

# **HEART FAILURE PROJECTION**

**CSE3020 – DATA VISUALISATION**

**PROJECT BASED COMPONENT REPORT**

*By*

TEAM: DATA VISION

SANJITHA RAJESH 20BCE2541

THIRSHA SREE H 20BCE2518

SRISHTI ACHARYYA 20BCE2561

T HARSHITHA DEEPTHI 20BCE2019

**School of Computer Science and Engineering**



**VIT<sup>®</sup>**  
**Vellore Institute of Technology**  
(Deemed to be University under section 3 of UGC Act, 1956)

**May 2021**

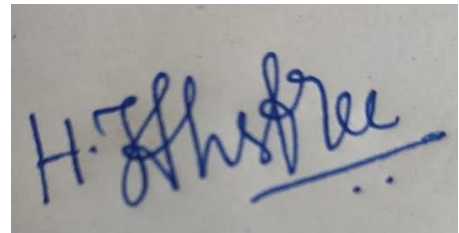
## **DECLARATION**

I hereby declare that the report entitle “**HEART FAILURE PROJECTION**” submitted by me, for the CSE3020 DATA VISUALISATION (EPJ) to VIT is a record of bonafide work carried out by me under the supervision of Dr. S.VENGADESWARAN .

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for any other courses in this institute or any other institute or university.

Place : Vellore

Date : 10/12/2021

A handwritten signature in blue ink, appearing to read 'H. J. J. J. J.', with a horizontal line underneath.

**Signature of the Candidate**

## CONTENTS

Title	P.No
<b>1. ABSTRACT</b>	
<b>2. INTRODUCTION TO THE PROJECT</b>	
• OBJECTIVE	
• PROBLEM STATEMENT	
• FUNCTIONAL REQUIREMENTS	
<b>3. DATA ABSTRACTION</b>	
<b>4. TASK ABSTRACTION</b>	
<b>5. DESIGN OF THE PROPOSED SYSTEM</b>	
<b>6. DASHBOARD IMPLEMENTATION</b>	
<b>7. CONCLUSION</b>	
<b>8. APPENDIX</b>	
• SCREEN SHOTS	
• SAMPLE CODING	

## **ABSTRACT:**

Our project explores and analyses records of patients who have had heart failure.

Heart failure means that the heart is unable to pump blood around the body properly. It usually occurs because the heart has become too weak or stiff. It's sometimes called congestive heart failure, although this name is not widely used nowadays. Heart failure does not mean your heart has stopped working.

The four types of heart failure are:

- Left-sided heart failure (the most common one)
- Right-sided heart failure (usually triggered by left side heart failure)
- Diastolic (when heart muscle becomes stiffer than normal)
- Systolic (when heart muscle loses its ability to contract)

There are many factors involved when a person is in risk of, or has gotten heart failure. However, there are some practices we can implement in order to prevent heart failure.

Like:

- Control blood pressure.
- Stay at a healthy weight.
- Eat a healthy diet.
- Don't smoke.
- Make sure that you get enough sleep.

## **INTRODUCTION TO THE PROJECT:**

### **Objective:**

Through this project, we aim to find out and analyze what causes heart failure, what substances in the body could trigger possible heart problems and predict which situations occurring in a patient's body would not be healthy for them cardiology-wise.

### **Problem Statement:**

Development of visual idioms in order to predict and analyze heart failure.

### **Functional requirements:**

Tools used: Kaggle, RStudio, Flexdashboard

1. Kaggle is the website where we obtained our heart failure dataset from.
2. RStudio: We have decided to code our project in R language and are using RStudio application for doing so.
3. Flexdashboard: Flexdashboard is a form of dashboard visualization. By using Flexdashboard, we will aim to visualize various charts and graphs that analyze key performance indicators, critical data points and other factors from which we can make valid inferences and analysis about our topic.

## DATA ABSTRACTION: (Heart Failure Prediction and analyse)

Data abstraction is the reduction of a particular body of data to a simplified representation of the whole. Abstraction, in general, is the process of taking away or removing characteristics from something in order to reduce it to a set of essential characteristics.

### 1. Dataset details:

**URL** – Where did you download it from – exact URL Link:

<https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>

### 2. Number of attributes and rows:

No. of attributes	13
No. of rows	300

### 3. Attributes types:

A piece of information which determines the properties of a field or tag in a database or a string of characters in a display.

- **Age** – The length of the time that a person had lived.
- **Anaemia** – It is a condition in which there is a deficiency of red cells or of haemoglobin in the blood, resulting in pallor and weariness.
- **Serum\_sodium** - A sodium blood test is a routine test that allows your doctor to see how much sodium is in your blood. It's also called a serum sodium test. A normal blood sodium level is between 135 and 145 milliequivalents per liter (mEq/L).
- **Diabetes** - a disease that occurs when your blood glucose is too high, also called blood sugar. Ejection\_fraction - a measurement of the percentage of blood leaving your heart each time it squeezes (contracts).
- **High\_blood\_pressure** – HBP is when your blood pressure, the force of your blood pushing against the walls of your blood vessels, is consistently too high.
- **Platelets** – these are small, colorless cell fragments in our blood that form clots and stop or prevent bleeding.
- **Serum\_creatinine** - Elevated creatinine level signifies impaired kidney function or kidney disease.

- **Sex:** Either of the two sexes (male and female), especially when considered with reference to social and cultural differences rather than biological ones.
- **Smoking:** It is the act of inhaling and exhaling the fumes of burning plant material. Death Event: In the final stages of heart failure, people feel breathless both during activity and at rest like persistent coughing or wheezing which causes the patient to death.

#### 4. Level of measurements:

- **Age:** It is a Discrete type of attribute because it is commonly expressed as an integer in units of years with no decimal to indicate days and presumably, hours, minutes, and seconds. It is a Ratio level of measurement.
- **Anaemia:** It is a Categorical type of attribute since it has either '0' or '1' as its data. It is a Nominal level of measurement.
- **Diabetes:** It is a Categorical type of attribute since it has '0' or '1' as its data. It is a Nominal level of measurement.
- **Ejection\_fraction:** It is a Continuous Data represents measurements and therefore their values can't be counted but they can be measured.
- **Serum\_sodium:** It is a Continuous Data represents measurements and therefore their values can't be counted but they can be measured.
- **High\_blood\_pressure:** It is a Categorical type of attribute since it has either '0' or '1' as its data.
- **Platelets:** It is a Continuous Data represents measurements and therefore their values can't be counted but they can be measured.
- **Serum\_creatinine:** It is a Continuous Data represents measurements and therefore their values can't be counted but they can be measured.
- **Sex:** It says that the patient is male or female. Hence, it is a Categorical type of attribute and it has '0' or '1' as its data. It is a Nominal level of measurement.
- **Smoking:** If the patient had a habit of smoking or not. Hence, it is a Categorical type of attribute and it has '0' or '1' as its data. It is a Nominal level of measurement. Death event: It is a Categorical type of attribute and it has '0' or '1' as its data. It is a Nominal level of measurement.

#### Table:

Many datasets come in the form of tables that are made up of rows and columns, a familiar form to anybody who has used a spreadsheet. Here we are using heart failure prediction table. This table have attributes, item and cell. The terms used in this book are that each row represents an item of data. Each column is an attribute of the dataset. Each cell in the table is fully specified by the combination of a row and a column.

age	anaemia	creatinine	phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	C
75	0		582	0	20	1	265000	1.9	130	1	0	4	
55	0		7861	0	38	0	263358	1.1	136	1	0	6	
65	0		146	0	20	0	162000	1.3	129	1	1	7	
50	1		111	0	20	0	210000	1.9	137	1	0	7	
65	1		160	1	20	0	327000	2.7	116	0	0	8	
90	1		47	0	40	1	204000	2.1	132	1	1	8	
75	1		246	0	15	0	127000	1.2	137	1	0	10	
60	1		315	1	60	0	454000	1.1	131	1	1	10	
65	0		157	0	65	0	263358	1.5	138	0	0	10	
80	1		123	0	35	1	388000	9.4	133	1	1	10	
75	1		81	0	38	1	368000	4	131	1	1	10	
62	0		231	0	25	1	253000	0.9	140	1	1	10	
45	1		981	0	30	0	136000	1.1	137	1	0	11	
50	1		168	0	38	1	276000	1.1	137	1	0	11	
49	1		80	0	30	1	427000	1	138	0	0	12	
82	1		379	0	50	0	47000	1.3	136	1	0	13	
87	1		149	0	38	0	262000	0.9	140	1	0	14	
45	0		582	0	14	0	166000	0.8	127	1	0	14	
70	1		125	0	25	1	237000	1	140	0	0	15	
48	1		582	1	55	0	87000	1.9	121	0	0	15	
65	1		52	0	25	1	276000	1.3	137	0	0	16	
65	1		178	1	30	1	387000	1.6	136	0	0	20	

### Dataset availability:

In visualization we will be having static file of dataset which is already predetermined for attributes and items.

### Scalar field:

A scalar field is univariate, with a single value attribute at each point in space.

**Example:** age is a scalar attribute for heart failure prediction and analyse.



## **TASK ABSTRACTION:**

1. How does the creatinine serum level in our body affect the occurrence of diabetes and ultimately heart failure?

High level: analyze the creatinine serum levels and produce this information on a scatterplot by investigating whether low or high creatinine levels will cause diabetes, and where on the graph there is more density.

Low level: Querying the targets by comparing the possibility of diabetes (1.00 represents having diabetes, while 0.00 represents otherwise on our plot) when the creatinine levels increase or decrease.

2. How does the creatinine phosphokinase level affect anemia and ultimately heart failure?

High level: analyze the creatinine phosphokinase levels and present it on a violin plot to find out how much creatinine phosphokinase will cause anemia and eventually heart failure.

Low level: compare and divide the effect of CP on anemia by the death event of the patients.

3. How does the age of a person affect whether that person dies by heart failure?

High level: analyze the ages using a bar plot and check which age group suffers more from heart failure and whether they die or not.

Low level: identify which age group has a higher probability of dying from heart failure.

4. What is the relationship between the levels of sodium serum and the age of the patient, and how does it relate to heart failure?

High level: analyze how the levels of sodium serum affect the patient.

Low level: identify the appropriate age group.

5. Does gender have an effect on the death event due to heart failure?

Low level: identify which gender is affected more from heart failures.

6. Does the ejection fraction have an effect on the death event due to heart failure?

High level: analyze whether a lower or higher ejection fraction leads to a higher percentage of death event and produce the results using a density distribution.

Targets: Trends- Peaks, to represent whether a greater or less ejection fraction leads to more deaths.

7. Do the creatinine phosphokinase levels have an effect on the event of death due to heart failure?

High level: Analyze whether a lower or higher level of CP leads to a higher percentage of death event and produce the results using a density distribution that will show the mean values.

Targets: Trends- Peaks, to represent whether a greater or less CP level leads to more deaths.

8. Does smoking have an effect on the ejection fraction of a patient's heart?

High level: Analyze whether smokers or non-smokers have a greater effect on the ejection fraction which ultimately causes heart failure by use of a scatterplot.

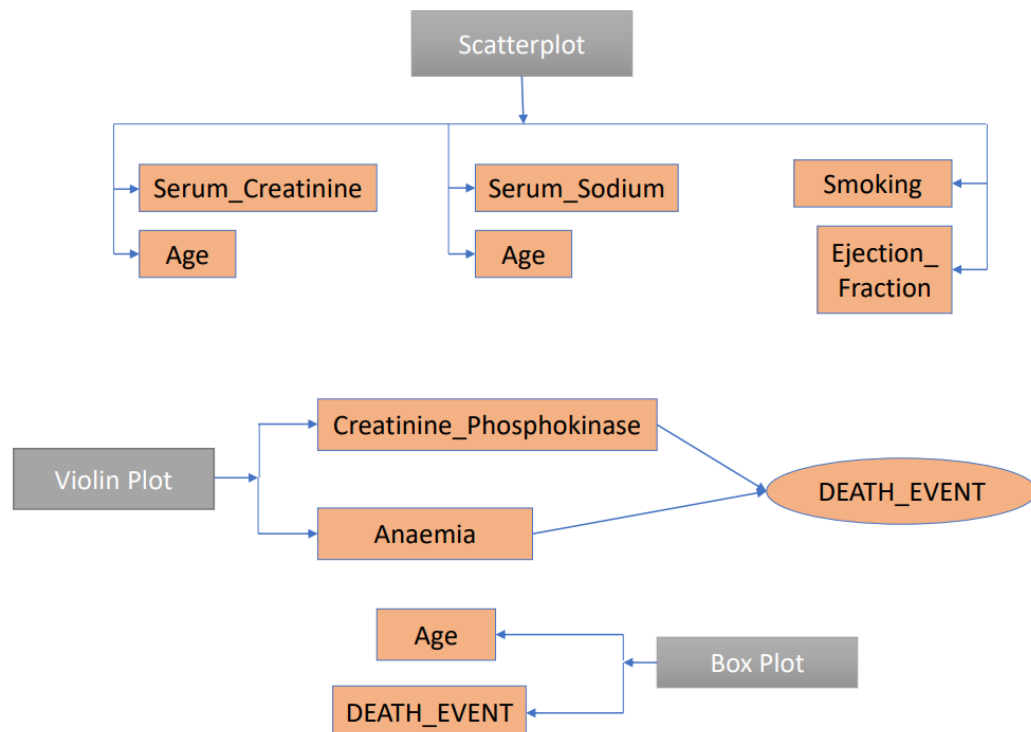
Low level: compare the ejection fraction levels based on whether those patients have habits of smoking or not.

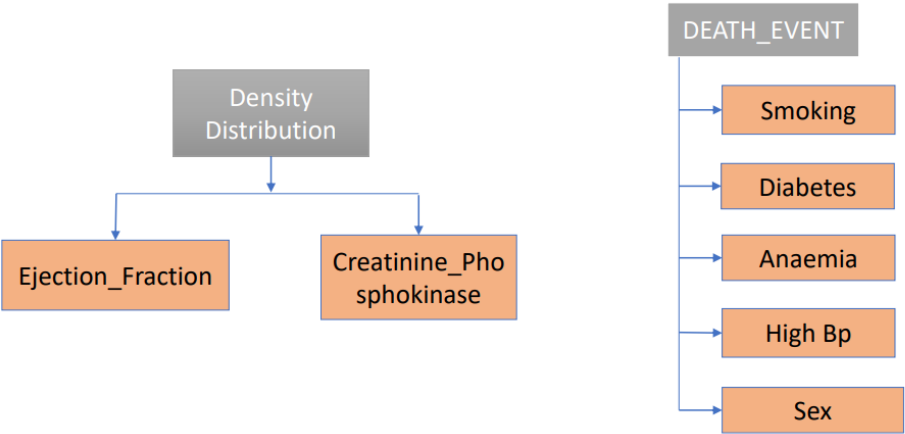
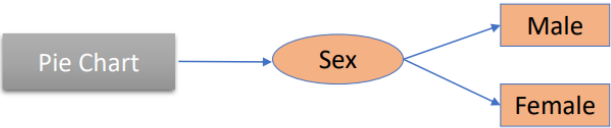
9. What effects do smoking, diabetes, anemia, high blood pressure, and sex have on the event of death due to a heart failure?

Low level: compare the results of the death event(or not) using barplots.

## DESIGN OF THE PROPOSED SYSTEM:

- Our dataset is about the Heart Failure Clinical Records. We made our Visualization design format into 5 tables. Which makes the viewer easy to understand about the dataset.
- We have used different types of graphs and charts like Scatterplots, Bar plots, Line graphs, Pie charts to show the attributes in the dataset.
- We used different colours to differentiate the best and most accurate representation of quantitative variables.
- In the plots it had been clearly represented to depict the number of people affected by heart failure.
- In the dashboard we used different libraries like
  - ggplot2
  - plotly
  - gridExtra
  - grid
  - Lattice

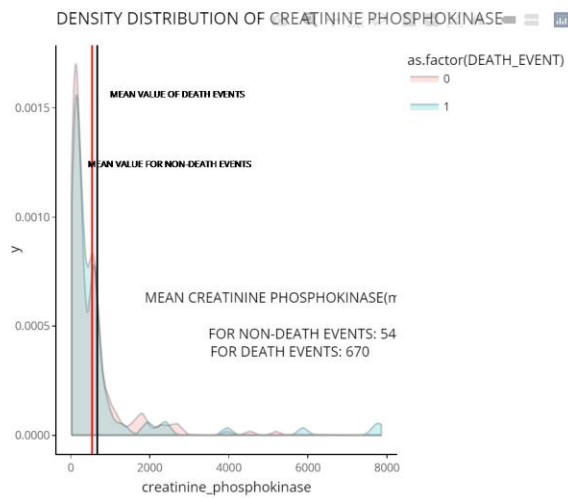




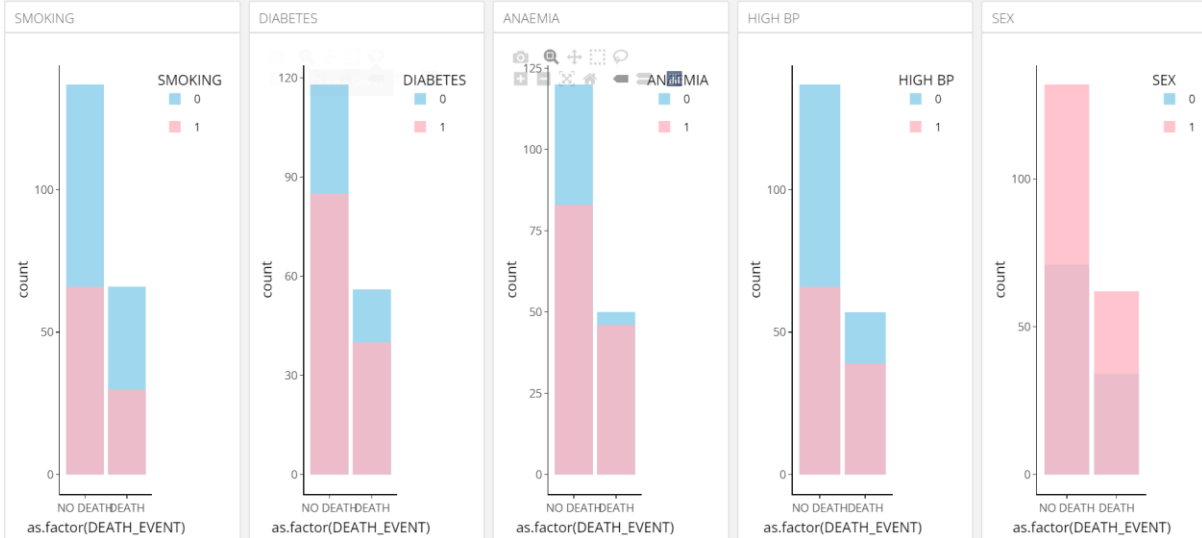
# DASHBOARD IMPLEMENTATION:



DENSITY DISTRIBUTION OF CREATININE PHOSPHOKINASE



SCATTERPLOT: SMOKING VS EJECTION FRACTION LEVELS



## summary

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			

## CONCLUSION:

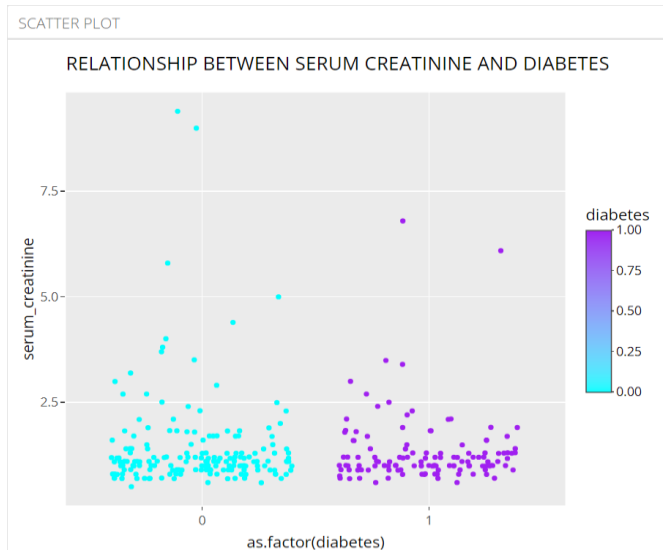
After analyzing the graphs made from the given dataset, we can conclude that:

- a higher proportion of patients who died had diabetes. Thus, low levels of serum creatinine leads to diabetes which in turn leads to heart attacks which can be fatal. (from serum creatinine vs diabetes graph)
- the lower amount of creatinine phosphokinase more is the chance of a person suffering from anemia, which means tht the heart pumps more blood to make up for the lack of oxygen in the blood. This can lead to an enlarged heart or heart failure. (from Creatinine phosphokinase for anaemia state separated by death event)
- death due to heart failure increases with increase in age. (from Age vs Death graph)
- serum sodium levels does not decrease with age, It stays in the range of (130-145) mg/L (Age vs Serum sodium graph)
- The ratio of males affected by heart failure is greater than those of females (male female piechart).
- There is a higher chance of death due to heart failure for lower value of ejection fraction. (Ejection Fraction Density Distribution).
- The chances of death due to heart failure increases with the increase in the levels of creatinine phosphokinase. (Density distribution of creatinine phosphokinase).
- Ejection fraction seems to have a higher range for people who don't smoke, ejection fraction and death event have an inverse relationship hence we can say that patients with lower ejection fraction have higher chances of death because of heart attack. (Smoking vs ejection fraction)
- There is a higher chance of death due to heart failure for patients who Smoke, have anemia, diabetes and high bp (smoking, diabetes, anemia and sex graphs).

## APPENDIX:

The following are the appendix of codes and graphs in our dashboard

### **1. SCATTER PLOT: RELATIONSHIP BETWEEN SERUM CREATININE AND DIABETES**



```
heart_failure.data= read.csv("C:\\Users\\ADMIN\\Desktop\\DV  
DASHBOARD\\heart_failure_clinical_records_dataset.csv",stringsAsFactors =  
FALSE)
```

```
splot1<-  
ggplot(heart_failure.data,aes(x=as.factor(diabetes),y=serum_creatinine,colour=diabetes)  
) +geom_point(position = "jitter",size=0.8)+ggtitle("RELATIONSHIP BETWEEN  
SERUM CREATININE AND DIABETES")
```

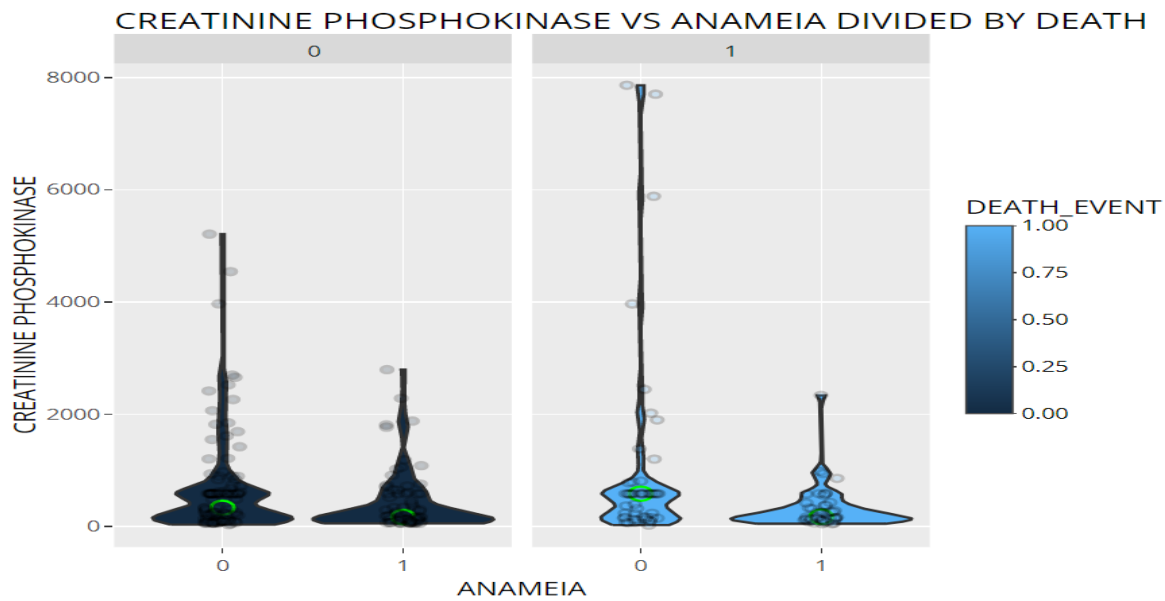
```
splot1=splot1+scale_color_gradient(low="cyan", high="purple")
```

```
ggplotly(splot1)
```



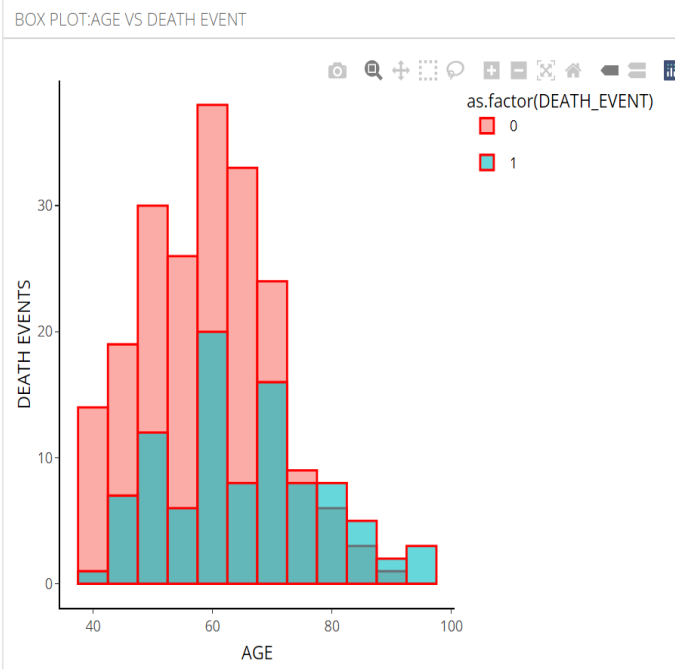
## **2.VIOLIN PLOT: CREATININE PHOSPHOKINASE VS ANAMEIA DIVIDED BY DEATH**

VIOLIN PLOT

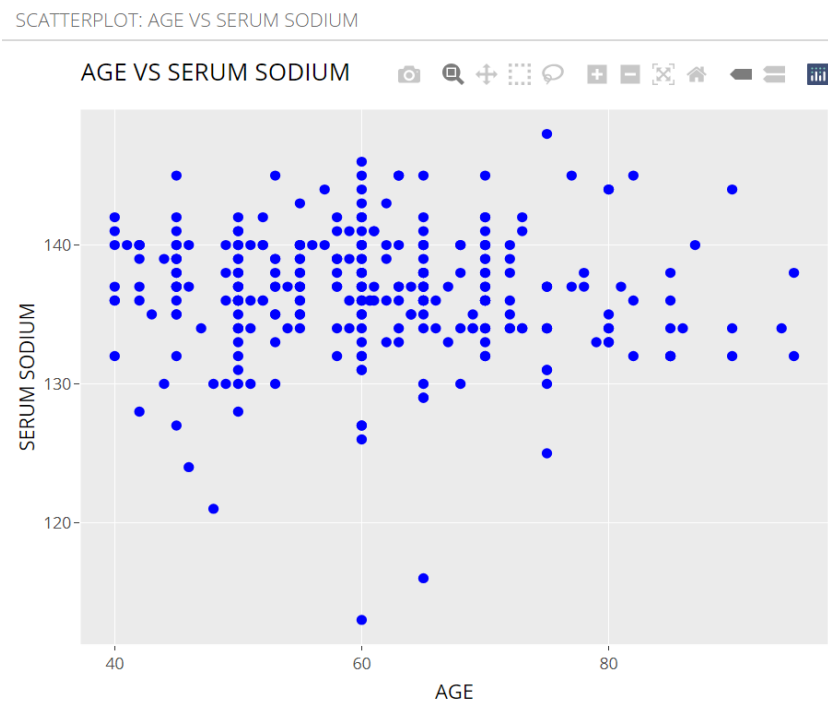


## **3.BOX PLOT:AGE VS DEATH EVENT**

```
p<-ggplot(heart_failure.data, aes(x=age,fill=as.factor(DEATH_EVENT))) +  
  geom_histogram(binwidth = 5, position = "identity",alpha = 0.6,color = "red") +  
  xlab("AGE") + ylab("DEATH EVENTS") + theme_classic() +  
  labs(caption = "AGE DISTRIBUTION OF HEART FAILURE WITH DEATH  
EVENT")  
ggplotly(p)
```



## **4. SCATTERPLOT: AGE VS SERUM SODIUM**



## 5. PIE CHART: RATIO OF MALES AND FEMALES

```
heart_failure.data= read.csv("C:\\Users\\ADMIN\\Desktop\\DV  
DASHBOARD\\heart_failure_clinical_records_dataset.csv",stringsAsFactors =  
FALSE)
```

```
male=sum(heart_failure.data$sex)
```

```
female=300-male
```

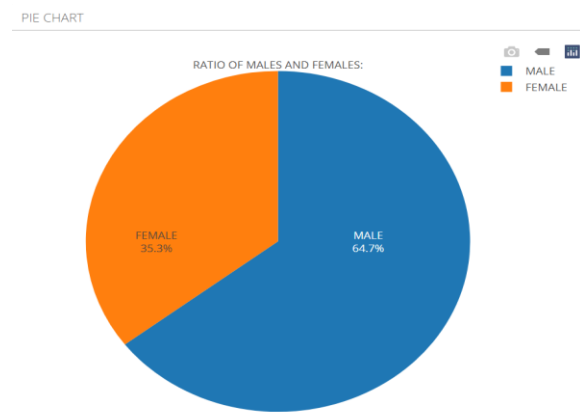
```
x<-c(male,female)
```

```
labels = c('MALE','FEMALE')
```

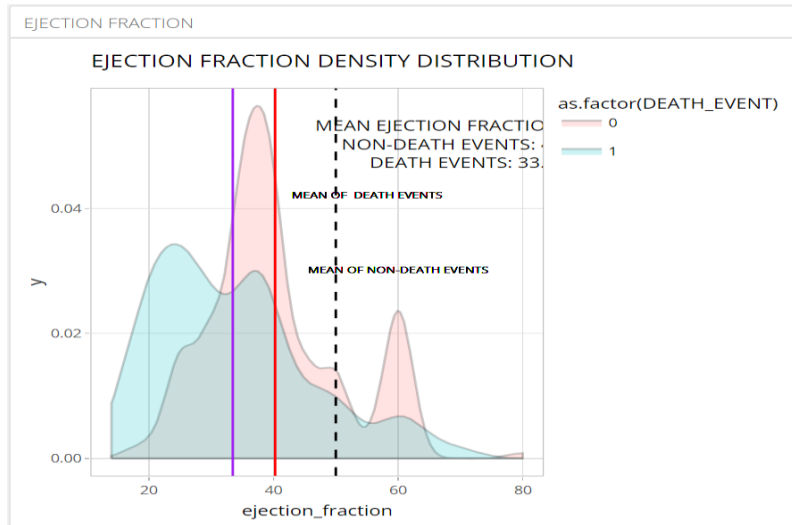
```
values = c(male, female)
```

```
fig <- plot_ly(title= " RATIO OF MALES AND FEMALES: ", type='pie', labels=labels,  
values=values, textinfo='label+percent',  
insidetextorientation='radial')
```

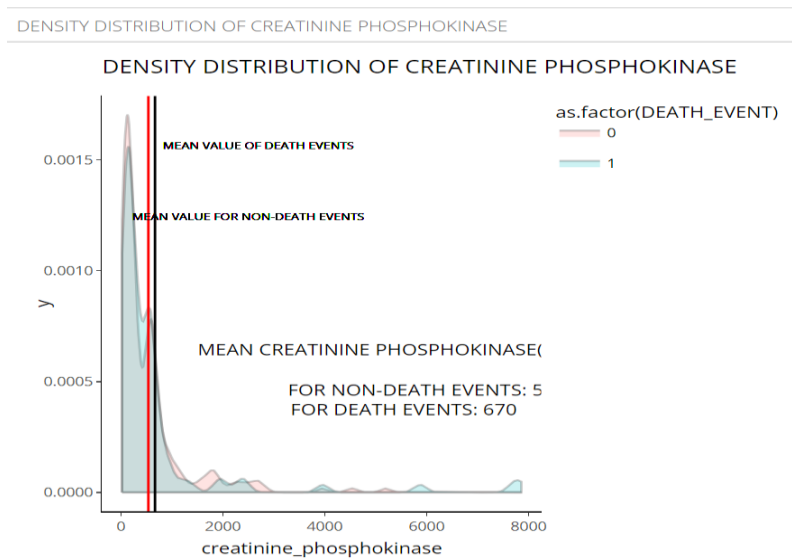
fig



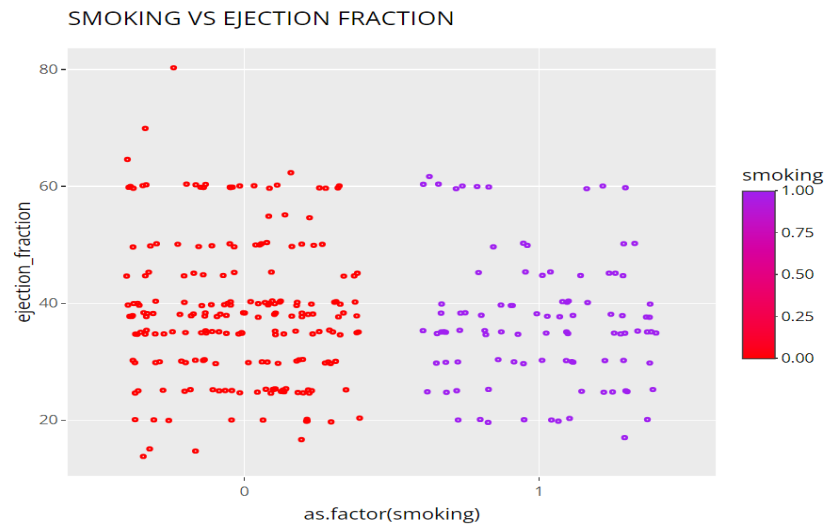
## 6.EJECTION FRACTION DENSITY DISTRIBUTION



## 7. DENSITY DISTRIBUTION OF CREATININE PHOSPHOKINASE



## 8.SCATTERPLOT:SMOKING VS EJECTION FRACTION LEVELS



## **9.BAR PLOT FOR SMOKING,DIABETES,ANAMEIA,HIP BP AND SEX**

### BAR PLOT: SMOKING

```
library(ggplot2)
```

```
library(plotly)
```

```
library(gridExtra)
```

```
library(grid)
```

```
library(lattice)
```

```
smoking = ggplot(heart_failure.data, aes(x = as.factor(DEATH_EVENT), fill =
as.factor(smoking))) +
geom_bar(position = "identity",alpha=0.8) +
theme_classic()+ scale_x_discrete(labels = c("NO DEATH","DEATH")) +
labs(subtitle = "SMOKING") +
scale_fill_manual(values = c("skyblue","lightpink"), name = "SMOKING",
labels = c("NEGATIVE","POSITIVE"))
ggplotly(smoking)
```

