In [3]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [4]:
```
hf_raw_df = pd.read_csv("heart_failure_clinical_records_dataset.csv")
hf_raw_df.head()
```

Out[4]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	p
0	75.0	0	582	0	20	1	26
1	55.0	0	7861	0	38	0	26
2	65.0	0	146	0	20	0	16
3	50.0	1	111	0	20	0	21
4	65.0	1	160	1	20	0	32

In [6]:
```
heart_failure_df = hf_raw_df.copy()
heart_failure_df.head()
```

Out[6]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	p
0	75.0	0	582	0	20	1	26
1	55.0	0	7861	0	38	0	26
2	65.0	0	146	0	20	0	16
3	50.0	1	111	0	20	0	21
4	65.0	1	160	1	20	0	32

In [7]:
```
heart_failure_df.shape
```

Out[7]: (299, 13)

In [9]:
```
heart_failure_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    int64  
 2   creatinine_phosphokinase  299 non-null    int64  
 3   diabetes                  299 non-null    int64  
 4   ejection_fraction         299 non-null    int64  
 5   high_blood_pressure       299 non-null    int64  
 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    int64  
 10  smoking                   299 non-null    int64  
 11  time                      299 non-null    int64  
 12  DEATH_EVENT               299 non-null    int64  
dtypes: float64(3), int64(10)
memory usage: 30.5 KB
```

In [11]:
```
heart_failure_df.drop_duplicates().any()
```

Out[11]:
```
age                         True
anaemia                     True
creatinine_phosphokinase    True
diabetes                    True
ejection_fraction           True
high_blood_pressure         True
platelets                   True
serum_creatinine            True
serum_sodium                True
sex                         True
smoking                     True
time                        True
DEATH_EVENT                 True
dtype: bool
```

In [16]:
```
## Renaming the columns
heart_failure_df.rename(columns={"DEATH_EVENT": "patient_dead"},inplace=True)
```

In [17]:
```
heart_failure_df.head(1)
```

Out[17]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	pl
0	75.0	0	582	0	20	1	26

In [18]:
```
heart_failure_df.drop(['time','creatinine_phosphokinase'],axis=1,inplace=True)
```

In [20]:
```
heart_failure_df.shape
```

Out[20]: (299, 11)

In [21]:
```
heart_failure_df.head()
```

Out[21]:

	age	anaemia	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine
0	75.0	0	0	20	1	265000.00	1.9
1	55.0	0	0	38	0	263358.03	1.1
2	65.0	0	0	20	0	162000.00	1.3
3	50.0	1	0	20	0	210000.00	1.9
4	65.0	1	1	20	0	327000.00	2.7

In [25]:
```
## FLOAT TO INT
heart_failure_df.age = heart_failure_df.age.astype(int)
```

In [30]:
```
heart_failure_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   age                   299 non-null    int32  
 1   anaemia               299 non-null    int64  
 2   diabetes              299 non-null    int64  
 3   ejection_fraction     299 non-null    int64  
 4   high_blood_pressure   299 non-null    int64  
 5   platelets             299 non-null    float64
 6   serum_creatinine      299 non-null    float64
 7   serum_sodium          299 non-null    int64  
 8   sex                   299 non-null    int64  
 9   smoking               299 non-null    int64  
 10  patient_dead          299 non-null    int64  
dtypes: float64(2), int32(1), int64(8)
memory usage: 24.7 KB
```

In [28]:
```
# Each type of integer has a different range of storage capacity

#    Type      Capacity

#    Int16 -- (-32,768 to +32,767)

#    Int32 -- (-2,147,483,648 to +2,147,483,647)

#    Int64 -- (-9,223,372,036,854,775,808 to +9,223,372,036,854,775,807)
```

In [31]:
```
### Convert Int32 to boolean only "0 & 1 " columns

heart_failure_df[['anaemia','diabetes','high_blood_pressure','smoking','patien
```

In [32]:
```
heart_failure_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   age                   299 non-null    int32
 1   anaemia               299 non-null    bool
 2   diabetes              299 non-null    bool
 3   ejection_fraction     299 non-null    int64
 4   high_blood_pressure   299 non-null    bool
 5   platelets             299 non-null    float64
 6   serum_creatinine      299 non-null    float64
 7   serum_sodium          299 non-null    int64
 8   sex                   299 non-null    int64
 9   smoking               299 non-null    bool
 10  patient_dead          299 non-null    bool
dtypes: bool(5), float64(2), int32(1), int64(3)
memory usage: 14.4 KB
```

In [33]:
```
heart_failure_df['sex'] = np.where(heart_failure_df['sex'] == 1,"Male","Female
```

In [35]:
```
heart_failure_df.head()
```

Out[35]:

	age	anaemia	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine
0	75	False	False	20	True	265000.00	1.9
1	55	False	False	38	False	263358.03	1.1
2	65	False	False	20	False	162000.00	1.3
3	50	True	False	20	False	210000.00	1.9
4	65	True	True	20	False	327000.00	2.7

In [36]:
```
heart_failure_df['platelets'] = (heart_failure_df.platelets/1000).astype(int)
```

In [37]:
```
heart_failure_df.head()
```

Out[37]:

	age	anaemia	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	s
0	75	False	False	20	True	265	1.9	
1	55	False	False	38	False	263	1.1	
2	65	False	False	20	False	162	1.3	
3	50	True	False	20	False	210	1.9	
4	65	True	True	20	False	327	2.7	

In [41]:
```
## Check the null values
heart_failure_df.isnull().sum()
# heart_failure_df.isnull().any()
```

Out[41]:
```
age                   0
anaemia               0
diabetes              0
ejection_fraction     0
high_blood_pressure   0
platelets             0
serum_creatinine      0
serum_sodium          0
sex                   0
smoking               0
patient_dead          0
dtype: int64
```

In [43]:
```
len(heart_failure_df.columns)
```

Out[43]:
```
11
```

In [44]:
```
!pip install lxml
```

```
Requirement already satisfied: lxml in c:\users\dhruv\appdata\local\programs
\python\python38\lib\site-packages (4.9.3)
```

In [45]:
```
column_deatils_df = pd.read_html("https://bmcmedinformdecismak.biomedcentral.c
```

In [46]:
```
column_deatils_df
```

Out[46]:

	Feature	Explanation	Measurement	Range
0	Age	Age of the patient	Years	[40,..., 95]
1	Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
2	High blood pressure	If a patient has hypertension	Boolean	0, 1
3	Creatinine phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
4	(CPK)	NaN	NaN	NaN
5	Diabetes	If the patient has diabetes	Boolean	0, 1
6	Ejection fraction	Percentage of blood leaving	Percentage	[14,..., 80]
7	NaN	the heart at each contraction	NaN	NaN
8	Sex	Woman or man	Binary	0, 1
9	Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,..., 850.00]
10	Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
11	Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
12	Smoking	If the patient smokes	Boolean	0, 1
13	Time	Follow-up period	Days	[4,...,285]
14	(target) death event	If the patient died during the follow-up period	Boolean	0, 1

In [47]:
```
column_deatils_df.drop('Range',axis=1,inplace=True)
```

In [48]:
```
column_deatils_df.drop([3,4,7,13],axis=0,inplace=True)
```

In [49]:
```
column_deatils_df.columns = ['feature','explanation','measurement_unit']
```

In [51]:
```
column_deatils_df
```

Out[51]:

	feature	explanation	measurement_unit
0	Age	Age of the patient	Years
1	Anaemia	Decrease of red blood cells or hemoglobin	Boolean
2	High blood pressure	If a patient has hypertension	Boolean
5	Diabetes	If the patient has diabetes	Boolean
6	Ejection fraction	Percentage of blood leaving	Percentage
8	Sex	Woman or man	Binary
9	Platelets	Platelets in the blood	kiloplatelets/mL
10	Serum creatinine	Level of creatinine in the blood	mg/dL
11	Serum sodium	Level of sodium in the blood	mEq/L
12	Smoking	If the patient smokes	Boolean
14	(target) death event	If the patient died during the follow-up period	Boolean

In [53]:
```
heart_failure_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 299 entries, 0 to 298
Data columns (total 11 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   age                   299 non-null    int32  
 1   anaemia               299 non-null    bool   
 2   diabetes              299 non-null    bool   
 3   ejection_fraction     299 non-null    int64  
 4   high_blood_pressure   299 non-null    bool   
 5   platelets             299 non-null    int32  
 6   serum_creatinine      299 non-null    float64
 7   serum_sodium          299 non-null    int64  
 8   sex                   299 non-null    object 
 9   smoking               299 non-null    bool   
 10  patient_dead          299 non-null    bool   
dtypes: bool(5), float64(1), int32(2), int64(2), object(1)
memory usage: 13.3+ KB
```

In [54]:
```
column_deatils_df = column_deatils_df.reindex([0,1,5,6,2,9,10,11,8,12,14])
```

In [55]:
```
column_deatils_df
```

Out[55]:

	feature	explanation	measurement_unit
0	Age	Age of the patient	Years
1	Anaemia	Decrease of red blood cells or hemoglobin	Boolean
5	Diabetes	If the patient has diabetes	Boolean
6	Ejection fraction	Percentage of blood leaving	Percentage
2	High blood pressure	If a patient has hypertension	Boolean
9	Platelets	Platelets in the blood	kiloplatelets/mL
10	Serum creatinine	Level of creatinine in the blood	mg/dL
11	Serum sodium	Level of sodium in the blood	mEq/L
8	Sex	Woman or man	Binary
12	Smoking	If the patient smokes	Boolean
14	(target) death event	If the patient died during the follow-up period	Boolean

In [56]:
```
column_deatils_df.feature= heart_failure_df.columns
```

In [57]:
```
column_deatils_df.feature
```

Out[57]:
```
0                    age
1                anaemia
5               diabetes
6      ejection_fraction
2    high_blood_pressure
9              platelets
10      serum_creatinine
11          serum_sodium
8                    sex
12              smoking
14          patient_dead
Name: feature, dtype: object
```