**PYTHON FINAL TERM PROJECT**

**Title: Heart Failure Prediction Data: Exploring Insights and Design**

**BSAN 710: PYTHON FOR BUSINESS**

**Thursday 7.00 PM**

**Dr. Justin Keeler**

**May 4, 2023**

**GROUP 2:**

**NARAYANAM NAGA VAMSI SANDEEP**

**DECHARAJU PHANINDRA**

**SAI KOUSHIK REDDY**

**AARTHI AWASTHI**

**RAGHU VAMSI**

**INTRODUCTION:**

Heart failure is a major global source of condition and mortality, and it is a severe public health issue. There has been a rise in interest in bringing machine learning methods to clinical data analysis to forecast heart failure and improve patient outcomes in recent years.

The open-access heart failure dataset includes age, gender, serum creatinine, ejection fraction, and other clinical information for people who have been diagnosed with heart failure. Machine learning models can be built and tested using this dataset, which has been widely used in the research community, with the goal of forecasting heart failure and detecting probable risk factors.

By analyzing this dataset, scientists and medical professionals can discover more about the root cause and mechanisms of heart failure. They can also create better more customized techniques for identifying and treating people with heart failure.

**Objective:**

The main objective of this project is to predict the occurrence of death due to heart failure in patients by including clinical parameters and patient health status as well as visualizing the data.

**DATA SOURCE DESCRIPTION:**

[**https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data**](https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data)

Firstly, we have downloaded the file of Heart Failure and changed the path of Jupyter notebook.

os.chdir("D:\\WICHITA\\Semister-1\\Python for Business Analytics\\Project")copied the path to the jupyter notebook.

The dataset consists of information on patients with cardiovascular disease. There are 299 patients in total, and each row represents one patient.

**Age:** the age of the patient in years (numerical)

**Anaemia:** whether the patient has anemia (0 = no, 1 = yes) (categorical)

**Creatinine phosphokinase:** level of the CPK enzyme in the blood (numerical)

**Diabetes:** whether the patient has diabetes (0 = no, 1 = yes) (categorical)

**Ejection fraction:** percentage of blood leaving the heart at each contraction (numerical)

**High blood pressure:** whether the patient has high blood pressure (0 = no, 1 = yes) (categorical)

**Platelets:** platelets in the blood (numerical)

**Serum creatinine:** level of creatinine in the blood (numerical)

**Serum sodium:** level of sodium in the blood (numerical)

**Gender:** gender of the patient (0 = female, 1 = male) (categorical)

**Smoking:** whether the patient smokes (0 = no, 1 = yes) (categorical)

**Time:** follow-up period in days (numerical)

**Death event:** whether the patient died during the follow-up period (0 = no, 1 = yes) (categorical)

**Geography:** the country where the patient is from (categorical)

**Load the dataset into Pandas Data Frame.**

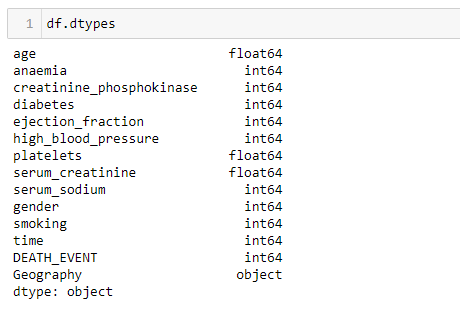


**EXPLORATORY DATA ANALYSIS:**

**Explore the dataset using Pandas to answer the following questions:**

* **Data type identification**

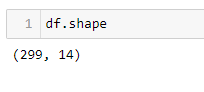
df.dtypes



Using this function, we can get the datatypes of all the columns present in the dataset.

* **To calculate Number of rows and columns**

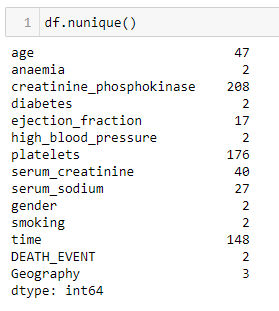
df.shape



This function is used to tell the number of rows and columns in the dataset.

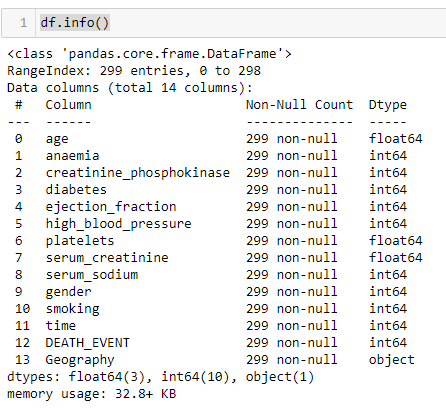
* **To calculate number of unique values, present in each column**

df.nunique()



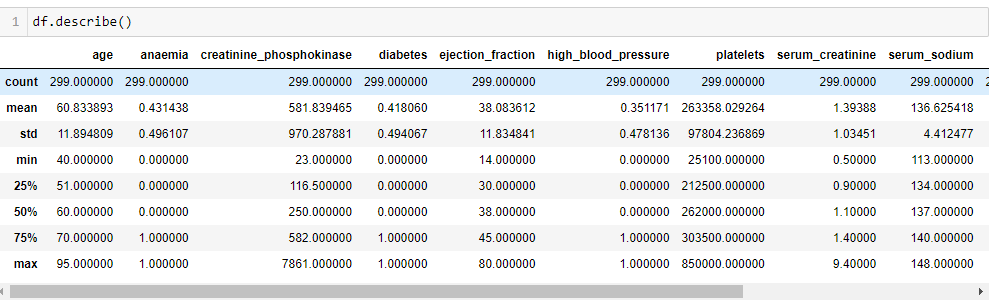
This function is used to calculate number of unique values present in each column.

df.info()



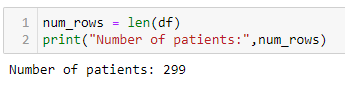
This function gives information about the dataset which include Datatype, Number of rows of each column and memory utilization.

df.describe()



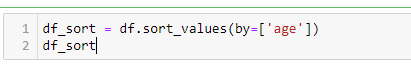
By using this function, we get a table describing mean, standard deviation, count. etc.

* **Calculating Number of Patients in the Dataset**



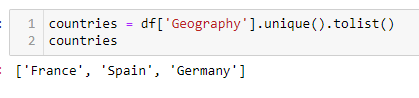
with mean and age by using loc function we can get the average age of the patients = 60.83 years

* **Sorting based on Age**

****

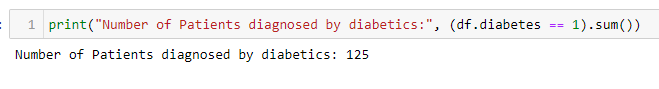
Here, we are using sort function to sort the values of the dataset based on the age

* **Number of countries in the dataset?**



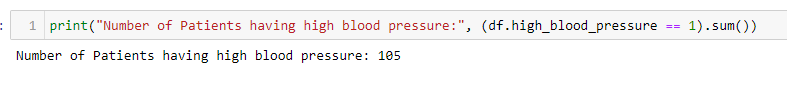
This function adds a unique value from the geography column to a list.

* **Number of patients diagnosed as diabetics?**

By using this function, we are getting the total number of patients diagnosed by diabetics if diabetes ==1 and counting the values

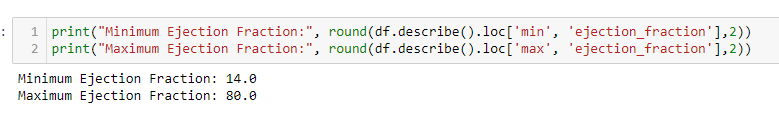
Number of Patients diagnosed by diabetics: 125.

* **How many patients in the dataset have high blood pressure?**

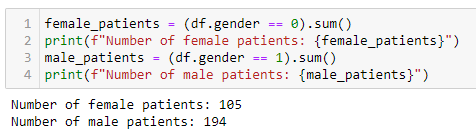
Using this function, we are getting the total number of patients having high blood pressure, blood pressure==1.

Number of Patients having high blood pressure: 105

* **What is the minimum and maximum ejection fraction for patients in the dataset?**

Using the location function of min and ejection\_fraction from the describe table and rounding it to 2 we get min and max ejection fraction values.

* **How many male and female patients are in the dataset?**

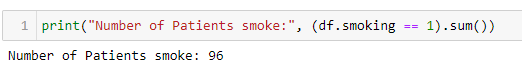


Here, we calculated Number of Male and Female patients and counted Number of Male and Female patients moreover we take female patients as gender == 0 and male patients as == 1.

Number of female patients: 105

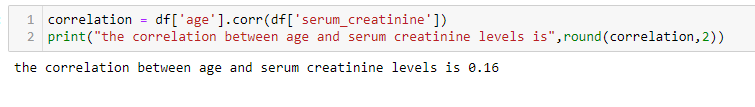
Number of male patients: 194

* **How many patients in the dataset are smokers?**



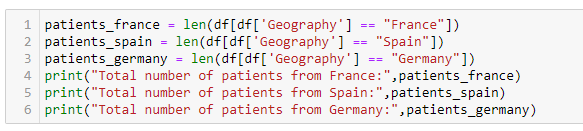
By using this function, we are getting how many patients are smokers if smoking ==1 and calculating using sum function.

* **What is the correlation between age and serum creatinine levels?**

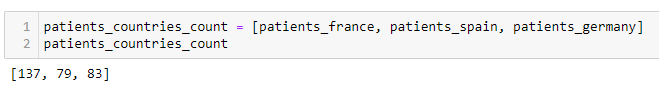
Here we calculate the correlation between age and serum creatinine levels using correlation function.

the correlation between age and serum creatinine levels is 0.16

* **The number of patients is from different countries.**

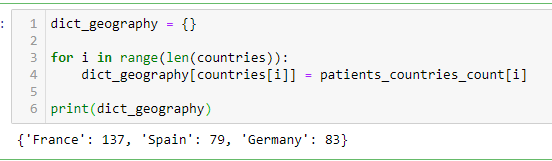
To find which patient belongs to which county we use geography column from the dataset and count their values using len function. Total number of patients from France: 137, Total number of patients from Spain: 79, Total number of patients from Germany: 83.

* **Considering the above values into the list**



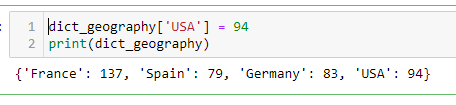
We are taking the results based on the number of patients from different countries and storing them into a list.

* **Create a Dictionary based on Number of Patients based on Geography**

Here, we are creating a dictionary named dict\_geography and key as Country name and value as Number of patients.

{'France': 137, 'Spain': 79, 'Germany': 83}

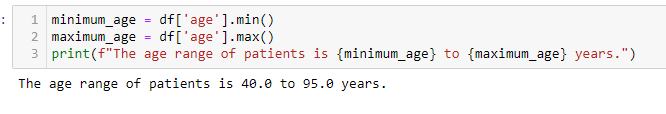
* **Add Number of patients in usa is 94 to the above dictionary**



here, we are adding number of patients in usa to the dictionary.

{'France': 137, 'Spain': 79, 'Germany': 83, 'USA': 94}

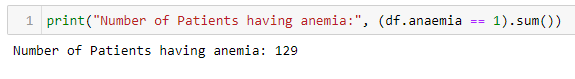
* **What is the age range of patients in this dataset?**



The range of patients is calculated by the minimum and maximum function to find out the minimum and maximum age.

The age range of patients is 40.0 to 95.0 years.

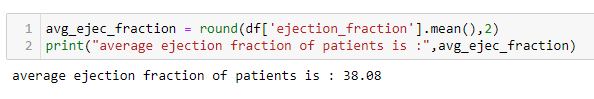
* **How many patients in this dataset have anemia?**



By using this function, we are getting how many patients having anemia if anemia==1 and counting those values

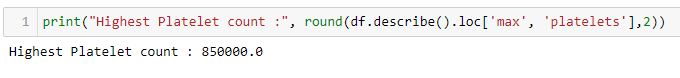
Number of Patients having anemia: 129

* **What is the average ejection fraction of patients in this dataset?**

with mean by using round function, we can get the average ejection fraction=38.08

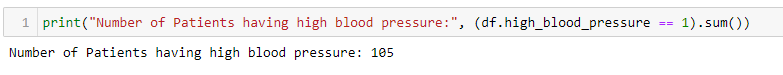
average ejection fraction of patients is : 38.08

* **What is the highest platelet count in this dataset?**

Using the location function of max from the described table and rounding it to 2 we get highest platelet count.

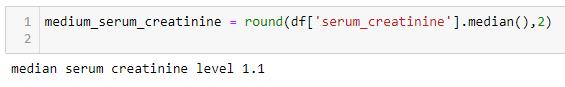
average ejection fraction of patients is : 38.08

* **How many patients in this dataset have high blood pressure?**

By using this function, we are getting how many patients having high blood pressure if blood pressure ==1.

Number of Patients having high blood pressure: 105

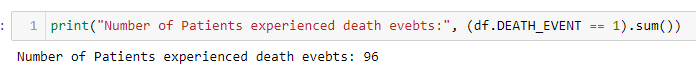
* **What is the median serum creatinine level in this dataset?**



Here we calculate the medium serum creatinine levels using median function and rounding its value up to 2.

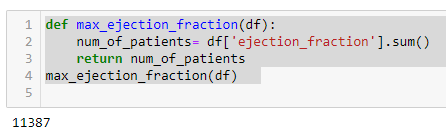
median serum creatinine level 1.1

* **How many patients in the dataset experienced death events?**

By using this function, we are how many patients experienced death events if death event ==1.

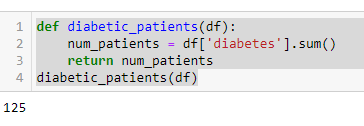
Number of Patients experienced death events: 96

* **create a function that calculates the maximum ejection fraction in the dataset?**

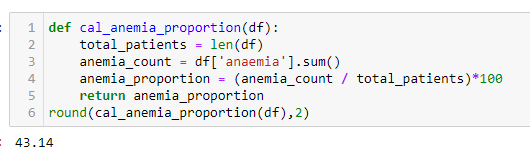


The Maximum ejection fraction is 11387.

* **How about a function that counts the number of patients with diabetes?**

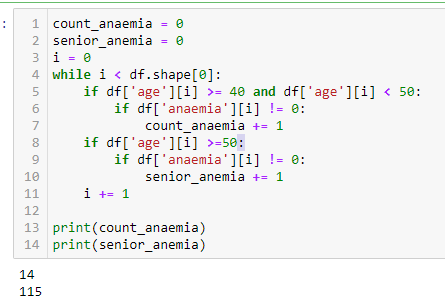
Number of Patients having diabetics is 125.

* **Create a function that determines proportion of patients having Anemia?**



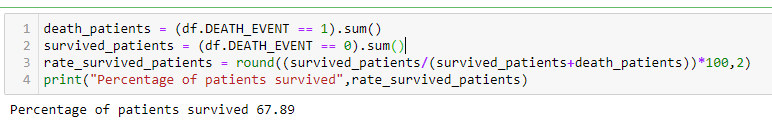
Here, we calculated the number of patients having anemia by the total number of patients and multiplied by 100 to get the proportion and we got 43.14 % of patients are with Anemia.

* **count the number of patients having anemia between age 40 to 50 and age greater than equal to 50?**



Here, we are taking the initial variable values of count\_anaemia, senior\_anemia as 0. And using while loop and if else we are initialing the loop with i and considering the number of patients with age between 40 and 50 into the count\_anamia and calculated as 14 and age greater than 50 into senior\_anemia and 115 patients are present with age greaterthan 50.

* **Calculate the survival rate of the patients?**

Here, we are calculating the number of patients survived and number of patients died and storing the values in the death\_patients and survived\_patients and calculating the survival rate.

If the Death\_EVENT column value is 1 then we are counting the values into death\_patients. And if the Death\_EVENT column value is 0 it is taken into survived\_patients.

* **Drop the value from the Dataset if the age = 60.667?**



We are storing the dataset into df1 dataframe and dropping the value using drop function.

* **Show the 1st 5 rows and columns from the dataset?**

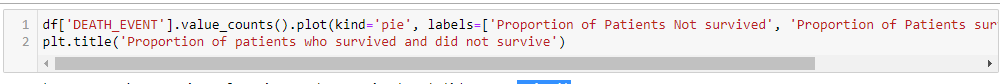


Using iloc function we are taking the first 5 rows and columns.

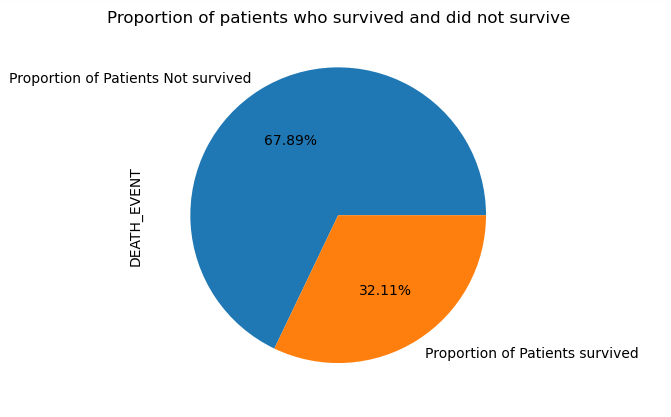
* **Draw a pie chart showing Survival rate of Patients?**

df['DEATH\_EVENT'].value\_counts().plot(kind='pie', labels=['Proportion of Patients Not survived', 'Proportion of Patients survived'], autopct='%1.2f%%')

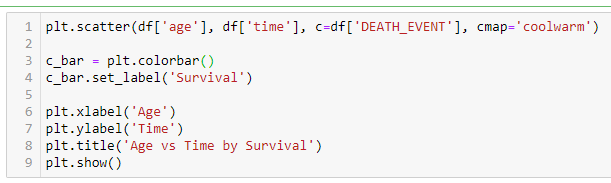
plt.title('Proportion of patients who survived and did not survive')



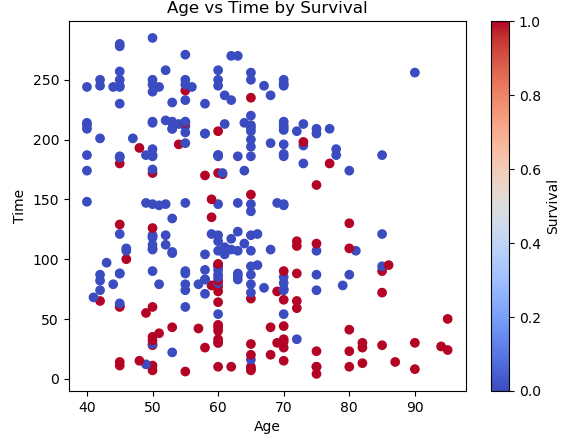
Proportion of Patients Not survived', 'Proportion of patients who survived and did not survive'. We created a pie chart using(kind=pie) function and labeled as 'Proportion of Patients Not survived', 'Proportion of Patients survived’. We found 67.89% of people did not survive and 32.11% of people survived.



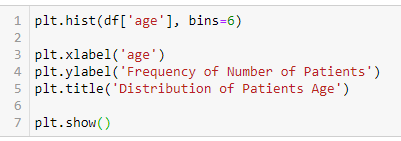
* **Draw a Scatterplot showing age and time based on Death event?**

Here we plotted scatter plot between age of the patients and time based on their Death event. we took the age and time column from the data set and mapped with death data column and assigned color as cool warm and created scatter plot function Using scatter plot function (plt. scatter)

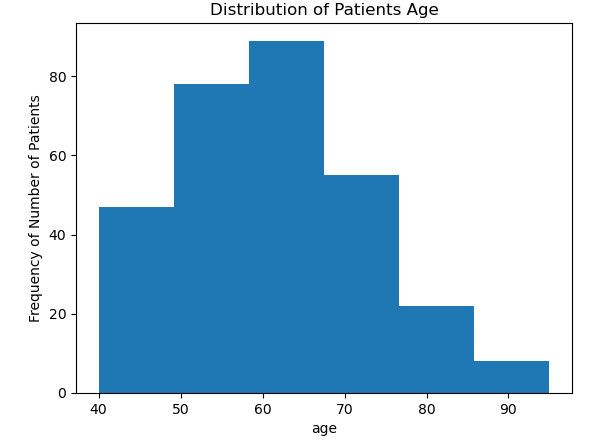
X-axis is labeled as age and y-axis is labeled as fair and title is given as Age Vs Time by Survival.



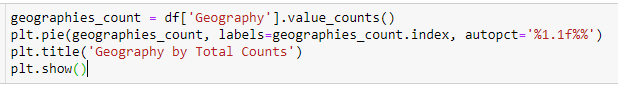
* **Histogram**

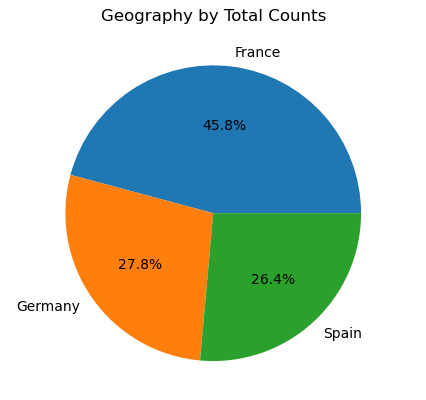


We created a Histogram based on the age of the patients. We took bin size as 6 and X-axis is labeled as Age and Y-axis as frequency of number of patients and titled as Distribution of Patients Age. We created Histogram function (plt. hits)



* **Draw a pie chart showing Percentage of patients based on Geography?**

We plotted the pie based on the number of patients from each country.



**ASSIGNMENT DESIGN FOR FUTURE STUDENTS:**

Expanding the Analysis of Heart Failure Dataset Using Python

In this assignment, you will use Python to analyze the Heart Failure dataset, which contains information about the patients who were affected with Heart Problem. You will use dictionaries, Pandas, and visualization using Matplotlib to explore and analyze this dataset.

* What is the shape of the dataset?
* Which variables are categorical, and which are numerical?
* What is the average value of "serum\_creatinine"?
* What is the median value of "platelets"?
* How many people have diabetes?
* Survival rate of male and female patients?
* Which country has the highest number of deaths?
* What is the maximum value of "serum\_creatinine" for people who died?
* What is the minimum value of "ejection\_fraction" for people who survived?
* How many people are smokers based on gender?
* How many people have an ejection fraction less than or equal to 20?
* How many people have an ejection fraction greater than or equal to 60 and a serum creatinine level greater than 1.5?
* Is there a correlation between high blood pressure and the ejection fraction?

Hints:

1. The Pandas **read\_excel** function can be used to load the xlsx file into a DataFrame.

2. The Pandas **describe** function can be used to get summary statistics for the numerical variables.

3. The Pandas **mean, sum, isnull** function can be used to check or calculate the average, summation, and number of null values.

4. The Matplotlib bar function can be used to create bar charts, the scatter function can be used to create scatter plots, hist function can be used to create histograms, pie function.

**Assessment Criteria (Rubric):**

Work Quality:

• Excellent (10 points): The code is well-organized, clearly documented, and easy to read. All deliverables are met, and the insights gained from the data analysis are insightful and well-presented in the report.

• Good (7 points): The code is mostly well-organized, documented, and easy to read. Most of the deliverables are met, and the insights gained from the data analysis are presented adequately in the report.

• Fair (4 points): The code is poorly organized, documented, and difficult to read. Some of the deliverables are not met, and the insights gained from the data analysis are presented inadequately in the report.

Task Completion:

• Excellent (40 points): All tasks are completed with a high level of accuracy, and the code is written efficiently and effectively. All deliverables are met, and the insights gained from the data analysis are comprehensive, detailed, and accurate.

• Good (30 points): All tasks are mostly completed accurately, but some code may be redundant or inefficient. Most of the deliverables are met, and the insights gained from the data analysis are presented adequately.

• Fair (20 points): Some tasks are not completed accurately or not completed at all, and the code may be very inefficient. Some of the deliverables are not met, and the insights gained from the data analysis are presented inadequately.

Visualization:

• Excellent (25 points): The visualizations are well-designed, accurately represent the data, and are effectively integrated into the report. The visualizations provide valuable insights into the data.

• Good (20 points): The visualizations are mostly well-designed and accurately represent the data, but they may be lacking in some aspects, such as labeling or color coding. The visualizations are adequately integrated into the report and provide some insights into the data.

• Fair (14 points): The visualizations are poorly designed, do not accurately represent the data, and are not effectively integrated into the report. The visualizations do not provide valuable insights into the data.

Overall Assessment:

• Excellent (81-100 points): The code is well-written, well-organized, and efficient. All tasks are completed accurately, and all deliverables are met. The report is well-written, comprehensive, and insightful.

• Good (59-79 points): The code is mostly well-written, and most tasks are completed accurately. The report is adequately written, but some aspects may be lacking in detail or insight.

• Fair (30-58 points): The code is poorly written, and many tasks are not completed accurately. The report is poorly written, and the insights gained from the data analysis are presented inadequately.

**CONCLUSION:**

To get insights into the data, we used pandas to run several analyses on the heart failure dataset. We determined the rows and columns counts as well as the data types of the columns and the number of distinct values included in each column. Additionally, we determined some summary statistics for the dataset, including the median and maximum ejection fractions as well as the mean age of the patients. We also tallied the number of patients with diabetes and high blood pressure diagnoses and sorted the dataset by age. Finally, to better understand the relationships between the variables, we visualized the data using a variety of graphs. In general, the data analysis allowed us to better comprehend the heart failure dataset and its characteristics.

**Team Weightage:**

NARAYANAM NAGA VAMSI SANDEEP ---21%

DECHARAJU PHANINDRA ---21%

SAI KOUSHIK REDDY ---20%

AARTHI AWASTHI ----20%

RAGHU VAMSI ----18%