In [1]:
```
import numpy as np
import pandas as pd
df = pd.read_csv("https://raw.githubusercontent.com/datasciencedojo/datasets/
print(df.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                Name     Sex   Age  SibSp  \
0                            Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                             Heikkinen, Miss. Laina  female  26.0      0
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                           Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin Embarked
0      0         A/5 21171   7.2500   NaN        S
1      0          PC 17599  71.2833   C85        C
2      0  STON/O2. 3101282   7.9250   NaN        S
3      0            113803  53.1000  C123        S
4      0            373450   8.0500   NaN        S
```

In [2]:
```
df_cat = df.select_dtypes(exclude=[np.number])
print(df_cat.head())
```

```
                                                Name     Sex  \
0                            Braund, Mr. Owen Harris    male
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female
2                             Heikkinen, Miss. Laina  female
3       Futrelle, Mrs. Jacques Heath (Lily May Peel)  female
4                           Allen, Mr. William Henry    male

            Ticket Cabin Embarked
0         A/5 21171   NaN        S
1          PC 17599   C85        C
2  STON/O2. 3101282   NaN        S
3            113803  C123        S
4            373450   NaN        S
```

In [3]:
```
print(df_cat['Sex'].unique())
```

```
['male' 'female']
```

In [4]:
```
print(df_cat['Sex'].value_counts())
```

```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [5]:
```
df_cat['Sex'] = df_cat['Sex'].replace(['male', 'female'], [0, 1])
print(df_cat['Sex'].head())
```

```
0    0
1    1
```

```
2    1
3    1
4    0
Name: Sex, dtype: int64
```

```
/tmp/ipykernel_5532/3209713335.py:1: FutureWarning: Downcasting behavior in `
replace` is deprecated and will be removed in a future version. To retain the
old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in t
o the future behavior, set `pd.set_option('future.no_silent_downcasting', Tru
e)`
  df_cat['Sex'] = df_cat['Sex'].replace(['male', 'female'], [0, 1])
```

In [6]:
```
df_cat['Embarked'] = df_cat['Embarked'].replace(['S', 'C', 'Q'], [0, 1, 2])
print(df_cat['Embarked'].head())
```

```
0    0.0
1    1.0
2    0.0
3    0.0
4    0.0
Name: Embarked, dtype: float64
```

```
/tmp/ipykernel_5532/1980851516.py:1: FutureWarning: Downcasting behavior in `
replace` is deprecated and will be removed in a future version. To retain the
old behavior, explicitly call `result.infer_objects(copy=False)`. To opt-in t
o the future behavior, set `pd.set_option('future.no_silent_downcasting', Tru
e)`
  df_cat['Embarked'] = df_cat['Embarked'].replace(['S', 'C', 'Q'], [0, 1, 2])
```

In [7]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [8]:
```
df['Sex_encoded'] = le.fit_transform(df['Sex'])
```

In [10]:
```
df['Embarked_encoded'] = le.fit_transform(df['Embarked'].astype(str))
print(df[['Sex', 'Sex_encoded', 'Embarked', 'Embarked_encoded']].head())
```

```
      Sex  Sex_encoded Embarked  Embarked_encoded
0    male            1        S                 2
1  female            0        C                 0
2  female            0        S                 2
3  female            0        S                 2
4    male            1        S                 2
```

In [11]:
```
lst = ['First', 'Second', 'Third', 'Crew']
le.fit(lst)
print(le.transform(lst))
```

```
[1 2 3 0]
```

In [12]:
```
print(le.inverse_transform([0, 1, 2, 3]))
```

```
['Crew' 'First' 'Second' 'Third']
```

In [13]:
```
df_ohe = pd.get_dummies(df, columns=['Sex'])
print(df_ohe.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
```

```
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                 Name   Age  SibSp  Parch  \
0                             Braund, Mr. Owen Harris  22.0      1      0
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  38.0      1      0
2                              Heikkinen, Miss. Laina  26.0      0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  35.0      1      0
4                            Allen, Mr. William Henry  35.0      0      0

             Ticket     Fare Cabin Embarked  Sex_encoded  Embarked_encoded  \
0         A/5 21171   7.2500   NaN        S            1                 2
1          PC 17599  71.2833   C85        C            0                 0
2  STON/O2. 3101282   7.9250   NaN        S            0                 2
3            113803  53.1000  C123        S            0                 2
4            373450   8.0500   NaN        S            1                 2

   Sex_female  Sex_male
0       False      True
1        True     False
2        True     False
3        True     False
4       False      True
```

In [14]:

```
df_ohe = pd.get_dummies(df, columns=['Sex', 'Embarked'])
print(df_ohe.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                 Name   Age  SibSp  Parch  \
0                             Braund, Mr. Owen Harris  22.0      1      0
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  38.0      1      0
2                              Heikkinen, Miss. Laina  26.0      0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  35.0      1      0
4                            Allen, Mr. William Henry  35.0      0      0

             Ticket     Fare Cabin  Sex_encoded  Embarked_encoded  Sex_female
\
0         A/5 21171   7.2500   NaN            1                 2       False
1          PC 17599  71.2833   C85            0                 0        True
2  STON/O2. 3101282   7.9250   NaN            0                 2        True
3            113803  53.1000  C123            0                 2        True
4            373450   8.0500   NaN            1                 2       False

   Sex_male  Embarked_C  Embarked_Q  Embarked_S
0      True       False       False        True
1     False        True       False       False
2     False       False       False        True
3     False       False       False        True
4      True       False       False        True
```

In [15]:

```
df_ohe = pd.get_dummies(df, columns=['Embarked'], prefix='Port')
print(df_ohe.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
```

```
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin  Sex_encoded  Embarked_encoded  \
0      0         A/5 21171   7.2500   NaN            1                 2
1      0          PC 17599  71.2833   C85            0                 0
2      0  STON/O2. 3101282   7.9250   NaN            0                 2
3      0            113803  53.1000  C123            0                 2
4      0            373450   8.0500   NaN            1                 2

   Port_C  Port_Q  Port_S
0   False   False    True
1    True   False   False
2   False   False    True
3   False   False    True
4   False   False    True
```

In [16]:

```
df_ohe = pd.get_dummies(df, columns=['Embarked'], drop_first=True)
print(df_ohe.head())
```

```
   PassengerId  Survived  Pclass  \
0            1         0       3
1            2         1       1
2            3         1       3
3            4         1       1
4            5         0       3

                                                 Name     Sex   Age  SibSp  \
0                             Braund, Mr. Owen Harris    male  22.0      1
1  Cumings, Mrs. John Bradley (Florence Briggs Th...  female  38.0      1
2                              Heikkinen, Miss. Laina  female  26.0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0      1
4                            Allen, Mr. William Henry    male  35.0      0

   Parch            Ticket     Fare Cabin  Sex_encoded  Embarked_encoded  \
0      0         A/5 21171   7.2500   NaN            1                 2
1      0          PC 17599  71.2833   C85            0                 0
2      0  STON/O2. 3101282   7.9250   NaN            0                 2
3      0            113803  53.1000  C123            0                 2
4      0            373450   8.0500   NaN            1                 2

   Embarked_Q  Embarked_S
0       False        True
1       False       False
2       False        True
3       False        True
4       False        True
```

In [17]:

```
from sklearn.preprocessing import OneHotEncoder

ohe = OneHotEncoder(sparse=False)

ohe_data = ohe.fit_transform(df[['Sex']])
print(ohe_data[:5])
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
/tmp/ipykernel_5532/3949029170.py in <module>
      1 from sklearn.preprocessing import OneHotEncoder
      2
----> 3 ohe = OneHotEncoder(sparse=False)
      4
      5 ohe_data = ohe.fit_transform(df[['Sex']])

TypeError: __init__() got an unexpected keyword argument 'sparse'
```

In [18]:

```
df['Age_group'] = pd.cut(
    df['Age'],
    bins=[0, 12, 30, 60, 100],
    labels=['Child', 'Young', 'Adult', 'Senior']
)
print(df[['Age', 'Age_group']].head(10))
```

```
    Age Age_group
0  22.0     Young
1  38.0     Adult
2  26.0     Young
3  35.0     Adult
4  35.0     Adult
5   NaN       NaN
6  54.0     Adult
7   2.0     Child
8  27.0     Young
9  14.0     Young
```

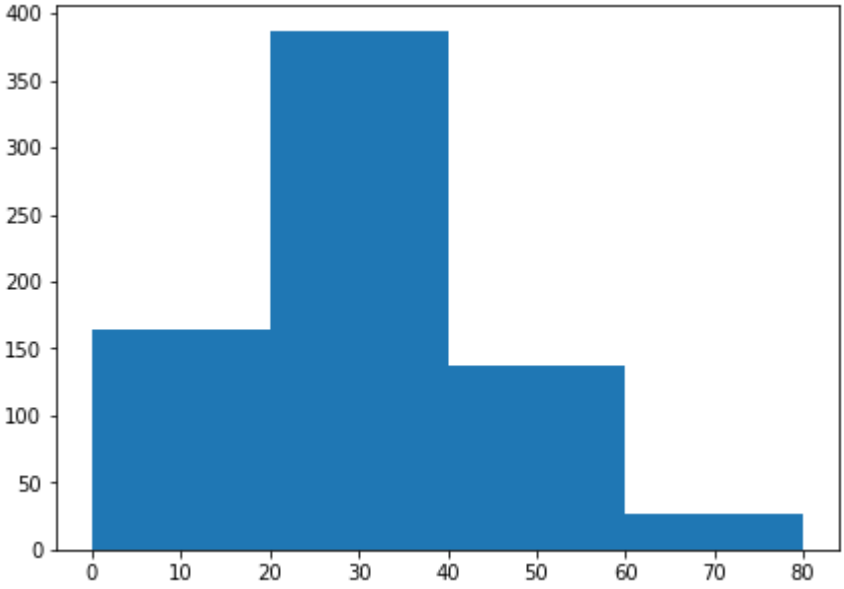
In [19]:

```
df['Fare_group'] = pd.cut(df['Fare'], bins=2, labels=['Low', 'High'])
print(df[['Fare', 'Fare_group']].head())
```

```
      Fare Fare_group
0   7.2500        Low
1  71.2833        Low
2   7.9250        Low
3  53.1000        Low
4   8.0500        Low
```

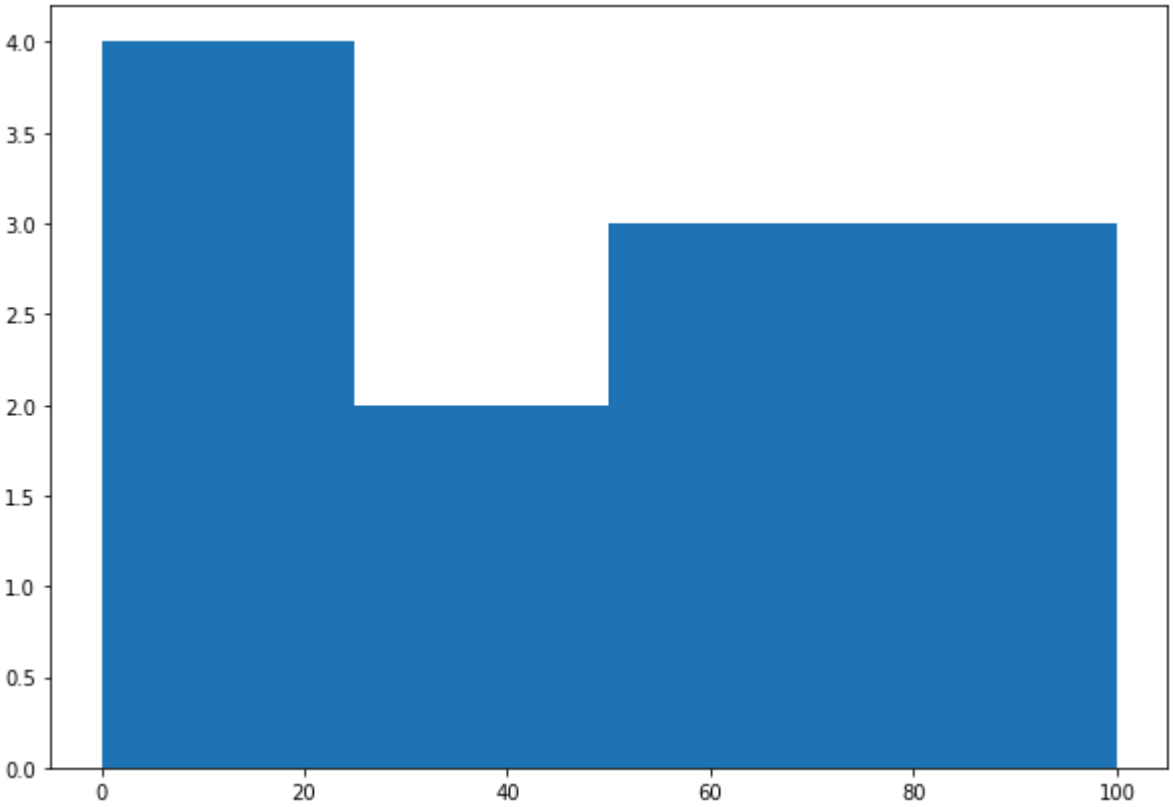
In [20]:

```
from matplotlib import pyplot as plt

fig, ax = plt.subplots(figsize=(7,5))
ax.hist(df['Age'].dropna(), bins=[0, 20, 40, 60, 80])
plt.show()
```



In [21]:

```
a = np.array([12, 45, 67, 23, 89, 90, 34, 55, 66, 78, 21, 10])

fig, ax = plt.subplots(figsize=(10,7))
ax.hist(a, bins=[0, 25, 50, 75, 100])
plt.show()
```



In [ ]: