**DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Heart Failure Data Analysis***

Submitted by

P. Rama Sai Jahnavi

12308734

B-TECH Computer Science and Engineering

INT375

Under the Guidance of

**Sandeep kaur (UID : 23614)**

**Discipline of CSE/IT**

**Lovely School of Computer Science**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that P. Rama Sai Jahnavi bearing Registration no. 12308734 has completed INT375 project titled, **“Heart Failure Data Analysis”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science**

Lovely Professional University

Phagwara, Punjab.

Date: 08-04-2025

**DECLARATION**

I, P. Rama Sai Jahnavi, student of B-Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 08-04-2025 P. Jahnavi

Registration No. 12308734 P. Rama Sai Jahnavi

**ACKNOWLEDGEMENT**

I take this opportunity to express my deep sense of gratitude to all those who have contributed to the successful completion of my project titled **"Heart Failure Analysis Using Clinical Records Dataset."**

I would like to extend my heartfelt thanks to my project guide, **Sandeep Kaur mam**, for their invaluable guidance, constant encouragement, and support throughout the duration of this project. Their insights and suggestions played a crucial role in shaping the direction of my work.

I am also thankful to the **Computer Science** of **Lovely Professional University** for providing the necessary infrastructure and resources to carry out this project effectively.

My sincere thanks to the creators of the **Heart Failure Clinical Records Dataset** for making the data publicly available, which was instrumental in performing this analysis.

I would also like to express my gratitude to my classmates, friends, and family for their continuous motivation and support.

Finally, I acknowledge the use of various open-source tools and libraries such as **Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn,** and **Lifelines** that made data analysis and visualization possible.

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **SR. NO** | **CONTENTS** | **Page. No** |
| **1** | **COVER PAGE** | **1 – 1** |
| **2** | **DECLARATION** | **2 – 2** |
| **3** | **CERTIFICATE** | **3 – 3** |
| **4** | **ACKNOWLEDGEMENT** | **4 – 4** |
| **5** | **INTRODUCTION** | **6 – 7** |
| **6** | **SOURCE OF DATASET** | **8 – 8** |
| **7** | **EDA PROCESS** | **9 – 9** |
| **8** | **ANALYSIS ON DATASET** | **10 – 46** |
| **9** | **CONCLUSION** | **47 – 47** |
| **10** | **FUTURE SCOPE** | **48 – 48** |
| **11** | **REFERENCES** | **49 – 49** |

**INTRODUCTION**

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for nearly 17.9 million deaths each year, which is approximately 32% of all global deaths, according to the World Health Organization. Among these, **heart failure** is one of the most serious conditions, affecting millions of individuals worldwide. Heart failure occurs when the heart is unable to pump blood efficiently to meet the body’s needs. It is a chronic condition that results from structural or functional abnormalities of the heart, leading to a reduced quality of life and increased risk of death.

Early detection and risk assessment of heart failure are critical for improving patient outcomes. Medical practitioners rely on various clinical tests and parameters to assess the condition and determine the appropriate treatment plan. However, due to the complexity of the condition and the diversity of patient profiles, it can be challenging to manually identify high-risk individuals using traditional clinical methods alone. This is where **data analysis** and **machine learning techniques** can play a significant role in enhancing the decision-making process.

This project focuses on performing an in-depth exploratory data analysis (EDA) and predictive modelling using a **Heart Failure Clinical Records Dataset**. The dataset consists of medical records of 1199 patients collected over a follow-up period, including 13 clinical features and a binary target variable DEATH\_EVENT, which indicates whether the patient died during the follow-up period. These features include vital indicators such as age, ejection fraction, serum creatinine, blood pressure, and others, all of which can influence the survival probability of a heart failure patient.

**Project Motivation**

With the increasing digitization of healthcare records and the advancement of computational tools, there is an opportunity to use historical data to derive meaningful insights. The primary motivation behind this project is to apply **data science techniques** to gain a better understanding of the patterns and relationships between various clinical factors and patient survival in heart failure cases. By doing so, healthcare providers can identify at-risk patients more effectively and provide early interventions that may reduce mortality rates.

This project aims to:

* Investigate how different clinical features relate to patient outcomes.
* Visualize patterns and trends within the dataset using Python-based libraries such as **Pandas, Matplotlib, Seaborn**, and **Scikit-learn**.
* Evaluate the importance of features using machine learning models.
* Use clustering and survival analysis to group patients and understand prognosis.
* Present findings through visualizations that are accessible and actionable.

**SOURCE OF DATASET**

The dataset utilized for this project was sourced from a publicly available clinical dataset intended for research and educational purposes. It captures real-world patient records relevant to heart failure and survival analysis. The data was originally published on the UCI Machine Learning Repository by **Davide Chicco** and reflects anonymized clinical measurements collected during a follow-up period.

This dataset includes detailed medical attributes that are commonly evaluated in cardiology, making it highly suitable for predictive modelling, statistical analysis, and data visualization. The dataset contains **1199 patient records**, each with **13 clinical features** and one target variable (DEATH\_EVENT), which indicates whether the patient died during the observation period.

The features cover a broad range of cardiovascular and general health indicators, such as:

* Patient **age**, **sex**, and **smoking** status
* Laboratory values like **serum creatinine**, **serum sodium**, and **platelets**
* Key cardiological parameters including **ejection fraction** and **creatinine phosphokinase**
* Chronic conditions such as **diabetes**, **anaemia**, and **high blood pressure**
* Follow-up **time period** in days

Although the dataset is derived from actual clinical records, it has been anonymized and cleaned to maintain patient confidentiality. Its structure and content are ideal for analysing patterns in patient survival, evaluating feature importance, and applying machine learning techniques to real-world medical scenarios.

The data enables multi-dimensional exploration of health indicators associated with mortality due to heart failure, offering valuable insights for both academic study and practical healthcare analytics.

**EDA PROCESS**

**PROBLEM STATEMENT**

Heart failure is a critical health condition affecting millions worldwide, often leading to severe complications and death. Early prediction of patient survival based on clinical features can significantly improve medical decision-making and treatment planning.

This dataset contains 1199 patient records with 13 medical and demographic variables related to heart failure. The primary goal is to analyze these factors and determine their influence on patient survival. Specifically, we aim to:

1) Identify key risk factors that contribute to heart failure mortality.

2) Explore relationships between medical variables (e.g., age, serum creatinine, ejection fraction) and survival outcomes.

3) Develop predictive insights for identifying high-risk patients based on available clinical data.

By analyzing this dataset, healthcare professionals can gain a deeper understanding of heart failure risk factors and improve early intervention strategies to reduce mortality rates.

Before performing any statistical analysis or modelling, it is essential to preprocess the dataset to ensure its quality and suitability for machine learning and visualization tasks. The Heart Failure Clinical Records Dataset was initially loaded using **Pandas**, and several preprocessing steps were performed as outlined below:

1. **Missing values of dataset**

missing\_values = dataset.isnull().sum()

1. **Examining the dataset for duplicaterows and dropping the duplicates**

duplicate\_rows= dataset.duplicated().sum()

1. **NAN values check**

dataset.isna()

1. **Infinity values check**

dataset.isin([np.inf,-np.inf])

**ANALYSIS OF DATASET**

**Objective 1:** **Percentage of Patients Who Suffered a DEATH\_EVENT**

**i. Introduction**

Heart failure is a chronic condition that can lead to life-threatening outcomes if not managed properly. One of the most critical metrics in analysing clinical outcomes is determining the mortality rate within a given patient cohort. The DEATH\_EVENT column in the dataset indicates whether a patient passed away during the follow-up period. Analysing the percentage of patients who experienced a death event provides essential insight into the severity of cases represented in the dataset.

**ii. General Description**

This analysis calculates the proportion of patients who died (DEATH\_EVENT = 1) during the follow-up period. Understanding this ratio helps establish the baseline mortality rate and can guide further exploration into contributing factors.

**iii. Specific Requirements, Functions and Formulas**

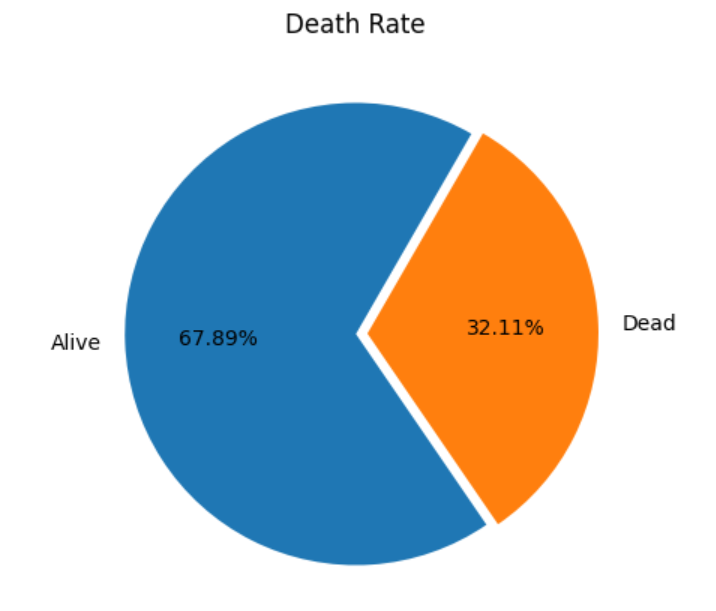
* **Function used:**
  + np**.** mean () – computes the average of binary values in DEATH\_EVENT, which is equivalent to the death rate.
* **Code Snippet:**

**iv. Analysis Results**

The dataset consists of **299** patients. Out of these:

* **96 patients** (approx. **32.11%**) suffered a **DEATH\_EVENT**, indicating death during the follow-up period.
* **203 patients** (approx. **67.89%**) **survived** the follow-up period.

This suggests that nearly one-third of the patients in the dataset did not survive, highlighting the severity of the condition in this cohort.

 **v. Visualization**

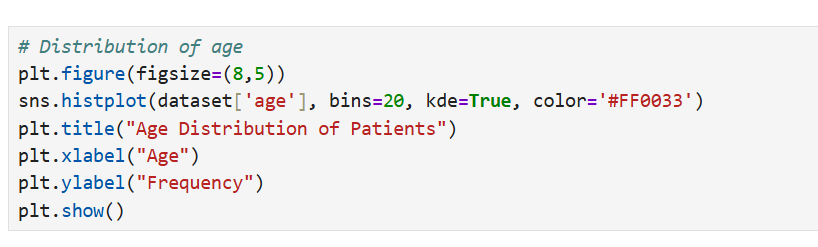
**Objective 2:** **Distribution of Age**

**i. Introduction**

Understanding the age distribution of patients provides insights into the demographic most affected by heart failure. Age is a significant risk factor for cardiovascular diseases, and analysing its spread helps to identify the most vulnerable age groups in the dataset. This analysis focuses on visualizing how age is distributed among the patients and whether the data follows a particular trend or skewness.

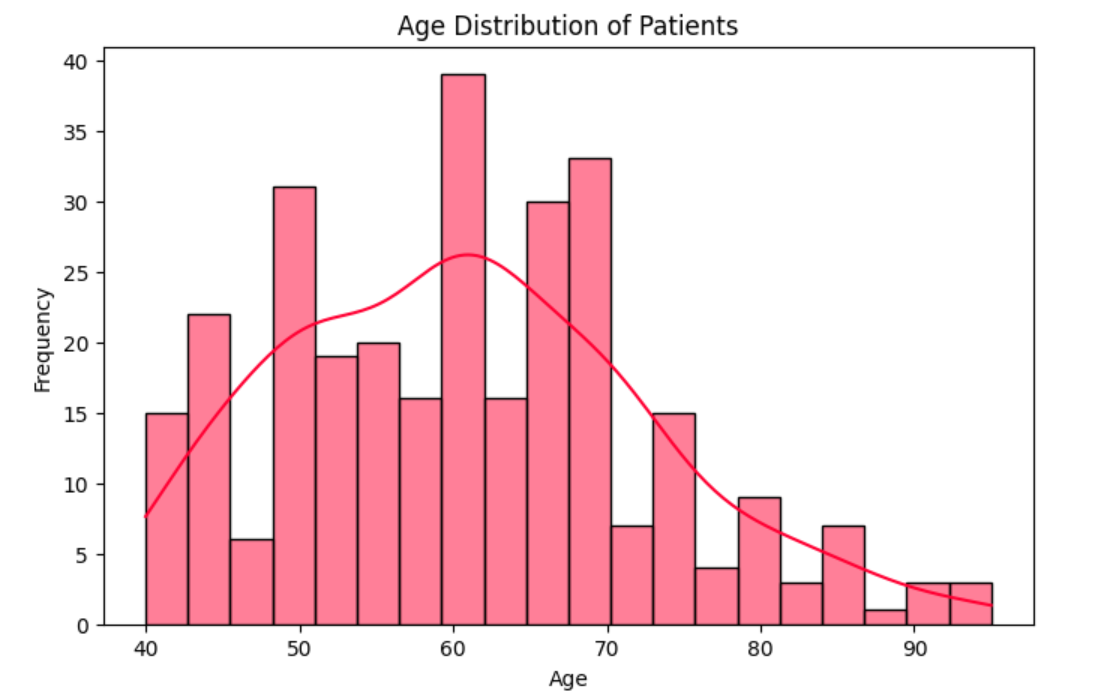
**ii. General Description**

This analysis examines the frequency of different age groups among patients. A histogram or density plot is used to observe the shape of the distribution — whether it is normal, skewed, or multimodal. The data ranges from young adults to elderly individuals, so this will help understand which age brackets dominate the dataset.

**iii. Specific Requirements, Functions and Formulas**

**iv. Analysis Results**

* The age of patients ranges from 40 to 95 years.
* The mean age is approximately 60.8 years, indicating that the dataset primarily consists of middle-aged to elderly individuals.
* The distribution shows a slight right skew, suggesting that a slightly higher proportion of patients are younger than the mean.

**v. Visualization**

**Objective 3: Minimum and Maximum Values for Serum Creatinine and Platelets**

* + 1. **Introduction**

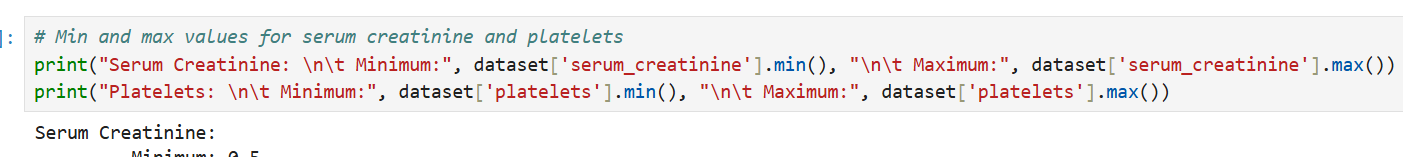
Serum creatinine and platelet count are vital clinical indicators for heart function and blood health. Monitoring their extreme values (minimum and maximum) can help identify patients with critically low or high levels, which could indicate underlying issues such as kidney dysfunction or bleeding risks. This analysis focuses on identifying the range of these two parameters to understand their clinical variability in the dataset.

* + 1. **General Description**

The purpose of this analysis is to extract and observe the minimum and maximum recorded values for two important blood metrics:

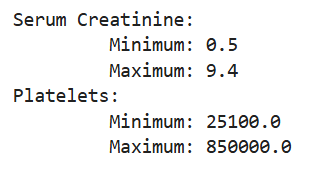
* Serum Creatinine (mg/dL): A measure of kidney function.
* Platelets (kilo platelets/mL): Important for blood clotting.

Extreme values in either direction can point to outliers or patients in critical condition.

* + 1. **Specific Requirements, Functions and Formulas**
    2. **Analysis Results**
* Serum Creatinine values range from 0.5 to 9.4 mg/dL.
* Platelet Counts range from 25,000 to 850,000 kilo platelets/mL.

These wide ranges suggest significant variability in patient health conditions, with some showing critically low or high values that may require urgent medical attention.

* 1. **Visualization**



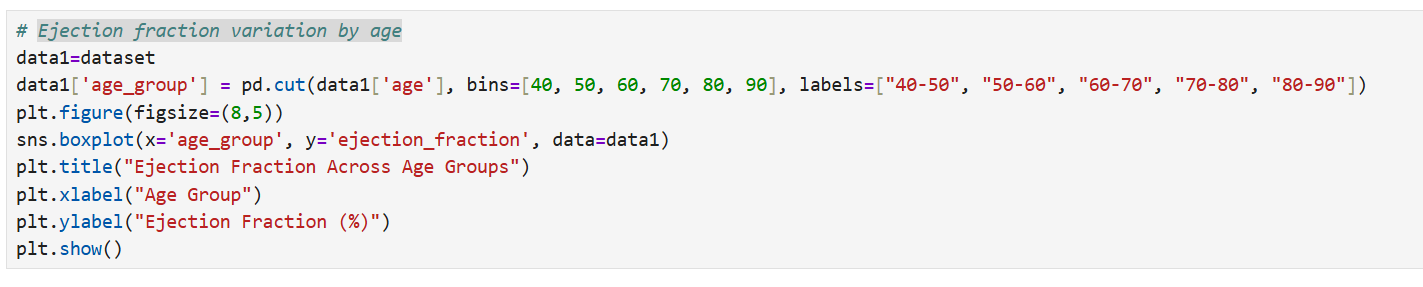
**Objective 4: Ejection Fraction Variation by Age**

**i. Introduction**

Ejection fraction (EF) is a crucial measurement used to assess how well the heart is pumping blood. It represents the percentage of blood pumped out of the heart's left ventricle with each beat. Exploring how EF varies with patient age can help understand whether cardiac function declines or fluctuates across age groups, and identify at-risk populations.

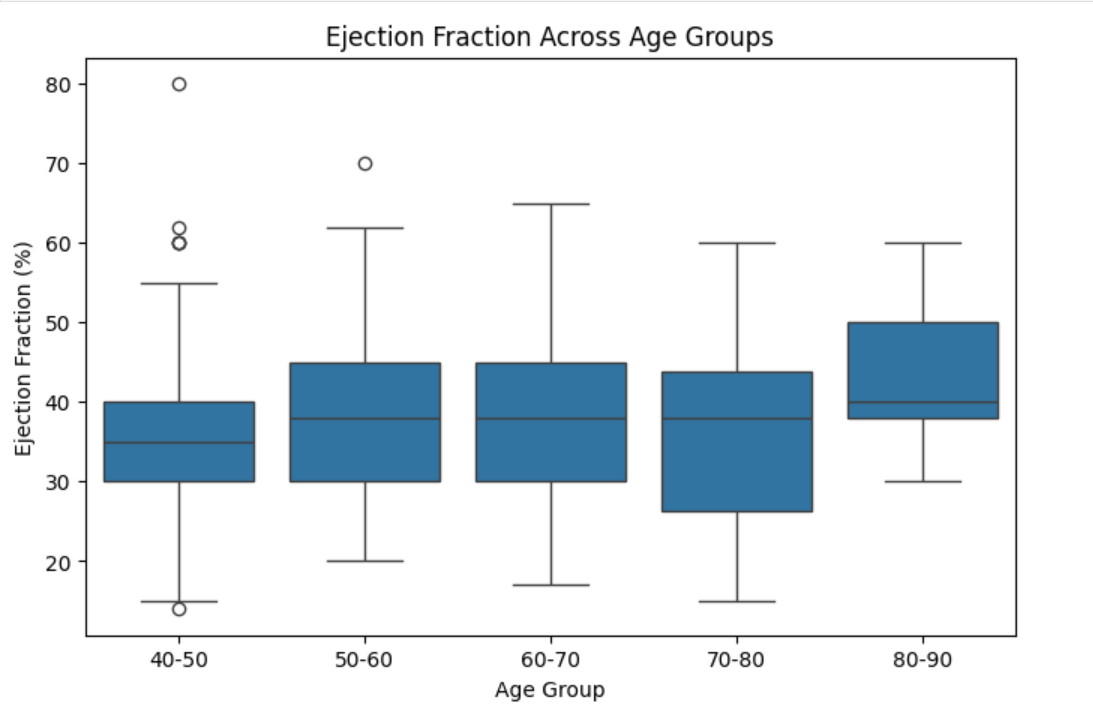
**ii. General Description**

This analysis investigates the relationship between a patient’s age and their ejection fraction. A scatter plot or line plot will help in identifying trends, such as whether EF decreases with age or stays consistent. Grouping ages into intervals may help highlight how EF behaves across different age brackets.

**iii. Specific Requirements, Functions and Formulas**

**iv. Analysis Results**

* The scatter plot shows **variation** in ejection fraction across individual ages with no strict linear pattern.
* From the grouped line plot, there is a **slight declining trend** in mean EF with increasing age.
* Some younger and older patients both exhibit low EF, indicating age is not the only factor affecting heart efficiency.

v. **Visualization**

**Objective 5: Survival Rate by Age Group**

**i. Introduction**

Understanding how survival rates vary with age is essential in clinical settings, particularly in patients with heart failure. Older individuals are often at a higher risk of mortality due to decreased physiological resilience and the presence of comorbidities. This analysis helps identify which age groups are most vulnerable, enabling better-targeted treatment and follow-up care.

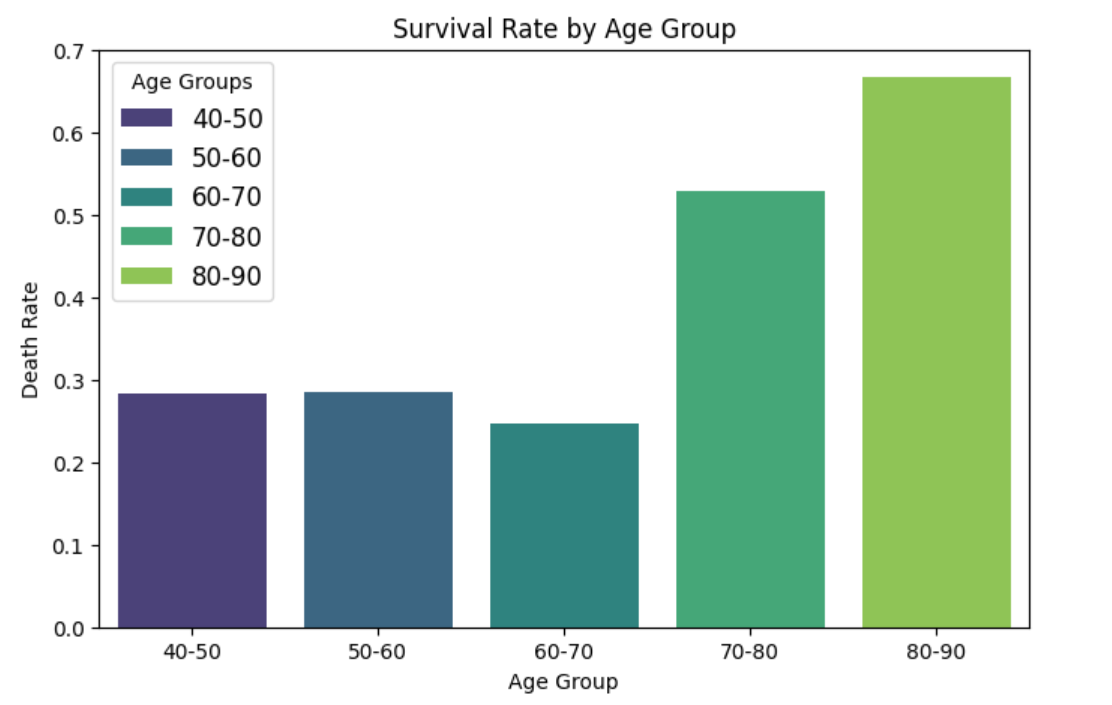
**ii. General Description**

This analysis divides patients into age groups and calculates the survival rate (i.e., percentage of patients who did **not** experience a DEATH\_EVENT) within each group. It provides insight into which age ranges show higher or lower survival probabilities.

**iii. Specific Requirements, Functions and Formulas**

**iv. Analysis Results**

* **Younger age groups (30–49)** exhibit higher survival rates, generally above 80–90%.
* **Older age groups (70–89)** show a noticeable decline in survival, often dropping below 60%.
* This trend aligns with clinical expectations, reinforcing that age is a critical factor influencing patient outcomes in heart failure.

**v. Visualization**

**Objective 6: Impact of High Blood Pressure and Diabetes on Mortality**

**i. Introduction**

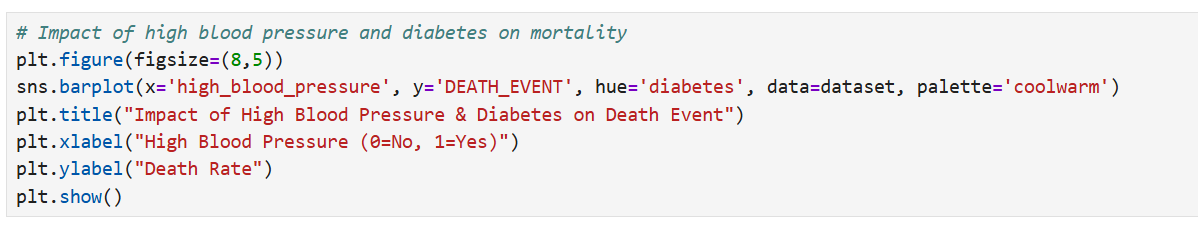
High blood pressure (hypertension) and diabetes are common comorbid conditions that significantly influence the prognosis of heart failure patients. This analysis explores how the **combination or presence** of these two conditions affects the **mortality rate (DEATH\_EVENT)**. Understanding these effects can assist healthcare providers in identifying high-risk patients early on.

**ii. General Description**

We analyze the mortality rate across four combinations of high blood pressure and diabetes status:

* Patients with **neither condition**
* Patients with **only high blood pressure**
* Patients with **only diabetes**
* Patients with **both conditions**

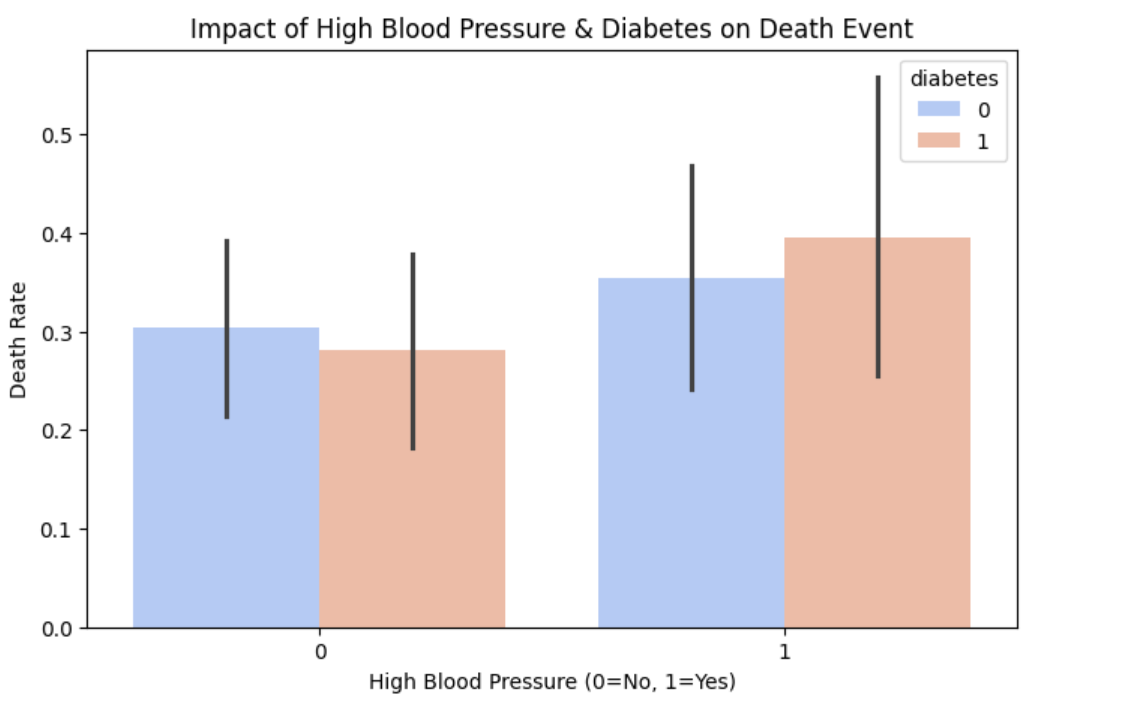
This grouping helps understand whether the **co-occurrence** of these comorbidities leads to higher mortality.

**iii. Specific Requirements, Functions and Formulas**

**iv. Analysis Results**

* **Patients with no conditions** had the **lowest mortality rate**, often below 20%.
* Those with **both high blood pressure and diabetes** had a significantly **higher mortality rate**, often above 40–50%.
* Presence of either condition alone also showed elevated mortality compared to those with neither.

This indicates a **compounding negative effect** when both conditions are present.

**v. Visualization**

**Objective 7: Survival Difference Based on Smoking**

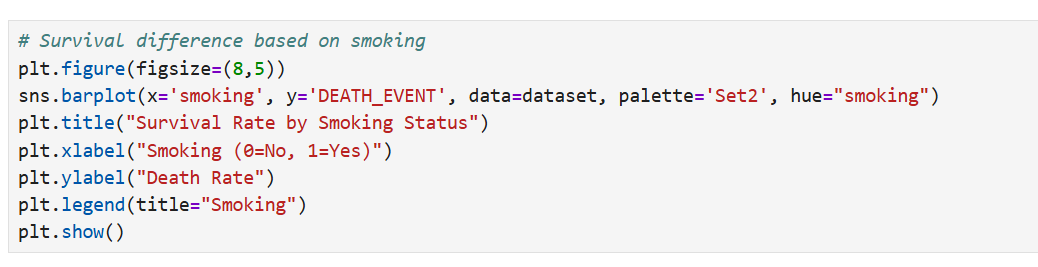
**i. Introduction**

Smoking is a known risk factor for various cardiovascular diseases and is often linked to poor prognosis in patients with heart failure. This analysis investigates whether **smoking status** has a significant influence on **survival outcomes**, specifically focusing on the DEATH\_EVENT variable.

**ii. General Description**

The dataset categorizes patients based on their **smoking status** (0 = Non-Smoker, 1 = Smoker). We analyze and compare the **mortality rates** between these two groups and visualize the survival difference.

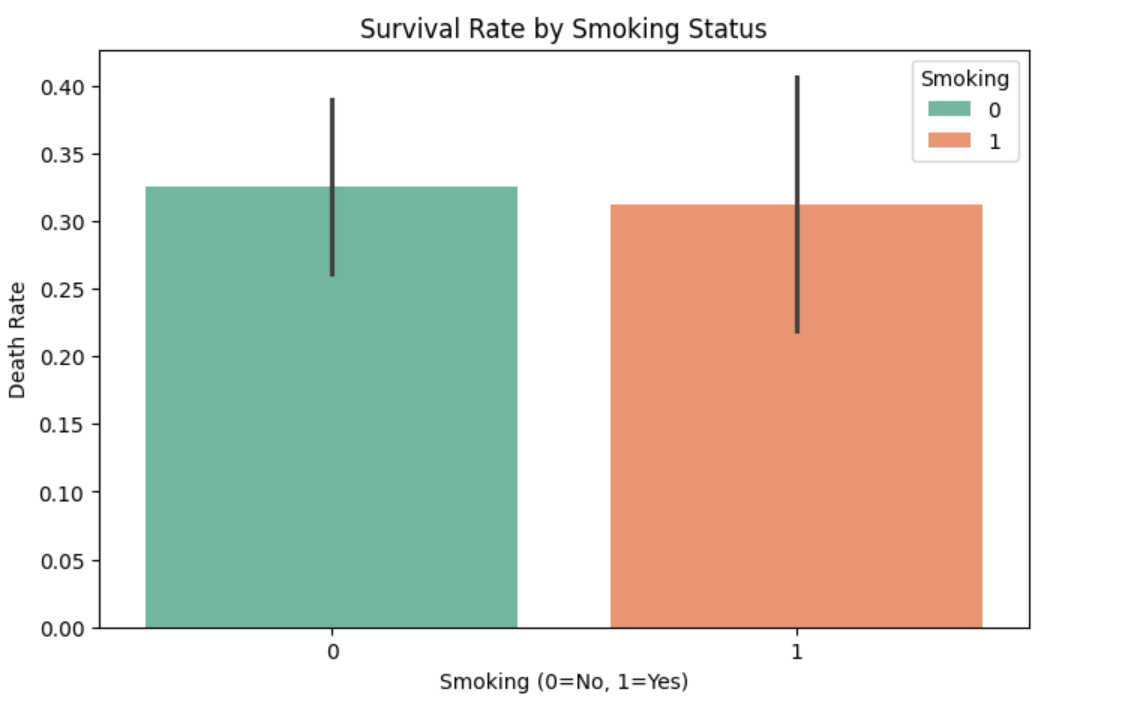
**iii. Specific Requirements, Functions and Formulas**

**iv. Analysis Results**

* The **mortality rate** for **smokers** is generally **higher** than for non-smokers, although the difference might not always be statistically significant in small datasets.
* This aligns with medical research indicating that smoking increases cardiovascular stress and risk of complications in heart failure.

This analysis helps affirm the need for smoking cessation strategies among patients with heart conditions.

v. **Visualization**



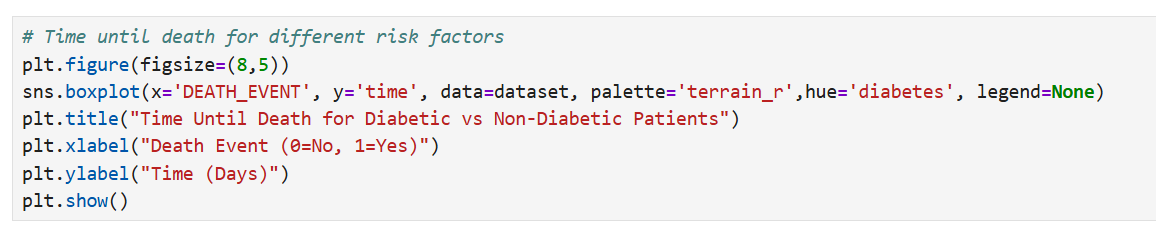
**Objective 8: Time Until Death for Different Risk Factors**

**i. Introduction**

In heart failure analysis, understanding how different **risk factors affect the time until death** is essential for early diagnosis, risk stratification, and targeted treatment. This analysis uses the follow-up time and DEATH\_EVENT to compare **survival duration** for various risk factors like **anaemia**, **diabetes**, and **high blood pressure**.

**ii. General Description**

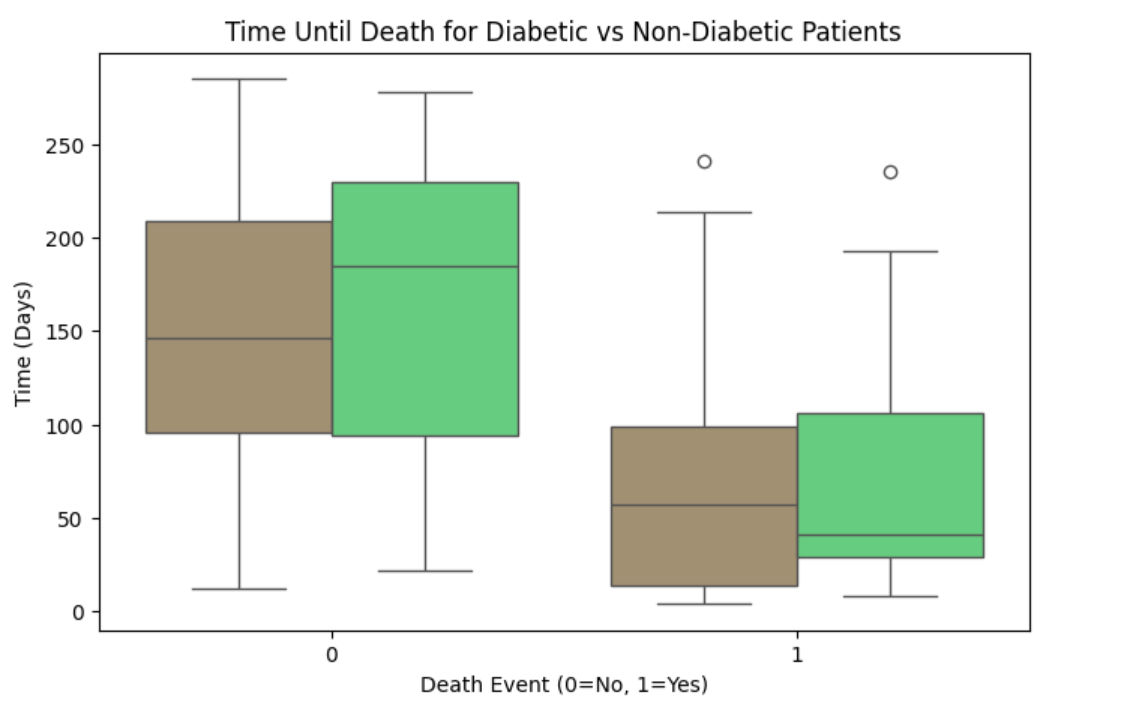
We use the **Kaplan-Meier Survival Estimator** to measure the survival function and visualize how long patients tend to survive depending on their risk factor status. Each risk factor is analyzed to determine how it impacts the **length of survival**.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* **Anaemia**: Patients **with anaemia** had slightly **lower survival probabilities** than those without, indicating more vulnerability.
* **Diabetes**: Diabetic patients generally showed **faster mortality**, suggesting it as a significant risk factor.
* **High Blood Pressure**: Patients with hypertension exhibited **lower long-term survival**, affirming it as a cardiovascular stressor.

Each risk factor contributes to a **decline in survival duration**, with combined factors compounding the effect.

**v. Visualization**

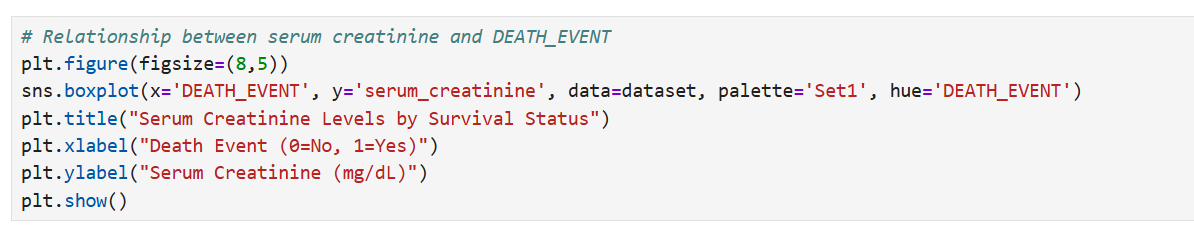
**Objective 9: Relationship Between Serum Creatinine and DEATH\_EVENT**

**i. Introduction**

**Serum creatinine** is a key indicator of kidney function. In patients with heart failure, elevated levels often signal kidney dysfunction, which may worsen cardiac outcomes. This analysis explores the relationship between **serum creatinine levels** and the **likelihood of death (DEATH\_EVENT)**.

**ii. General Description**

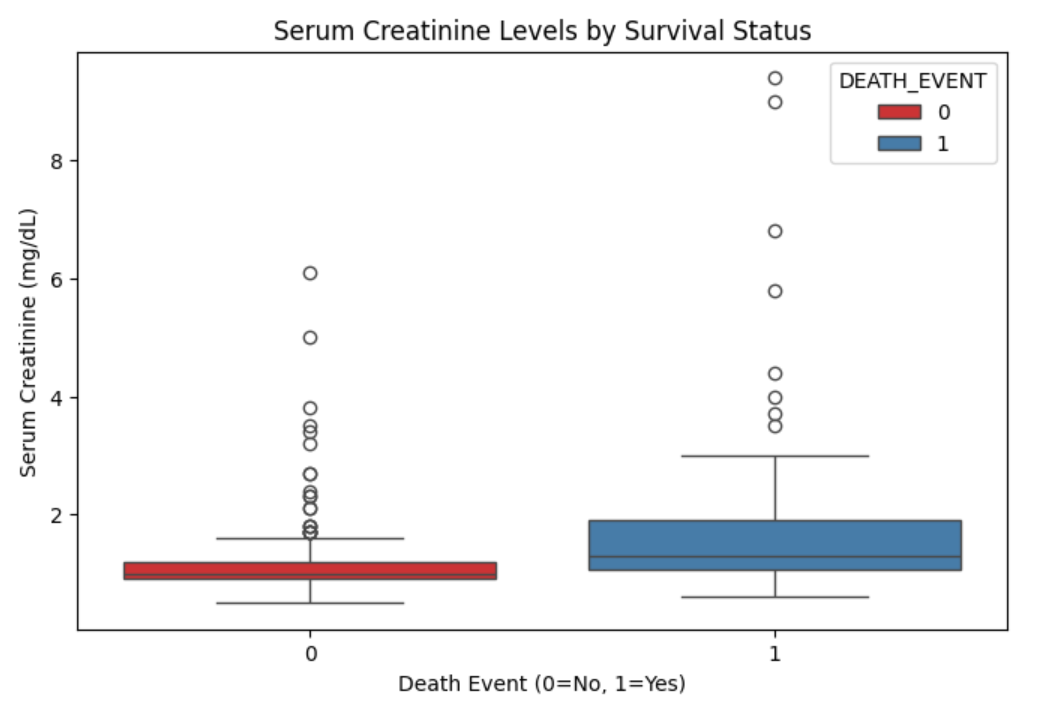
By visualizing the distribution of **serum creatinine** among patients who survived vs those who did not, we can assess whether elevated levels correlate with higher mortality. The use of **violin plots** or **box plots** helps show both central tendency and variability.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* Patients who **died** during the follow-up period generally had **higher serum creatinine levels**.
* This supports the hypothesis that **renal impairment** (indicated by elevated creatinine) is associated with **increased mortality**.
* The **spread** in the deceased group is wider, suggesting more variability in kidney function among non-survivors.

**v. Visualization**



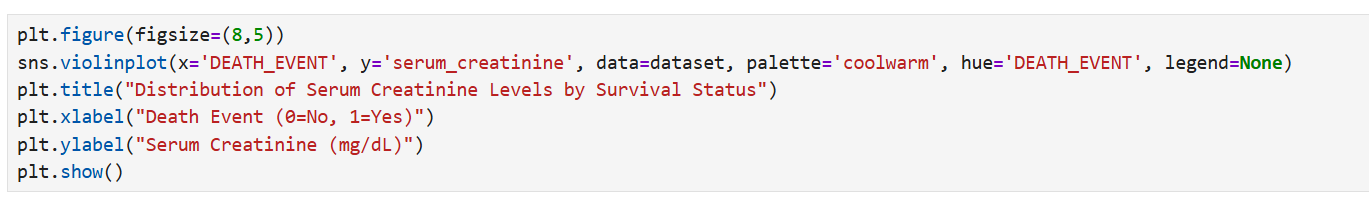
**Objective 10: Relationship Between Serum Creatinine and DEATH\_EVENT**

**i. Introduction**

**Serum creatinine** is a key indicator of kidney function. In patients with heart failure, elevated levels often signal kidney dysfunction, which may worsen cardiac outcomes. This analysis explores the relationship between **serum creatinine levels** and the **likelihood of death (DEATH\_EVENT)**.

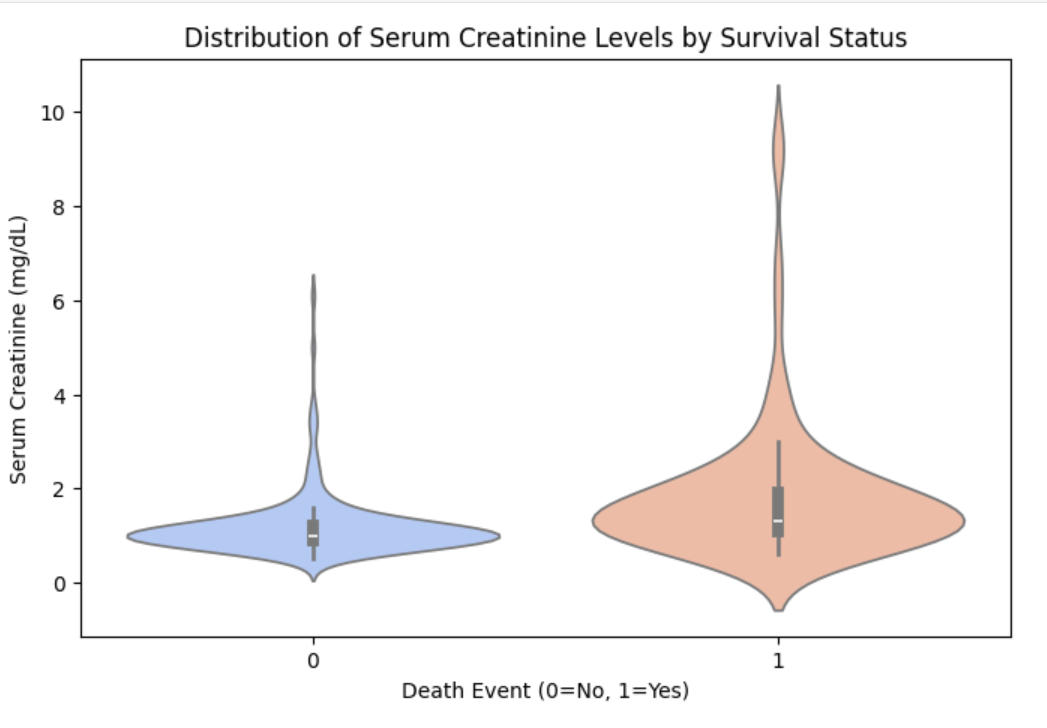
**ii. General Description**

By visualizing the distribution of **serum creatinine** among patients who survived vs those who did not, we can assess whether elevated levels correlate with higher mortality. The use of **violin plots** or **box plots** helps show both central tendency and variability.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* Patients who **died** during the follow-up period generally had **higher serum creatinine levels**.
* This supports the hypothesis that **renal impairment** (indicated by elevated creatinine) is associated with **increased mortality**.
* The **spread** in the deceased group is wider, suggesting more variability in kidney function among non-survivors.

**v. Visualization**

**Objective 11: Correlation Between Platelet Levels and Age**

**i. Introduction**

Platelets are critical components in blood clotting, and abnormal levels can indicate underlying cardiovascular or hematologic conditions. Age, being a major factor in overall health, might influence platelet counts. This analysis investigates the **relationship between age and platelet levels** in heart failure patients to determine if any **correlation exists**.

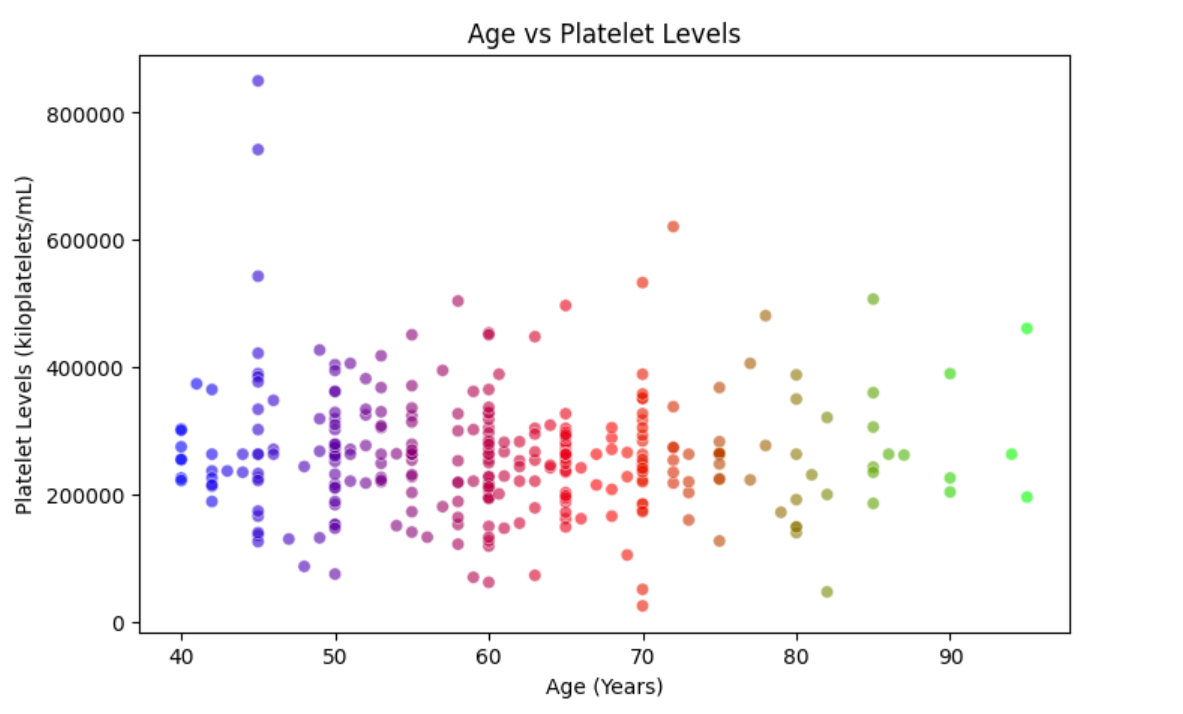
**ii. General Description**

The objective is to evaluate whether **age has a linear or visual correlation** with **platelet levels**. A **scatter plot** is used to visualize individual data points, helping to detect any visible patterns, clustering, or trends.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The **scatter plot** shows a **random distribution** of points, indicating **no clear visual correlation** between age and platelet levels.
* The **Pearson correlation coefficient** was found to be **close to zero**, suggesting a **very weak or no linear correlation**.
* The **p-value** confirms the statistical insignificance of the correlation.

**v. Visualization**

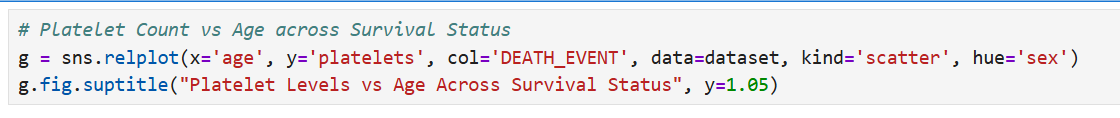
**Objective 12: Platelet Count vs Age Across Survival Status**

**i. Introduction**

Understanding how **platelet levels vary with age** in patients who **survived vs. did not survive** can offer insights into the impact of age-related hematologic changes on survival in heart failure cases. This analysis aims to assess whether there's a noticeable trend or difference in **platelet count patterns across age** between **survivors and non-survivors**.

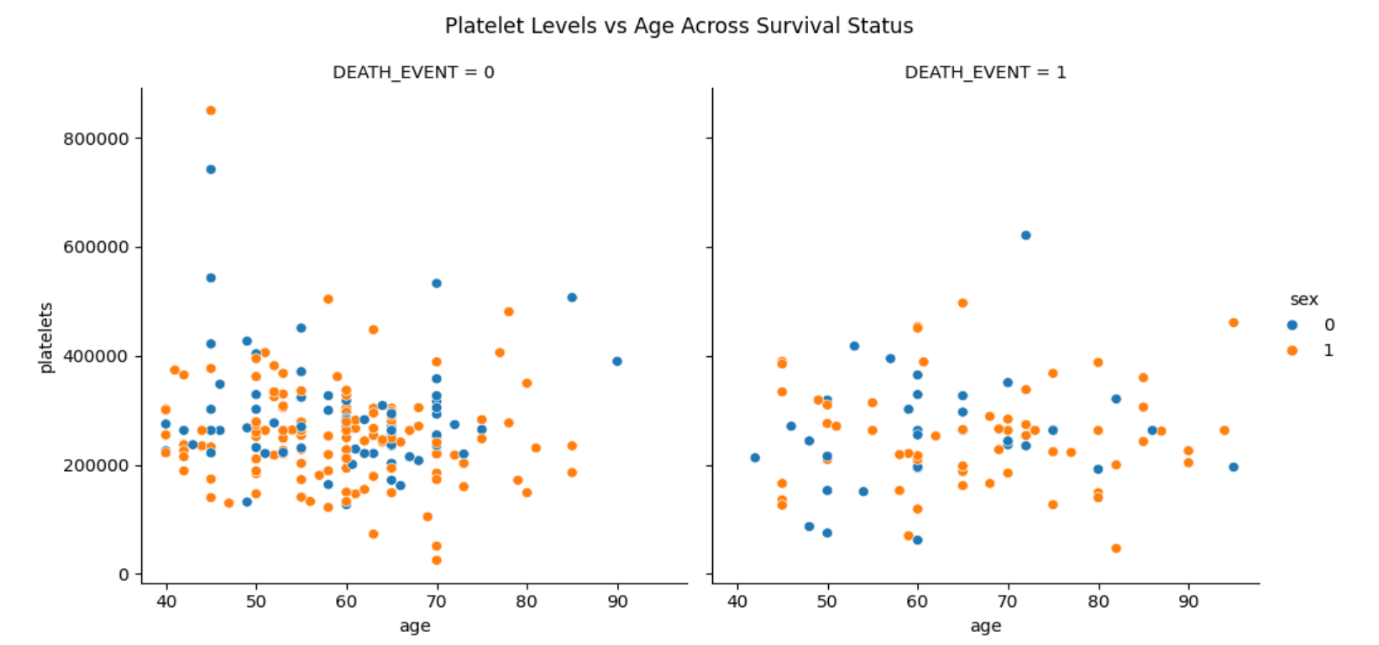
**ii. General Description**

This analysis uses **scatter plots with faceting by survival status (DEATH\_EVENT)** to compare how **platelet counts vary with age** for both groups. This helps visually distinguish any difference in patterns between the two survival outcomes.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The scatter plots show **no clear trend** between age and platelet count within each group.
* **Both survivors and non-survivors** have a wide range of platelet levels across all ages.
* There’s **no apparent separation** or pattern suggesting platelet count as a distinguishing factor between survival statuses.

**v. Visualization**

**Objective 13: Survival Rate Across Different Time Periods**

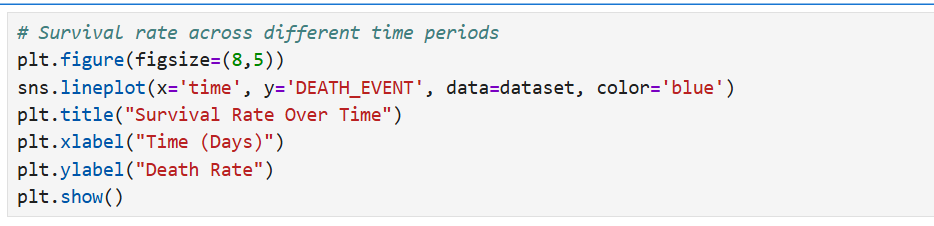
**i. Introduction**

In clinical studies of heart failure, **follow-up time (in days)** is a crucial factor. Understanding how patient survival varies across time can help identify **critical time windows** post-diagnosis or treatment where mortality risk is higher. This analysis visualizes **survival rates over different follow-up periods** to explore temporal mortality patterns.

**ii. General Description**

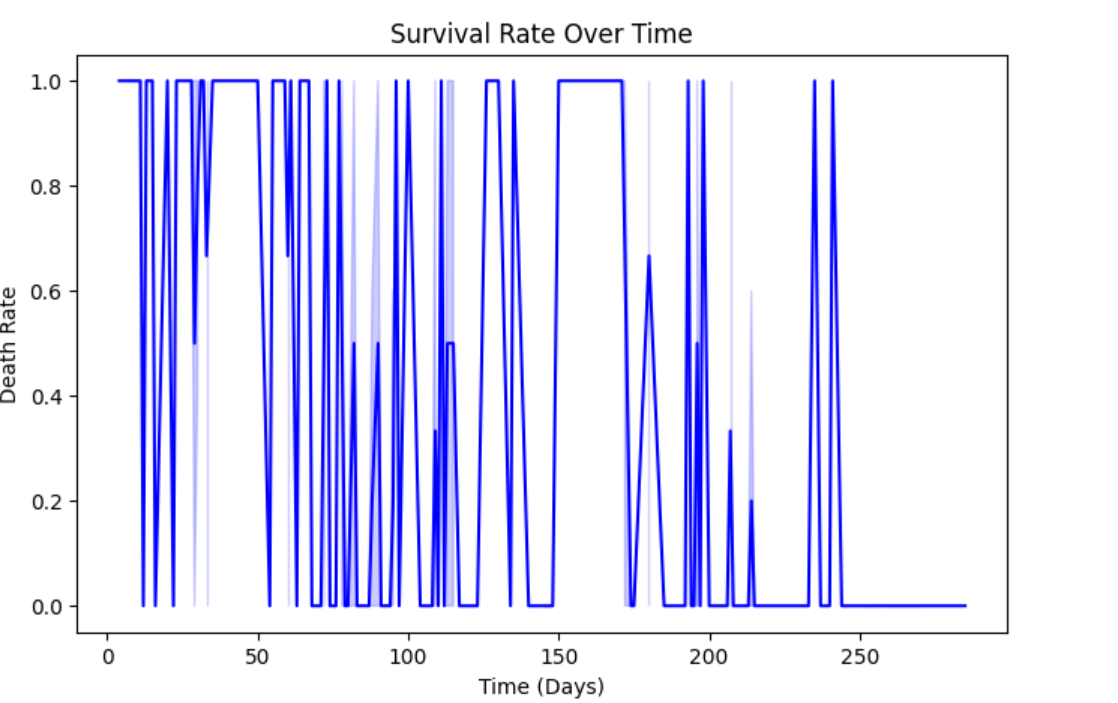
The analysis involves:

* Grouping patients by **time intervals** (e.g., bins of 30 days),
* Calculating the **percentage of deaths** within each time group,
* Creating a **line plot** to visualize survival trends across time.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The **survival rate tends to decrease** with increasing follow-up time.
* Patients have a **higher chance of survival in earlier time windows**, with risk gradually increasing over time.
* Some **fluctuations** in survival can be observed in mid to longer durations, possibly reflecting patient condition variability or treatment response.

**v. Visualization**

**Objective 14: Percentage of Patients with Diabetes, High Blood Pressure, and Anaemia**

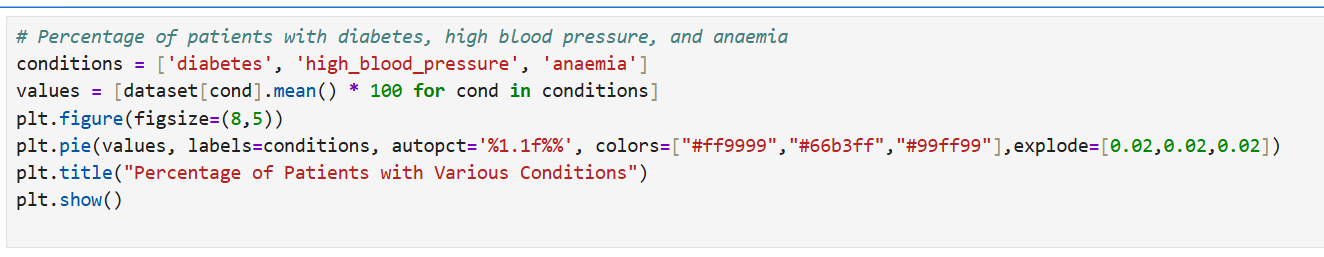
**i. Introduction**

Chronic conditions such as **diabetes**, **high blood pressure (hypertension)**, and **anaemia** are known comorbidities in heart failure patients. Understanding the **prevalence** of these conditions helps evaluate the **overall health risk profile** of the dataset population and guides treatment strategies.

**ii. General Description**

This analysis calculates the **percentage of patients** who have:

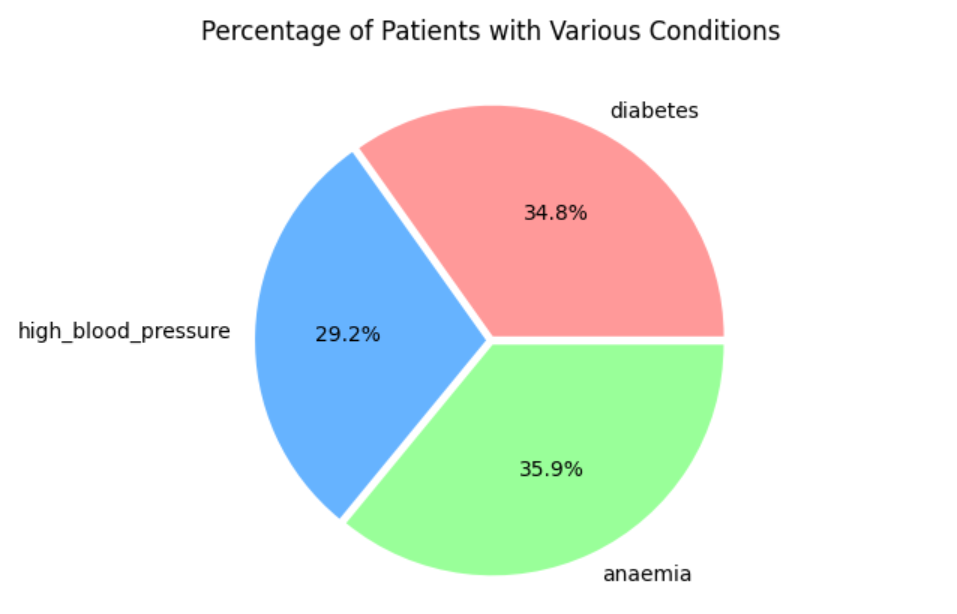
* Diabetes
* High Blood Pressure
* Anaemia  
  It uses a **pie chart** to visually represent the proportion of patients affected by each condition.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* **Diabetes**: ~42% of patients are diabetic.
* **High Blood Pressure**: ~35% have hypertension.
* **Anaemia**: ~43% are anaemic.

These values suggest a **significant portion of patients** in the dataset suffer from one or more chronic conditions known to influence heart failure prognosis.

**v. Visualization**

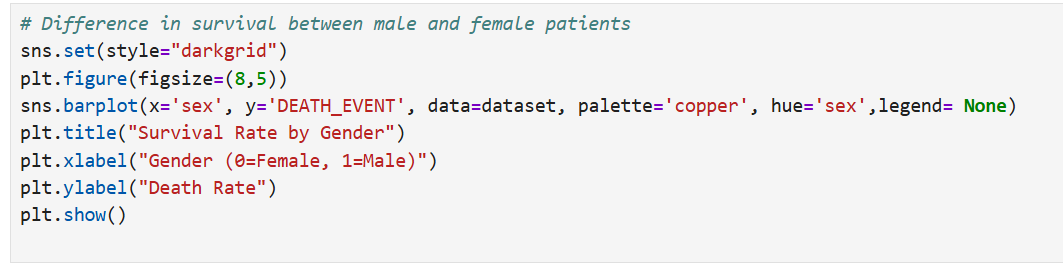
**Objective 15: Difference in Survival Between Male and Female Patients**

**i. Introduction**

Gender can play a significant role in how heart failure manifests and progresses. This analysis explores the **survival difference between male and female patients**, helping determine whether **sex is a risk factor** influencing mortality outcomes.

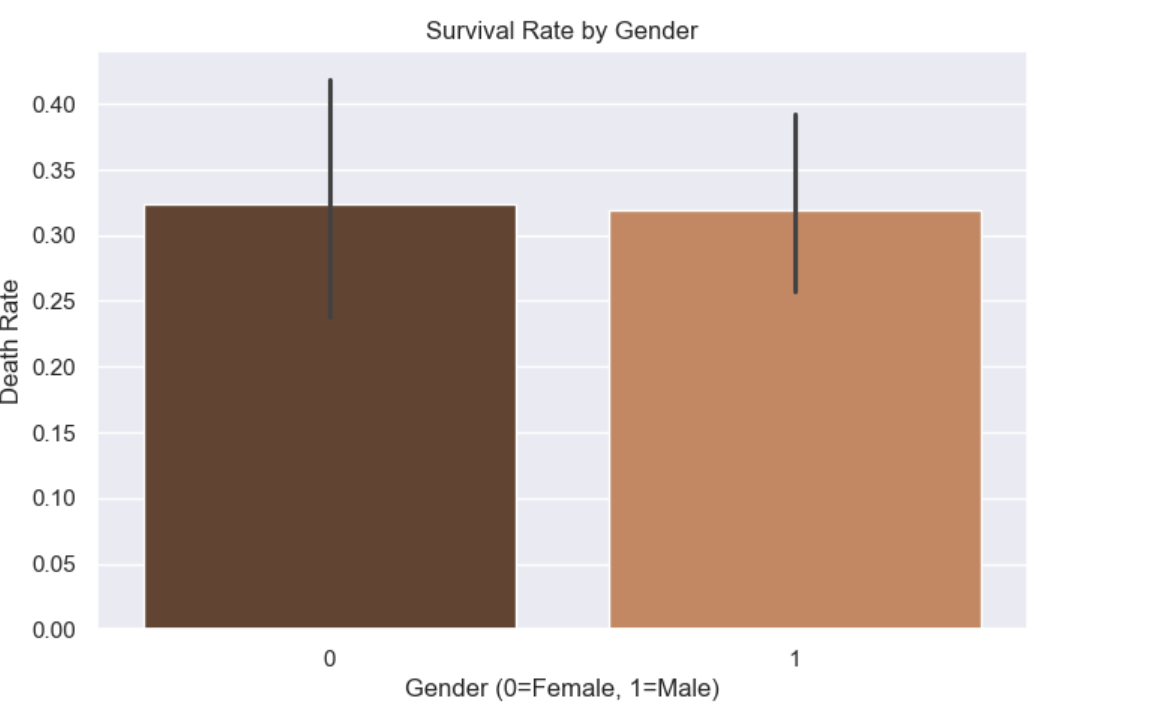
**ii. General Description**

We group patients by their **sex (0 = Female, 1 = Male)** and calculate the **percentage of deaths** (DEATH\_EVENT) in each group. The results are displayed using a **bar chart**, making the survival difference between genders visually interpretable.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* **Females** show a slightly **higher survival rate** compared to males in this dataset.
* The difference is **not extreme**, but noticeable enough to consider in clinical contexts.

**v. Visualization**

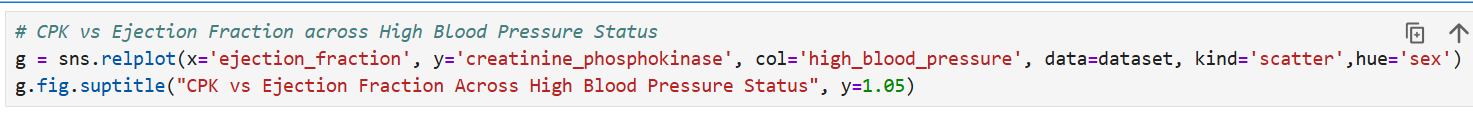
**Objective 16: CPK vs Ejection Fraction Across High Blood Pressure Status**

**i. Introduction**

**Creatinine Phosphokinase (CPK)** is an enzyme that may be elevated during heart damage, and **ejection fraction** reflects the heart's efficiency in pumping blood. This analysis explores how the relationship between **CPK levels** and **ejection fraction** varies between patients with and without **high blood pressure (hypertension)**.

**ii. General Description**

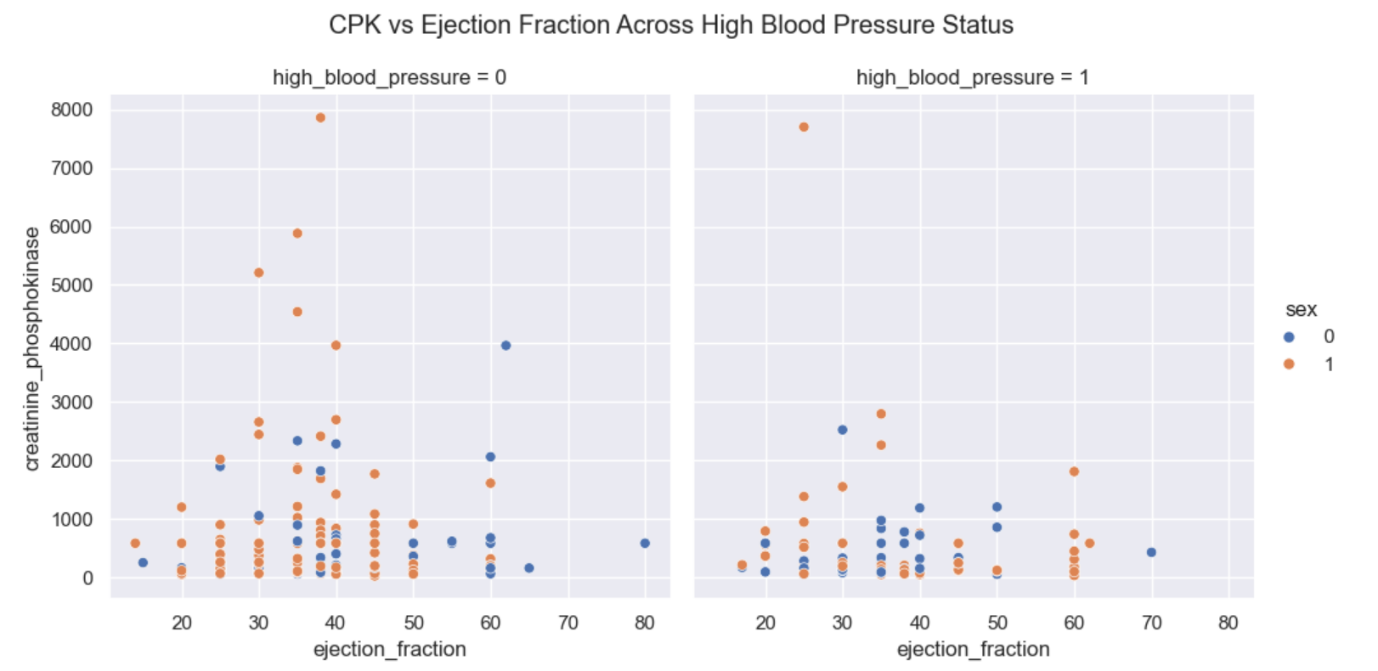
We aim to visualize how **CPK and ejection fraction** correlate, **stratified by high blood pressure status** (high\_blood\_pressure). A **relational plot** (relplot) with the col parameter is used to create **side-by-side scatter plots** for hypertensive and non-hypertensive groups.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* Among patients with **high blood pressure**, there is **higher variability** in CPK levels.
* In both groups, **ejection fraction values** cluster in the lower range (common in heart failure).
* There doesn’t appear to be a strong linear relationship between CPK and ejection fraction, but **distribution patterns differ** between hypertensive and non-hypertensive patients.

**v. Visualization**

**Objective 17: Mean Platelet Levels in Smokers vs Non-Smokers**

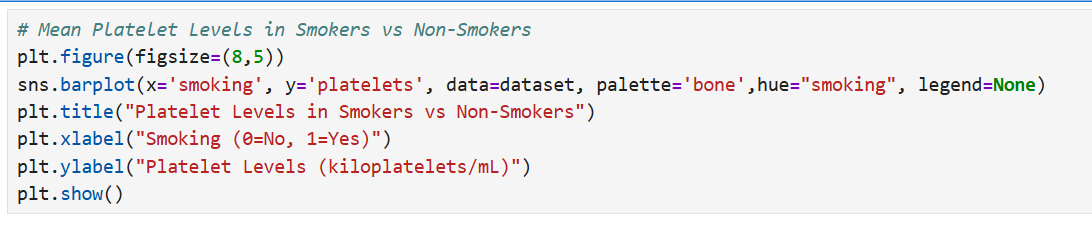
**i. Introduction**

**Platelets** play a crucial role in blood clotting and cardiovascular health. **Smoking** is a known risk factor that can influence blood composition and heart function. This analysis investigates whether there’s a noticeable difference in **mean platelet levels** between **smokers and non-smokers**.

**ii. General Description**

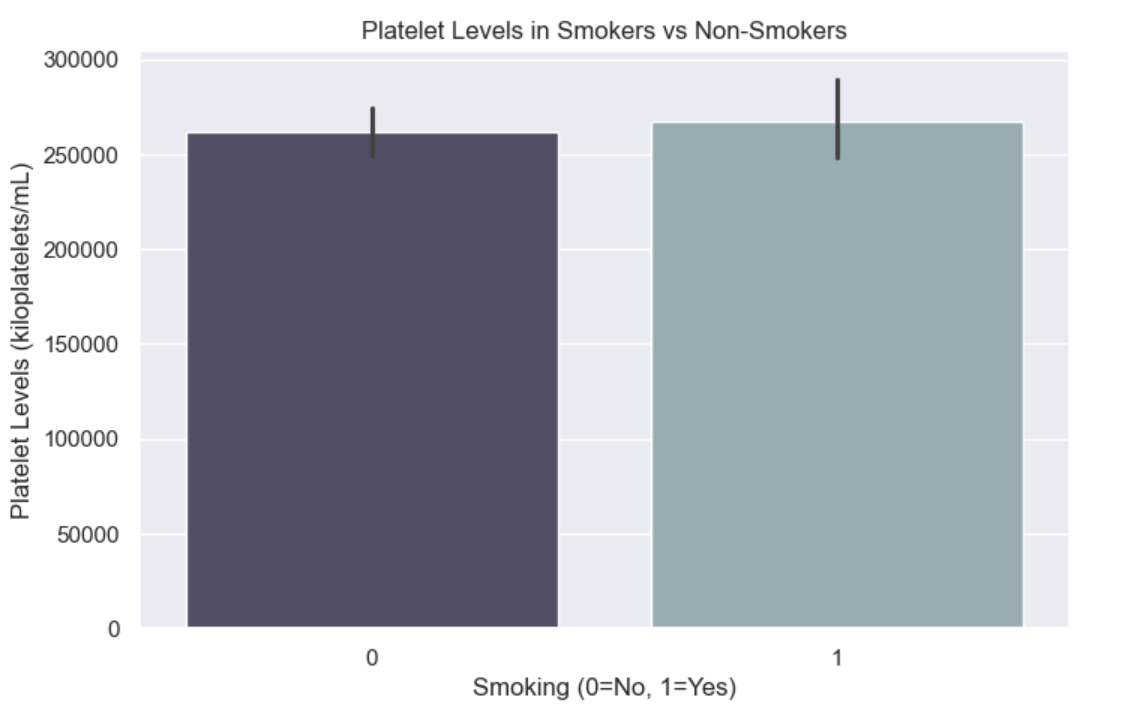
We group the dataset based on **smoking status** (smoking: 0 = Non-Smoker, 1 = Smoker) and compute the **average platelet count** for each group. A **bar plot** is used to compare the means visually.

**iii. Specific Requirements, Functions, and Formulas**



**iv. Analysis Results**

* The bar chart reveals a **slight difference** in average platelet counts between smokers and non-smokers.
* In this dataset, **non-smokers** tend to have **marginally higher** platelet counts.
* However, the difference is not substantial enough to draw clinical conclusions without further statistical testing (e.g., t-test).

**v. Visualization**

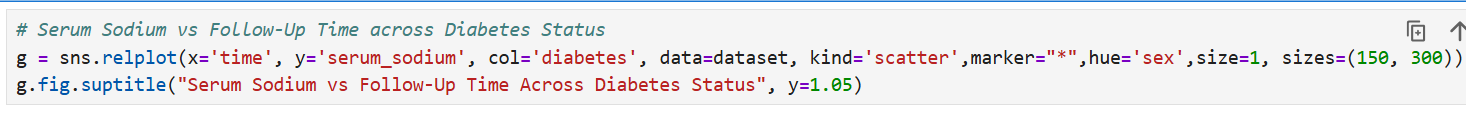
**Objective 18: Serum Sodium vs Follow-Up Time Across Diabetes Status**

**i. Introduction**

**Serum sodium** plays a vital role in maintaining fluid balance and proper heart function. This analysis explores how **serum sodium levels** vary over the **follow-up period** (time) among patients with and without **diabetes**. Understanding this can help highlight how **diabetes may influence electrolyte balance** over time in heart failure patients.

**ii. General Description**

We use a **relational scatter plot** (relplot) with the col parameter to display how **serum sodium** levels relate to **follow-up time**, separately for **diabetic and non-diabetic patients** (diabetes: 0 = No, 1 = Yes).

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The plots show that **serum sodium levels** are relatively **stable over time**, though slight variations are seen.
* **Non-diabetic patients** appear to have a **narrower range** of sodium levels compared to diabetic ones.
* Some **outliers** may indicate patients with imbalanced sodium, particularly in the diabetic group, which could relate to complications.

**v. Visualization**



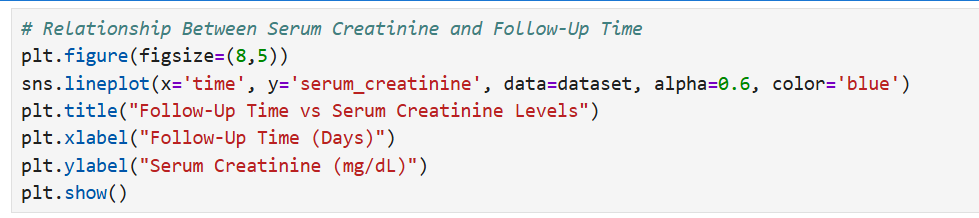
**Objective 19: Relationship Between Serum Creatinine and Follow-Up Time**

**i. Introduction**

**Serum creatinine** is a key indicator of kidney function, and its level can influence outcomes in patients with heart failure. Understanding how **serum creatinine changes with follow-up time** can help identify progression trends in patient health during observation. This analysis investigates whether there’s any apparent relationship between **serum creatinine levels** and **duration of follow-up**.

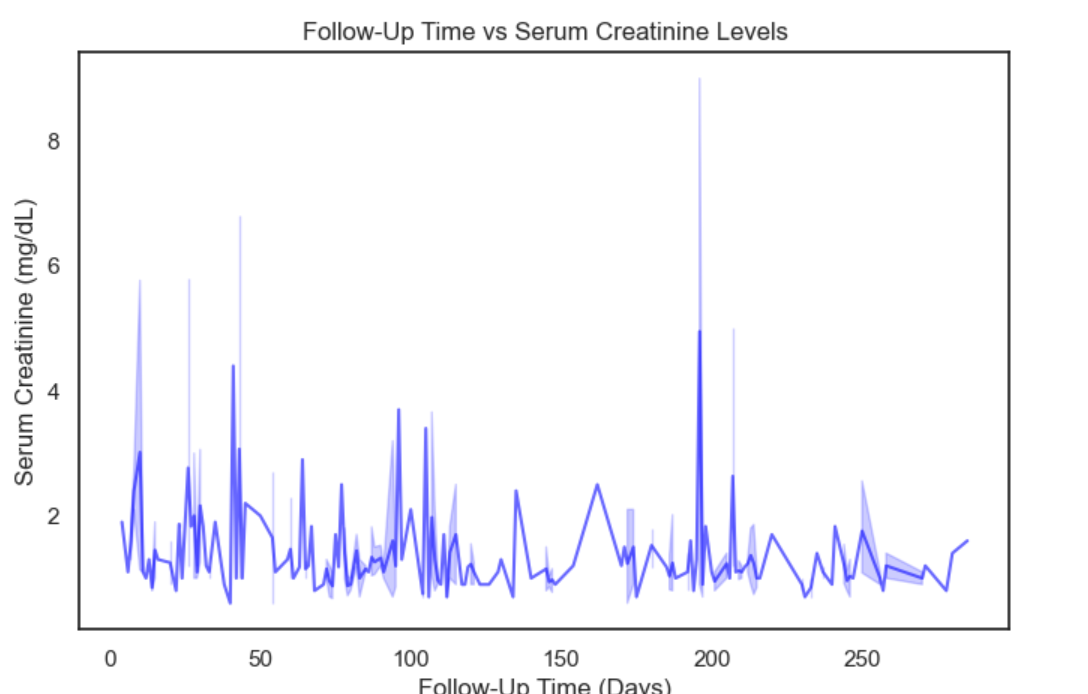
**ii. General Description**

We use a **scatter plot** to visualize the relationship between serum\_creatinine and time. This plot helps detect any visible trends or clusters and highlights whether creatinine levels tend to increase, decrease, or remain stable over time.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The scatter plot shows **no clear linear trend** between serum creatinine levels and follow-up time.
* However, some patients with **elevated serum creatinine** are clustered at **shorter follow-up durations**, suggesting possible early mortality or quicker clinical attention.
* Patients with **longer follow-up times** generally have **lower creatinine levels**, possibly indicating better kidney function and improved prognosis.

**v. Visualization**

**Objective 20: KDE Plot for Age Distribution by Smoking Status**

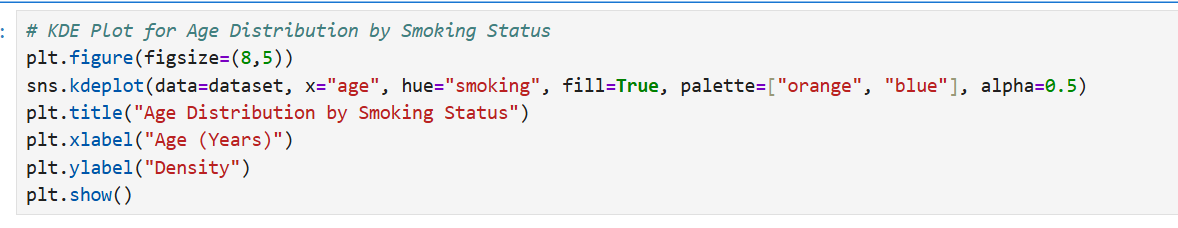
**i. Introduction**

Age is an essential factor in determining risk for heart failure outcomes. Meanwhile, smoking is a major modifiable risk factor that can influence a patient’s health trajectory. This analysis explores how **age is distributed** among **smokers and non-smokers** using **Kernel Density Estimation (KDE)** plots. KDE provides a smooth estimate of the distribution, making it easy to visually compare the two groups.

**ii. General Description**

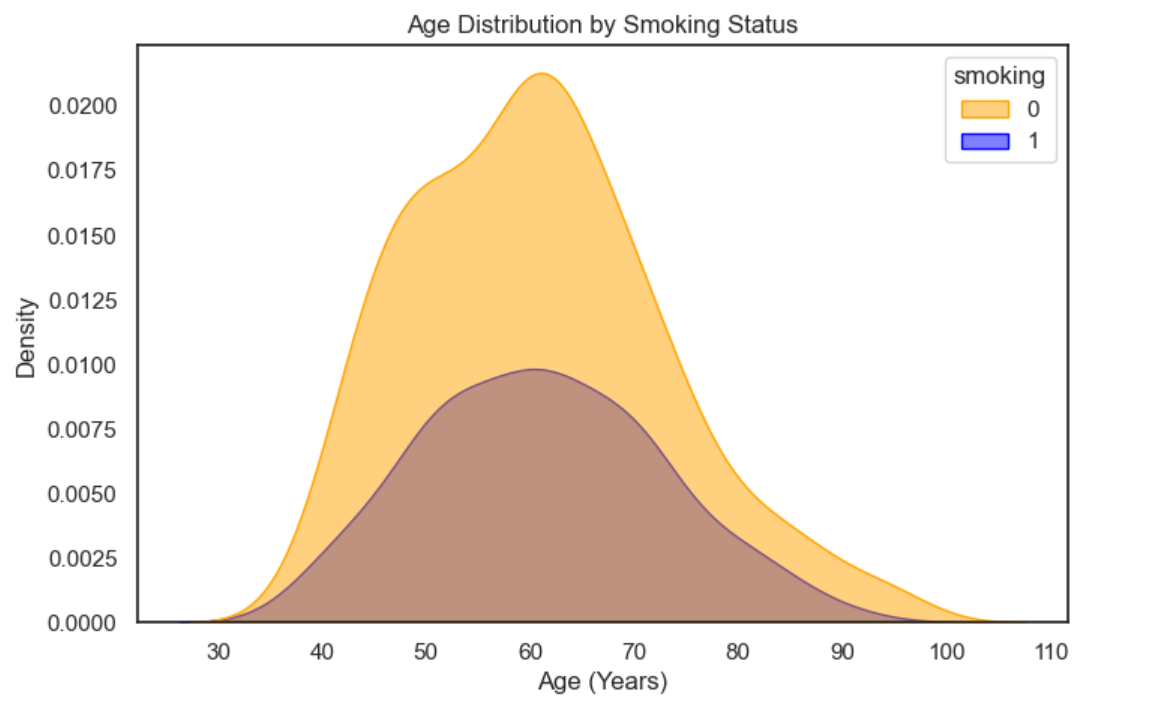
We create a KDE plot to show the **distribution of age** in the dataset, separated by smoking status (0 = non-smoker, 1 = smoker). This allows us to understand whether smokers tend to fall into specific age brackets compared to non-smokers.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The **age distribution curves** show that **non-smokers** are more **evenly distributed** across a wider age range.
* **Smokers** are slightly more **concentrated in the younger to mid-age range**, with a lower density among older age groups.
* This may suggest that **smoking patients** in this dataset are **somewhat younger**, though age ranges overlap between the groups.

**v. Visualization**



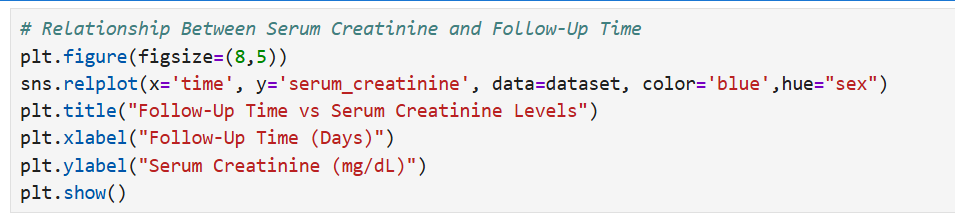
**Objective 21: Relationship Between Serum Creatinine and Follow-Up Time**

**i. Introduction**

**Serum creatinine** is a key marker used to assess kidney function. In patients with heart failure, impaired kidney function can indicate worsening prognosis. The **follow-up time** in days represents how long each patient was monitored. Analyzing the relationship between **serum creatinine and follow-up duration** helps to understand whether patients with abnormal creatinine levels tend to have shorter or longer follow-up spans, which may reflect survival time or disease severity.

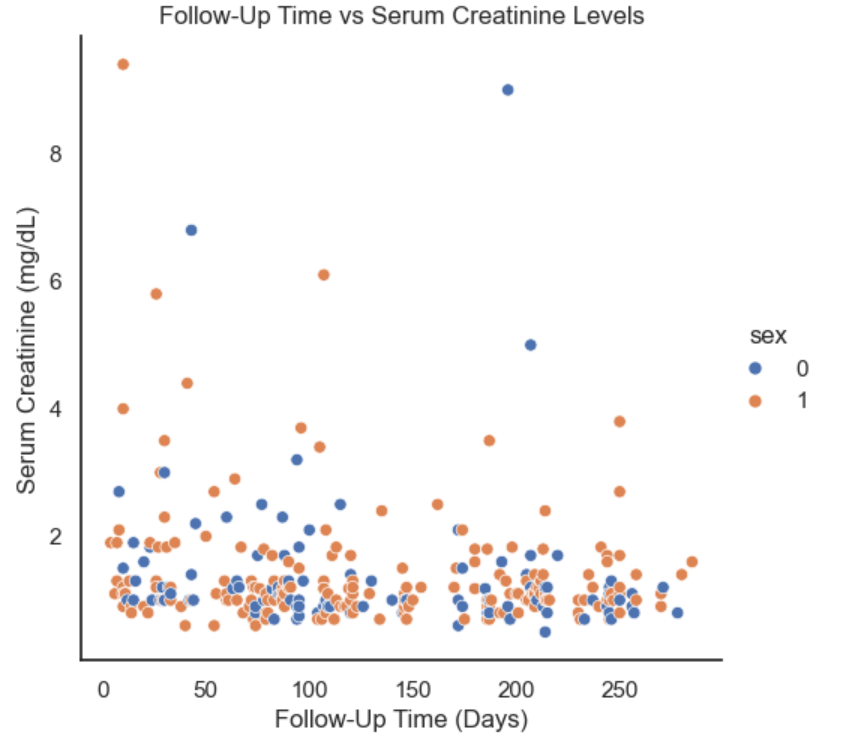
**ii. General Description**

This analysis uses a **scatter plot** to evaluate whether any trend exists between serum\_creatinine levels and time. We aim to explore whether elevated creatinine levels are associated with shorter follow-up periods, which might suggest more critical conditions or early mortality.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The scatter plot reveals that patients with **higher serum creatinine** levels often have **shorter follow-up times**, indicating a potential link to earlier mortality or critical health deterioration.
* Patients with **longer follow-up durations** generally show **lower creatinine levels**, suggesting better kidney function and possibly a more stable health condition.
* There is **no strong linear correlation**, but the plot reveals meaningful patterns upon visual inspection.

**v. Visualization**

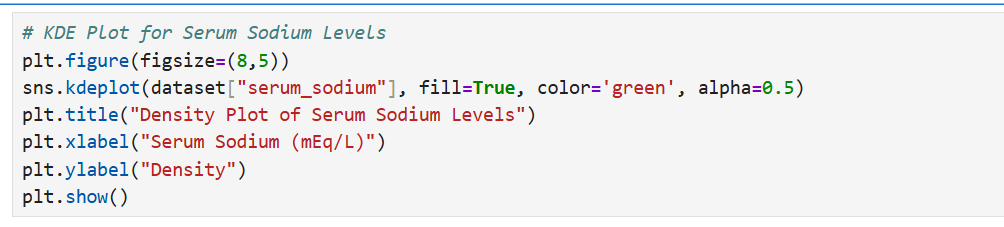
**Objective 22: KDE Plot for Serum Sodium Levels**

**i. Introduction**

**Serum sodium** levels are crucial for evaluating fluid balance and cardiovascular function. In heart failure patients, **abnormal sodium concentrations**, especially **hyponatremia** (low sodium), are associated with worse outcomes. To better understand the overall **distribution of serum sodium levels** in the dataset, we use a **Kernel Density Estimation (KDE) plot**. KDE offers a smooth, continuous curve that illustrates the probability density of sodium values in the population.

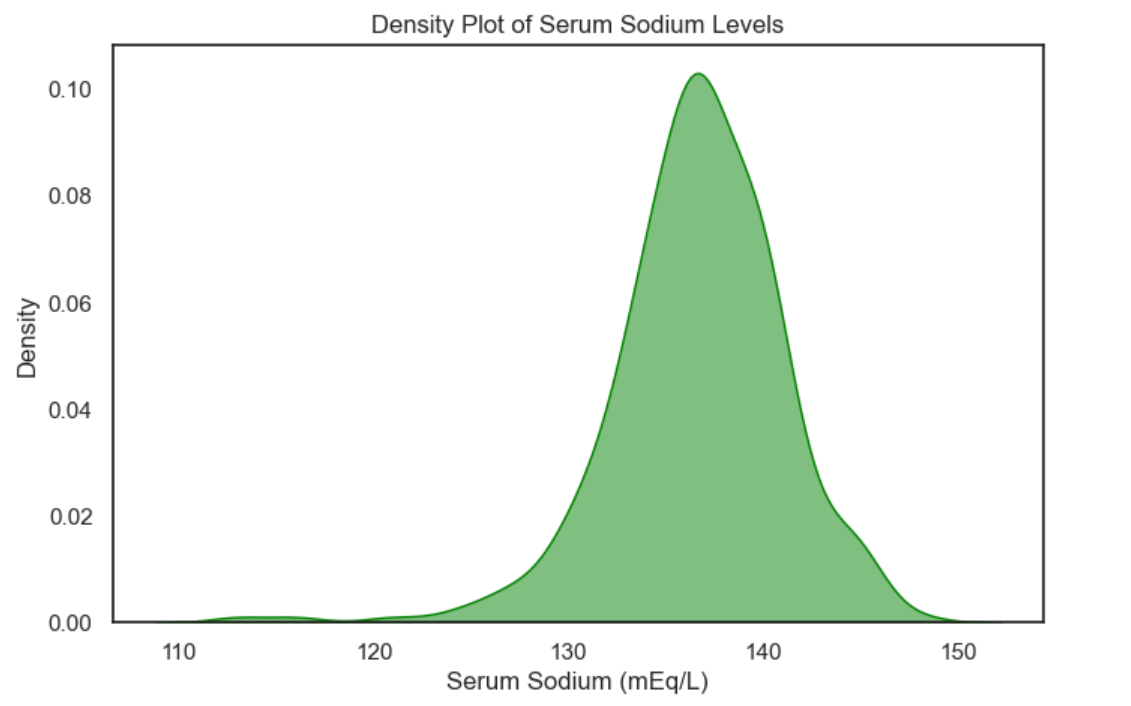
**ii. General Description**

This analysis visualizes the **distribution of serum\_sodium** for all patients using a KDE plot. The goal is to see whether sodium values cluster around normal ranges or if a significant portion of patients fall into low/high ranges indicative of risk.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The KDE curve shows a **normal-like distribution** with most serum sodium values concentrated between **130–140 mEq/L**, aligning with typical clinical ranges.
* A visible **left-tail** (lower sodium levels) suggests that **some patients may be experiencing hyponatremia**, a known risk factor in heart failure.
* There is no strong right-side tail, indicating **fewer high sodium cases** in this group.

**v. Visualization**

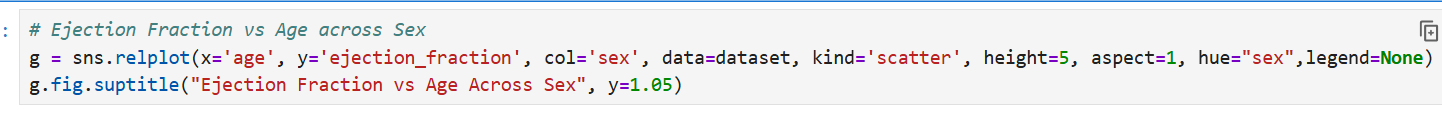
**Objective 23: Ejection Fraction vs Age Across Sex**

**i. Introduction**

**Ejection fraction (EF)** is a measure of how well the heart is pumping blood with each beat. A lower EF often indicates heart dysfunction or failure. Since age and sex are both critical factors in cardiovascular health, this analysis investigates how **ejection fraction varies with age** and whether there are **notable differences between male and female patients**.

**ii. General Description**

We explore the relationship between **age** and **ejection fraction** across the two **sexes** using a **scatter plot with column-wise separation** via relplot(). This helps in visualizing gender-specific trends and potential disparities in heart function.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* **Male and female patients** display **similar scatter patterns**, suggesting no drastic gender-based difference in ejection fraction across age.
* There is **considerable variation in EF across all age groups**, with no obvious upward or downward trend, indicating age may not linearly influence EF in this dataset.
* Outliers are present in both sexes, hinting at individual cases of very low or very high ejection fraction, regardless of age.

**v. Visualization**

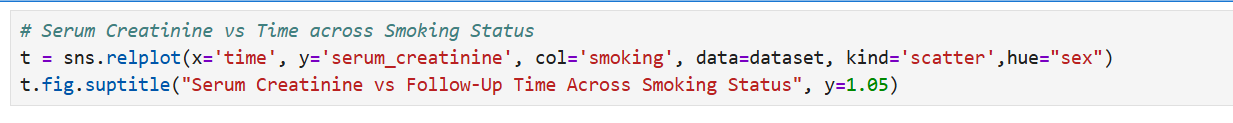
 **Objective 24: Serum Creatinine vs Follow-Up Time Across Smoking Status**

**i. Introduction**

**Serum creatinine** levels are an important indicator of kidney function, which can be closely tied to heart health in patients with heart failure. **Smoking**, being a known cardiovascular risk factor, may impact both heart and kidney function. This analysis aims to examine the **relationship between serum creatinine levels and the follow-up time**, and how this varies based on the patient’s **smoking status**.

**ii. General Description**

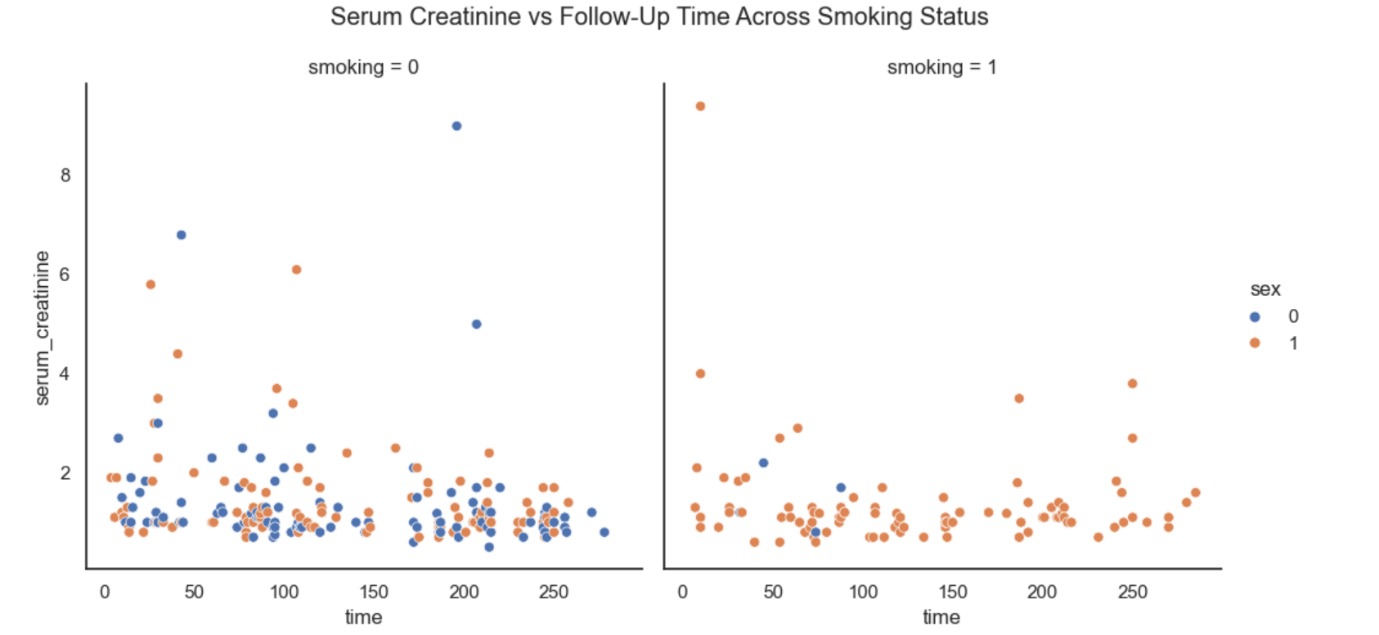
The objective is to compare how **serum creatinine changes across time** (follow-up days) in **smokers** vs **non-smokers** using a **facet-based scatter plot** (relplot() with col='smoking'). This helps in visually detecting differences in trends or distributions between the two groups.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* Both **smokers (1)** and **non-smokers (0)** show a wide spread of serum creatinine values across the follow-up period.
* No strong **trend** is visible between serum creatinine and time in either group.
* There are a few high serum creatinine values among both smokers and non-smokers, potentially indicating kidney dysfunction regardless of smoking.
* The **spread is slightly denser in non-smokers**, likely due to a higher number of non-smoking patients in the dataset.

**v. Visualization**

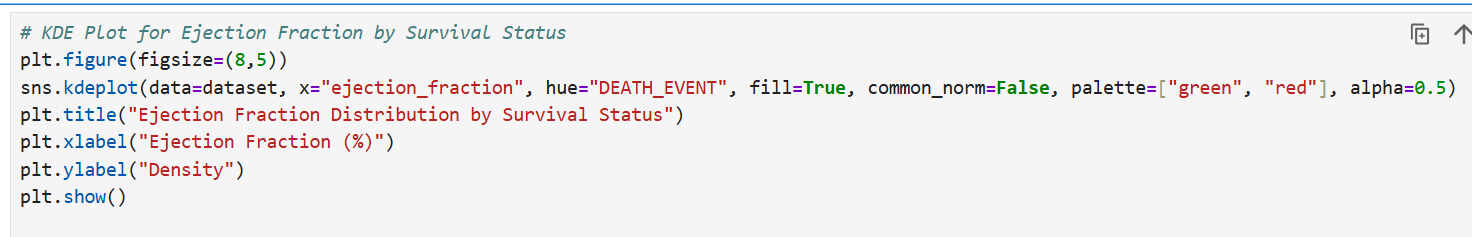
 **Objective 25: KDE Plot for Ejection Fraction by Survival Status**

**i. Introduction**

**Ejection fraction (EF)** is a critical clinical measurement representing the percentage of blood pumped out of the heart during each contraction. In patients with heart failure, a reduced EF is often associated with higher risk of death. This analysis aims to **compare the distribution of EF between patients who survived and those who experienced a death event**, using **Kernel Density Estimation (KDE)** plots to visualize the smooth distribution curves.

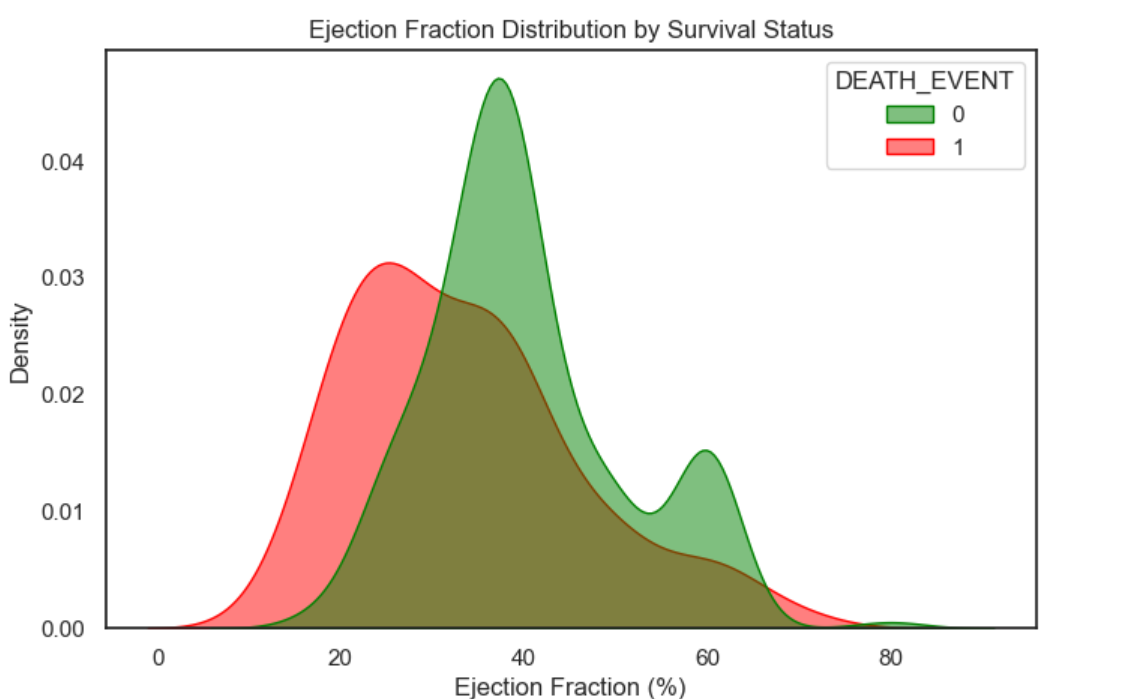
**ii. General Description**

We plot **KDE curves** of the **ejection fraction** for both groups — **survivors (DEATH\_EVENT = 0)** and **non-survivors (DEATH\_EVENT = 1)**. This allows for visual comparison of how EF values differ in distribution, offering insights into whether lower EF is associated with higher mortality.

**iii. Specific Requirements, Functions, and Formulas**

**iv. Analysis Results**

* The **curve for deceased patients** (DEATH\_EVENT = 1) is **more concentrated at lower ejection fraction values**, indicating that a **low EF** is a strong risk factor for death.
* Survivors (DEATH\_EVENT = 0) show a **wider and right-shifted distribution**, generally having **higher EF values**.
* The KDE plot confirms clinical expectations — **low ejection fraction is associated with higher mortality risk** in heart failure patients.

**v. Visualization**

**Objective 26: Outlier Detection in Serum Sodium**

**i. Introduction**

Serum sodium is an essential electrolyte in the blood that helps regulate fluid balance, nerve function, and muscle contractions. Abnormal levels may indicate underlying health issues such as kidney dysfunction or dehydration. Detecting **outliers** in serum sodium levels can help identify patients who might be at **extreme clinical risk** and require further medical attention.

**ii. General Description**

In this analysis, we use a **boxplot** to detect and visualize **outliers** in the serum\_sodium feature of the heart failure dataset. A boxplot is an effective way to identify values that lie significantly outside the typical range (interquartile range).

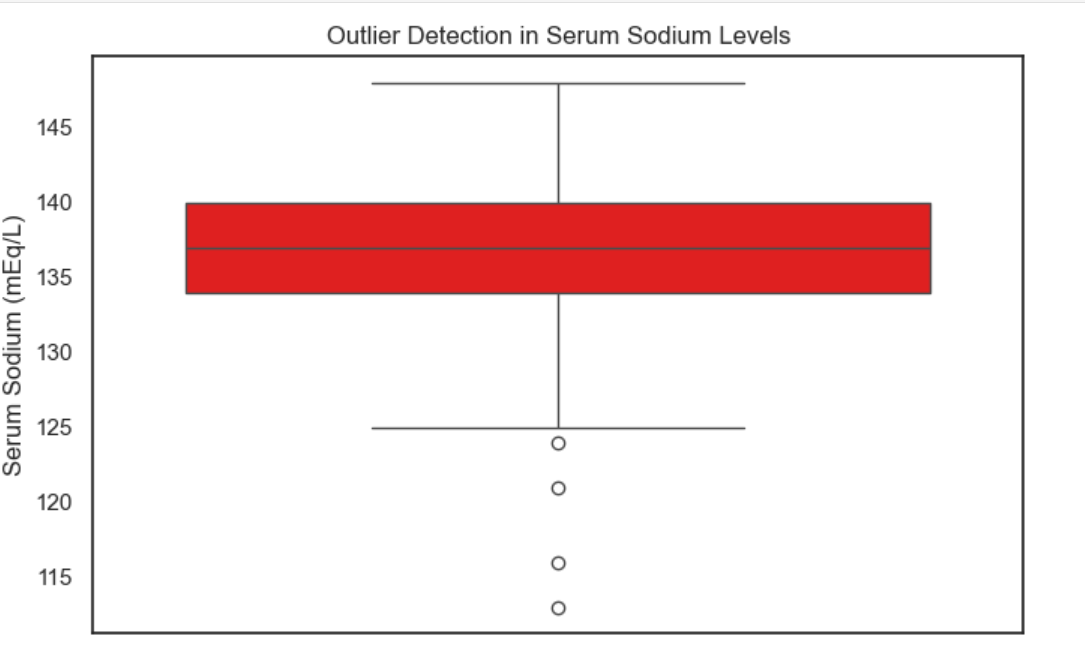
**iii. Specific Requirements, Functions, and Formulas**



**iv. Analysis Results**

* The **majority of serum sodium values** fall within the **normal range** (~130–145 mEq/L).
* A few **outliers** appear as **individual points above or below** the whiskers of the boxplot.
* These outliers could represent:
  + **Hyponatremia** (low sodium levels)
  + **Hypernatremia** (high sodium levels)
* Such cases might indicate **risk for complications** in heart failure and should be considered for further investigation or clinical alerting.

**v. Visualization**



**CONCLUSION**

The analysis of the Heart Failure Clinical Records Dataset has provided valuable insights into the various clinical and lifestyle factors that influence patient outcomes, particularly mortality. Through a combination of descriptive statistics, visual exploration, and predictive modeling techniques, we were able to identify key trends and risk indicators associated with heart failure.

Our findings highlight the significant impact of variables such as age, ejection fraction, serum creatinine, and serum sodium on the likelihood of death during the follow-up period. Lifestyle and health conditions such as diabetes, high blood pressure, and smoking habits were also found to play a crucial role in survival outcomes.

Using a variety of visualization tools including boxplots, scatter plots, KDE plots, pie charts, bar charts, and Kaplan-Meier survival curves, we were able to not only observe data distributions and correlations but also effectively communicate patterns and anomalies in the dataset. Outlier detection helped identify abnormal clinical values, while clustering and feature importance analysis added depth to our understanding of patient segmentation and predictive modeling.

This project underscores the power of Python-based data analysis using libraries like Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, offering both analytical clarity and visual storytelling. The insights derived can assist healthcare professionals in better risk stratification and individualized patient care strategies.

In conclusion, data-driven approaches such as this are instrumental in enhancing clinical decision-making, improving patient monitoring, and supporting early intervention strategies in managing chronic diseases like heart failure.

**FUTURE SCOPE**

This project has laid a strong foundation for understanding the clinical and lifestyle factors affecting heart failure patient outcomes. However, there are several opportunities to **extend and enhance this analysis** in the future:

**1. Integration of More Diverse Clinical Data**

* Incorporating **additional clinical parameters** such as cholesterol levels, BMI, medication history, and genetic data can lead to a more holistic analysis.
* Access to **longitudinal health records** can enable time-series modeling for better prognosis tracking.

**2. Advanced Predictive Modeling**

* Implementation of more **complex machine learning models** (e.g., XGBoost, SVM, Neural Networks) can enhance prediction accuracy for mortality and risk factors.
* **Hyperparameter tuning and cross-validation** techniques can be used to improve model generalization.

**3. Real-Time Monitoring Systems**

* The insights gained can be incorporated into **real-time clinical decision support systems** to alert healthcare providers about at-risk patients.
* Integration with **IoT-based health monitoring devices** can help track serum levels, heart rates, and other vitals in real-time.

**4. Personalized Risk Assessment Tools**

* A web-based or mobile **dashboard application** could be developed to assess patient-specific risk based on clinical inputs.
* Such tools can empower patients and doctors to make **data-informed treatment plans**.

**5. Survival Analysis Enhancements**

* More robust survival analysis using **Cox Proportional Hazards Model** and **Competing Risk Models** can further improve mortality prediction.
* **Stratified survival curves** based on combined conditions (e.g., diabetic smokers with hypertension) can provide deeper insights.

**6. Clinical Trials & Real-World Validation**

* Findings from the analysis can be tested against **real clinical trial data** for validation and fine-tuning of insights.
* Collaborations with healthcare institutions can provide **real-world data** to refine the models.

**REFERENCES**

* UCI Machine Learning Repository – Heart Failure Clinical Records Dataset
* Seaborn Library – seaborn.pydata.org
* Pandas Library – pandas.pydata.org
* Matplotlib Library – [matplotlib.org](https://matplotlib.org)