shadowfox-advanced

August 22, 2025

[1]:
```
import zipfile
with zipfile.ZipFile("titanic.zip", 'r') as zip_ref:
    zip_ref.extractall("titanic_data")
```

[10]:
```
import pandas as pd
titanic = pd.read_csv("titanic_data/train.csv")
titanic.head()
```

[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th…	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

[9]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
titanic = pd.read_csv("titanic_data/train.csv")
titanic.head()
```

[9]:

	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
```

[4]:
```
print("Dataset Info:")
print(titanic.info())
print("\nSummary Statistics:")
print(titanic.describe(include="all"))
```

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None

Summary Statistics:
```

```
        PassengerId    Survived      Pclass                     Name    Sex  \
count    891.000000  891.000000  891.000000                      891    891
unique          NaN         NaN         NaN                      891      2
top             NaN         NaN         NaN  Braund, Mr. Owen Harris   male
freq            NaN         NaN         NaN                        1    577
mean     446.000000    0.383838    2.308642                      NaN    NaN
std      257.353842    0.486592    0.836071                      NaN    NaN
min        1.000000    0.000000    1.000000                      NaN    NaN
25%      223.500000    0.000000    2.000000                      NaN    NaN
50%      446.000000    0.000000    3.000000                      NaN    NaN
75%      668.500000    1.000000    3.000000                      NaN    NaN
max      891.000000    1.000000    3.000000                      NaN    NaN

                Age       SibSp       Parch  Ticket        Fare    Cabin  \
count   714.000000  891.000000  891.000000     891  891.000000      204
unique         NaN         NaN         NaN     681         NaN      147
top            NaN         NaN         NaN  347082         NaN  B96 B98
freq           NaN         NaN         NaN       7         NaN        4
mean     29.699118    0.523008    0.381594     NaN   32.204208      NaN
std      14.526497    1.102743    0.806057     NaN   49.693429      NaN
min       0.420000    0.000000    0.000000     NaN    0.000000      NaN
25%      20.125000    0.000000    0.000000     NaN    7.910400      NaN
50%      28.000000    0.000000    0.000000     NaN   14.454200      NaN
75%      38.000000    1.000000    0.000000     NaN   31.000000      NaN
max      80.000000    8.000000    6.000000     NaN  512.329200      NaN

       Embarked
count       889
unique        3
top           S
freq        644
mean        NaN
std         NaN
min         NaN
25%         NaN
50%         NaN
75%         NaN
max         NaN
```

[11]:
```
titanic = titanic.drop(columns=["PassengerId", "Name", "Ticket", "Cabin"])
titanic["Age"].fillna(titanic["Age"].median(), inplace=True)
titanic["Embarked"].fillna(titanic["Embarked"].mode()[0], inplace=True)
print("Missing values after cleaning:")
print(titanic.isnull().sum())
titanic.head()
```

```
Missing values after cleaning:
Survived    0
```

```
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Fare        0
Embarked    0
dtype: int64
```

```
C:\Users\KOUSHITHA KETHINENI\AppData\Local\Temp\ipykernel_22152\1403935889.py:2:
FutureWarning: A value is trying to be set on a copy of a DataFrame or Series
through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  titanic["Age"].fillna(titanic["Age"].median(), inplace=True)
C:\Users\KOUSHITHA KETHINENI\AppData\Local\Temp\ipykernel_22152\1403935889.py:3:
FutureWarning: A value is trying to be set on a copy of a DataFrame or Series
through chained assignment using an inplace method.
The behavior will change in pandas 3.0. This inplace method will never work
because the intermediate object on which we are setting values always behaves as
a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using
'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value)
instead, to perform the operation inplace on the original object.


  titanic["Embarked"].fillna(titanic["Embarked"].mode()[0], inplace=True)
```

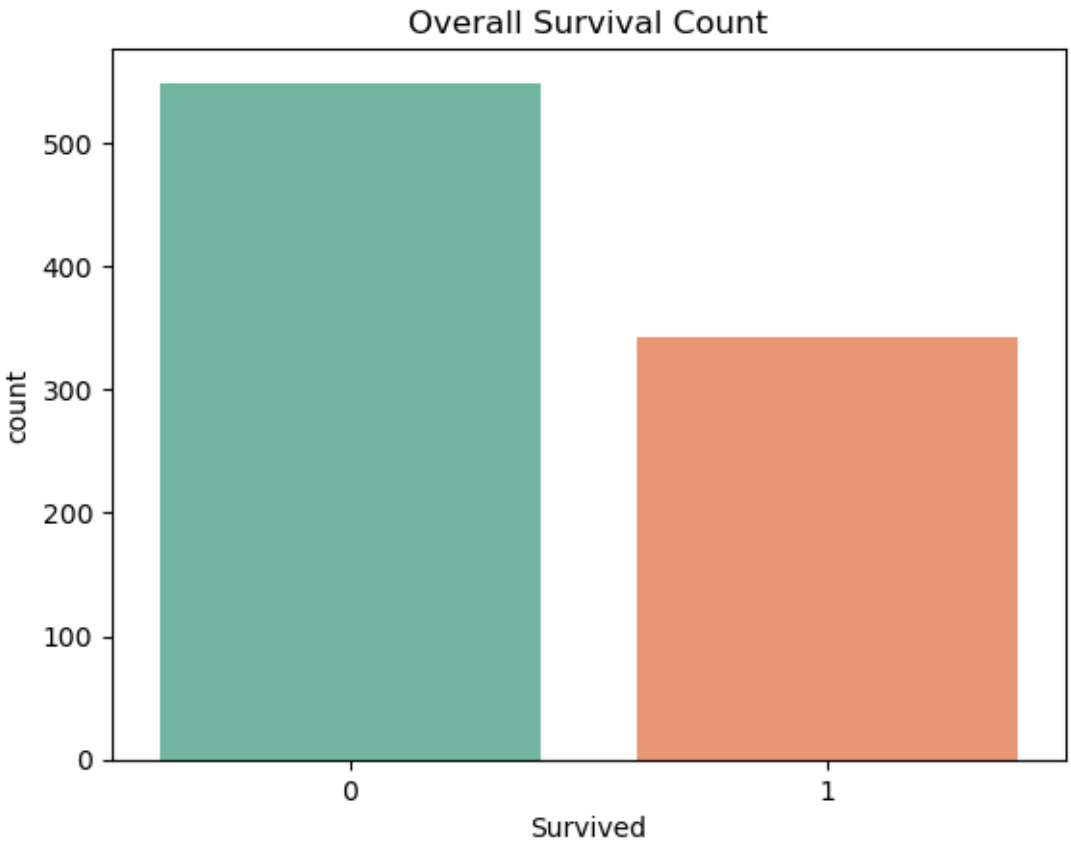
[11]:

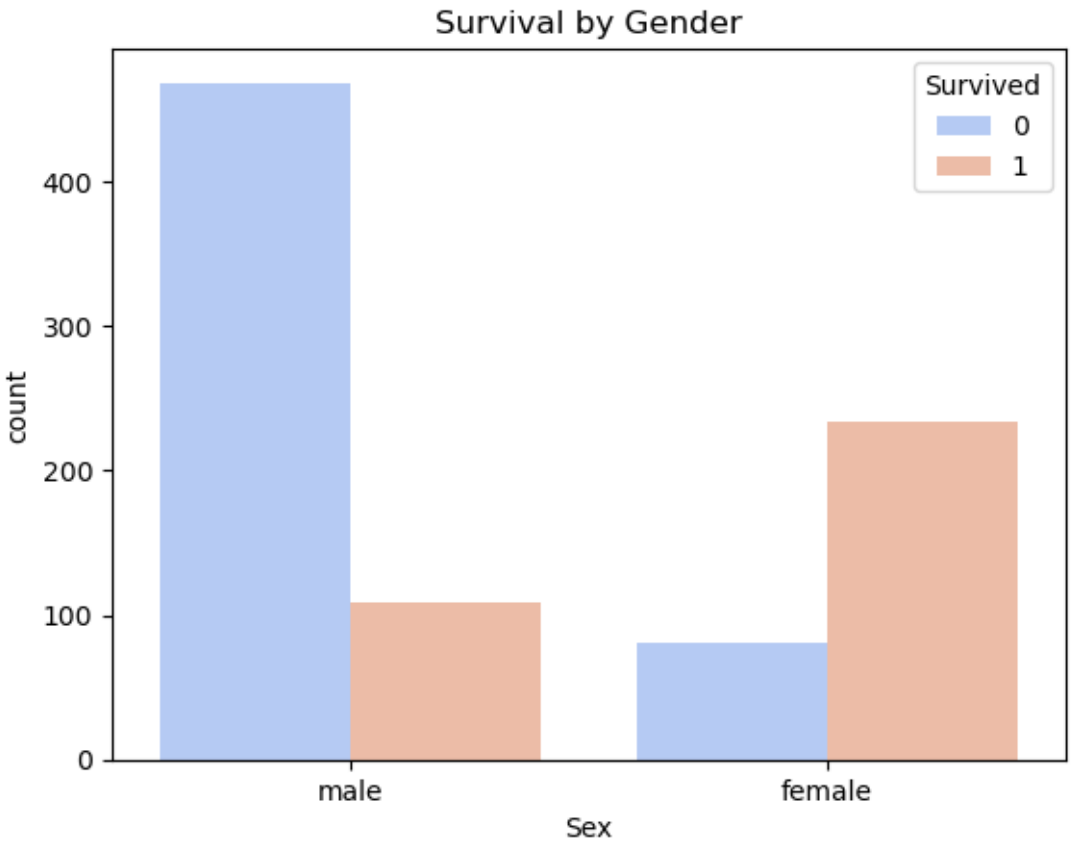
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.0	1	0	7.2500	S
1	1	1	female	38.0	1	0	71.2833	C
2	1	3	female	26.0	0	0	7.9250	S
3	1	1	female	35.0	1	0	53.1000	S
4	0	3	male	35.0	0	0	8.0500	S

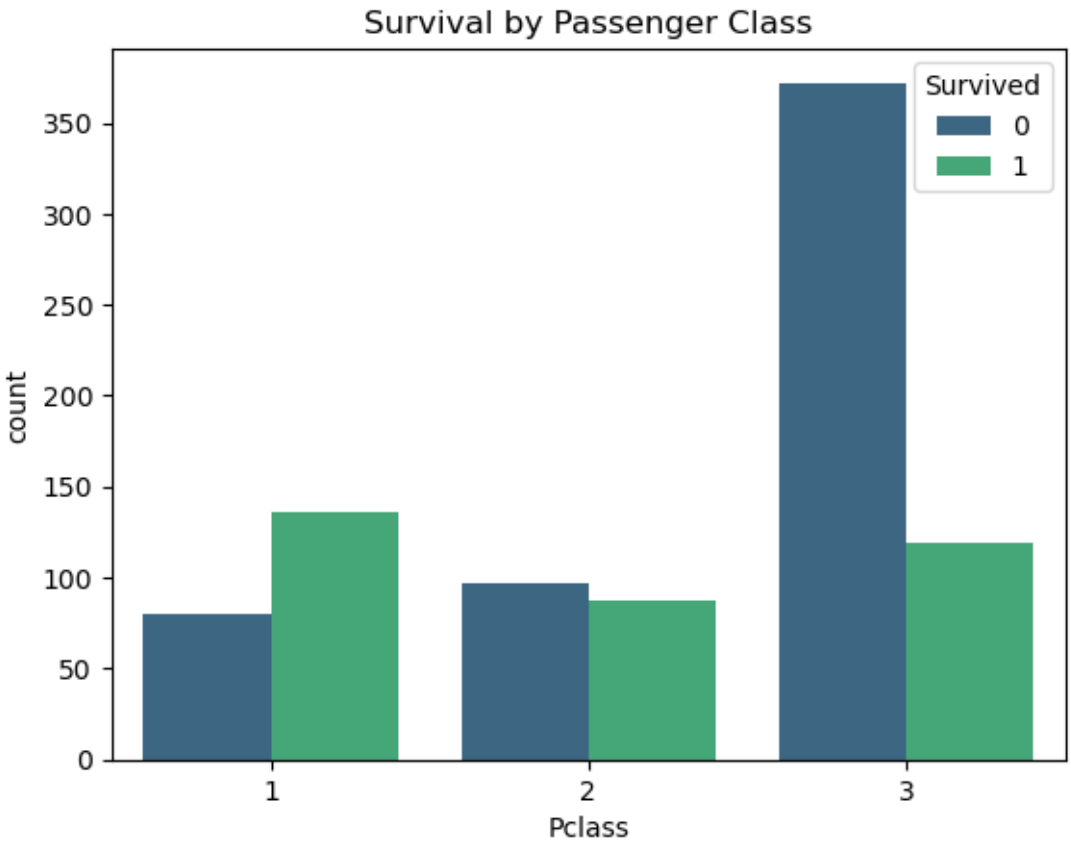
[7]:
```
# ------------------------------
# Step 4: Exploratory Data Analysis (EDA)
# ------------------------------

# 1. Survival Count
```

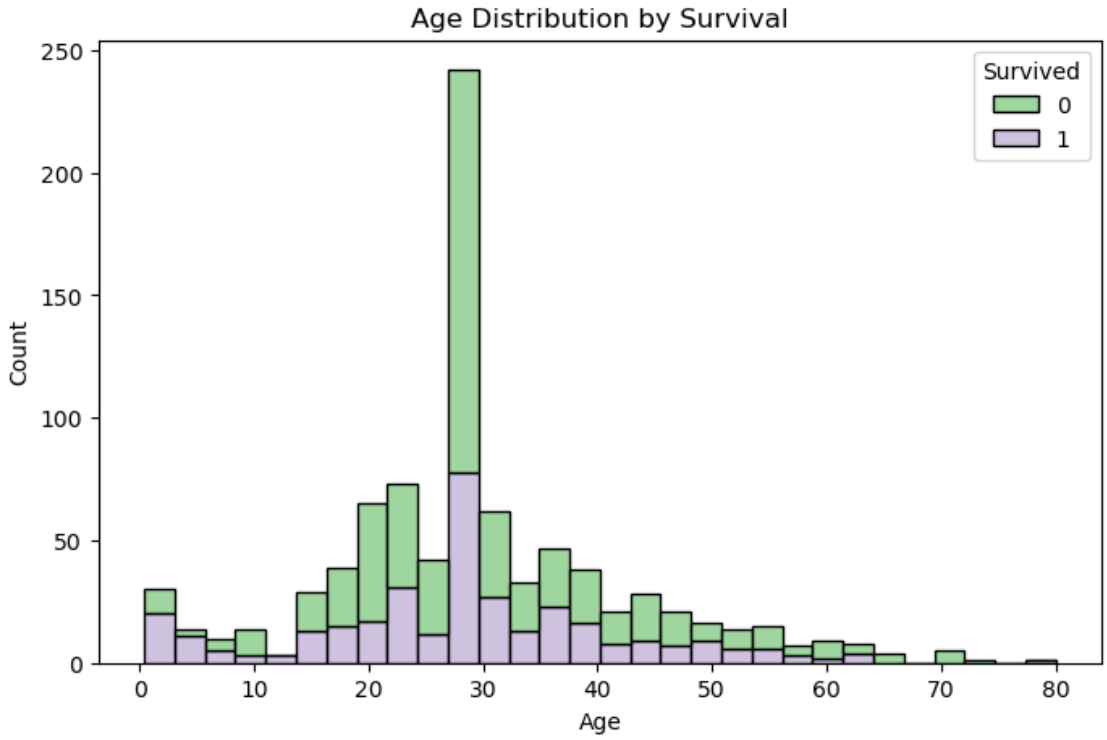
```
sns.countplot(x="Survived", data=titanic, palette="Set2")
plt.title("Overall Survival Count")
plt.show()

# 2. Survival by Gender
sns.countplot(x="Sex", hue="Survived", data=titanic, palette="coolwarm")
plt.title("Survival by Gender")
plt.show()

#  Survival by Passenger Class
sns.countplot(x="Pclass", hue="Survived", data=titanic, palette="viridis")
plt.title("Survival by Passenger Class")
plt.show()

#  Age Distribution by Survival
plt.figure(figsize=(8,5))
sns.histplot(data=titanic, x="Age", hue="Survived", multiple="stack", bins=30,␣
 ↪palette="Accent")
plt.title("Age Distribution by Survival")
plt.show()

#  Correlation Heatmap
plt.figure(figsize=(8,5))
sns.heatmap(titanic.corr(numeric_only=True), annot=True, cmap="coolwarm")
plt.title("Correlation Heatmap")
plt.show()
```

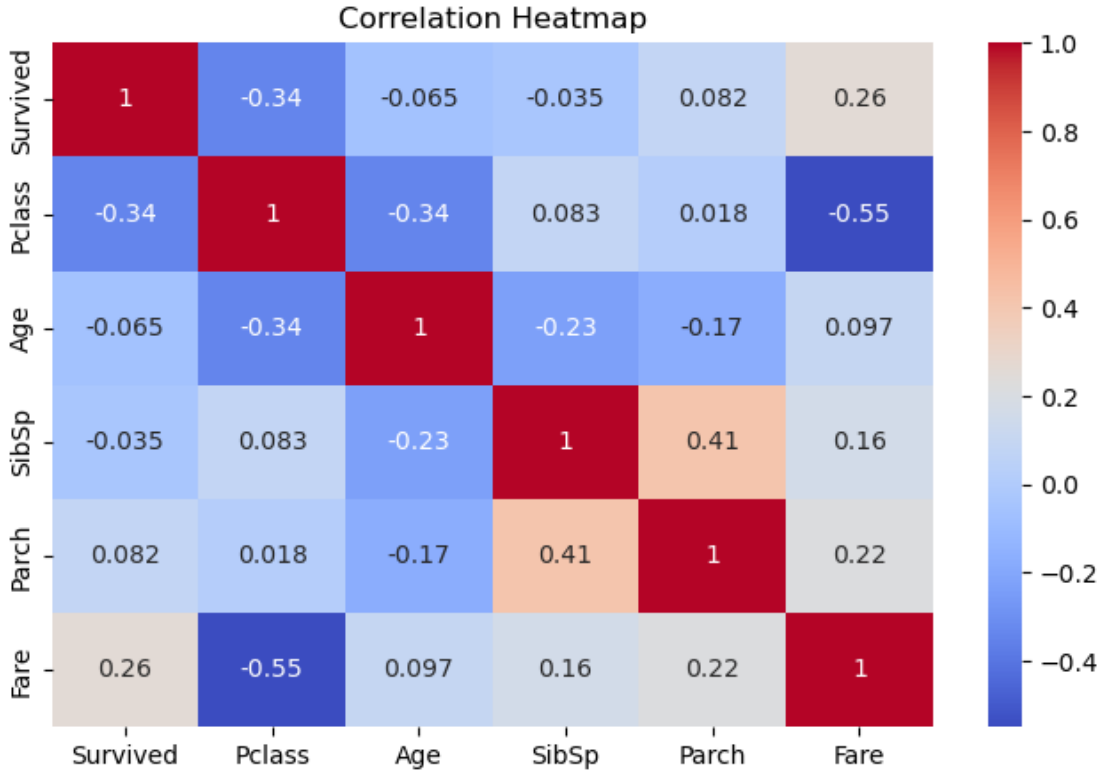
```
C:\Users\KOUSHITHA KETHINENI\AppData\Local\Temp\ipykernel_22152\2192871814.py:6:
FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(x="Survived", data=titanic, palette="Set2")
```


Overall Survival Count
500
400
300
200
100
0
count
0
1
Survived


Survival by Gender
Survived
0
1
400
300
200
100
0
count
male
female
Sex


Survival by Passenger Class
Survived
0
1
count
350
300
250
200
150
100
50
0
1
2
3
Pclass


Age Distribution by Survival
Survived
0
1
Count
Age
0
50
100
150
200
250
0
10
20
30
40
50
60
70
80


Correlation Heatmap
Survived
Pclass
Age
SibSp
Parch
Fare
1
-0.34
-0.065
-0.035
0.082
0.26
-0.34
1
-0.34
0.083
0.018
-0.55
-0.065
-0.34
1
-0.23
-0.17
0.097
-0.035
0.083
-0.23
1
0.41
0.16
0.082
0.018
-0.17
0.41
1
0.22
0.26
-0.55
0.097
0.16
0.22
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

[8]:
```
print("\nKey Insights:")
print("1. Females had a much higher survival rate compared to males.")
print("2. Passengers in 1st class had better survival chances than 2nd and 3rd␣
 ↪class.")
print("3. Younger passengers (children) had higher survival rates.")
print("4. Gender and Passenger Class were the strongest factors affecting␣
 ↪survival.")
```

```
Key Insights:
1. Females had a much higher survival rate compared to males.
2. Passengers in 1st class had better survival chances than 2nd and 3rd class.
3. Younger passengers (children) had higher survival rates.
4. Gender and Passenger Class were the strongest factors affecting survival.
```

[ ]: