

AICS P-4

April 9, 2025

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

```
[2]: df = sns.load_dataset('titanic')
df.head()
```

```
[2]:   survived  pclass    sex  age  sibsp  parch   fare embarked  class \
0         0        3   male  22.0     1     0   7.2500         S  Third
1         1        1  female  38.0     1     0  71.2833         C  First
2         1        3  female  26.0     0     0   7.9250         S  Third
3         1        1  female  35.0     1     0  53.1000         S  First
4         0        3   male  35.0     0     0   8.0500         S  Third
```

```
   who  adult_male deck  embark_town  alive  alone
0  man         True  NaN  Southampton    no  False
1 woman        False   C   Cherbourg   yes  False
2 woman        False  NaN  Southampton   yes   True
3 woman        False   C   Southampton   yes  False
4  man         True  NaN  Southampton    no   True
```

```
[3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 15 columns):
#   Column          Non-Null Count  Dtype
---  -
0   survived        891 non-null   int64
1   pclass          891 non-null   int64
2   sex             891 non-null   object
3   age             714 non-null   float64
4   sibsp           891 non-null   int64
5   parch           891 non-null   int64
6   fare            891 non-null   float64
7   embarked        889 non-null   object
8   class           891 non-null   category
9   who             891 non-null   object
10  adult_male       891 non-null   bool
```

```

11  deck          203 non-null    category
12  embark_town  889 non-null    object
13  alive        891 non-null    object
14  alone        891 non-null    bool
dtypes: bool(2), category(2), float64(2), int64(4), object(5)
memory usage: 80.7+ KB

```

```
[4]: df.isnull().sum()
```

```

[4]: survived          0
    pclass            0
    sex              0
    age             177
    sibsp           0
    parch           0
    fare            0
    embarked         2
    class           0
    who             0
    adult_male       0
    deck           688
    embark_town       2
    alive            0
    alone            0
    dtype: int64

```

```
[5]: df['age'] = df['age'].fillna(df['age'].median())
```

```
[6]: df['embarked'] = df['embarked'].fillna(df['embarked'].mode()[0])
```

```

[7]: if 'Unknown' not in df['deck'].cat.categories:
      df['deck'] = df['deck'].cat.add_categories('Unknown')

df['deck'] = df['deck'].fillna('Unknown')

```

```
[8]: df['embark_town'] = df['embark_town'].fillna(df['embark_town'].mode()[0])
```

```
[9]: df['family_size'] = df['sibsp'] + df['parch'] + 1
```

```

[10]: df['title'] = df['who'].map({
      'man': 'Mr',
      'woman': 'Mrs',
      'child': 'Master'
    })

```

```

[11]: df['age_bin'] = pd.cut(df['age'], bins=[0, 12, 18, 35, 60, 100],
    ↪ labels=['Child', 'Teen', 'Young_Adult', 'Adult', 'Senior'])

```

```
[12]: df.head()
```

```
[12]:
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	\
0	0	3	male	22.0	1	0	7.2500	S	Third	
1	1	1	female	38.0	1	0	71.2833	C	First	
2	1	3	female	26.0	0	0	7.9250	S	Third	
3	1	1	female	35.0	1	0	53.1000	S	First	
4	0	3	male	35.0	0	0	8.0500	S	Third	

	who	adult_male	deck	embark_town	alive	alone	family_size	title	\
0	man	True	Unknown	Southampton	no	False	2	Mr	
1	woman	False	C	Cherbourg	yes	False	2	Mrs	
2	woman	False	Unknown	Southampton	yes	True	1	Mrs	
3	woman	False	C	Southampton	yes	False	2	Mrs	
4	man	True	Unknown	Southampton	no	True	1	Mr	

	age_bin
0	Young_Adult
1	Adult
2	Young_Adult
3	Young_Adult
4	Young_Adult