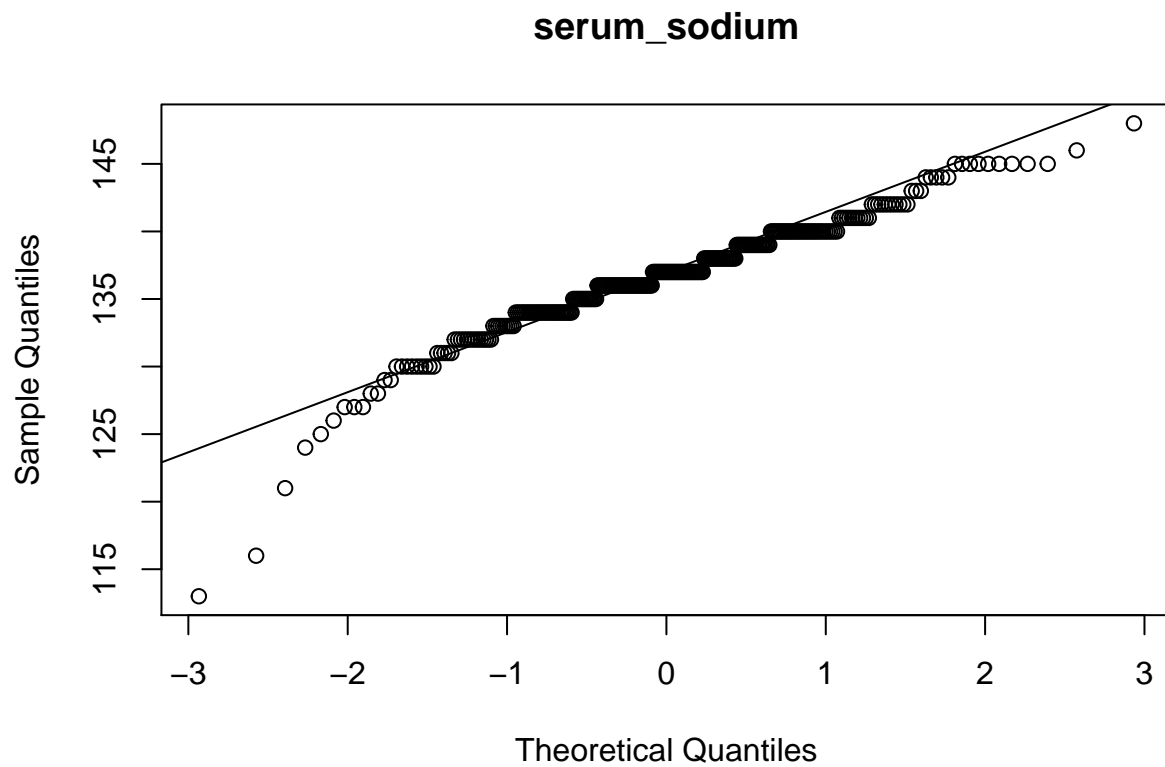
```
qqnorm(data[, "serum_creatinine"], main = "serum_creatinine")
qqline(data[, "serum_creatinine"])
```



serum\_creatinine doesn't look normal but skewed

serum\_sodium

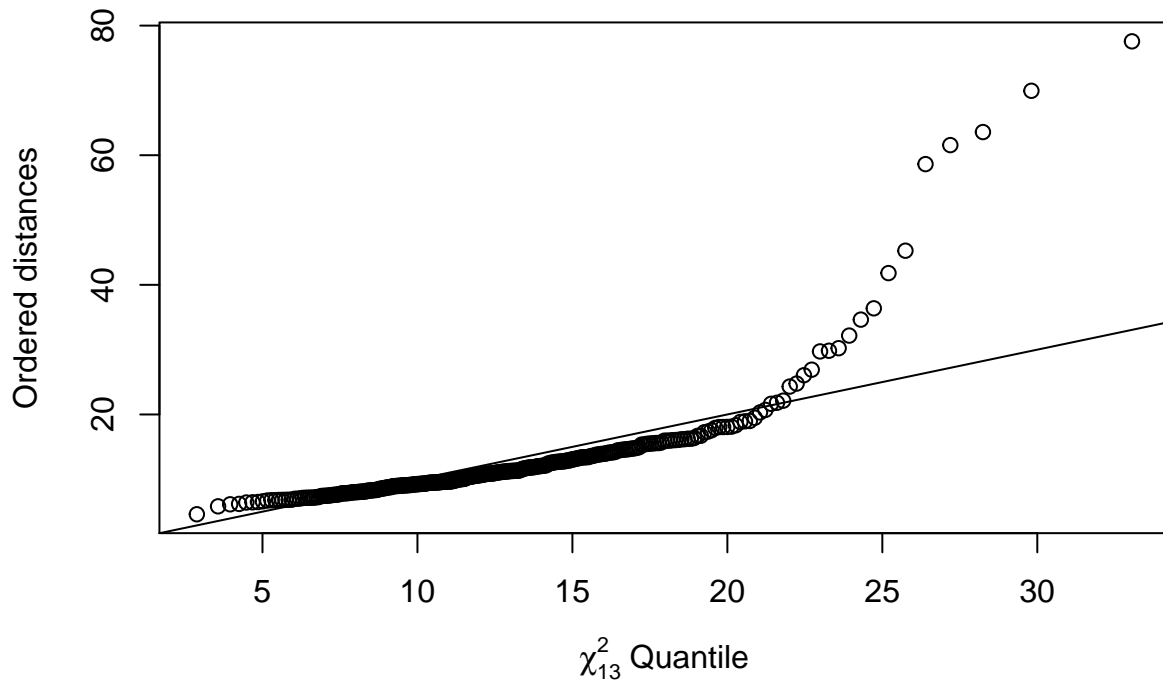
```
qqnorm(data[, "serum_sodium"], main = "serum_sodium")  
qqline(data[, "serum_sodium"])
```



serum\_sodium looks normal but is slightly skewed on lower end

#### Multi-variate normality

```
plot(qchisq((1:nrow(data) - 1/2) / nrow(data), df = 13), sort(d),  
     xlab = expression(paste(chi[13]^2, " Quantile")),  
     ylab = "Ordered distances")  
abline(a = 0, b = 1)
```



While plotting for multivariate normality, we see that data is non normal and has some skewness towards positive side

t-tests for death events vs not for each variable

```
# age
with(data,t.test(age[DEATH_EVENT=="1"],age[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: age[DEATH_EVENT == "1"] and age[DEATH_EVENT == "0"]
## t = 4.5206, df = 297, p-value = 8.917e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.643992 9.262758
## sample estimates:
## mean of x mean of y
##  65.21528  58.76191

# anaemia
with(data,t.test(anaemia[DEATH_EVENT=="1"],anaemia[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: anaemia[DEATH_EVENT == "1"] and anaemia[DEATH_EVENT == "0"]
## t = 1.1446, df = 297, p-value = 0.2533
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.05057162 0.19117096
## sample estimates:
```

```

## mean of x mean of y
## 0.4791667 0.4088670

# creatinine_phosphokinase
with(data,t.test(creatinine_phosphokinase[DEATH_EVENT=="1"],
                 creatinine_phosphokinase[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: creatinine_phosphokinase[DEATH_EVENT == "1"] and creatinine_phosphokinase[DEATH_EVENT == "0"]
## t = 1.0832, df = 297, p-value = 0.2796
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -106.3109 366.5984
## sample estimates:
## mean of x mean of y
## 670.1979 540.0542

# diabetes
with(data,t.test(diabetes[DEATH_EVENT=="1"],
                 diabetes[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: diabetes[DEATH_EVENT == "1"] and diabetes[DEATH_EVENT == "0"]
## t = -0.033483, df = 297, p-value = 0.9733
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1226917 0.1185866
## sample estimates:
## mean of x mean of y
## 0.4166667 0.4187192

# ejection_fraction
with(data,t.test(ejection_fraction[DEATH_EVENT=="1"],
                 ejection_fraction[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: ejection_fraction[DEATH_EVENT == "1"] and ejection_fraction[DEATH_EVENT == "0"]
## t = -4.8056, df = 297, p-value = 2.453e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -9.580849 -4.013671
## sample estimates:
## mean of x mean of y
## 33.46875 40.26601

# high_blood_pressure
with(data,t.test(high_blood_pressure[DEATH_EVENT=="1"],
                 high_blood_pressure[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test

```

```

##
## data: high_blood_pressure[DEATH_EVENT == "1"] and high_blood_pressure[DEATH_EVENT == "0"]
## t = 1.3718, df = 297, p-value = 0.1711
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.03525453 0.19750823
## sample estimates:
## mean of x mean of y
## 0.4062500 0.3251232

# platelets
with(data,t.test(platelets[DEATH_EVENT=="1"],
                  platelets[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: platelets[DEATH_EVENT == "1"] and platelets[DEATH_EVENT == "0"]
## t = -0.84787, df = 297, p-value = 0.3972
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -34129.06 13576.17
## sample estimates:
## mean of x mean of y
## 256381.0 266657.5

# serum_creatinine
with(data,t.test(serum_creatinine[DEATH_EVENT=="1"],
                  serum_creatinine[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: serum_creatinine[DEATH_EVENT == "1"] and serum_creatinine[DEATH_EVENT == "0"]
## t = 5.3065, df = 297, p-value = 2.19e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## 0.409539 0.892374
## sample estimates:
## mean of x mean of y
## 1.835833 1.184877

# serum_sodium
with(data,t.test(serum_sodium[DEATH_EVENT=="1"],
                  serum_sodium[DEATH_EVENT=="0"],var.equal=TRUE))

##
## Two Sample t-test
##
## data: serum_sodium[DEATH_EVENT == "1"] and serum_sodium[DEATH_EVENT == "0"]
## t = -3.4301, df = 297, p-value = 0.0006889
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -2.8984440 -0.7850535
## sample estimates:
## mean of x mean of y

```

```
## 135.3750 137.2167
```

```
# sex
with(data,t.test(sex[DEATH_EVENT=="1"],
                  sex[DEATH_EVENT=="0"],var.equal=TRUE))
```

```
##
## Two Sample t-test
##
## data: sex[DEATH_EVENT == "1"] and sex[DEATH_EVENT == "0"]
## t = -0.074388, df = 297, p-value = 0.9408
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1211614 0.1123355
## sample estimates:
## mean of x mean of y
## 0.6458333 0.6502463
```

```
# smoking
with(data,t.test(smoking[DEATH_EVENT=="1"],
                  smoking[DEATH_EVENT=="0"],var.equal=TRUE))
```

```
##
## Two Sample t-test
##
## data: smoking[DEATH_EVENT == "1"] and smoking[DEATH_EVENT == "0"]
## t = -0.21756, df = 297, p-value = 0.8279
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.1268080 0.1015617
## sample estimates:
## mean of x mean of y
## 0.3125000 0.3251232
```

```
# time
with(data,t.test(time[DEATH_EVENT=="1"],
                  time[DEATH_EVENT=="0"],var.equal=TRUE))
```

```
##
## Two Sample t-test
##
## data: time[DEATH_EVENT == "1"] and time[DEATH_EVENT == "0"]
## t = -10.686, df = 297, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -103.5612 -71.3478
## sample estimates:
## mean of x mean of y
## 70.88542 158.33990
```

p-value is below 0.05 for -

1. age
2. serum\_sodium
3. serum\_creatinine
4. ejection\_fraction
5. time

so we may conclude that death event does differ by these variables



## t-test multi-variate

```
# Hotelling's T2 test. Comparing multivariate means between death events and non-death event
t2testdata <- hotelling.test(age + anaemia + creatinine_phosphokinase +
                             diabetes + ejection_fraction + high_blood_pressure +
                             platelets + serum_creatinine + serum_sodium +
                             sex + smoking + time
                             ~ DEATH_EVENT, data)

cat("T2 statistic =", t2testdata$stat[[1]], "\n")
```

```
## T2 statistic = 212.2908
```

```
print(t2testdata)
```

```
## Test stat: 17.036
## Numerator df: 12
## Denominator df: 286
## P-value: 0
```

The difference in means in the two groups taken together is significant as well

## Homoskedasticity check

```
data$DEATH_EVENT <- factor(data$DEATH_EVENT)
leveneTest(age ~ DEATH_EVENT, data=data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value  Pr(>F)
## group 1  7.1338 0.007981 **
##      297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(anaemia ~ DEATH_EVENT, data=data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  1.3101 0.2533
##      297
```

```
leveneTest(creatinine_phosphokinase ~ DEATH_EVENT, data=data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  1.0303 0.3109
##      297
```

```
leveneTest(diabetes ~ DEATH_EVENT, data=data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  0.0011 0.9733
```

```

##          297
leveneTest(ejection_fraction ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group    1  3.7021 0.0553 .
##          297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

leveneTest(high_blood_pressure ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group    1  1.8819 0.1711
##          297

leveneTest(platelets ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group    1  1.085 0.2984
##          297

leveneTest(serum_creatinine ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value    Pr(>F)
## group    1 16.242 7.087e-05 ***
##          297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

leveneTest(serum_sodium ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value  Pr(>F)
## group    1  5.274 0.02234 *
##          297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

leveneTest(sex ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group    1  0.0055 0.9408
##          297

leveneTest(smoking ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df F value Pr(>F)
## group    1  0.0473 0.8279
##          297

leveneTest(time ~ DEATH_EVENT, data=data)

## Levene's Test for Homogeneity of Variance (center = median)

```

```
##           Df F value   Pr(>F)
## group      1  7.9512 0.005129 **
##           297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is below 0.05 for -

1. age
2. serum\_creatinine
3. ejection\_fraction
4. time
5. serum\_sodium

so we may conclude that variance between the two groups differ in them

### One-way ANOVA tests: comparing univariate means

```
aov_age <- aov(age ~ DEATH_EVENT, data)
summary(aov_age)
```

```
##           Df Sum Sq Mean Sq F value   Pr(>F)
## DEATH_EVENT  1    2714    2714.4    20.44 8.92e-06 ***
## Residuals    297   39449     132.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov_anaemia <- aov(anaemia ~ DEATH_EVENT, data)
summary(aov_anaemia)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT  1     0.32   0.3221     1.31 0.253
## Residuals    297   73.02   0.2459
```

```
aov_creatinine_phosphokinase <- aov(creatinine_phosphokinase ~ DEATH_EVENT, data)
summary(aov_creatinine_phosphokinase)
```

```
##           Df      Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT  1   1103933 1103933     1.173  0.28
## Residuals    297 279450722   940912
```

```
aov_diabetes <- aov(diabetes ~ DEATH_EVENT, data)
summary(aov_diabetes)
```

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT  1    0.00 0.00027   0.001  0.973
## Residuals    297   72.74 0.24492
```

```
aov_ejection_fraction <- aov(ejection_fraction ~ DEATH_EVENT, data)
summary(aov_ejection_fraction)
```

```
##           Df Sum Sq Mean Sq F value   Pr(>F)
## DEATH_EVENT  1    3011    3011.4    23.09 2.45e-06 ***
## Residuals    297   38728     130.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov_high_blood_pressure <- aov(high_blood_pressure ~ DEATH_EVENT, data)
summary(aov_high_blood_pressure)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT   1    0.43  0.4290   1.882  0.171
## Residuals  297   67.70  0.2279
```

```
aov_platelets <- aov(platelets ~ DEATH_EVENT, data)
summary(aov_platelets)
```

```
##              Df      Sum Sq   Mean Sq F value Pr(>F)
## DEATH_EVENT   1 6.883e+09 6.883e+09   0.719  0.397
## Residuals  297 2.844e+12 9.575e+09
```

```
aov_serum_creatinine <- aov(serum_creatinine ~ DEATH_EVENT, data)
summary(aov_serum_creatinine)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## DEATH_EVENT   1  27.62  27.618   28.16 2.19e-07 ***
## Residuals  297 291.30   0.981
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov_serum_sodium <- aov(serum_sodium ~ DEATH_EVENT, data)
summary(aov_serum_sodium)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## DEATH_EVENT   1    221  221.08   11.77 0.000689 ***
## Residuals  297   5581   18.79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov_sex <- aov(sex ~ DEATH_EVENT, data)
summary(aov_sex)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT   1    0.00  0.00127   0.006  0.941
## Residuals  297   68.13  0.22938
```

```
aov_smoking <- aov(smoking ~ DEATH_EVENT, data)
summary(aov_smoking)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT   1    0.01  0.01039   0.047  0.828
## Residuals  297   65.17  0.21942
```

```
aov_time <- aov(time ~ DEATH_EVENT, data)
summary(aov_time)
```

```
##              Df   Sum Sq Mean Sq F value Pr(>F)
## DEATH_EVENT   1  498494  498494   114.2 <2e-16 ***
## Residuals  297 1296647   4366
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value is below 0.05 for -

1. age
2. serum\_creatinine
3. ejection\_fraction

4. time  
 5. serum\_sodium  
 so we may conclude that means between the two groups differ in them

## Comparing multivariate means (One-way MANOVA)

```
mnv <- manova(as.matrix(data[, -13]) ~ DEATH_EVENT, data)
summary(mnv)

##              Df  Pillai approx F num Df den Df      Pr(>F)
## DEATH_EVENT    1  0.41684    17.036     12   286 < 2.2e-16 ***
## Residuals     297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We observe from MANOVA that estimated effects may be unbalanced indicating that mean between groups may be different

## Multi-collinearity check

```
# Let us also look at Multicollinearity check
# Earlier we saw the correlation plot
# We will check VIF for this purpose
# In classification, although linear regression isn't to be used
# For VIF, a rudimentary model lets us know the association
# between continuous and categorical variables

data <- read.csv('/Users/mac/Downloads/heart_failure_clinical_records_dataset.csv')
mod <- lm( DEATH_EVENT ~ age+anaemia+creatinine_phosphokinase+
  diabetes+ejection_fraction+high_blood_pressure+platelets+
  serum_creatinine+serum_sodium+sex+smoking+time, data)
summary(mod)

##
## Call:
## lm(formula = DEATH_EVENT ~ age + anaemia + creatinine_phosphokinase +
##      diabetes + ejection_fraction + high_blood_pressure + platelets +
##      serum_creatinine + serum_sodium + sex + smoking + time, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.80866 -0.28041 -0.04205  0.24742  0.96983
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.664e+00  6.954e-01   2.392  0.01738 *
## age             5.767e-03  1.867e-03   3.088  0.00221 **
## anaemia        -2.766e-03  4.438e-02  -0.062  0.95035
## creatinine_phosphokinase 3.427e-05  2.247e-05   1.525  0.12840
## diabetes        1.928e-02  4.410e-02   0.437  0.66236
## ejection_fraction -9.834e-03  1.844e-03  -5.333 1.96e-07 ***
```

```
## high_blood_pressure      -1.430e-02  4.565e-02  -0.313  0.75438
## platelets                -8.370e-08  2.208e-07  -0.379  0.70492
## serum_creatinine         8.527e-02  2.123e-02  4.017  7.54e-05 ***
## serum_sodium             -7.599e-03  5.024e-03  -1.513  0.13149
## sex                      -6.369e-02  5.108e-02  -1.247  0.21353
## smoking                  -5.733e-03  5.119e-02  -0.112  0.91091
## time                     -2.733e-03  2.903e-04  -9.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3646 on 286 degrees of freedom
## Multiple R-squared:  0.4168, Adjusted R-squared:  0.3924
## F-statistic: 17.04 on 12 and 286 DF,  p-value: < 2.2e-16
```

```
vif(mod)
```

```
##                age                anaemia creatinine_phosphokinase
##          1.106067                1.087163                1.066014
##          diabetes      ejection_fraction      high_blood_pressure
##          1.064324                1.067758                1.068377
##          platelets      serum_creatinine      serum_sodium
##          1.045809                1.081241                1.101927
##          sex                smoking                time
##          1.337716                1.285049                1.138009
```

We see that most VIF values are below 1.5  
This indicates absence of multi-collinearity in our data

**This concludes our initial EDA for the data**