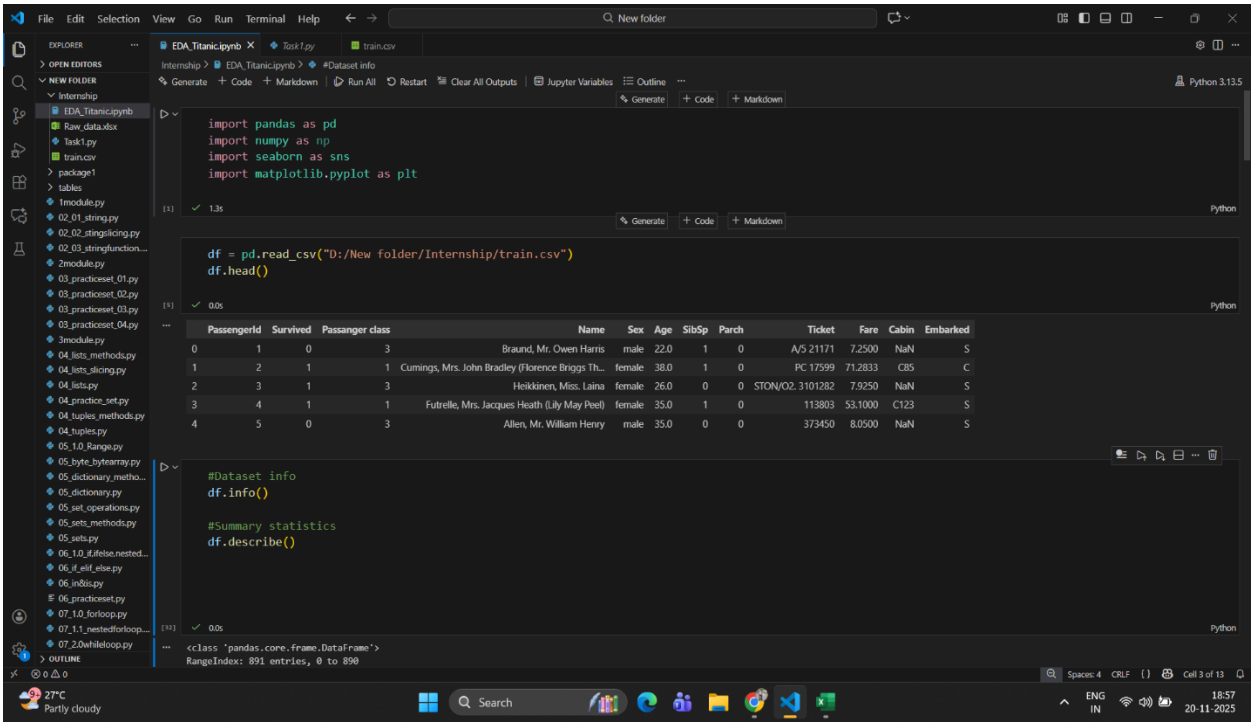
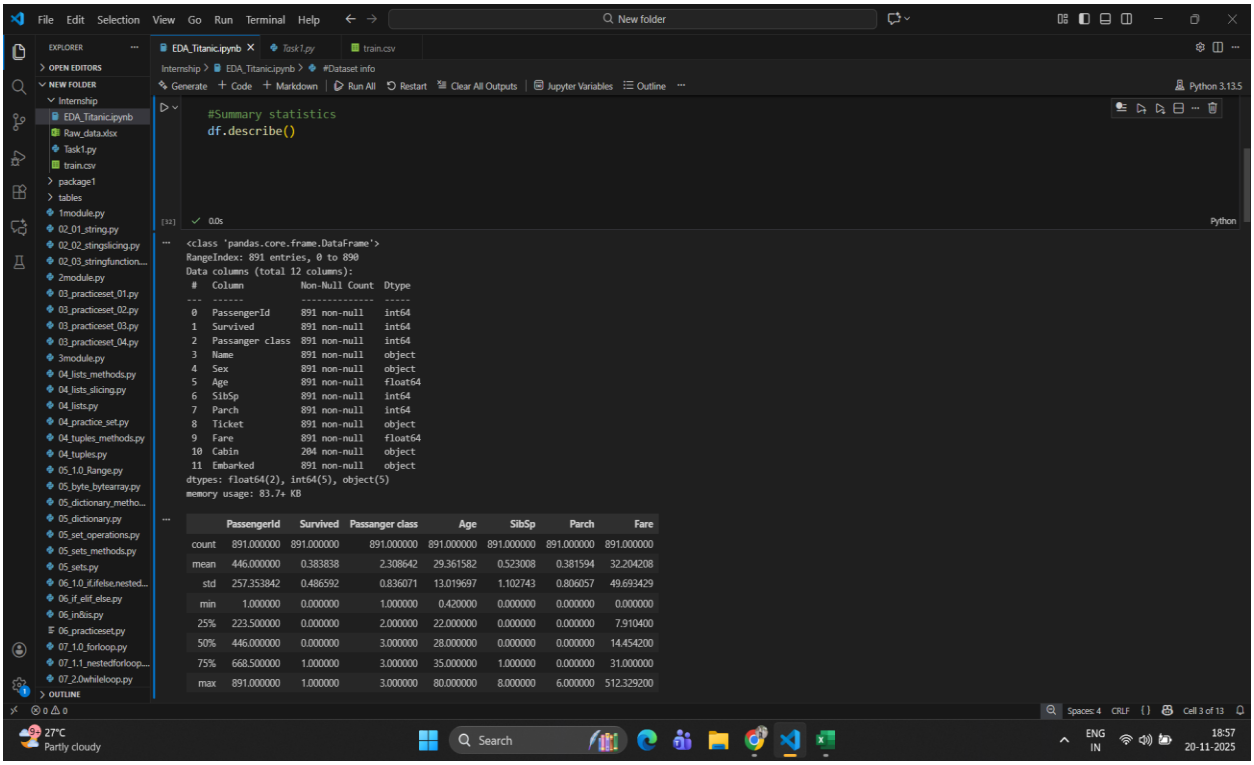
Titanic Dataset Exploratory Data Analysis (EDA) Using Jupyter Notebook

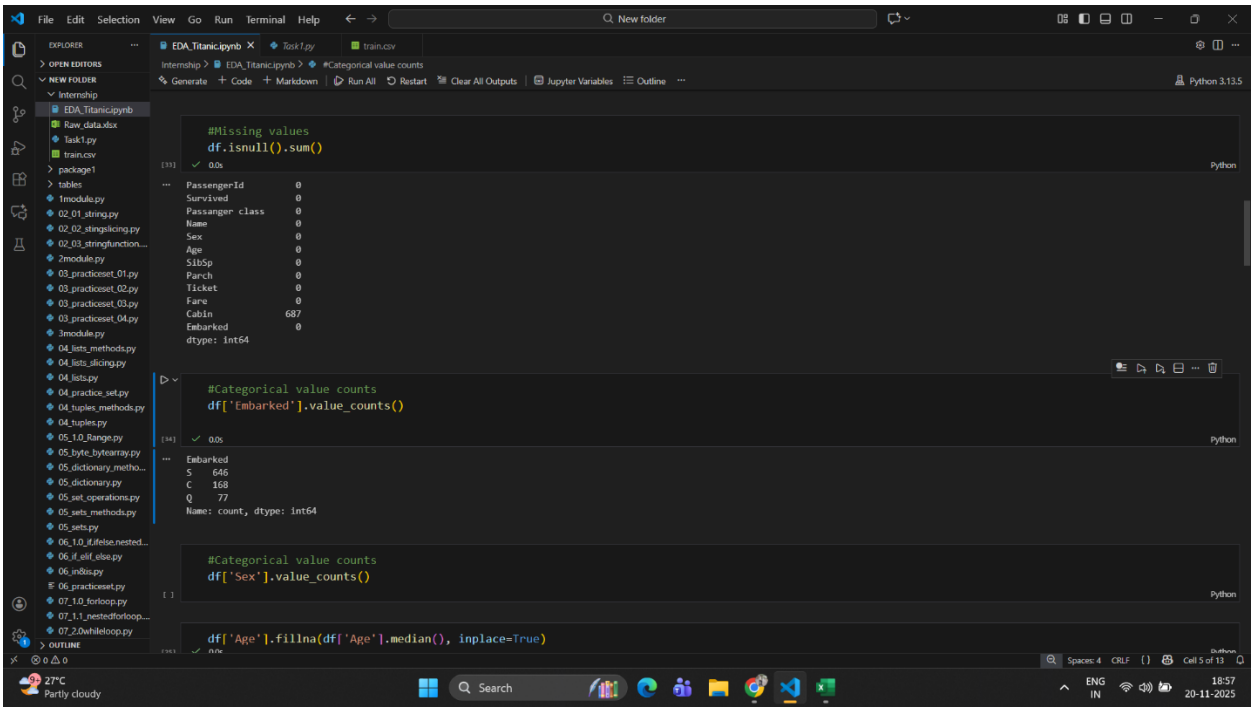
I created a Jupyter Notebook file named **EDA_Titanic.ipynb** in VS Code and wrote Python scripts to perform exploratory data analysis on the Titanic dataset (**train.csv**). I imported essential libraries such as Pandas, NumPy, Matplotlib, and Seaborn to process and visualize the data.

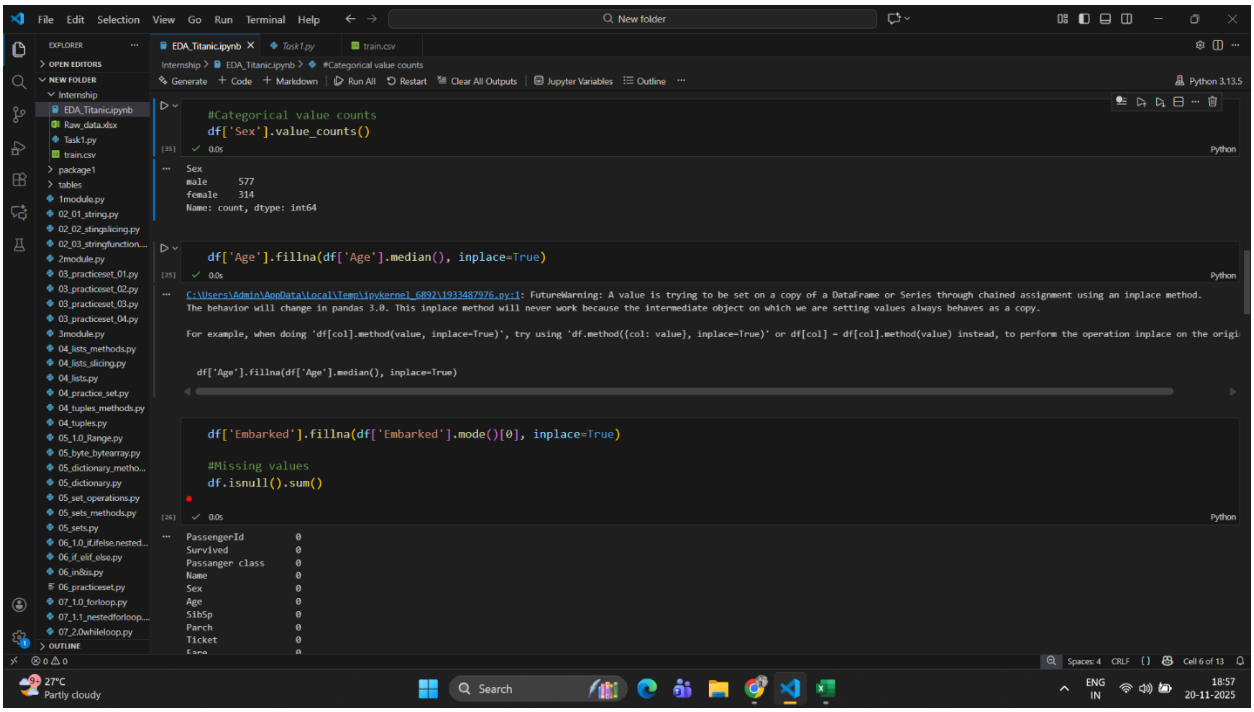
I began the analysis by loading the dataset and using functions like head(), info(), and isnull().sum() to understand the structure of the data and identify missing values. I handled null values by applying appropriate data-cleaning techniques: **missing Age values were replaced with the median**, and the **Embarked column was filled using the mode**. I also examined categorical value counts for features such as **Sex** and **Embarked** to understand their distributions.

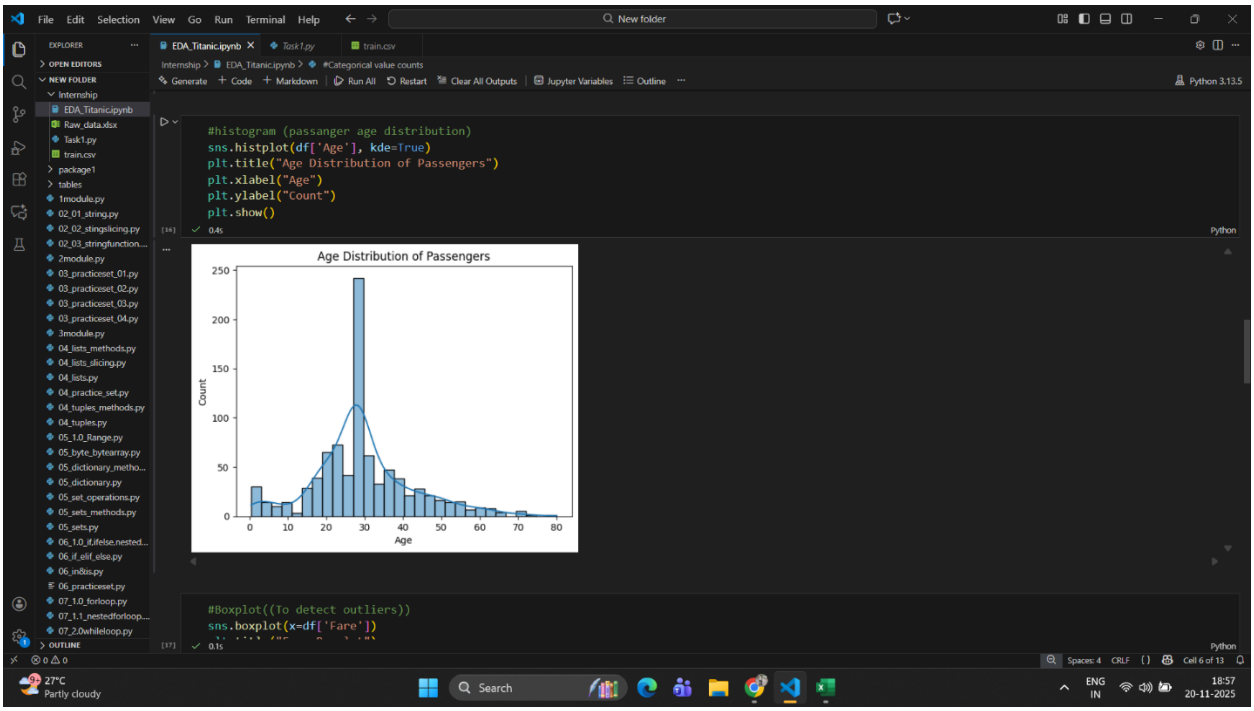
After data cleaning, I performed a series of visualisations to explore patterns and trends within the dataset. These included a **Histogram** for passenger age distribution, a **Box Plot** to detect outliers, a **Scatter Plot** for Age vs Fare, a **Count Plot** for Gender vs Survival, and a **Bar Plot** to compare survival rates across passenger classes. These analyses helped uncover key insights about passenger demographics, fare variation, and survival patterns.

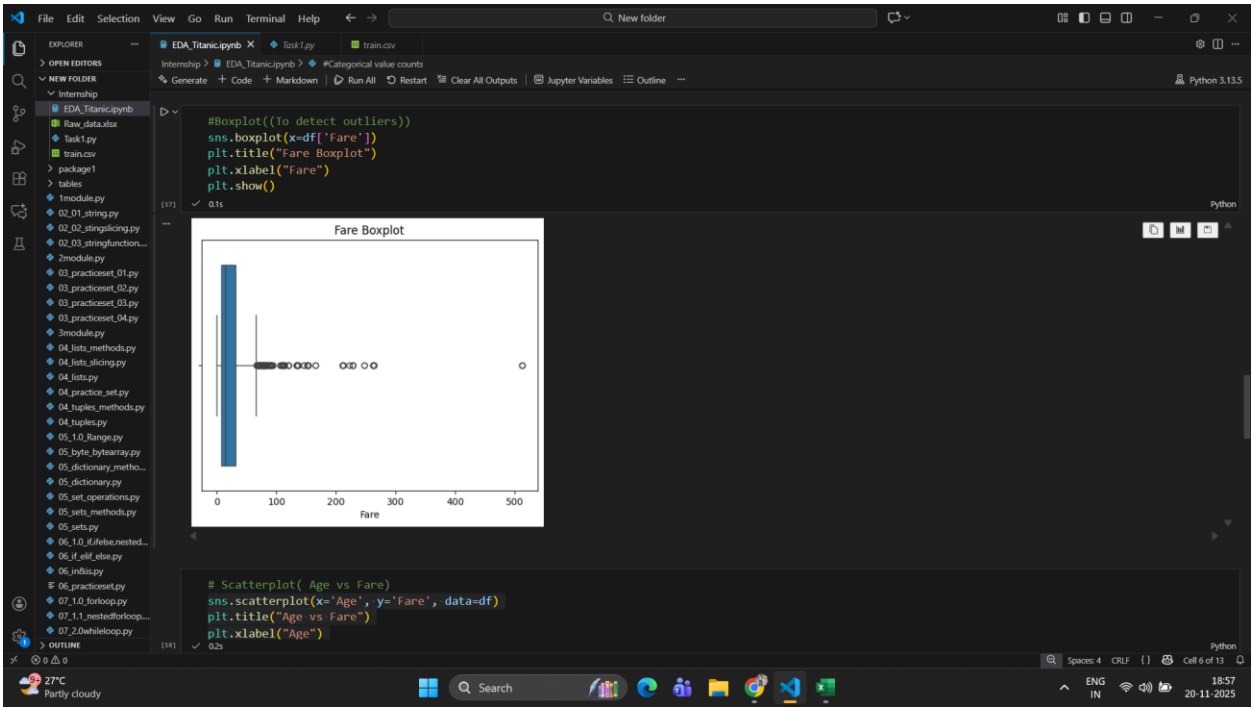

File Edit Selection View Go Run Terminal Help
New folder
EDA_Titanic.ipynb
Task1.py
train.csv
Internship > EDA_Titanic.ipynb > #Dataset info
Generate + Code + Markdown | Run All | Restart | Clear All Outputs | Jupyter Variables | Outline
Python 3.13.5
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
1.3s
Python
df = pd.read_csv("D:/New folder/Internship/train.csv")
df.head()
0.0s
Python
PassengerId Survived Passanger class 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
#Dataset info
df.info()
#Summary statistics
df.describe()
0.0s
Python
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Spaces: 4 CRLF Cell 3 of 13
27°C
Partly cloudy
Search
ENG
IN
18:57
20-11-2025

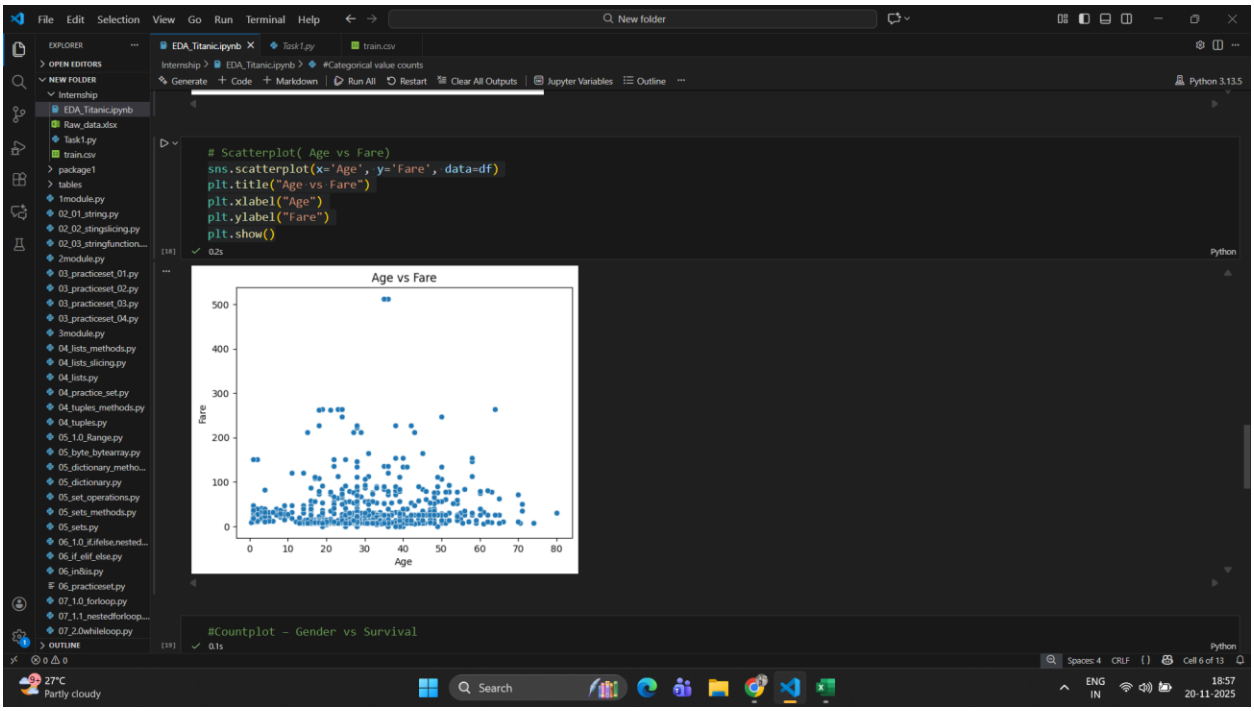

File Edit Selection View Go Run Terminal Help
New folder
EDA_Titanic.ipynb
Task1.py
train.csv
Internship > EDA_Titanic.ipynb > #Dataset info
Generate + Code + Markdown | Run All | Restart | Clear All Outputs | Jupyter Variables | Outline
Python 3.13.5
#Summary statistics
df.describe()
0.0s
Python
<class '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 Passanger class 891 non-null int64
3 Name 891 non-null object
4 Sex 891 non-null object
5 Age 891 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 891 non-null object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
PassengerId Survived Passanger class Age SibSp Parch Fare
count 891.000000 891.000000 891.000000 891.000000 891.000000 891.000000 891.000000
mean 446.000000 0.383838 2.308642 29.361582 0.523008 0.381594 32.204208
std 257.353842 0.486592 0.836071 13.019697 1.102743 0.806057 49.693429
min 1.000000 0.000000 1.000000 0.420000 0.000000 0.000000 0.000000
25% 223.500000 0.000000 2.000000 22.000000 0.000000 0.000000 7.910400
50% 446.000000 0.000000 3.000000 28.000000 0.000000 0.000000 14.454200
75% 668.500000 1.000000 3.000000 35.000000 1.000000 0.000000 31.000000
max 891.000000 1.000000 3.000000 80.000000 8.000000 6.000000 512.329200
Spaces: 4 CRLF Cell 3 of 13
27°C
Partly cloudy
Search
ENG
IN
18:57
20-11-2025

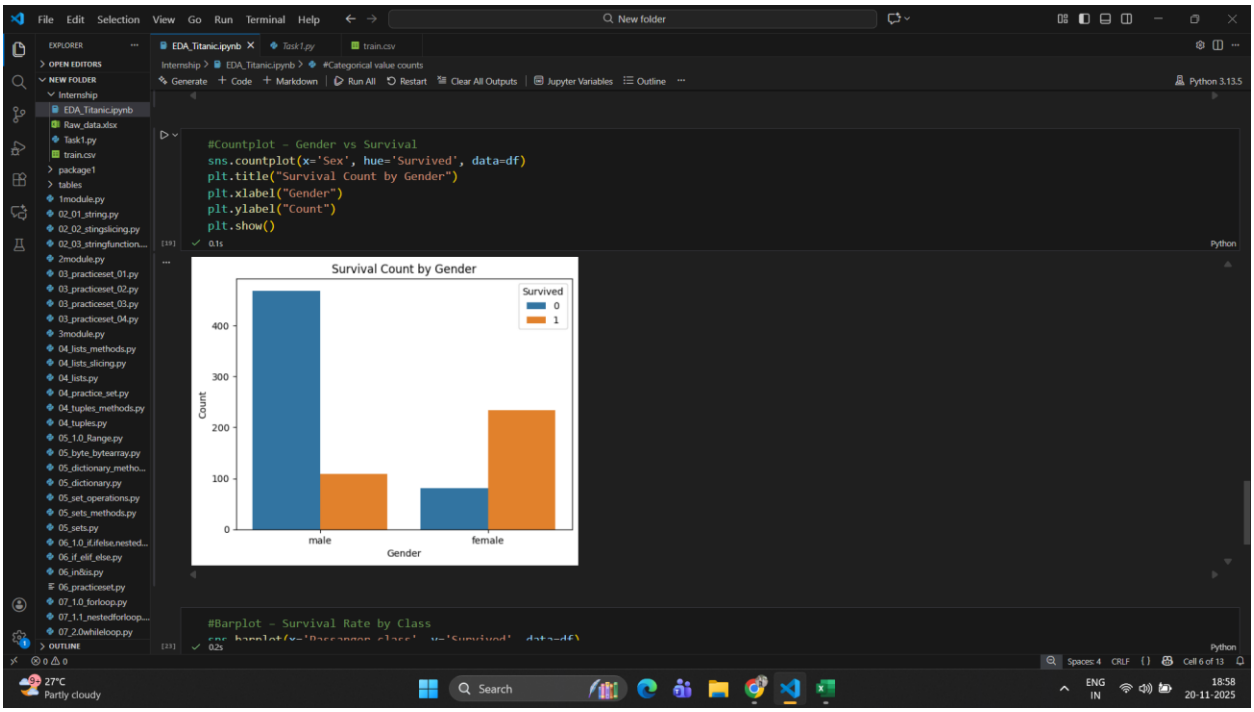

#Missing values
df.isnull().sum()
PassengerId 0
Survived 0
Passanger class 0
Name 0
Sex 0
Age 0
SibSp 0
Parch 0
Ticket 0
Fare 0
Cabin 687
Embarked 0
dtype: int64
#Categorical value counts
df['Embarked'].value_counts()
Embarked
S 646
C 168
Q 77
Name: count, dtype: int64
#Categorical value counts
df['Sex'].value_counts()
df['Age'].fillna(df['Age'].median(), inplace=True)

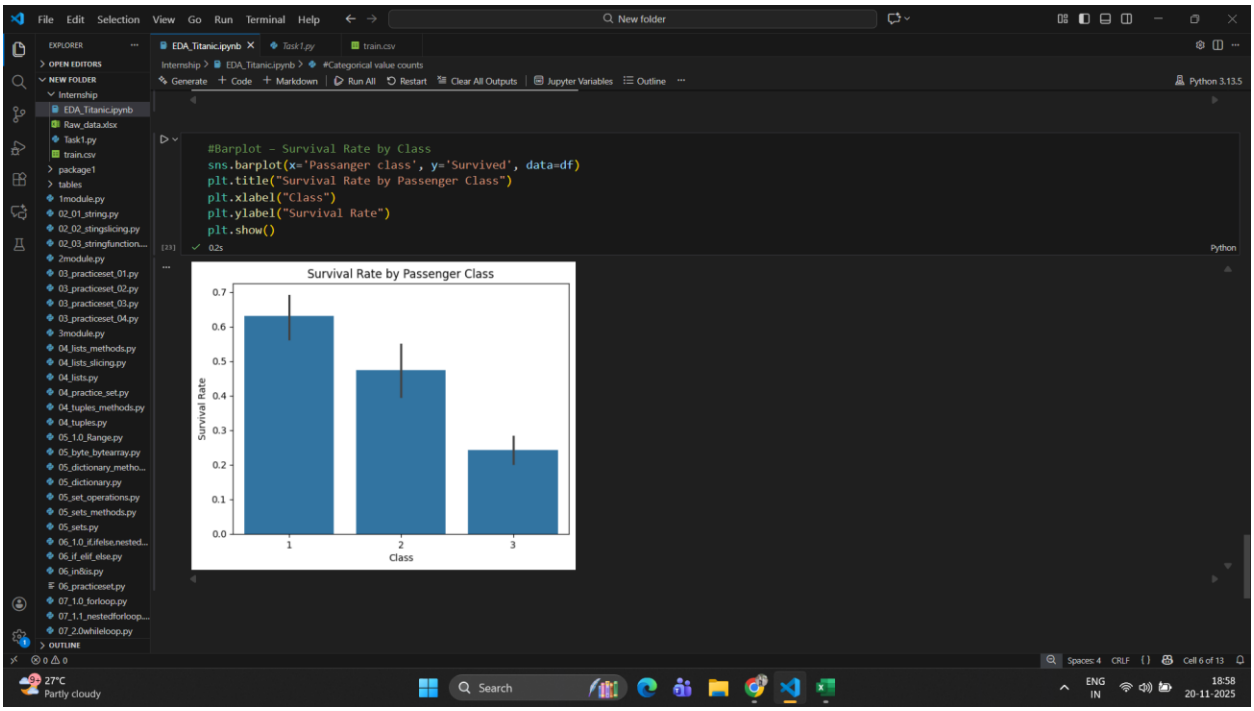

#Categorical value counts
df['Sex'].value_counts()
Sex
male 577
female 314
Name: count, dtype: int64
df['Age'].fillna(df['Age'].median(), inplace=True)
C:\Users\Admin\AppData\Local\Temp\ipykernel_6892\1933487976.py:1: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through 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 origi
df['Age'].fillna(df['Age'].median(), inplace=True)
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
#Missing values
df.isnull().sum()
PassengerId 0
Survived 0
Passanger class 0
Name 0
Sex 0
Age 0
SibSp 0
Parch 0
Ticket 0


#histogram (passanger age distribution)
sns.histplot(df['Age'], kde=True)
plt.title("Age Distribution of Passengers")
plt.xlabel("Age")
plt.ylabel("Count")
plt.show()
Age Distribution of Passengers
Count
Age
#Boxplot((To detect outliers))
sns.boxplot(x=df['Fare'])


#Boxplot((To detect outliers))
sns.boxplot(x=df['Fare'])
plt.title("Fare Boxplot")
plt.xlabel("Fare")
plt.show()
Fare Boxplot
Fare
# Scatterplot(Age vs Fare)
sns.scatterplot(x='Age', y='Fare', data=df)
plt.title("Age vs Fare")
plt.xlabel("Age")


File Edit Selection View Go Run Terminal Help
New folder
EDA_Titanic.ipynb
Task1.py
train.csv
Internship > EDA_Titanic.ipynb > #Categorical value counts
Generate + Code + Markdown Run All Restart Clear All Outputs Jupyter Variables Outline
Python 3.13.5
# Scatterplot(Age vs Fare)
sns.scatterplot(x='Age', y='Fare', data=df)
plt.title("Age vs Fare")
plt.xlabel("Age")
plt.ylabel("Fare")
plt.show()
0.2s
Python
Age vs Fare
Fare
Age
0
100
200
300
400
500
0
10
20
30
40
50
60
70
80
#Countplot - Gender vs Survival
0.1s
Python
Spaces: 4 CRLF Cell 6 of 13
27°C
Partly cloudy
Search
ENG
IN
18:57
20-11-2025


File Edit Selection View Go Run Terminal Help
New folder
EDA_Titanic.ipynb
Task1.py
train.csv
Internship > EDA_Titanic.ipynb > #Categorical value counts
Generate + Code + Markdown Run All Restart Clear All Outputs Jupyter Variables Outline
Python 3.13.5
#Countplot - Gender vs Survival
sns.countplot(x='Sex', hue='Survived', data=df)
plt.title("Survival Count by Gender")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.show()
0.1s
Python
Survival Count by Gender
Survived
0
1
Count
Gender
male
female
0
100
200
300
400
#Barplot - Survival Rate by Class
0.2s
Python
Spaces: 4 CRLF Cell 6 of 13
27°C
Partly cloudy
Search
ENG
IN
18:58
20-11-2025


File Edit Selection View Go Run Terminal Help
New folder
EDA_Titanic.ipynb
Task1.py
train.csv
Internship > EDA_Titanic.ipynb > #Categorical value counts
Generate + Code + Markdown Run All Restart Clear All Outputs Jupyter Variables Outline
Python 3.13.5
#Barplot – Survival Rate by Class
sns.barplot(x='Passanger class', y='Survived', data=df)
plt.title("Survival Rate by Passenger Class")
plt.xlabel("Class")
plt.ylabel("Survival Rate")
plt.show()
0.2s
Python
Survival Rate by Passenger Class
Survival Rate
Class
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
Spaces: 4 CRLF Cell 6 of 13
27°C
Partly cloudy
Search
ENG
IN
18:58
20-11-2025