In [1]:
```
import pandas as pd
```

In [2]:
```
pddf = pd.read_csv("heart_failure_clinical_records_dataset.csv")
```

In [3]:
```
pddf.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 [4]:
```
pdDF = pddf.copy()
pdDF
```

Out[4]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	plate
0	75.0	0	582	0	20	1	26500(
1	55.0	0	7861	0	38	0	26335(
2	65.0	0	146	0	20	0	16200(
3	50.0	1	111	0	20	0	21000(
4	65.0	1	160	1	20	0	32700(
...	...	...	...	...	...	...	
294	62.0	0	61	1	38	1	15500(
295	55.0	0	1820	0	38	0	27000(
296	45.0	0	2060	1	60	0	74200(
297	45.0	0	2413	0	38	0	14000(
298	50.0	0	196	0	45	0	39500(

299 rows × 13 columns

In [5]:
```
pdDF.equals(pddf)
```

Out[5]: True

In [6]:
```
pdDF = pdDF.drop_duplicates(keep = False)
pdDF
```

Out[6]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	plate
0	75.0	0	582	0	20	1	265000
1	55.0	0	7861	0	38	0	263358
2	65.0	0	146	0	20	0	162000
3	50.0	1	111	0	20	0	210000
4	65.0	1	160	1	20	0	327000
...	...	...	...	...	...	...	
294	62.0	0	61	1	38	1	155000
295	55.0	0	1820	0	38	0	270000
296	45.0	0	2060	1	60	0	742000
297	45.0	0	2413	0	38	0	140000
298	50.0	0	196	0	45	0	395000

299 rows × 13 columns

In [7]:
```
def remove_whitespace(s):
    return s.strip()
pdDF = pdDF.applymap(lambda x: remove_whitespace(x) if isinstance(x, str) else x)
pdDF
```

Out[7]:

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	plate
0	75.0	0	582	0	20	1	265000
1	55.0	0	7861	0	38	0	263358
2	65.0	0	146	0	20	0	162000
3	50.0	1	111	0	20	0	210000
4	65.0	1	160	1	20	0	327000
...	...	...	...	...	...	...	
294	62.0	0	61	1	38	1	155000
295	55.0	0	1820	0	38	0	270000
296	45.0	0	2060	1	60	0	742000
297	45.0	0	2413	0	38	0	140000
298	50.0	0	196	0	45	0	395000

299 rows × 13 columns

In [10]:
```
pdDF.drop(['anaemia', 'DEATH_EVENT'], axis=1, inplace=True)
pdDF
```

Out[10]:

	age	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serur
0	75.0	582	0	20	1	265000.00	
1	55.0	7861	0	38	0	263358.03	
2	65.0	146	0	20	0	162000.00	
3	50.0	111	0	20	0	210000.00	
4	65.0	160	1	20	0	327000.00	
...	...	...	...	...	...	...	
294	62.0	61	1	38	1	155000.00	
295	55.0	1820	0	38	0	270000.00	
296	45.0	2060	1	60	0	742000.00	
297	45.0	2413	0	38	0	140000.00	
298	50.0	196	0	45	0	395000.00	

299 rows × 11 columns

In [11]:
```
pdDF= pdDF.dropna()
```

In [12]:
```
pdDF.shape
```

Out[12]:
```
(299, 11)
```

In [13]:
```
pdDF.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 299 entries, 0 to 298
Data columns (total 11 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   age                       299 non-null    float64
 1   creatinine_phosphokinase  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  time                      299 non-null    int64  
dtypes: float64(3), int64(8)
memory usage: 28.0 KB
```

In [14]:
```
pdDF['age'] = pd.to_numeric(pdDF['age'], errors='coerce').fillna(0).astype(int)
pdDF['platelets'] = pd.to_numeric(pdDF['platelets'], errors='coerce').fillna(0).astype
pdDF['serum_creatinine'] = pd.to_numeric(pdDF['serum_creatinine'], errors='coerce').fi
```

In [15]:
```
pdDF.describe()
```

Out[15]:

	age	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	
count	299.000000	299.000000	299.000000	299.000000	299.000000	29
mean	60.829431	581.839465	0.418060	38.083612	0.351171	26335
std	11.894997	970.287881	0.494067	11.834841	0.478136	9780
min	40.000000	23.000000	0.000000	14.000000	0.000000	2510
25%	51.000000	116.500000	0.000000	30.000000	0.000000	21250
50%	60.000000	250.000000	0.000000	38.000000	0.000000	26200
75%	70.000000	582.000000	1.000000	45.000000	1.000000	30350
max	95.000000	7861.000000	1.000000	80.000000	1.000000	85000

In [16]:
```
pd.get_dummies(pdDF)
```

Out[16]:

	age	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum
0	75	582	0	20	1	265000	
1	55	7861	0	38	0	263358	
2	65	146	0	20	0	162000	
3	50	111	0	20	0	210000	
4	65	160	1	20	0	327000	
...	...	...	...	...	...	...	
294	62	61	1	38	1	155000	
295	55	1820	0	38	0	270000	
296	45	2060	1	60	0	742000	
297	45	2413	0	38	0	140000	
298	50	196	0	45	0	395000	

299 rows × 11 columns

In [17]:
```
pdDF.equals(pddf)
```

Out[17]:
```
False
```

In [18]:
```
pdDF.to_csv('NewHeartFailure.csv', index = False)
```

LinkedIn Ismael (Ishmael T.) Ngobeni