Jahangirnagar University
Department of Statistics and Data Science

Assignment

WM-ASDS02 : Sampling Methodology

Factors Affecting Heart Failure: Evidence from Clinical Data

Submitted by:

Zinia Sultana Joti

ID: 20231278

Section: B

Semester: Fall 2023

Submitted to:

Dr. Mohd. Muzibur Rahman

Professor

May 15, 2024

Factors Affecting Heart Failure: Evidence from Clinical Data.

Objective of the Study:

From the "Clinical Heart Failure" data set of 300 records take a random sample of 20 records separately. Carry out simple analysis. Based on this findings, make conclusion taking into consideration of the above title.

Sample Selection:

Sampling is a statistical technique where you select a subset (sample) from a larger population to gather information and make inferences about the entire population. It's a practical and cost-effective way to conduct research.

Sampling Frame: The "Clinical Heart Failure" data set has 300 records of patients' medical history(Diabetes, High Blood Pressure, Smoking history), physiological measurements like Ejection Fraction(heart pumping efficiency), Serum creatinine levels (kidney function) etc.

Taking a sample of 20 records using Simple Random Sampling where each member has an equal chance of being selected.

Serial No	Age	Anaemia	Creatinine Phosphokinase	Diabetes	Ejection Fraction	High Blood Pressure	Platelets	Serum Creatinine	Serum Sodium	Sex	Smoking	Time	DEATH EVENT
173	50	1	115	0	20	0	189000	0.8	139	1	0	146	0
287	45	0	582	1	55	0	543000	1	132	0	0	250	0
51	53	1	91	0	20	1	418000	1.4	139	0	0	43	1
146	52	0	132	0	30	0	218000	0.7	136	1	1	112	0
214	65	1	135	0	35	1	290000	0.8	134	1	0	194	0
197	65	0	167	0	30	0	259000	0.8	138	0	0	186	0
274	60	1	257	1	30	0	150000	1	137	1	1	245	0
62	55	0	109	0	35	0	254000	1.1	139	1	1	60	0
27	70	0	122	1	45	1	284000	1.3	136	1	1	26	1
201	45	0	308	1	60	1	377000	1	136	1	0	186	0
112	50	0	369	1	25	0	252000	1.6	136	1	0	90	0
119	86	0	582	0	38	0	263358.03	1.83	134	0	0	95	1
11	62	0	231	0	25	1	253000	0.9	140	1	1	10	1
244	54	0	582	1	38	0	264000	1.8	134	1	0	213	0
110	85	0	129	0	60	0	306000	1.2	132	1	1	90	1
295	55	0	1820	0	38	0	270000	1.2	139	0	0	271	0
73	65	0	224	1	50	0	149000	1.3	137	1	1	72	0
105	72	1	328	0	30	1	621000	1.7	138	0	1	88	1
265	50	1	298	0	35	0	362000	0.9	140	1	1	240	0
107	45	1	1876	1	35	0	226000	0.9	138	1	0	88	0

Exploratory Data Analysis :

Check for Null/Missing values:
Using command: sample.isnull().sum(), there are 0 missing values in the sample data set.

```
age                         0
anaemia                     0
creatinine_phosphokinase    0
diabetes                    0
ejection_fraction           0
high_blood_pressure         0
platelets                   0
serum_creatinine            0
serum_sodium                0
sex                         0
smoking                     0
time                        0
DEATH_EVENT                 0
dtype: int64
```

Descriptive statistics:

Variable Name	Description	Level of Measurement	Appropriate Measures
Age	Patient's age in years	Ratio	Mean
Anemia	Indicator variable for anaemia (0=no, 1=yes)	Nominal	Mode
Creatinine Phosphokinase	Level of a blood enzyme (CPK)	Ratio	Mean
Diabetes	Indicator variable for diabetes (0=no, 1=yes)	Nominal	Mode
Ejection Fraction	Heart pumping efficiency (percentage)	Ratio	Mean
High Blood Pressure	Indicator variable for high blood pressure (coding may vary)	Nominal	Mode
Platelets	Blood platelet count	Ratio	Mean
Serum Creatinine	Level of creatinine in the blood (marker of kidney function)	Ratio	Mean
Serum Sodium	Level of sodium in the blood	Ratio	Mean
Sex	Patient's sex (coding may vary)	Nominal	Mode
Smoking	Indicator variable for smoking status (0=non-smoker, 1=smoker)	Nominal	Mode
Time	Time variable (might be time of data collection or a time interval)	Ratio	Mean
Death Event	Indicator variable for death event (0=no death, 1=death)	Nominal	Mode

Measure of central tendency, spread and potential skewness of the data:

	Age	Anaemia	Creatinine Phosphokinase	Diabetes	Ejection Fraction	High Blood Pressure	Platelets	Serum Creatinine	Serum Sodium	Sex	Smoking	Time	DEATH EVENT
count	20	20	20	20	20	20	20	20	20	20	20	20	20
mean	59.2	N/A	422.85	N/A	36.7	N/A	297417.90	1.16	136.7	N/A	N/A	135.25	N/A
std	12.10	0.48	512.85	0.50	11.91	0.47	118702.70	0.35	2.47	0.47	0.51	81.40	0.47
min	45	0	91	0	20	0	149000	0.7	132	0	0	10	0
25%	50	0	131.25	0	30	0	245500	0.9	135.5	0	0	84	0
50%	55	0	244	0	35	0	263679.01	1.05	137	1	0	103.5	0
75%	65	1	422.25	1	39.75	1	320000	1.32	139	1	1	198.75	1
max	86	1	1876	1	60	1	621000	1.83	140	1	1	271	1

Highlights:

- STD for Platelets is very high. Creatinine Phosphokinase is also high. For standard deviation, higher value represents widely spread data.
- Age, Time, Platelets data are slightly skewed towards right.
- Creatinine Phosphokinase is skewed to the right.

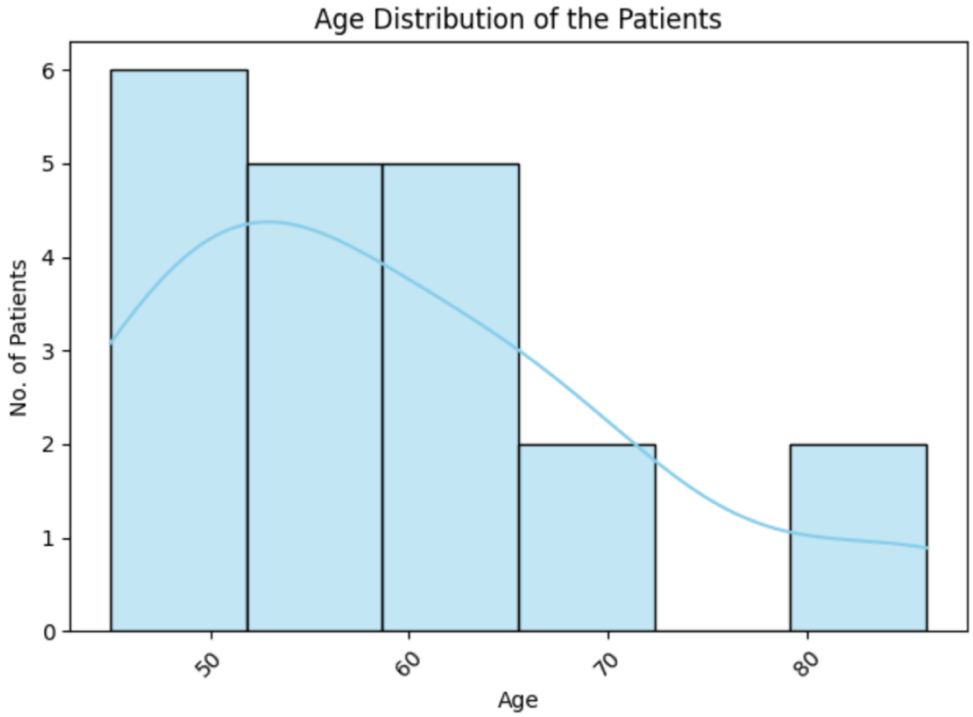
Visualization:

Here are some visualization to explore relationships betweem features and identify the relevence in the data that might influence the outcome DEATH_EVENT. Common factors to include: Age, Diabetes, High Blood Pressure, Sex, Smoking History, Ejection Fraction and Serum Creatinine levels.

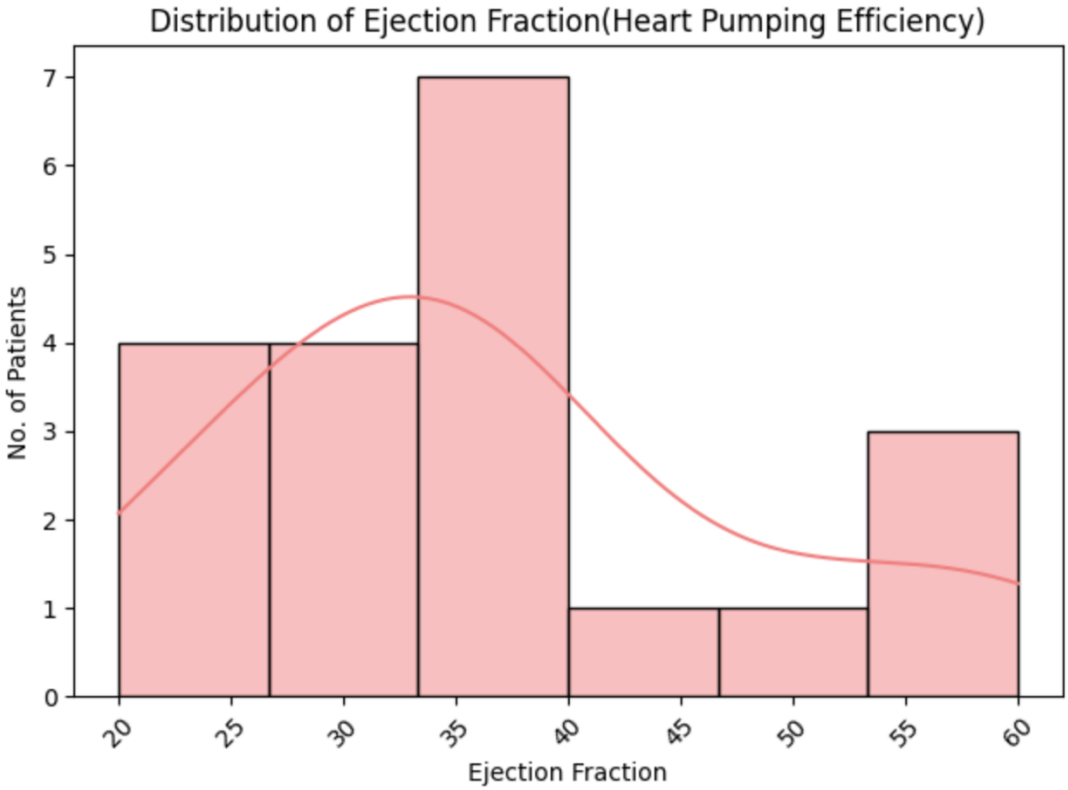
Univariat Analysis:

Histogram distributions for different variables(numerical) related Heart Failure:

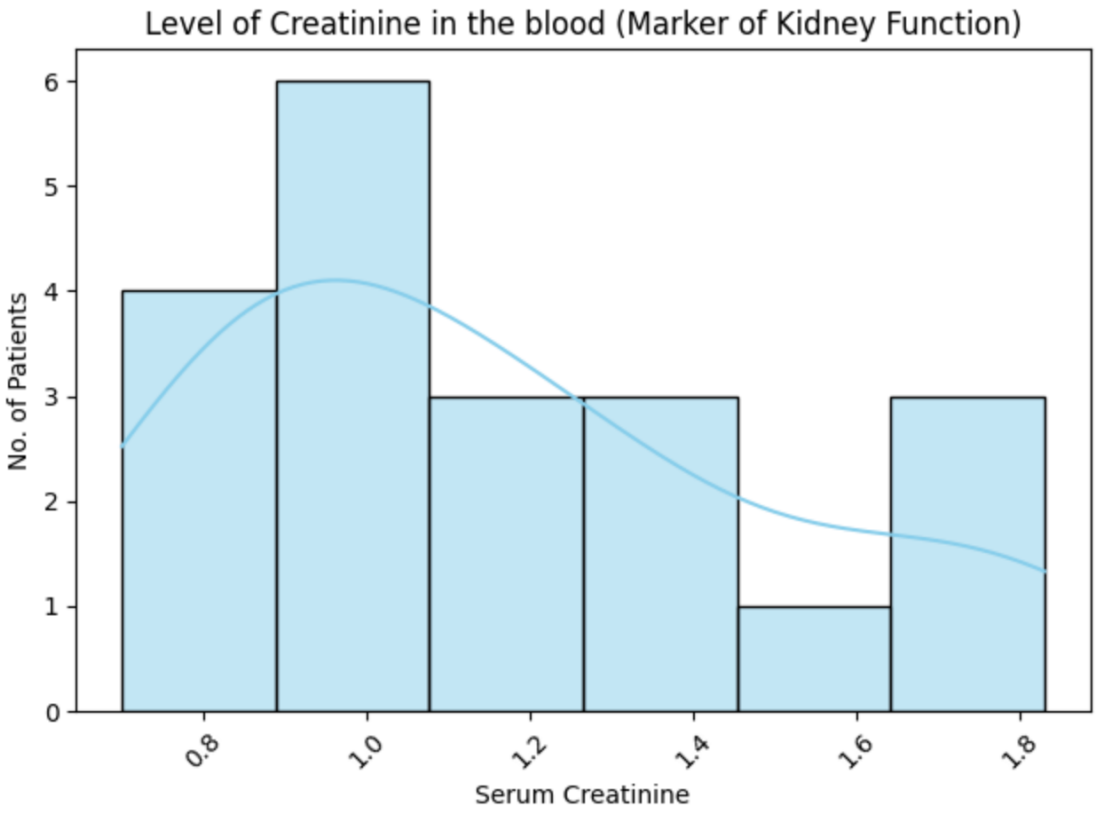
1. Age Distribution:



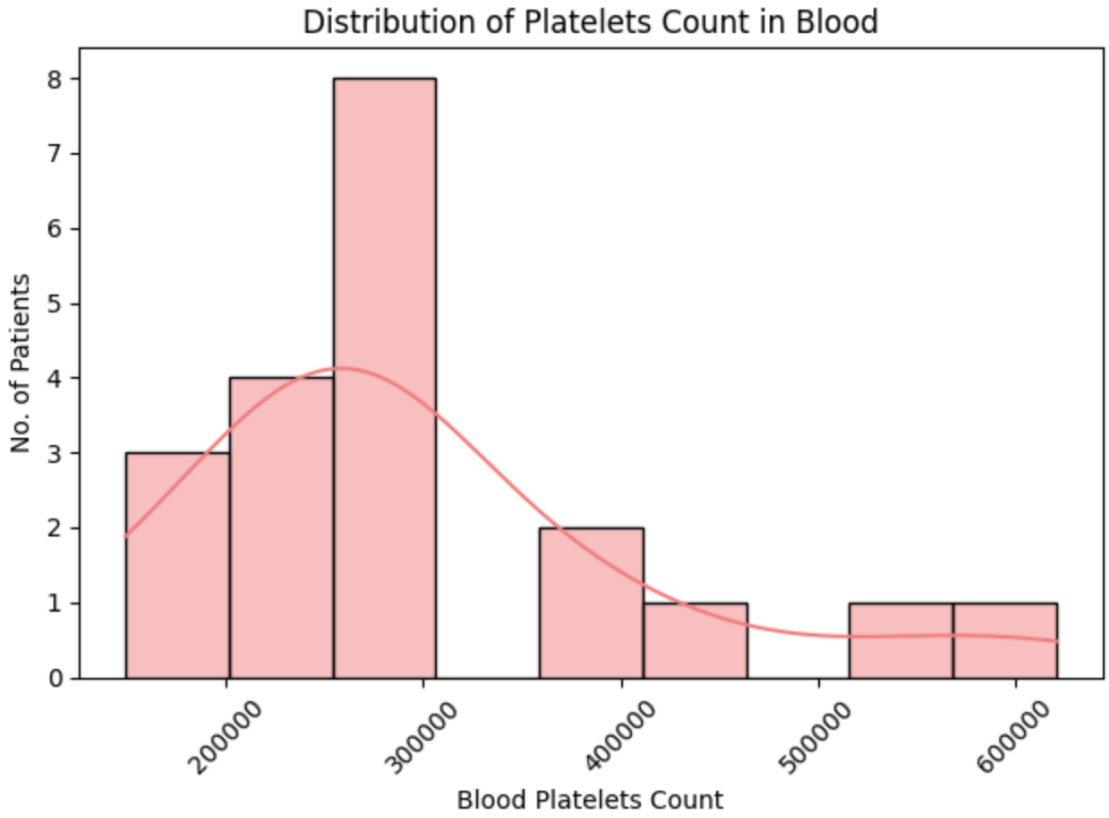
2. Ejection Fraction



3. Serum Creatinine Level

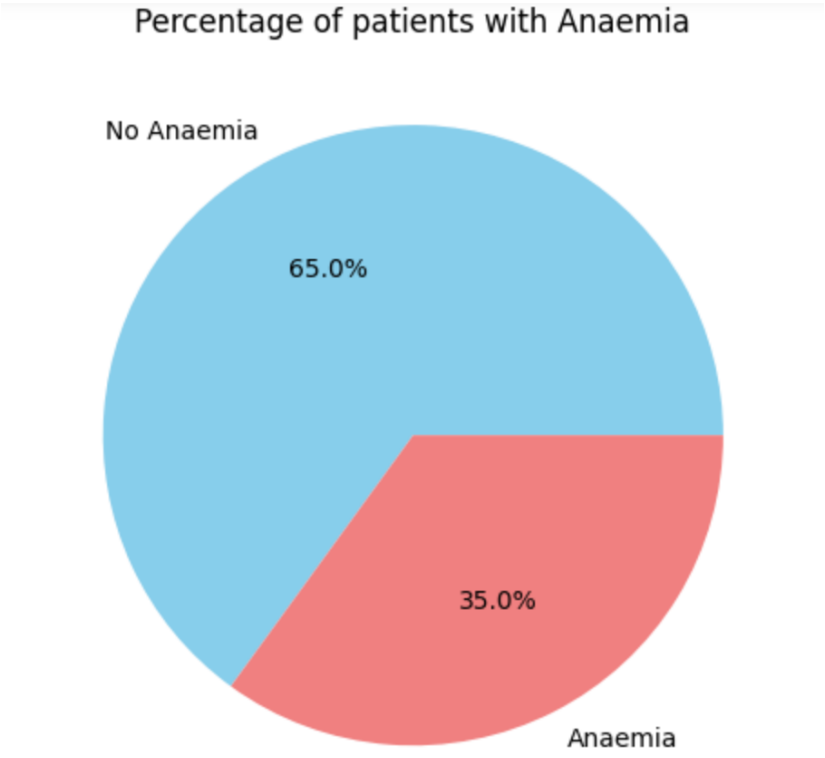


4. Platelets

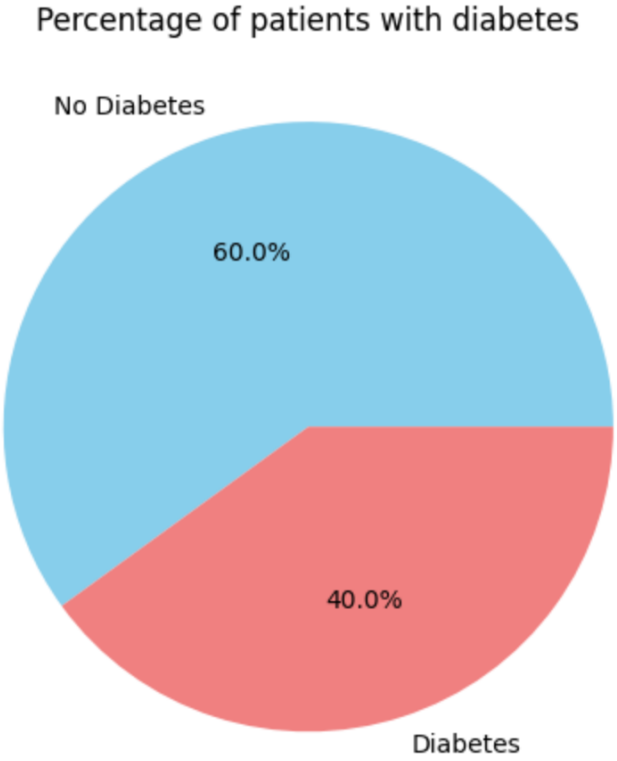


Pie charts for different variables(categorical) related Heart Failure

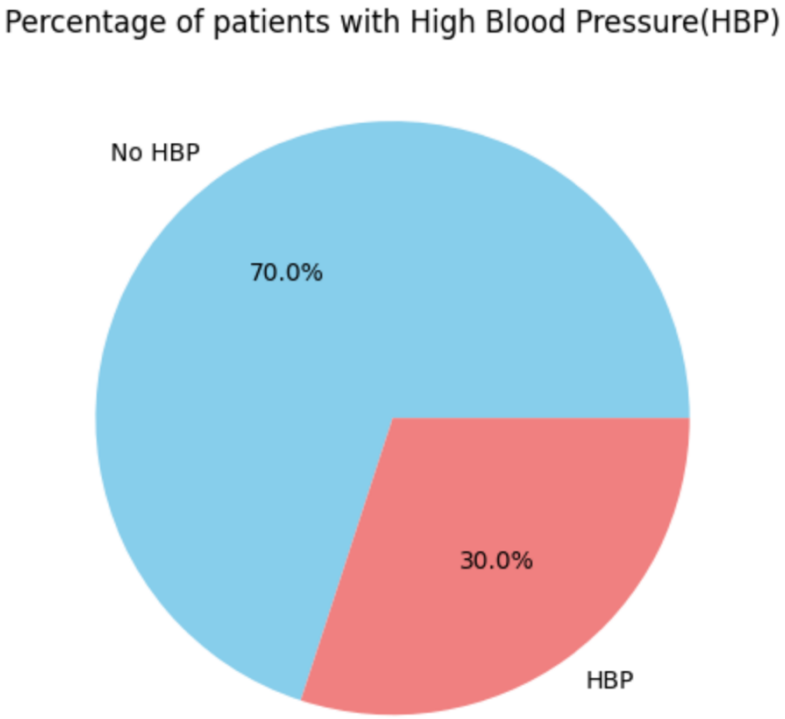
1. Anaemia



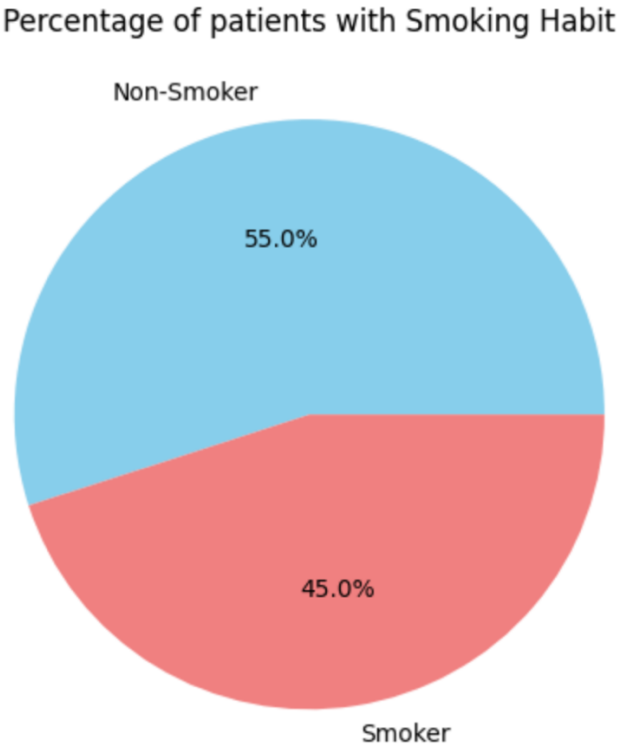
2. Diabetes



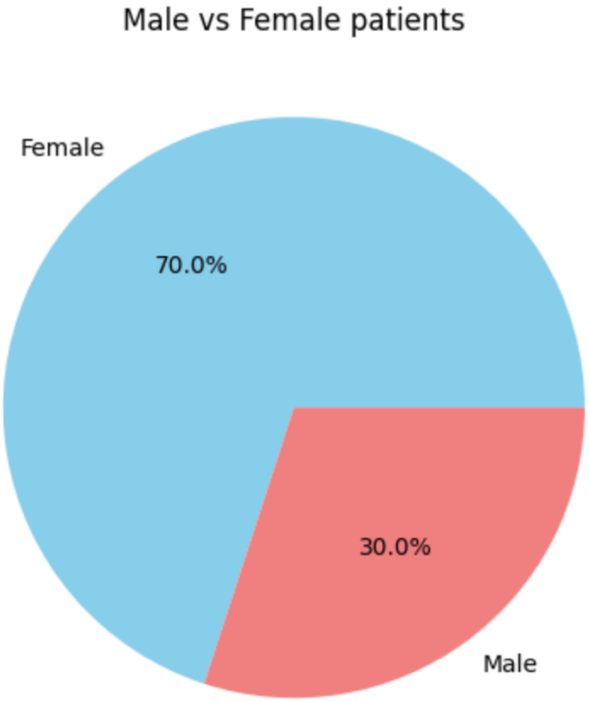
3. High Blood Pressure



4. Smoking Habit

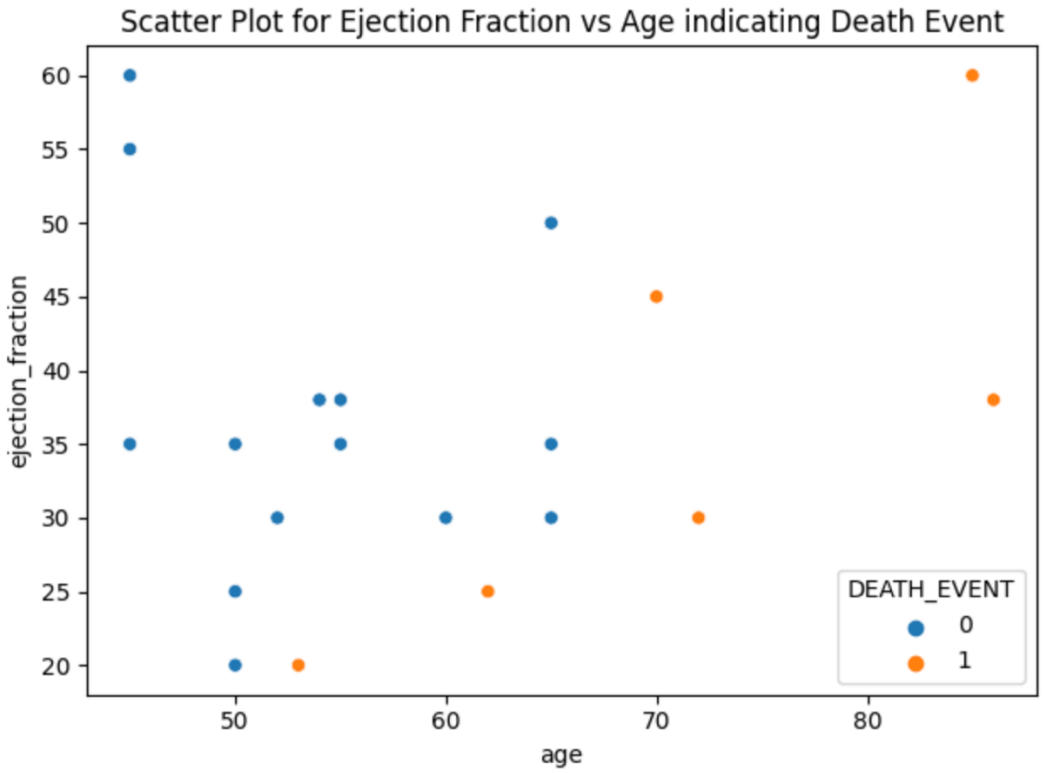


5. Gender

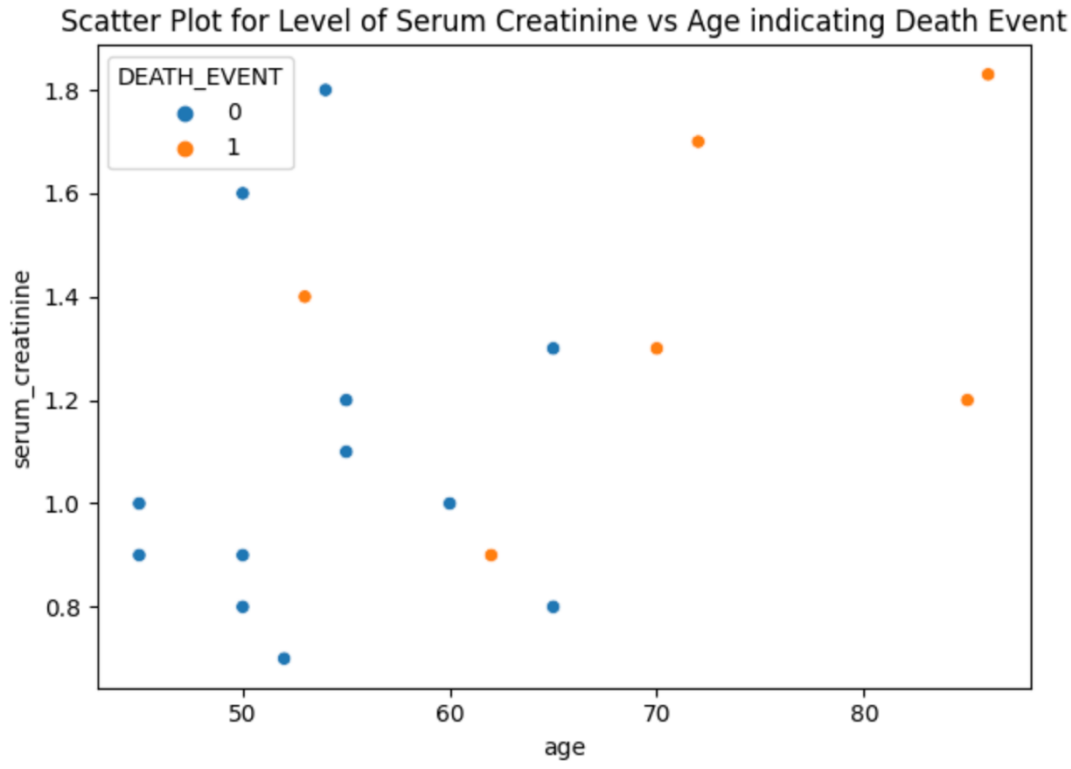


Bivariate Analysis:

1. Scatter Plot for Ejection Fraction vs Age indicating Death Event



2. Scatter Plot for Serum Creatinine vs Age indicating Death Event



Discussion :

Lower serum creatinine levels, mostly below 1.4 serum level. Additionally, the graph suggests that individuals with no Death Event are younger and have generally lower serum creatinine levels, which is typically associated with better kidney function. On the other hand, the red points are scattered at older ages with higher serum creatinine levels indicating compromised kidney functions for older people

There might be a positive correlation between ejection fraction and no death event. In other words, patients with higher ejection fractions (better heart function) seem to be more likely to survive in the records. Lower ejection fractions could be a potential risk factor for death events in heart failure patients. However, there are some red points (death event) have higher ejection fractions, suggesting other factors might also be at play. Age doesn't seem to have a strong visual relationship with the death event outcome in this scatter plot. There are patients who experienced death events (red) across the entire age range.

Conclusion :

This report is based on limited information (20 samples from 300 records of patients). A larger sample size and more details about the data would provide a more robust analysis. Serum creatinine, Ejection Fraction are just few indicators of kidney function and heart pumping efficiency. Other factors could influence the relationship between age, serum creatinine, ejection fraction and death events.

Association doesn't imply causation. While higher creatinine might be associated with heart failure, in addition to lower ejection fraction were also identified as potential risk for heart failure, but it might not directly cause them. Further analysis with statistical tests is needed to confirm any observed trends.