

01_data_cleaning

July 20, 2025

01_data_cleaning.ipynb

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

Load dataset

```
[2]: df = pd.read_csv('E:/hospital_analytics_project/data/raw/
↳heart_failure_clinical_records_dataset.csv')
```

Preview

```
[3]: print("Shape:", df.shape)
display(df.head())
```

Shape: (299, 13)

	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	\
0	75.0	0	582	0	20	
1	55.0	0	7861	0	38	
2	65.0	0	146	0	20	
3	50.0	1	111	0	20	
4	65.0	1	160	1	20	

	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	\
0		1	265000.00	1.9	130	1
1		0	263358.03	1.1	136	1
2		0	162000.00	1.3	129	1
3		0	210000.00	1.9	137	1
4		0	327000.00	2.7	116	0

	smoking	time	DEATH_EVENT
0	0	4	1
1	0	6	1
2	1	7	1
3	0	7	1
4	0	8	1

Check types and missing values

```
[4]: print("\nInfo:")
df.info()
```

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

```
[5]: print("\nMissing values:")
      print(df.isnull().sum())
```

Missing values:

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

Rename columns (optional, to make it cleaner for visuals)

```
[6]: df.rename(columns={
      'DEATH_EVENT': 'death_event',
      'serum_creatinine': 'serum_creatinine_mg/dL',
```

```
        'serum_sodium': 'serum_sodium_mEq/L'  
    }, inplace=True)
```

Save cleaned data

```
[7]: df.to_csv('E:/hospital_analytics_project/data/processed/heart_failure_cleaned.  
    ↪csv', index=False)
```