Heart Disease Data Set

# Introduction

Cardiovascular diseases are disorders of the heart and blood vessels including Heart failure. About 26 million people in United States had died due to heart failure in past. In 2018 heart failure was mentioned on 379,800 death certificates which account 13.4 % of total. The factors causing the heart failure are Diabetes, High blood pressure, Coronary artery disease, Obesity and Valvular Heart disease. There are other unhealthy behaviour’s which increases the risk of heart failure namely smoking, eating foods high in fat, Cholesterol and Sodium, Lack of physical activity and excessive alcohol consumption. The goal of this project is to predict survival of patients with heart failure with other variables by using R.

Dataset:-  
The data set containing the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan). These data set is taken from UCI Machine learning Repository(<https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>). The dataset contains 13 characteristics which cover the information on lifestyle of each patient. Some characteristics are binary and other are numerical.Each row explains the underlying health condition in the patient (like gender,smoking,diabetes,anaemia and High Blood Pressure) and value of attributes associated with patient(like, platelets, Serum Creatinine, ejection fraction etc).I am conducting an observation study to analyse the variables behaviour in different patients and predict survival pattern.The Creatinine Phosphokinase (CPK) is the enzyme produced in blood when a blood vessel is damaged, so I want to study whether the damage to blood vessel is more in aged people and does it cause deaths due to heart failure. A Serum Creatinine is generally measured in kidney failures but now a days there is heart-kidney interaction are increasingly recognized by researches involved in study of heart failures and kidney disease. Ejection Fraction is the proportion of blood pumped out of heart during a single contraction. There are different types of heart failure due to rate of ejection fraction namely 1) Heart failure with reduced ejection fraction (HFrEF): EF less than or equal to 40%( systolic heart failure), 2) Heart failure with preserved EF (HFpEF): EF is greater than or equal to 50%(diastolic heart failure). The Characteristics are listed in below table

|  |  |  |  |
| --- | --- | --- | --- |
| **Characteristic** | **Description** | **Measurement** | **Range** |
| Age | Age of the patient | Years | 40 to 95 |
| Anaemia | Decrease of red blood cells or haemoglobin | Boolean | 0 = false  1 = true |
| Creatinine Phosphokinase (CPK) | Level of CPK enzyme in the blood | mcg/L (micrograms per litre) | 23 to7861 |
| Diabetes | If the patient has diabetes | Boolean | 0 = false  1 = true |
| Ejection Fraction | Percentage of blood leaving the heart at each contraction | Percentage | 14 to 80 |
| High Blood Pleasure | If the patient has hypertension | Boolean | 0 = false  1 = true |
| Gender | Woman or Man | Binary | 0 = Woman  1 = Man |
| Platelets | Platelets in the blood | Kilo platelets/mL | 25.01 to 850.0 |
| Serum Creatinine | Level of Creatinine in the blood | mg/dL | 0.50 to 9.40 |
| Serum Sodium | Level of Sodium in the Blood | mEq/L (milliequivalents per litre) | 114 to 148 |
| Smoking | If Patient smokes | Boolean | 0 = false  1 = true |
| Time | Follow up period | Days | 4 to 285 |
| Death Event | If the patient dies during the follow-up period | Boolean | 0 = false  1 = true |

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.3 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

data1<-read.csv("C:/Users/Vasu/Documents/R Programing/heart\_failure\_clinical\_records\_dataset.csv")  
# Checking weather the data is in tibble format or not and converting it into tibble format  
is\_tibble(data1)

## [1] FALSE

data2<-as\_tibble(data1)  
is\_tibble(data2)

## [1] TRUE

head(data2)

## # A tibble: 6 x 13  
## age anaemia creatinine\_phosphok~ diabetes ejection\_fracti~ high\_blood\_press~  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 75 0 582 0 20 1  
## 2 55 0 7861 0 38 0  
## 3 65 0 146 0 20 0  
## 4 50 1 111 0 20 0  
## 5 65 1 160 1 20 0  
## 6 90 1 47 0 40 1  
## # ... with 7 more variables: platelets <dbl>, serum\_creatinine <dbl>,  
## # serum\_sodium <dbl>, sex <dbl>, smoking <dbl>, time <dbl>, DEATH\_EVENT <dbl>

summary(data2)

## age anaemia creatinine\_phosphokinase diabetes   
## Min. :40.00 Min. :0.0000 Min. : 23.0 Min. :0.0000   
## 1st Qu.:51.00 1st Qu.:0.0000 1st Qu.: 116.5 1st Qu.:0.0000   
## Median :60.00 Median :0.0000 Median : 250.0 Median :0.0000   
## Mean :60.83 Mean :0.4314 Mean : 581.8 Mean :0.4181   
## 3rd Qu.:70.00 3rd Qu.:1.0000 3rd Qu.: 582.0 3rd Qu.:1.0000   
## Max. :95.00 Max. :1.0000 Max. :7861.0 Max. :1.0000   
## ejection\_fraction high\_blood\_pressure platelets serum\_creatinine  
## Min. :14.00 Min. :0.0000 Min. : 25100 Min. :0.500   
## 1st Qu.:30.00 1st Qu.:0.0000 1st Qu.:212500 1st Qu.:0.900   
## Median :38.00 Median :0.0000 Median :262000 Median :1.100   
## Mean :38.08 Mean :0.3512 Mean :263358 Mean :1.394   
## 3rd Qu.:45.00 3rd Qu.:1.0000 3rd Qu.:303500 3rd Qu.:1.400   
## Max. :80.00 Max. :1.0000 Max. :850000 Max. :9.400   
## serum\_sodium sex smoking time   
## Min. :113.0 Min. :0.0000 Min. :0.0000 Min. : 4.0   
## 1st Qu.:134.0 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 73.0   
## Median :137.0 Median :1.0000 Median :0.0000 Median :115.0   
## Mean :136.6 Mean :0.6488 Mean :0.3211 Mean :130.3   
## 3rd Qu.:140.0 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:203.0   
## Max. :148.0 Max. :1.0000 Max. :1.0000 Max. :285.0   
## DEATH\_EVENT   
## Min. :0.0000   
## 1st Qu.:0.0000   
## Median :0.0000   
## Mean :0.3211   
## 3rd Qu.:1.0000   
## Max. :1.0000

# Checking weather there are any Missing values in th data set  
which(is.na(data2))

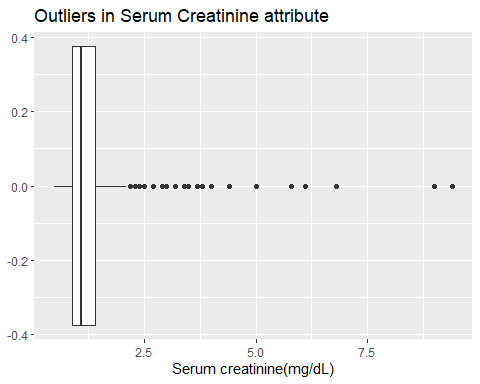
## integer(0)

dataf<-data2

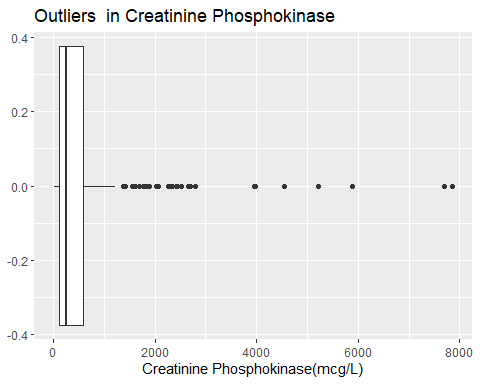
## Data Cleaning

The missing values have been checked and removed using na.omit command. Outliers some time lead to misleading conclusions but in some cases outliers cannot be removed because they are needed in the analysis to arrive at conclusions. So,I am cleaning these variables Creatinine Phosphokinase for my analysis. i don’t want to remove outliers because in health the extreme high and low values gives signs of predicting something important so, eliminating any outliers lead to inaccurate conclusions. We are using boxplot to see if any outlier are present in the variable and if present they are cleansed with the filter command and boxplot is plotted again to check.From the boxplot we can see the points which lie after maximum(Q3+1.5\*IQR) these are outliers which needs to be taken care in data visualization and analysis to arrive at correct conclusion. So the box plot is again plotted by eliminating these outliers.

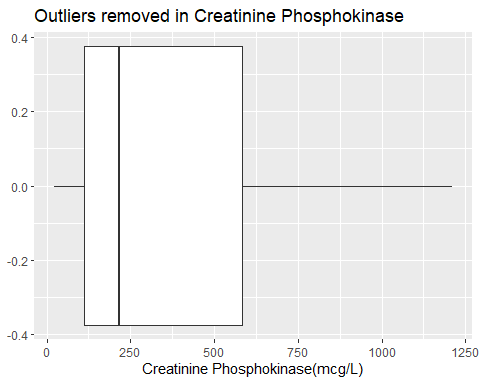
# Checking and cleaning the outlier in Serum Creatinine  
ggplot(dataf)+geom\_boxplot(aes(x=serum\_creatinine))+labs(title = "Outliers in Serum Creatinine attribute", x="Serum creatinine(mg/dL)")



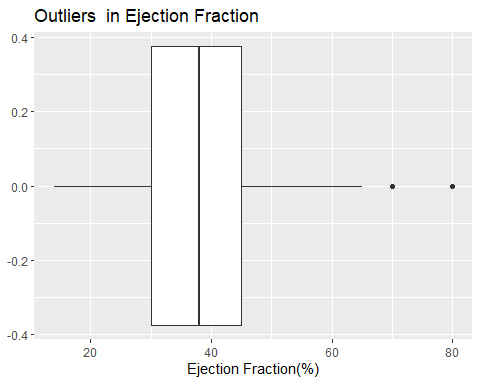
# Checking and cleaning the outlier in Creatinine Phosphokinase  
ggplot(data2)+geom\_boxplot(aes(x=creatinine\_phosphokinase))+labs(title = "Outliers in Creatinine Phosphokinase", x="Creatinine Phosphokinase(mcg/L)")



b<-filter(dataf,creatinine\_phosphokinase<1300)  
ggplot(b)+geom\_boxplot(aes(x=creatinine\_phosphokinase))+labs(title = "Outliers removed in Creatinine Phosphokinase", x="Creatinine Phosphokinase(mcg/L)")



d<-filter(b,serum\_creatinine<2)  
ggplot(dataf)+geom\_boxplot(aes(x=ejection\_fraction))+labs(title = "Outliers in Ejection Fraction", x="Ejection Fraction(%)")



# Visualization

These are the following question which I am interested to analyse from the data set:-  
1)What age group patients are dying more from heart failure?  
2)which gender suffering from heart failure diseases have more survival rate?  
3)Is there any relation between Ejection Fraction, age, and death event?  
4)Is there any relation between Serum creatinine, age, and death event?  
5)Whether high level of creatinine phosphokinase increase with age and increase mortality in patients suffering from heart failure diseases?

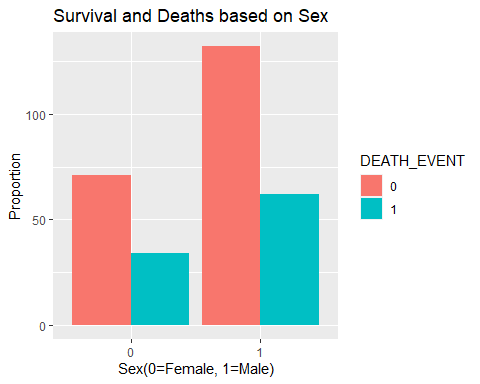
ggplot(dataf, aes(x = age)) + geom\_density() + geom\_vline(aes(xintercept = mean(age)), linetype = "dashed", size = 0.6)



d1<-filter(dataf,DEATH\_EVENT == '1')  
ggplot(d1, aes(x = age)) + geom\_density() + geom\_vline(aes(xintercept = mean(age)), linetype = "dashed", size = 0.6)



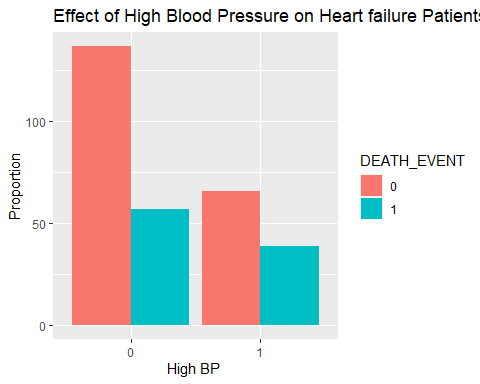
#tbl<-with(dataf, table(age,DEATH\_EVENT))  
#ggplot(as.data.frame(tbl), aes(factor(age), Freq,fill = DEATH\_EVENT))+geom\_col(position = 'dodge')+labs(title = "Age vs Death Event", x="Age of Patients",y="Frequency of Death Event")+coord\_flip()  
gender<-with(dataf, table(sex,DEATH\_EVENT))  
ggplot(as.data.frame(gender), aes(factor(sex), Freq,fill = DEATH\_EVENT))+geom\_col(position = 'dodge')+labs(title = "Survival and Deaths based on Sex", x="Sex(0=Female, 1=Male)",y="Proportion")



bp<-with(dataf, table(high\_blood\_pressure,DEATH\_EVENT))  
dataf%>%  
 count(sex,DEATH\_EVENT)

## # A tibble: 4 x 3  
## sex DEATH\_EVENT n  
## <dbl> <dbl> <int>  
## 1 0 0 71  
## 2 0 1 34  
## 3 1 0 132  
## 4 1 1 62

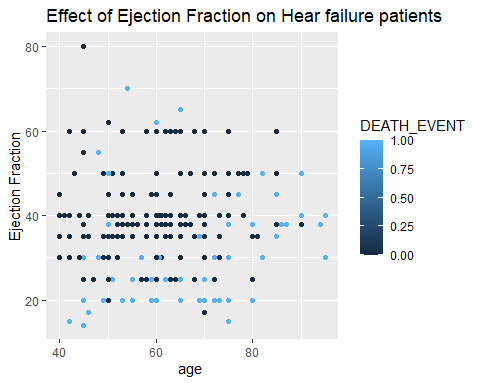
ggplot(as.data.frame(bp), aes(factor(high\_blood\_pressure), Freq,fill = DEATH\_EVENT))+geom\_col(position = 'dodge')+labs(title = "Effect of High Blood Pressure on Heart failure Patients", x="High BP ",y="Proportion")



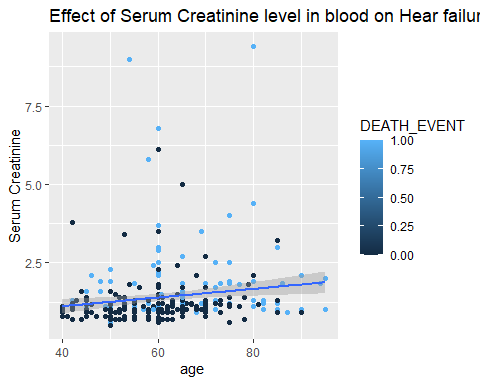
dataf%>%  
 count(high\_blood\_pressure,DEATH\_EVENT)

## # A tibble: 4 x 3  
## high\_blood\_pressure DEATH\_EVENT n  
## <dbl> <dbl> <int>  
## 1 0 0 137  
## 2 0 1 57  
## 3 1 0 66  
## 4 1 1 39

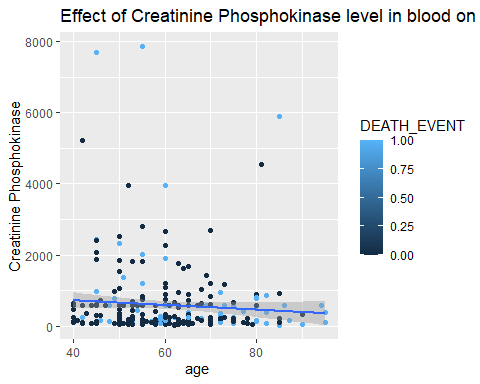
ggplot(dataf, aes(x= age, y= ejection\_fraction, color = DEATH\_EVENT))+geom\_point()+labs(title = "Effect of Ejection Fraction on Hear failure patients ", x="age ",y="Ejection Fraction")



ggplot(dataf, aes(x= age, y= serum\_creatinine, color = DEATH\_EVENT))+geom\_point()+geom\_smooth(formula = y ~ x, method = "lm")+labs(title = "Effect of Serum Creatinine level in blood on Hear failure patients ", x="age ",y="Serum Creatinine")



ggplot(dataf, aes(x= age, y= creatinine\_phosphokinase, color = DEATH\_EVENT))+geom\_point()+geom\_smooth(formula = y ~ x, method = "lm")+labs(title = "Effect of Creatinine Phosphokinase level in blood on Hear failure patients ", x="age ",y="Creatinine Phosphokinase")

 ## EDA  
From the Age density plot and drawing a line at the mean age we can conclude that most patient having heart failure condition are around 60 age. In second Age plot i have plotted age density plot of only patients died due to heart failure and mean age was around 65. I wanted to check whether there is any particular gender vulnerable to deaths caused by heart failure.So, i had plotted a bar graph “Survival and Deaths based on Sex” from that we can see that both male and female has equally probability of survival. The normal Ejection fraction of a healthy human being if 50 to 70 and 40 to 50 tells us that a patient is suffering from some thing and below 40 mean danger. From the “Effect of Ejection Fraction on Hear failure patients” graph we can say that patients with low ejection fraction rate i.e, < 40 have high mortality rate. Patients moderate to good ejection fraction rate i.e 40 to 70 have good chances of survival eventhough suffering from heart failure. From graph we can say Age has no effect on Ejection fraction. From the Graph “Effect of Serum Creatinine level in blood on Hear failure patients” we can visualize that patients with more serum creatinine level in the blood are more pron to death. Patients with less serum creatinine level have high chances of survival. There is slight increase in serum creatinine level as the age of patient increases. From th graph plotted Creatinine Phosphokinase vs age along with death we can say that as age increase there in no increase in Creatinine phosphokinase level in blood and both are independent. we can say that very high level of creatinine phosphokinase(CPK) increaeses mortality rate but can also see that patients with low and medium level CPK in blood are equally prone to death. so we cannot conclude that high CPK level in blood increase death in patients suffering from heart failure diseases. From all above Visualizations we can say that that paitents suffering from heart failure with low ejection fraction percent and high serum creatinine levels have high mortality rate.

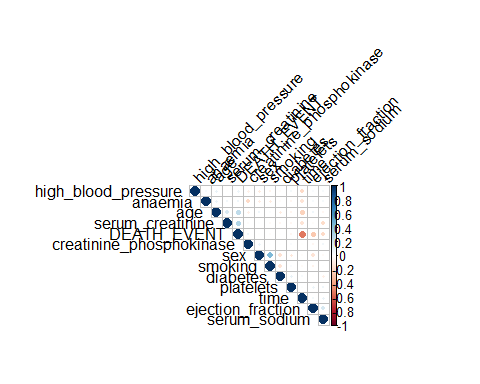
# Principle Component Analysis (PCA)

The most common application of PCA is to represent a multivariate data table as a smaller number of variables (summary indices) so that trends, jumps, clusters, and outliers may be observed. This overview may uncover the relationships between observations and variables, and among the variables.

#Performing PCA  
library(corrplot)

## corrplot 0.90 loaded

res<-cor(dataf[, unlist(lapply(dataf, is.numeric))])  
corrplot(res, type = "upper", order = "hclust",   
 tl.col = "black", tl.srt = 45)



p<-select(d, age,creatinine\_phosphokinase,ejection\_fraction,platelets,serum\_creatinine,serum\_sodium)  
p<-data.frame(p)  
cv<-cov(p)  
scaled\_cv <- apply(p, 2, scale)  
cr1<-cor(p)  
cr1

## age creatinine\_phosphokinase ejection\_fraction  
## age 1.00000000 0.015335388 0.06050369  
## creatinine\_phosphokinase 0.01533539 1.000000000 -0.08690512  
## ejection\_fraction 0.06050369 -0.086905118 1.00000000  
## platelets -0.09513779 -0.003686027 0.04285817  
## serum\_creatinine 0.22019908 0.111205836 -0.17547489  
## serum\_sodium 0.02405617 -0.180843971 0.10507995  
## platelets serum\_creatinine serum\_sodium  
## age -0.095137787 0.22019908 0.02405617  
## creatinine\_phosphokinase -0.003686027 0.11120584 -0.18084397  
## ejection\_fraction 0.042858166 -0.17547489 0.10507995  
## platelets 1.000000000 -0.01684866 0.06725604  
## serum\_creatinine -0.016848662 1.00000000 -0.25097735  
## serum\_sodium 0.067256041 -0.25097735 1.00000000

e2<-eigen(cr1)  
e2$values

## [1] 1.5010982 1.1298707 0.9849928 0.9359628 0.8415822 0.6064934

e2$vectors

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] 0.2329751 0.7684713 -0.24294026 0.02545441 -0.23801268 0.4887295  
## [2,] 0.3935316 -0.2766219 -0.16243591 -0.59098817 -0.61823179 -0.1036832  
## [3,] -0.3713185 0.3414443 -0.29167028 -0.66290692 0.36986343 -0.2902108  
## [4,] -0.1693608 -0.3488388 -0.87082230 0.22686877 0.02704928 0.1977268  
## [5,] 0.5891383 0.2026765 -0.26402196 0.28322377 0.19227944 -0.6518770  
## [6,] -0.5265083 0.2314772 -0.03852252 0.28098813 -0.62180175 -0.4495904

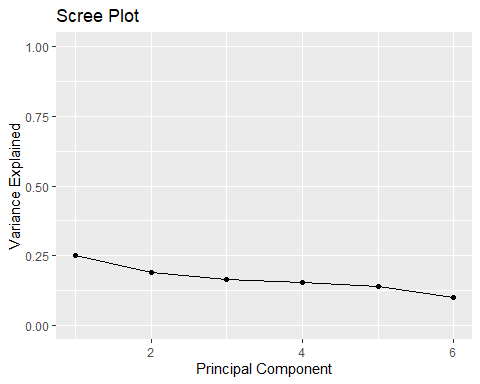
# calculating percentage variance  
pv2<-e2$values/sum(e2$values)  
pv2

## [1] 0.2501830 0.1883118 0.1641655 0.1559938 0.1402637 0.1010822

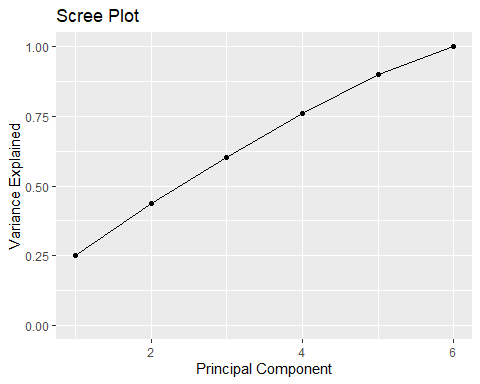
cumsum(pv2)

## [1] 0.2501830 0.4384948 0.6026603 0.7586541 0.8989178 1.0000000

# plotting scree plot for both PVE and cumulative PVE values  
qplot(c(1:6), pv2) + geom\_line() + xlab("Principal Component") + ylab("Variance Explained") +ggtitle("Scree Plot") +ylim(0, 1)



qplot(c(1:6), cumsum(pv2)) + geom\_line() + xlab("Principal Component") + ylab("Variance Explained") +ggtitle("Scree Plot") +ylim(0, 1)



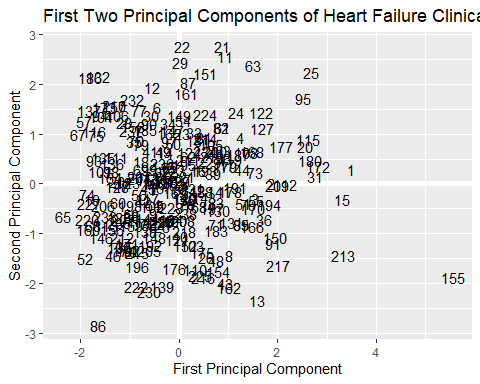
eig\_v = e2$vectors[,1:5]  
colnames(eig\_v) = c("Pcomp\_1", "Pcomp\_2", "Pcomp\_3", "Pcomp\_4", "Pcomp\_5")  
row.names(eig\_v) = colnames(p)  
eig\_v

## Pcomp\_1 Pcomp\_2 Pcomp\_3 Pcomp\_4  
## age 0.2329751 0.7684713 -0.24294026 0.02545441  
## creatinine\_phosphokinase 0.3935316 -0.2766219 -0.16243591 -0.59098817  
## ejection\_fraction -0.3713185 0.3414443 -0.29167028 -0.66290692  
## platelets -0.1693608 -0.3488388 -0.87082230 0.22686877  
## serum\_creatinine 0.5891383 0.2026765 -0.26402196 0.28322377  
## serum\_sodium -0.5265083 0.2314772 -0.03852252 0.28098813  
## Pcomp\_5  
## age -0.23801268  
## creatinine\_phosphokinase -0.61823179  
## ejection\_fraction 0.36986343  
## platelets 0.02704928  
## serum\_creatinine 0.19227944  
## serum\_sodium -0.62180175

PrComp\_1 <- as.matrix(scaled\_cv) %\*% eig\_v[,1]  
PrComp\_2 <- as.matrix(scaled\_cv) %\*% eig\_v[,2]  
PrComp\_3 <- as.matrix(scaled\_cv) %\*% eig\_v[,3]  
PrComp\_4 <- as.matrix(scaled\_cv) %\*% eig\_v[,4]  
PrComp\_5 <- as.matrix(scaled\_cv) %\*% eig\_v[,5]  
#plotting Principle component 1 and 2 of covarience matrix  
PComp\_v <- data.frame(model = row.names(p),PrComp\_1, PrComp\_2, PrComp\_3, PrComp\_4, PrComp\_5)  
head(PComp\_v)

## model PrComp\_1 PrComp\_2 PrComp\_3 PrComp\_4 PrComp\_5  
## 1 1 3.4905355 0.291283006 -0.5648852 0.7787931 0.09130702  
## 2 2 1.8759301 0.009502044 1.3554830 0.8269617 1.01180233  
## 3 3 1.5537861 -0.293442589 0.6706749 2.0539686 0.58741958  
## 4 4 1.2455612 0.930031516 1.5454045 1.2758781 -0.82062148  
## 5 5 -0.3735161 -0.442823095 -2.1645131 -1.0976047 1.64315465  
## 6 6 -0.4430421 1.530441325 -0.9331574 -0.6644550 1.19695786

ggplot(PComp\_v, aes(PrComp\_1, PrComp\_2)) + modelr::geom\_ref\_line(h = 0) + modelr::geom\_ref\_line(v = 0) + geom\_text(aes(label=model), size = 4) + xlab("First Principal Component") + ylab("Second Principal Component") + ggtitle("First Two Principal Components of Heart Failure Clinical Records data set")



I had chosen correlation matrix to perform the principal component analysis. so PCA can be performend on continuous varible we have reduced our attribute and formed a new data set p. We have calculated the correlation matrix and calculated eigen vectors. From the correlation matrix we can tell that age and serum creatinine has high positive relation(0.22) among all and serum creatinine and serum sodium has highest negative relationship(-0.25).Using the eigen values percentage variance explained(PVE) and cumulative PVE also calculated. The Percentage Variance explained values tells us the variation of values and the first five account for 89 percentage of variance so the sixth principle component is value is excluded. From th Principle component matrix we can see serum creatinine is strongly related to Principle component 1 compared to others and age is strongly releated to principle component 2.

## Logistic Regression

fit <- aov(DEATH\_EVENT ~.,data = data2)  
anova(fit)

## Analysis of Variance Table  
##   
## Response: DEATH\_EVENT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## age 1 4.196 4.1960 31.5731 4.556e-08 \*\*\*  
## anaemia 1 0.127 0.1268 0.9543 0.32946   
## creatinine\_phosphokinase 1 0.569 0.5695 4.2852 0.03934 \*   
## diabetes 1 0.043 0.0433 0.3256 0.56870   
## ejection\_fraction 1 5.203 5.2030 39.1504 1.429e-09 \*\*\*  
## high\_blood\_pressure 1 0.275 0.2746 2.0664 0.15167   
## platelets 1 0.025 0.0254 0.1908 0.66255   
## serum\_creatinine 1 4.107 4.1070 30.9035 6.218e-08 \*\*\*  
## serum\_sodium 1 0.631 0.6310 4.7479 0.03015 \*   
## sex 1 0.201 0.2015 1.5160 0.21923   
## smoking 1 0.009 0.0091 0.0687 0.79346   
## time 1 11.781 11.7810 88.6465 < 2.2e-16 \*\*\*  
## Residuals 286 38.009 0.1329   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

final\_fit <- aov(DEATH\_EVENT ~ age+serum\_creatinine+creatinine\_phosphokinase+ejection\_fraction+serum\_sodium+time,data = data2)  
anova(final\_fit)

## Analysis of Variance Table  
##   
## Response: DEATH\_EVENT  
## Df Sum Sq Mean Sq F value Pr(>F)   
## age 1 4.196 4.1960 31.9466 3.768e-08 \*\*\*  
## serum\_creatinine 1 4.310 4.3105 32.8180 2.517e-08 \*\*\*  
## creatinine\_phosphokinase 1 0.467 0.4665 3.5518 0.06047 .   
## ejection\_fraction 1 4.960 4.9599 37.7622 2.614e-09 \*\*\*  
## serum\_sodium 1 0.593 0.5925 4.5112 0.03451 \*   
## time 1 12.299 12.2992 93.6406 < 2.2e-16 \*\*\*  
## Residuals 292 38.353 0.1313   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

fit1<-glm(`DEATH\_EVENT` ~ age+serum\_creatinine+creatinine\_phosphokinase+ejection\_fraction+serum\_sodium+time ,data = data2)  
summary(fit1)

##   
## Call:  
## glm(formula = DEATH\_EVENT ~ age + serum\_creatinine + creatinine\_phosphokinase +   
## ejection\_fraction + serum\_sodium + time, data = data2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.80648 -0.27111 -0.02274 0.26527 1.01124   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.657e+00 6.815e-01 2.431 0.01567 \*   
## age 5.471e-03 1.838e-03 2.977 0.00316 \*\*   
## serum\_creatinine 8.545e-02 2.104e-02 4.062 6.25e-05 \*\*\*  
## creatinine\_phosphokinase 3.224e-05 2.178e-05 1.480 0.13990   
## ejection\_fraction -9.477e-03 1.811e-03 -5.234 3.17e-07 \*\*\*  
## serum\_sodium -7.991e-03 4.942e-03 -1.617 0.10696   
## time -2.713e-03 2.804e-04 -9.677 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.1313448)  
##   
## Null deviance: 65.177 on 298 degrees of freedom  
## Residual deviance: 38.353 on 292 degrees of freedom  
## AIC: 250.49  
##   
## Number of Fisher Scoring iterations: 2

ANOVA test is used to find the best fit of variables for predicted variable the condition to be the best fit is for (p<0.05) and Residual Sum of squares reduced.The predicted Y values from a linear regression model may not be limited to 0 and 1. Here’s where logistic regression kicks in, giving you a probability score that reflects the likelihood of anything happening during the event.You can use logistic regression as an example of a classification technique to predict a qualitative response. More specifically, logistic regression models the likelihood that a person’s death falls into one of two categories.So, in our case we are trying to predict the Death Event which is qualitative response so we are using logarithmic regression.The p values of serum creatinine, ejection fraction, time and age are less than 0.05 so they more significant nad age comparatively less significant compared to others. The Variable creatinine phosphokinase and serum sodium level in blood are not significant because their p values are greater than 0.05. From the beta values in the summary we can say dependent variable (i.e. log(p/(1-p))) have positive linear relationship with age ,serum creatinine and negative relation with Ejection fraction and time. Accordingly, we can see if the value of ejection fraction is low then there is high probability of death occurrence (i.e., Death event=1).

# Hypothesis Testing

We answer the following questions using Hypothesis Testing 1) If a heart failure patient has diabetes, Weather the diabetes increases the chance of death in patient? Null Hypothesis: No difference in proportions of deaths in patients having diabetes and not having diabetes. Alternate Hypothesis: There is Significant Difference in Proportions of deaths in heart failure patients with diabetes and without.

q1<-select(dataf,diabetes,DEATH\_EVENT)  
# H0 = no difference between two proportion (deaths among both no BP and High BP)  
# H1 = There if significant Difference between two proportions  
q1%>%  
 count(diabetes,DEATH\_EVENT)

## # A tibble: 4 x 3  
## diabetes DEATH\_EVENT n  
## <dbl> <dbl> <int>  
## 1 0 0 118  
## 2 0 1 56  
## 3 1 0 85  
## 4 1 1 40

prop.test(x=c(56,40),n=c(174,125))

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(56, 40) out of c(174, 125)  
## X-squared = 6.9701e-31, df = 1, p-value = 1  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.1072650 0.1109432  
## sample estimates:  
## prop 1 prop 2   
## 0.3218391 0.3200000

As P value is 1 which is greater than 0.05 we cannot reject null hypothesis. the both proportions does not differ so there is no connection between diabetes and death event.

1. Does Smoking have any correlation with Death event in heart failure patients?

Null Hypothesis: No difference in proportions of deaths patients smoking and notsmoking.

Alternate Hypothesis: There is Significant Difference in Proportions of deaths in heart failure patients with diabetes and without.

q2<-select(dataf,smoking,DEATH\_EVENT)  
# H0 = no difference between two proportion (deaths among both no BP and High BP)  
# H1 = There if significant Difference between two proportions  
q2%>%  
 count(smoking,DEATH\_EVENT)

## # A tibble: 4 x 3  
## smoking DEATH\_EVENT n  
## <dbl> <dbl> <int>  
## 1 0 0 137  
## 2 0 1 66  
## 3 1 0 66  
## 4 1 1 30

prop.test(x=c(66,30),n=c(203,96))

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(66, 30) out of c(203, 96)  
## X-squared = 0.0073315, df = 1, p-value = 0.9318  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.1079604 0.1332067  
## sample estimates:  
## prop 1 prop 2   
## 0.3251232 0.3125000

As P value is 0.93 which is greater than 0.05, we cannot reject null hypothesis. Patient who smokes and who doesn’t smoke has no influence on chances of death

1. Does High Blood pressure plays an important role in Heart Failure patient’s death? Null Hypothesis: High blood pressure has no influence in occurrence of Death in Heart failure Patients.

Alternate Hypothesis: High blood pressure plays a significant role in occurrence of Death in Heart failure Patients. .

q3<-select(dataf,age,high\_blood\_pressure,DEATH\_EVENT)  
# H0 = no difference between two proportion (deaths among both no BP and High BP)  
# H1 = There if significant Difference between two proportions  
q3%>%  
 count(high\_blood\_pressure,DEATH\_EVENT)

## # A tibble: 4 x 3  
## high\_blood\_pressure DEATH\_EVENT n  
## <dbl> <dbl> <int>  
## 1 0 0 137  
## 2 0 1 57  
## 3 1 0 66  
## 4 1 1 39

prop.test(x=c(57,39),n=c(194,105))

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: c(57, 39) out of c(194, 105)  
## X-squared = 1.5435, df = 1, p-value = 0.2141  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.19742595 0.04219767  
## sample estimates:  
## prop 1 prop 2   
## 0.2938144 0.3714286

As P value is 0.21 which is greater than 0.05, we cannot reject null hypothesis. Both proportions does not differ so there is no connection between high blood pressure and death event.

## Conclusion

In this project my goal was to predict the survival pattern of patients suffering from heart failure condition. Performing the PCA has reduced the dimensionality was reduced 5 principal components has explained the 90 percent of variance. The Death Event in the data set was a categorical value so we have used log regression method which helped in predicting the death in patients. The variables which are significant in predicting the pattern using regression are serum creatinine and ejection fraction and time. Hypothesis testing has helped in predicting weather smoking, high blood pressure and diabetes in patients will decrease the survival chances but from results we can see they had no influence on death event. Finally, we can conclude that Ejection fraction and Serum Creatinine are the two variable which are most relevant in predicting the survival of the patient. There are many limitations in our research and some of them are, As the data set was small with less observation which makes difficult arrive at correct conclusion in predicting survival of patients. we need to use more advanced techniques to arrive at better conclusion. We can not arrive to conclusion based on data collected from a particular region, so we need to collect data from different parts of the world to conclude that predicted pattern is universal and applicable to all. We need more variables from patient like height, weight, body mass index and cholesterol which also play major role in heart diseases.