task5

June 9, 2025

```python
[2]: import pandas as pd
     import numpy as np
     import seaborn as sns
     import glob as gb
```

```python
[4]: titanic = gb.glob("C:/Users/Priya Dutta/Downloads/titanic.zip")
```

```python
[5]: titanic
```

```python
[5]: ['C:/Users/Priya Dutta/Downloads/titanic.zip']
```

```python
[7]: zip_path = "C:/Users/Priya Dutta/Downloads/titanic.zip"
```

```python
[9]: import zipfile as zf
```

```python
[12]: with zf.ZipFile(zip_path, 'r') as z:
          print(z.namelist())
```

```
['gender_submission.csv', 'test.csv', 'train.csv']
```

```python
[23]: zip_path = 'C:/Users/Priya Dutta/Downloads/titanic.zip'
      dfs = []
```

```python
[28]: with zf.ZipFile('C:/Users/Priya Dutta/Downloads/titanic.zip', 'r') as z:
          for file_name in ['train.csv', 'test.csv', 'gender_submission.csv']:
              with z.open(file_name) as f:
                  df = pd.read_csv(f)
                  dfs.append(df)
```

```python
[29]: combined_df = pd.concat(dfs, ignore_index=True)
      print(combined_df.head())
```

```
   PassengerId  Survived  Pclass  \
0            1       0.0     3.0
1            2       1.0     1.0
2            3       1.0     3.0
3            4       1.0     1.0
4            5       0.0     3.0
```

```
                                             Name     Sex   Age  SibSp  \
0                         Braund, Mr. Owen Harris    male  22.0    1.0
1  Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0    1.0
2                          Heikkinen, Miss. Laina  female  26.0    0.0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0    1.0
4                        Allen, Mr. William Henry    male  35.0    0.0

   Parch            Ticket     Fare Cabin Embarked
0    0.0         A/5 21171   7.2500   NaN        S
1    0.0          PC 17599  71.2833   C85        C
2    0.0  STON/O2. 3101282   7.9250   NaN        S
3    0.0            113803  53.1000  C123        S
4    0.0            373450   8.0500   NaN        S
```

[30]: `combined_df.to_csv('combined_titanic.csv', index=False)`

[31]: `combined_df.head()`

[31]:
```
   PassengerId  Survived  Pclass  \
0            1       0.0     3.0
1            2       1.0     1.0
2            3       1.0     3.0
3            4       1.0     1.0
4            5       0.0     3.0

                                             Name     Sex   Age  SibSp  \
0                         Braund, Mr. Owen Harris    male  22.0    1.0
1  Cumings, Mrs. John Bradley (Florence Briggs Th…  female  38.0    1.0
2                          Heikkinen, Miss. Laina  female  26.0    0.0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)  female  35.0    1.0
4                        Allen, Mr. William Henry    male  35.0    0.0

   Parch            Ticket     Fare Cabin Embarked
0    0.0         A/5 21171   7.2500   NaN        S
1    0.0          PC 17599  71.2833   C85        C
2    0.0  STON/O2. 3101282   7.9250   NaN        S
3    0.0            113803  53.1000  C123        S
4    0.0            373450   8.0500   NaN        S
```

[32]: `combined_df.info()`

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1727 entries, 0 to 1726
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  1727 non-null   int64
```

```
 1    Survived    1309 non-null    float64
 2    Pclass      1309 non-null    float64
 3    Name        1309 non-null    object
 4    Sex         1309 non-null    object
 5    Age         1046 non-null    float64
 6    SibSp       1309 non-null    float64
 7    Parch       1309 non-null    float64
 8    Ticket      1309 non-null    object
 9    Fare        1308 non-null    float64
 10   Cabin        295 non-null    object
 11   Embarked    1307 non-null    object
dtypes: float64(6), int64(1), object(5)
memory usage: 162.0+ KB
```

[33]:
```python
print("Dimension of the DF:", combined_df.ndim)
print("Shape of the DF:", combined_df.shape)
print("Rows of the DF:", combined_df.shape[0])
print("Columns of the DF:", combined_df.shape[1])
print("size of the DF:", combined_df.size)
print("Column data types of the DF:", combined_df.dtypes)
```

```
Dimension of the DF: 2
Shape of the DF: (1727, 12)
Rows of the DF: 1727
Columns of the DF: 12
size of the DF: 20724
Column data types of the DF: PassengerId      int64
Survived        float64
Pclass          float64
Name             object
Sex              object
Age             float64
SibSp           float64
Parch           float64
Ticket           object
Fare            float64
Cabin            object
Embarked         object
dtype: object
```

[35]: `combined_df.describe().T`

[35]:
```
                count         mean          std    min        25%        50%  \
PassengerId    1727.0   762.828025   385.032264   1.00   432.5000   864.0000
Survived       1309.0     0.377387     0.484918   0.00     0.0000     0.0000
Pclass         1309.0     2.294882     0.837836   1.00     2.0000     3.0000
Age            1046.0    29.881138    14.413493   0.17    21.0000    28.0000
```

```
SibSp          1309.0    0.498854    1.041658  0.00    0.0000    0.0000
Parch          1309.0    0.385027    0.865560  0.00    0.0000    0.0000
Fare           1308.0   33.295479   51.758668  0.00    7.8958   14.4542

                   75%        max
PassengerId  1093.500  1309.0000
Survived        1.000     1.0000
Pclass          3.000     3.0000
Age            39.000    80.0000
SibSp           1.000     8.0000
Parch           0.000     9.0000
Fare           31.275   512.3292
```

[ ]: