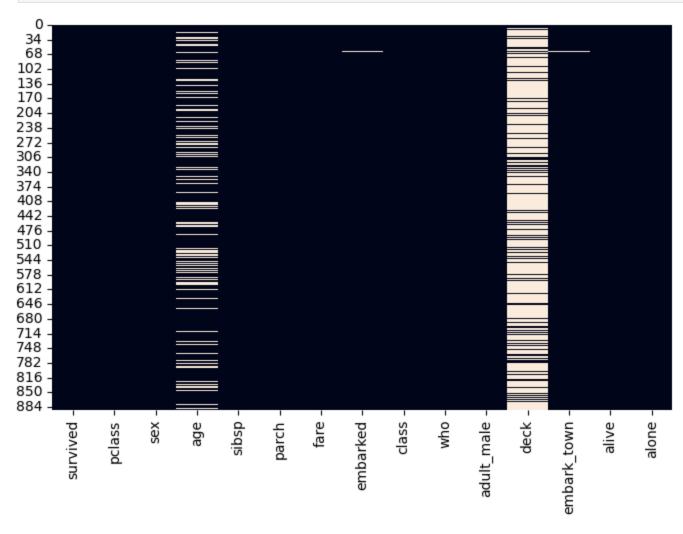
## **Dealing with Missing Values**

```
In [1]:
        # import libraries
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
In [2]:
        # Load the data
        data = sns.load_dataset('titanic')
        data.head()
Out[2]:
                              sex age sibsp parch
           survived pclass
                                                         fare embarked class
                                                                                 who adult male deck embark town alive alone
         0
                  0
                              male 22.0
                                                       7.2500
                                                                      S Third
                                                                                             True NaN
                                                                                                         Southampton
                                                                                                                             False
                                                                                  man
                                                                                                                        no
         1
                  1
                         1 female 38.0
                                                   0 71.2833
                                                                                                     C
                                                                          First woman
                                                                                             False
                                                                                                            Cherbourg
                                                                                                                        yes
                                                                                                                             False
         2
                  1
                         3 female 26.0
                                            0
                                                   0 7.9250
                                                                                                         Southampton
                                                                         Third
                                                                               woman
                                                                                             False
                                                                                                   NaN
                                                                                                                             Tru
                                                                                                                        yes
         3
                                                                         First woman
                         1 female 35.0
                                                   0 53.1000
                                                                                             False
                                                                                                         Southampton
                                                                                                                             False
                                                                                                                        yes
                              male 35.0
         4
                                            0
                                                       8.0500
                                                                      S Third
                                                                                             True NaN
                                                                                                         Southampton
                                                                                                                              Tru
                                                                                  man
                                                                                                                        no
In [3]:
        # Import libraries
        import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        import seaborn as sns
        # Load titanic dataset
        data = sns.load_dataset('titanic')
        # Visualize the data
        plt.figure(figsize=(8, 5))
```

```
sns.heatmap(data.isnull(), cbar=False)
plt.show()
```



In [4]: data.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
                                          object
        2
                          891 non-null
            sex
        3
            age
                          714 non-null
                                          float64
            sibsp
                          891 non-null
                                          int64
                          891 non-null
        5
            parch
                                          int64
            fare
                          891 non-null
                                          float64
            embarked
                          889 non-null
                                          object
        8
            class
                          891 non-null
                                          category
        9
            who
                          891 non-null
                                          object
            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
        data.isnull().sum().sort_values(ascending=False)
In [5]:
Out[5]:
        deck
                        688
         age
                        177
         embarked
                          2
         embark town
                          2
                          0
         survived
         pclass
                          0
                          0
         sex
         sibsp
                          0
                          0
         parch
                          0
         fare
                          0
         class
                          0
         who
         adult male
                          0
         alive
                          0
         alone
```

dtype: int64

```
round(data.isnull().sum() / len(data) * 100, 2).sort_values(ascending=False)
                     77.22
Out[6]:
       deck
                     19.87
        age
        embarked
                      0.22
                      0.22
        embark_town
        survived
                      0.00
                      0.00
        pclass
                      0.00
        sex
        sibsp
                      0.00
        parch
                      0.00
        fare
                      0.00
        class
                      0.00
        who
                      0.00
        adult_male
                      0.00
        alive
                      0.00
        alone
                      0.00
        dtype: float64
In [7]: # Import libraries
       import pandas as pd
       import numpy as np
        import seaborn as sns
        # load titanic dataset
       data = sns.load_dataset('titanic')
       # calculate missing values
       print("----")
       print(f"Missing values in each column:\n{data.isnull().sum().sort_values(ascending=False)}")
       print("----")
       print(f"Percentage of missing values in each column:\n{round(data.isnull().sum() / len(data) * 100, 2).sort_values(as)
```

```
Missing values in each column:
deck
               688
age
               177
embarked
                 2
embark_town
survived
pclass
sex
sibsp
parch
fare
class
who
adult_male
alive
                 0
alone
dtype: int64
Percentage of missing values in each column:
               77.22
deck
               19.87
age
                0.22
embarked
                0.22
embark_town
survived
                0.00
pclass
                0.00
                0.00
sex
sibsp
                0.00
                0.00
parch
fare
                0.00
                0.00
class
                0.00
who
adult_male
                0.00
alive
                0.00
alone
                0.00
dtype: float64
```

In [8]: data.head()

```
Out[8]:
            survived pclass
                                sex age sibsp parch
                                                           fare embarked class
                                                                                   who adult male deck embark town alive alone
                                                                        S Third
          0
                   0
                          3
                               male 22.0
                                                     0
                                                         7.2500
                                                                                                     NaN
                                                                                                           Southampton
                                                                                                                                False
                                              1
                                                                                               True
                                                                                                                           no
                                                                                    man
          1
                   1
                          1 female 38.0
                                                     0 71.2833
                                                                        C First woman
                                                                                               False
                                                                                                        C
                                                                                                              Cherbourg
                                                                                                                                False
                                                                                                                          yes
          2
                   1
                          3 female 26.0
                                              0
                                                         7.9250
                                                                           Third
                                                                                 woman
                                                                                               False
                                                                                                     NaN
                                                                                                           Southampton
                                                                                                                          yes
                                                                                                                                Tru
          3
                   1
                          1 female 35.0
                                              1
                                                     0 53.1000
                                                                           First woman
                                                                                               False
                                                                                                        C
                                                                                                           Southampton
                                                                                                                          yes
                                                                                                                                False
          4
                   0
                               male 35.0
                                              0
                                                         8.0500
                                                                                                                                Tru
                                                                        S
                                                                           Third
                                                                                    man
                                                                                               True
                                                                                                     NaN
                                                                                                           Southampton
                                                                                                                           no
         round(data['age'].mean(), 2)
 Out[9]: 29.7
         data['age'].median()
In [10]:
Out[10]:
         28.0
In [11]:
          # Mean of age to fill age missing values
          data['age'] = data['age'].fillna(data['age'].median())
          # drop deck column
          data.drop('deck', axis=1, inplace=True)
          data.head()
Out[11]:
            survived pclass
                                sex age sibsp parch
                                                           fare embarked class
                                                                                   who adult male embark town alive alone
          0
                   0
                               male 22.0
                                                         7.2500
                                                                          Third
                                                                                                     Southampton
                          3
                                              1
                                                     0
                                                                        S
                                                                                    man
                                                                                               True
                                                                                                                     no
                                                                                                                         False
          1
                   1
                          1 female 38.0
                                                     0 71.2833
                                                                           First woman
                                                                                               False
                                                                                                        Cherbourg
                                                                                                                         False
                                              1
                                                                                                                    yes
          2
                   1
                          3 female 26.0
                                              0
                                                        7.9250
                                                                           Third
                                                                                                     Southampton
                                                                                                                    yes
                                                                                 woman
                                                                                               False
                                                                                                                          True
          3
                   1
                          1 female 35.0
                                                     0 53.1000
                                                                            First woman
                                              1
                                                                                               False
                                                                                                     Southampton
                                                                                                                         False
                                                                                                                    yes
          4
                   0
                          3
                               male 35.0
                                              0
                                                     0
                                                         8.0500
                                                                        S
                                                                          Third
                                                                                                     Southampton
                                                                                    man
                                                                                               True
                                                                                                                          True
                                                                                                                     no
```

```
data['embark_town'].value_counts()
In [12]:
Out[12]:
         embark_town
          Southampton
                         644
          Cherbourg
                         168
                          77
          Queenstown
          Name: count, dtype: int64
In [13]: # replacing embarked missing values with mode
         data['embark_town'] = data['embark_town'].fillna(data['embark_town'].mode()[0])
         data['embarked'] = data['embarked'].fillna(data['embarked'].mode()[0])
         data.isnull().sum().sort_values(ascending=False)
Out[13]: survived
          pclass
                         0
          sex
                         0
          age
          sibsp
                         0
          parch
                         0
          fare
          embarked
          class
          who
                         0
          adult_male
          embark_town
          alive
          alone
                         0
          dtype: int64
```

#### we can also impute using sk learn

```
In [14]: # import Libraries
   import pandas as pd
   import numpy as np
   import seaborn as sns

from sklearn.impute import SimpleImputer
```

```
df = sns.load_dataset('titanic')
In [15]:
          # impute age column using simpleimputer from sklearn
          imputer = SimpleImputer(strategy='mean')
          df['age'] = imputer.fit_transform(df[['age']])
          df.isnull().sum().sort_values(ascending=False)
Out[15]:
          deck
                         688
          embarked
                            2
                            2
          embark_town
          survived
                            0
                            0
          pclass
                            0
          sex
                            0
          age
                            0
          sibsp
                            0
          parch
          fare
                            0
                            0
          class
          who
                            0
                            0
          adult_male
          alive
                            0
          alone
          dtype: int64
         df.head()
In [16]:
Out[16]:
             survived pclass
                                sex age sibsp parch
                                                                                   who adult male deck embark town alive alone
                                                           fare embarked class
                                                         7.2500
                                                                        S Third
          0
                   0
                               male 22.0
                                                                                                True NaN
                                                                                                            Southampton
                                                                                                                                False
                                                                                    man
                                                                                                                           no
                                                     0 71.2833
                          1 female 38.0
          1
                   1
                                                                            First woman
                                                                                               False
                                                                                                        C
                                                                                                              Cherbourg
                                              1
                                                                                                                                False
                                                                                                                           yes
                                                         7.9250
          2
                   1
                           3 female 26.0
                                              0
                                                                                                     NaN
                                                                                                            Southampton
                                                                           Third
                                                                                 woman
                                                                                               False
                                                                                                                                Tru
                                                                                                                          yes
          3
                                                                                                            Southampton
                   1
                           1 female 35.0
                                                     0 53.1000
                                                                            First woman
                                              1
                                                                                               False
                                                                                                                                False
                                                                                                                          yes
          4
                   0
                               male 35.0
                                              0
                                                         8.0500
                                                                        S Third
                                                                                                            Southampton
                                                                                    man
                                                                                                True NaN
                                                                                                                                Tru
                                                                                                                           no
                                                                                                                                 •
```

## multivariate imputation

```
In [17]: df = sns.load_dataset('titanic')
         from sklearn.experimental import enable_iterative_imputer
         from sklearn.impute import IterativeImputer
         # impute age column using iterative imputer from sklearn
         imputer = IterativeImputer(max_iter=20, n_nearest_features=5)
         df['age'] = imputer.fit_transform(df[['age']])
         df.isnull().sum().sort_values(ascending=False)
Out[17]: deck
                         688
          embarked
                           2
          embark_town
          survived
          pclass
          sex
          age
          sibsp
                           0
          parch
          fare
                           0
          class
          who
          adult_male
          alive
                           0
          alone
          dtype: int64
```

#### ffill and backward fill

```
In [18]: df = sns.load_dataset('titanic')
```

```
df.isnull().sum().sort_values(ascending=False)
Out[18]:
         deck
                        688
                        177
          age
                          2
         embarked
         embark_town
          survived
          pclass
          sex
                          0
         sibsp
          parch
         fare
          class
         who
                          0
         adult_male
          alive
          alone
         dtype: int64
In [19]: # using forward fill impute age column
         df['age'] = df['age'].bfill()
```

# using KNN imputer

```
In [20]: from sklearn.impute import KNNImputer
    # impute age column using KNNImputer from sklearn
    imputer = KNNImputer(n_neighbors=5)
    df['age'] = imputer.fit_transform(df[['age']])

# drop rows having missing values
    df.dropna(inplace=True)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
        Index: 201 entries, 1 to 889
        Data columns (total 15 columns):
             Column
                           Non-Null Count Dtype
             -----
         0
             survived
                           201 non-null
                                           int64
         1
             pclass
                           201 non-null
                                           int64
         2
                           201 non-null
                                           object
             sex
         3
                           201 non-null
                                           float64
             age
             sibsp
                           201 non-null
                                           int64
         5
             parch
                           201 non-null
                                           int64
             fare
                           201 non-null
                                           float64
             embarked
                           201 non-null
                                           object
         8
             class
                           201 non-null
                                           category
         9
             who
                           201 non-null
                                           object
             adult_male
                          201 non-null
                                           bool
             deck
         11
                           201 non-null
                                           category
         12 embark_town 201 non-null
                                           object
         13 alive
                           201 non-null
                                           object
         14 alone
                           201 non-null
                                           bool
        dtypes: bool(2), category(2), float64(2), int64(4), object(5)
        memory usage: 20.1+ KB
In [21]: df.isnull().sum().sort_values(ascending=False)
Out[21]: survived
                         0
                         0
          pclass
                         0
          sex
                         0
          age
          sibsp
                         0
                         0
          parch
          fare
                         0
          embarked
                         0
          class
                         0
          who
                         0
          adult_male
                         0
          deck
                         0
          embark_town
          alive
                         0
          alone
          dtype: int64
```

In [ ]: