DATA ANALYST INTERNSHIP

TASK 5

Exploratory Data Analysis (EDA)

Titanic Dataset

Submitted By,

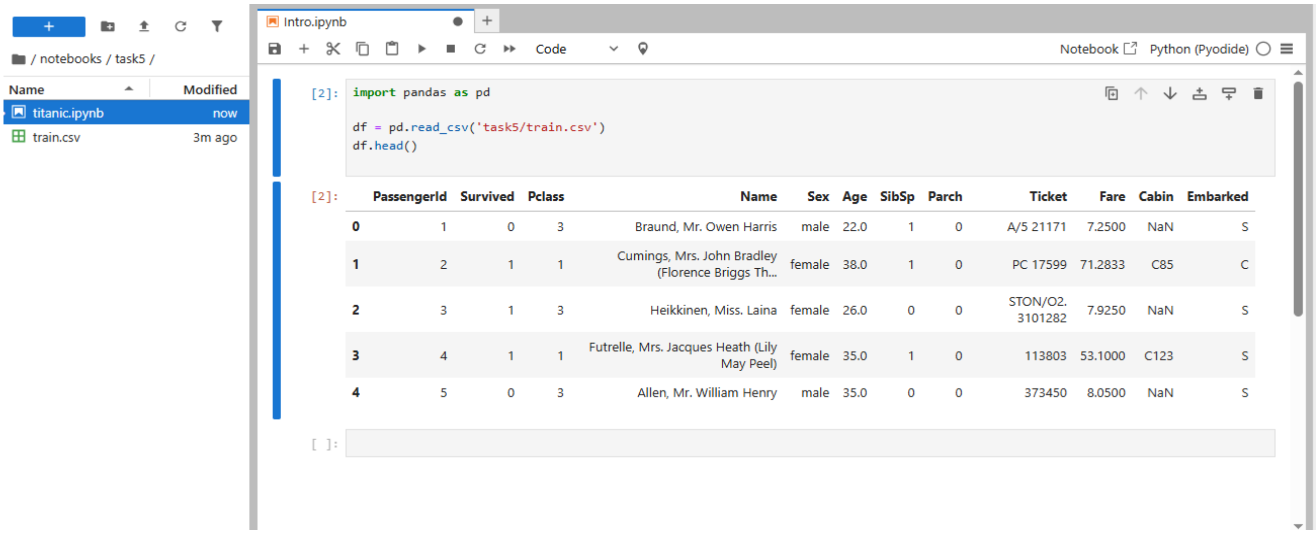
M.B.Arthi

OBJECTIVE:

The purpose of this exercise is to conduct Exploratory Data Analysis (EDA) of the Titanic data to reveal patterns, trends, and relationships in the data. Through the use of statistical summaries and visualizations, the aim is to:

- Observe the data structure and distribution.
- Determine the most significant factors affecting passenger survival.
- Identify missing values and data anomalies.
- Uncover relationships between variables through plots and correlation measures.
- Extract meaningful insights that can inform additional predictive modeling and feature selection.

This exercise assists in building core competencies in data exploration, visualization, and critical thinking—basics for any data scientist or data analyst.




```
import pandas as pd

df = pd.read_csv('task5/train.csv')
df.head()
```

	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

The Titanic data was loaded into a Jupyter Notebook successfully with the pandas library. The file was loaded from the relative path 'tasks/train.csv', and the first five rows were shown using df.head(). This preview verifies the dataset has necessary features like PassengerId, Survived, Pclass, Name, Sex, Age, Fare, and more. The data seems correctly structured and is ready to be further explored with data analysis.

```
[4]: df.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: 66.2+ KB
```

The df.info() method was employed to analyze the Titanic dataset structure. It shows that the dataset has 891 rows and 12 columns. The majority of the columns are filled with complete data, while Age, Cabin, and Embarked columns have missing values. The dataset consists of both numerical data types like Age, Fare, and SibSp and categorical data types like Sex, Embarked, and Name. This information is necessary to determine preprocessing requirements like missing value handling and data type conversion for analysis.

[5]:
```
df.describe()
```

[5]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	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	20.125000	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	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

The `df.describe()` result gives summary statistics for the numerical columns in the Titanic dataset.

The typical age of passengers is approximately 26.7, ranging from 80 years old to 0.42 years old.

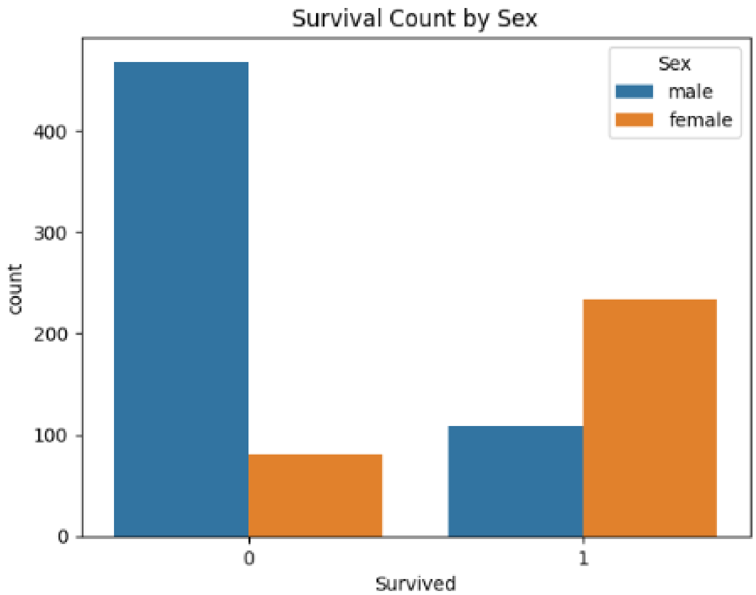
The mean fare is approximately 32.20, with a high of 512.33, reflecting a skewed distribution.

There are only 714 age values, pointing out missing data that can be imputed.

```
df.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

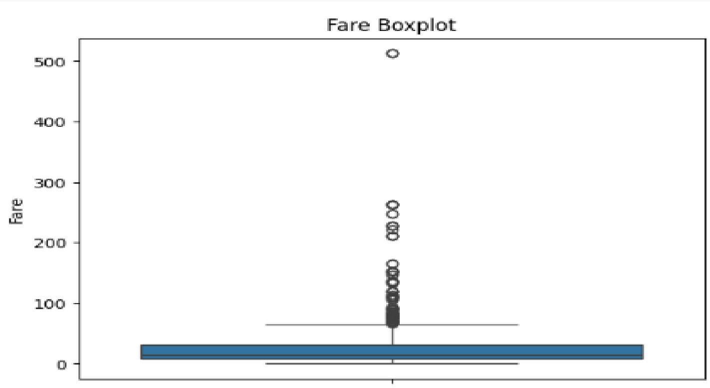
```python
sns.countplot(x='Survived', hue='Sex', data=df)
plt.title("Survival Count by Sex")
plt.show()
```



This bar plot shows the survival count of passengers categorized by sex.It uses `sns.countplot` with 'Survived' on the x-axis and 'Sex' as the hue from the DataFrame `df`.0 means not survived, and 1 means survived.More females survived compared to males, while more males did not survive.

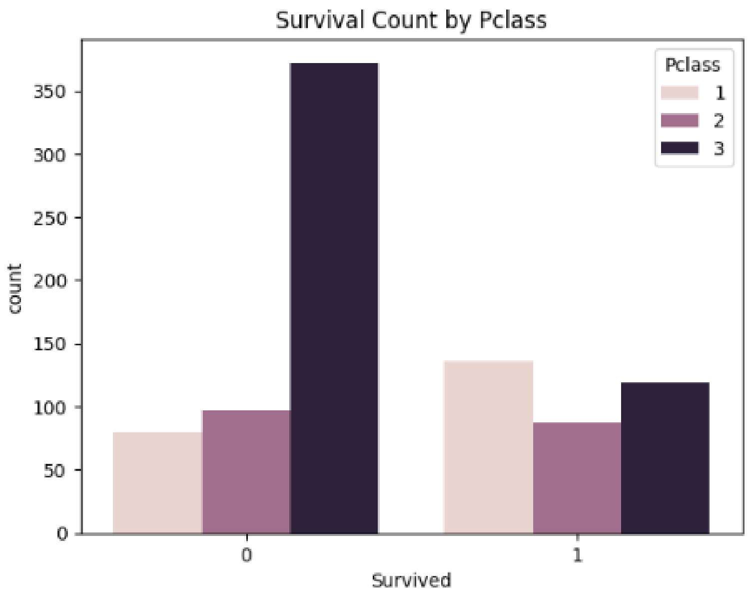
```python
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(y='Fare', data=df)
plt.title("Fare Boxplot")
plt.show()
```



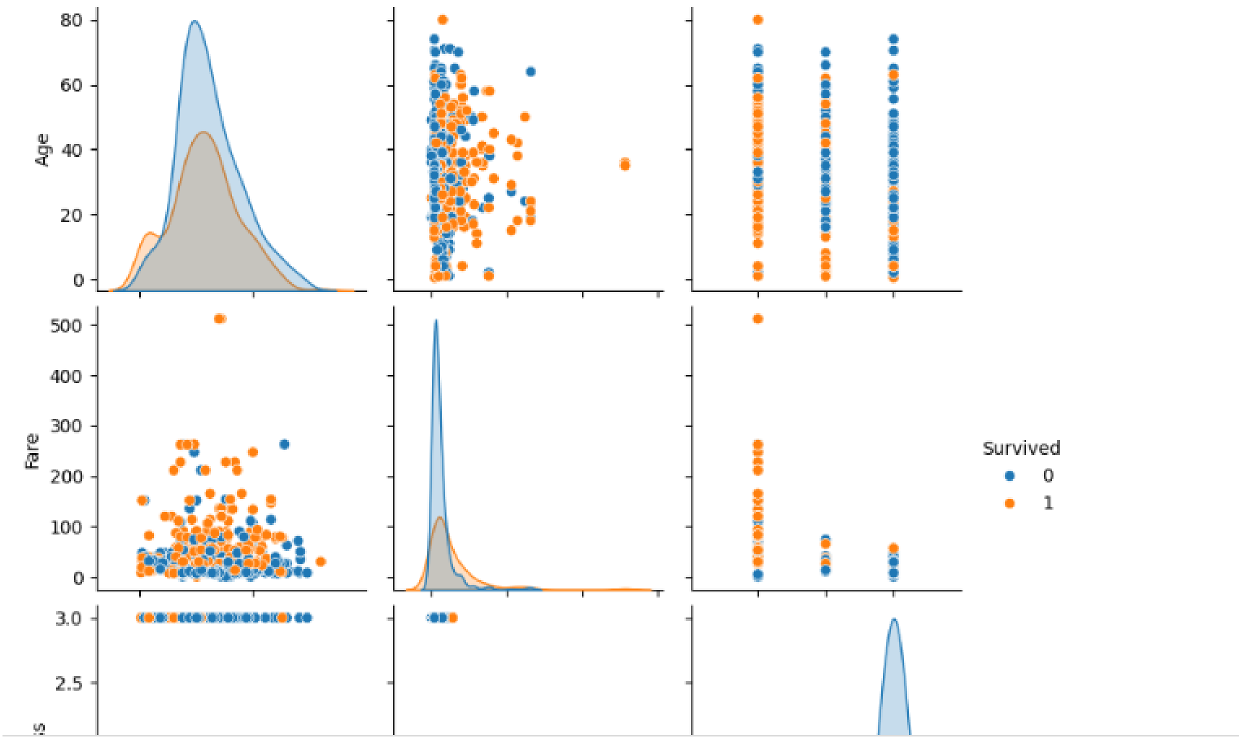
This boxplot displays the distribution of the 'Fare' column from the DataFrame df. It shows the median, interquartile range (IQR), and several outliers as dots above the box.Most fares are concentrated below 100, with a few extreme values above 200. The plot helps identify fare variability and detect potential outliers.

```
sns.countplot(x='Survived', hue='Pclass', data=df)
plt.title("Survival Count by Pclass")
plt.show()
```



This countplot shows survival counts grouped by passenger class (Pclass). Pclass 1 had the highest number of survivors, while Pclass 3 had the highest number Of non-survivors. It indicates a correlation between higher class and higher survival rate. The plot is created using Seaborn with 'Survived' on the x-axis and Pclass as the hue.

```python
# Pairplot for relationships
sns.pairplot(df[['Survived', 'Age', 'Fare', 'Pclass']].dropna(), hue='Survived')
plt.show()
```

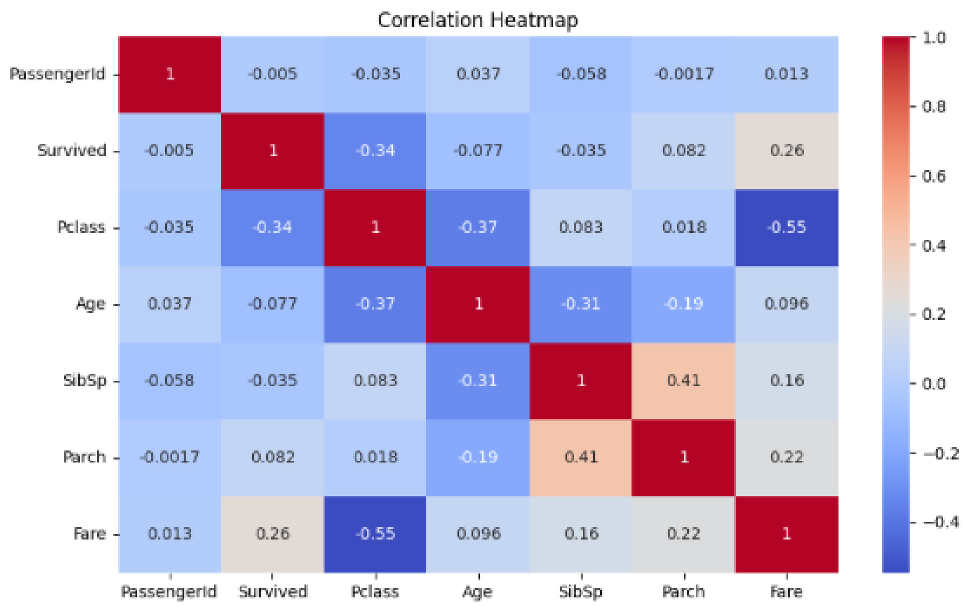


This pairplot visualizes relationships between 'Age', 'Fare', and 'Pclass' with survival status as hue.Orange dots represent survivors (1), and blue dots represent non-survivors (0).Survivors tend to be younger and often paid higher fares (likely higher class).Diagonal plots show distribution; off-diagonals show variable relationships by survival

```
import seaborn as sns
import matplotlib.pyplot as plt

# Select only numeric columns for correlation
numeric_df = df.select_dtypes(include=['number'])

# Create heatmap
plt.figure(figsize=(10,6))
sns.heatmap(numeric_df.corr(), annot=True, cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```



This heatmap shows correlation values between numeric features in the dataset. Strongest negative correlation is between Pclass and Fare (-0.55), meaning higher class passengers paid lower fares.Survived is positively correlated with Fare (0.26) and weakly with Parch and SibSp.Diagonal values are 1, showing perfect self-correlation; color intensity reflects strength and direction of correlations.

Tools Used:

- Python
- Pandas
- Matplotlib
- Seaborn
- Jupyter Notebook

ANALYSIS

- **Age Distribution**

The age distribution is right-skewed, with most passengers aged between 20 and 40**.**

There are also a significant number of children and a few elderly passengers.

- **Fare Distribution**

Fare values are highly skewed to the right, with most fares below 100 but some exceeding **500**, indicating a few high-paying passengers (likely first class).

- **Survival by Sex**

A higher percentage of females survived compared to males.

This aligns with the "women and children first" evacuation policy.

- **Survival by Passenger Class**

First-class passengers had a much higher survival rate than those in second and third class.

Third-class passengers had the lowest survival rate overall.

- **Correlation Heatmap**

Pclass and Fare show a moderate negative correlation, indicating higher class passengers paid more.

Fare and Survival show a weak positive correlation, suggesting wealthier passengers had slightly better chances.

SibSp and Parch show a mild correlation — larger families were often traveling together.

SUMMARY OF FINDINGS:

Here in this Exploratory Data Analysis of the Titanic dataset,We tried different attributes that had an impact on the survival of passengers.

I. Sex and Pclass were the most significant determinants of survival, as females and first-class passengers tended to survive more likely.

II. Fare distribution confirmed skewness, as the richer passengers tend to be in higher classes.

III. Age confirmed a broad distribution, with lower-age passengers (children) having relatively higher chances of survival.

IV. Correlation analysis showed moderate correlations of Fare, Pclass, and Survival, whereas most other features correlated lowly in a linear manner.

These findings indicate that survival was significantly influenced by social and ec onomic status, as well as gender and age.This analysis forms a foundation for feature selection and preprocessing of machine learning models for predicting Titanic survival.