Heart Failure Prediction Analysis

Heart failure is a condition in which the heart can't pump enough blood to meet the body's needs. In this EDA Project, I will analyse the Heart Failure Prediction dataset through tables and charts using numpy, pandas, matplotlib and seaborn.

NAME: PENDEM SANJAY

COLLEGE: INSTITUTE OF TECHNOLOGY, GURU GHASIDAS VISHWAVIDYALAYA

Importing the libraries:

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

Loading the Dataset:

In [2]:

```
hf_df= pd.read_csv('heart_failure_clinical_records_dataset.csv')
hf_df
```

Out[2]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_so
0	75.0	absent	582	absent	20	present	265000.00	1.9	
1	55.0	absent	7861	absent	38	absent	263358.03	1.1	
2	65.0	absent	146	absent	20	absent	162000.00	1.3	
3	50.0	present	111	absent	20	absent	210000.00	1.9	
4	65.0	present	160	present	20	absent	327000.00	2.7	
...	...	...	...	...	...	...	...	...	
294	62.0	absent	61	present	38	present	155000.00	1.1	
295	55.0	absent	1820	absent	38	absent	270000.00	1.2	
296	45.0	absent	2060	present	60	absent	742000.00	0.8	
297	45.0	absent	2413	absent	38	absent	140000.00	1.4	
298	50.0	absent	196	absent	45	absent	395000.00	1.6	

299 rows × 13 columns

In [3]:

```
hf_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 13 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   age                       299 non-null    float64
 1   anaemia                   299 non-null    object 
 2   creatinine_phosphokinase  299 non-null    int64  
 3   diabetes                  299 non-null    object 
 4   ejection_fraction         299 non-null    int64  
 5   high_blood_pressure       299 non-null    object 
 6   platelets                 299 non-null    float64
 7   serum_creatinine          299 non-null    float64
 8   serum_sodium              299 non-null    int64  
 9   sex                       299 non-null    object 
 10  smoking                   299 non-null    bool   
 11  time                      299 non-null    int64  
 12  DEATH_EVENT               299 non-null    object 
dtypes: bool(1), float64(3), int64(4), object(5)
memory usage: 28.4+ KB
```

There are 299 rows and 13 columns in the given dataset

Data Preparation and Cleaning:

In [4]:

```
hf_df.nunique()
```

Out[4]:

```
age                         47
anaemia                      2
creatinine_phosphokinase    208
diabetes                     2
ejection_fraction           17
high_blood_pressure          2
platelets                  176
serum_creatinine            40
serum_sodium                27
sex                          2
smoking                      2
time                       148
DEATH_EVENT                  2
dtype: int64
```

Checking for Null Values:

In [5]:

```
hf_df.isnull().values.any()
```

Out[5]:

```
False
```

In [6]:

```
hf_df.isnull().sum()
```

Out[6]:

```
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
```

As we can see, there are no missing values/NANs in the dataset

Exploratory Analysis and Visualization:

In [7]:

```
hf_df.head()
```

Out[7]:

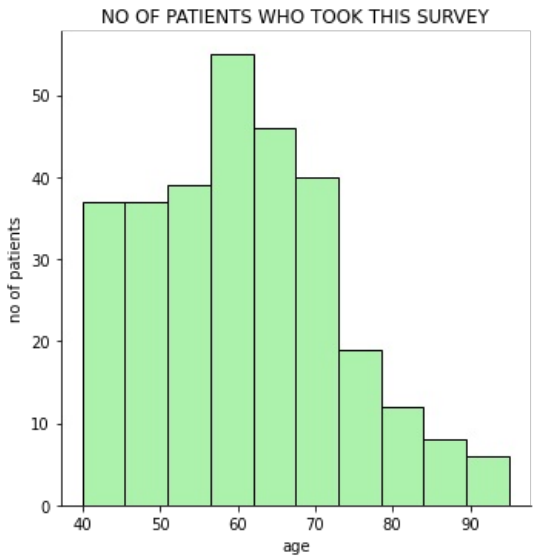
	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodiu
0	75.0	absent	582	absent	20	present	265000.00	1.9	1
1	55.0	absent	7861	absent	38	absent	263358.03	1.1	1
2	65.0	absent	146	absent	20	absent	162000.00	1.3	1
3	50.0	present	111	absent	20	absent	210000.00	1.9	1
4	65.0	present	160	present	20	absent	327000.00	2.7	1

NO OF PARTICIPANTS:

In [139]:

```
plt.figure(figsize=(12,6));
sns.displot(hf_df.age, color='lightgreen');
plt.title("NO OF PATIENTS WHO TOOK THIS SURVEY");
plt.ylabel("no of patients");
```

```
<Figure size 864x432 with 0 Axes>
```

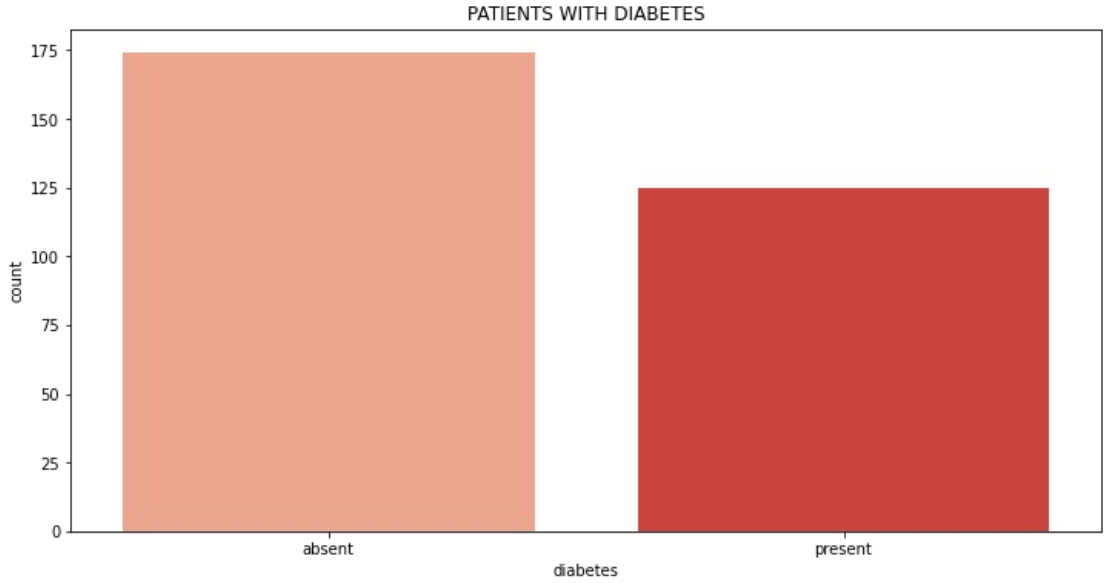


PATIENTS WITH DIABETES:

In [9]:

```
plt.figure(figsize=(12,6))

plt.title('PATIENTS WITH DIABETES')
sns.countplot(x=hf_df.diabetes, palette="Reds");
```



175 patients have diabetes and 125 of them don't have

PATIENTS WITH COMORBIDITIES:

In [12]:

```
comorbidities=hf_df[['age','anaemia','diabetes','high_blood_pressure']]
comorbidities
```

Out[12]:

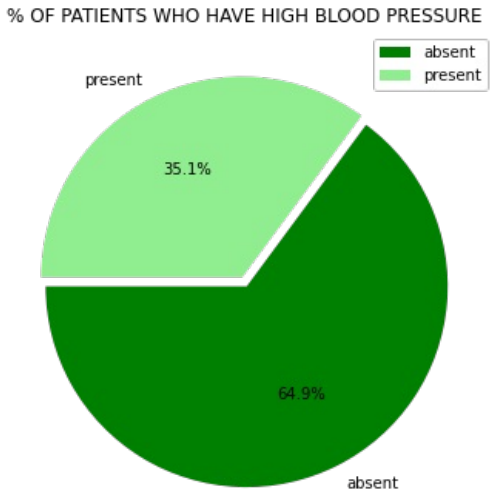
	age	anaemia	diabetes	high_blood_pressure
0	75.0	absent	absent	present
1	55.0	absent	absent	absent
2	65.0	absent	absent	absent
3	50.0	present	absent	absent
4	65.0	present	present	absent
...	...	...	...	...
294	62.0	absent	present	present
295	55.0	absent	absent	absent
296	45.0	absent	present	absent
297	45.0	absent	absent	absent
298	50.0	absent	absent	absent

299 rows × 4 columns

PATIENTS WITH HIGH BLOOD PRESSURE:

In [10]:

```
plt.figure(figsize=(12,6))
plt.title("% OF PATIENTS WHO HAVE HIGH BLOOD PRESSURE")
plt.pie(hf_df.high_blood_pressure       .value_counts(), explode=(0.025,0.025), labels=hf_df.high_blood_pressure
.value_counts().index, colors=['green','lightgreen'],autopct='%1.1f%%', startangle=180);
plt.legend();
```



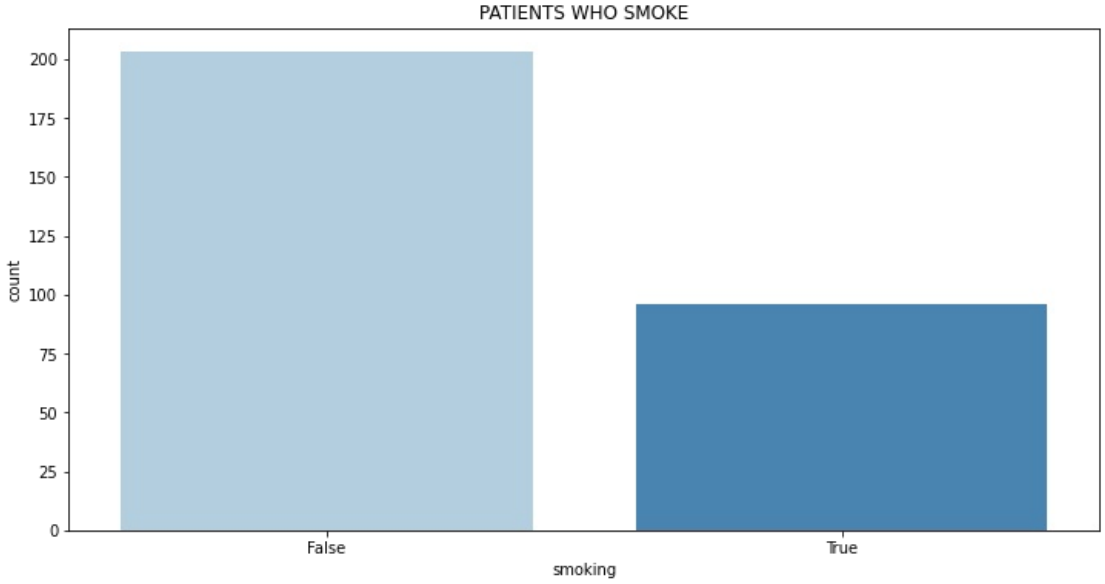
64.9% of the patients have High Blood Pressure

PATIENTS WHO SMOKE:

In [11]:

```
plt.figure(figsize=(12,6))

plt.title("PATIENTS WHO SMOKE")
sns.countplot(x=hf_df.smoking, palette="Blues");
```

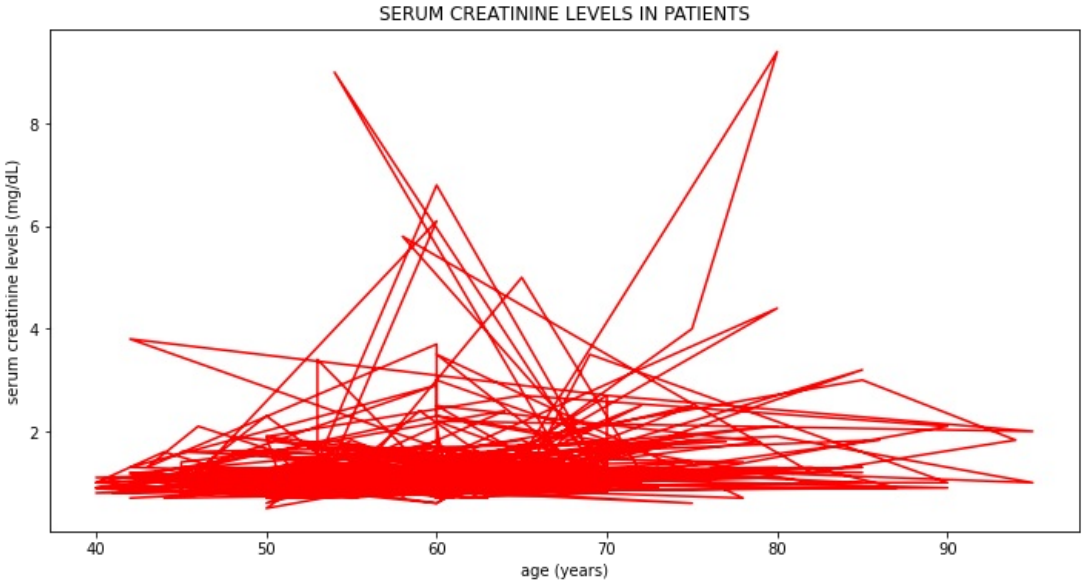


200 patients smoke and the rest 100 don't smoke

SERUM CREATININE LEVELS IN BLOOD:

In [13]:

```
plt.figure(figsize=(12,6))
plt.title("SERUM CREATININE LEVELS IN PATIENTS")
plt.xlabel("age (years)")
plt.ylabel("serum creatinine levels (mg/dL)")
plt.plot(hf_df.age,hf_df.serum_creatinine,"-r");
```

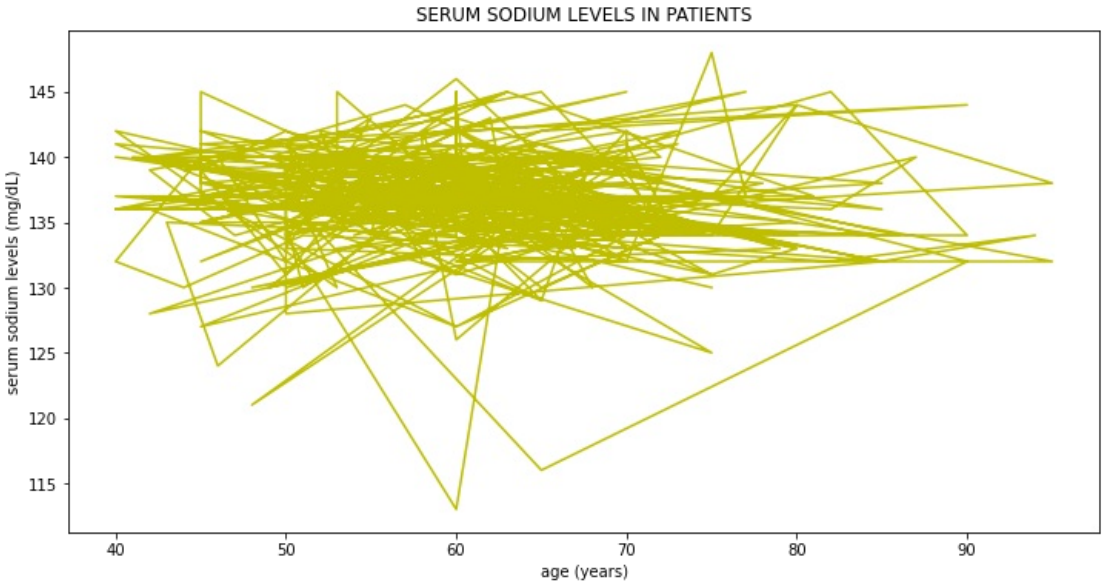


From the above chart, we can conclude that most of the patients have serum creatinine levels below 2 mg/dL

SERUM SODIUM LEVELS IN BLOOD:

In [14]:

```
plt.figure(figsize=(12,6))
plt.title("SERUM SODIUM LEVELS IN PATIENTS")
plt.xlabel("age (years)")
plt.ylabel("serum sodium levels (mg/dL)")
plt.plot(hf_df.age,hf_df.serum_sodium,"-y");
```

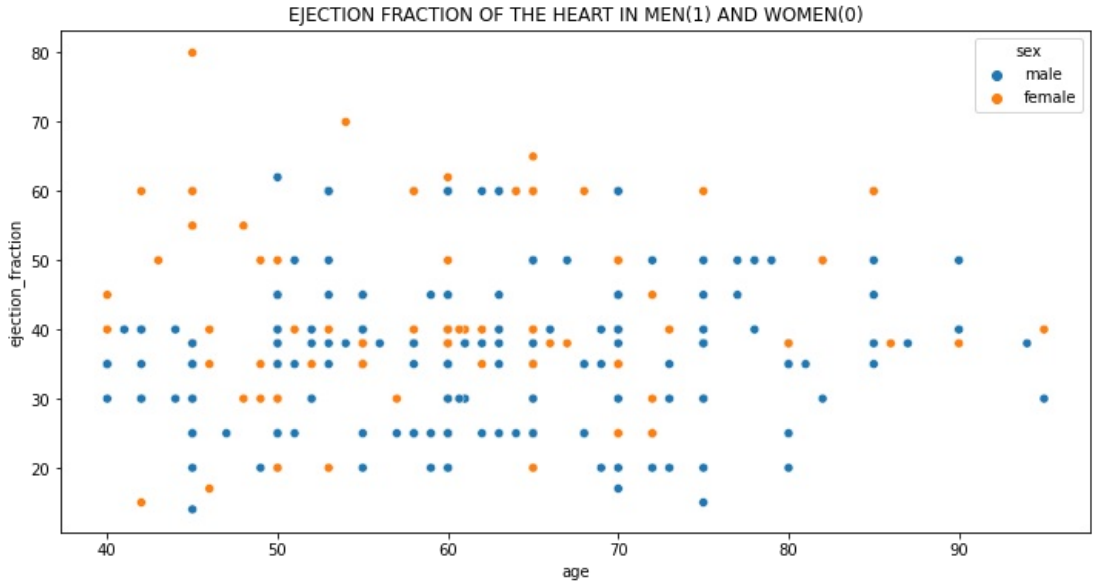


Most of the patients have sodium levels between 130-145 dg/mL

EJECTION FRACTION:

In [15]:

```
plt.figure(figsize=(12,6))
sns.scatterplot(x=hf_df.age, y=hf_df.ejection_fraction,hue=hf_df.sex,legend=True)
plt.title("EJECTION FRACTION OF THE HEART IN MEN(1) AND WOMEN(0)");
```



QUESTIONS AND ANSWERS:

Q1: How many patients have all three comorbidities?

In [70]:

```
a=hf_df.loc[(hf_df['diabetes'] == 'present') & (hf_df['anaemia'] == 'present') & (hf_df['high_blood_pressure'] ==
'present')]
counter=a.age.count()
print(counter)
```

Q2: What are the ages of youngest and oldest people who took part in this survey?

In [61]:

```
oldest=max(hf_df.age)
youngest=min(hf_df.age)
print("Oldest:", oldest)
print("Youngest:", youngest)
```

```
Oldest: 95.0
Youngest: 40.0
```

Q3: What conclusion can be drawn when we compare the smoking data with the platelets count?

In [97]:

```
hf_df[['smoking','platelets']]
```

Out[97]:

	smoking	platelets
0	False	265000.00
1	False	263358.03
2	True	162000.00
3	False	210000.00
4	False	327000.00
...	...	...
294	True	155000.00
295	False	270000.00
296	False	742000.00
297	True	140000.00
298	True	395000.00

299 rows × 2 columns

Those who smoke have less no of platelets count in their blood

Q4: What is the average levels of creatinine phosphokinase for those people in age group 40-50?

In [126]:

```
agecount=(hf_df.age>=40) &(hf_df.age<50)
average=hf_df[agecount].age.count()
average
sumofcreatinine=hf_df[agecount].creatinine_phosphokinase.sum()

average_levels=sumofcreatinine/average
print("Average:",average_levels )
```

```
Average: 802.1489361702128
```

Q5: Does the person with the highest platelet count have any comorbidities?

In [147]:

```
maximum= max(hf_df.platelets)
hf_df.loc[hf_df['platelets'] == maximum]
```

Out[147]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodi
109	45.0	absent	292	present	35	absent	850000.0	1.3	

yes, the patient has diabetes