Task 5: Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) of the Titanic dataset. The goal of this task is to uncover patterns, relationships, and insights that can inform future modeling efforts — particularly to understand the factors influencing passenger survival.

The dataset used is TRAIN_1.csv, which includes demographic, socioeconomic, and travel information for each passenger aboard the Titanic.

Dataset Overview

The dataset contains the following key variables:

Variable	Description
Survived	0 = No, 1 = Yes
Pclass	Ticket class (1st = Upper, 2nd = Middle, 3rd = Lower)
Sex	Gender
Age	Age in years
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Fare	Passenger fare
Embarked	Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Other columns like Ticket, Cabin, and Name were not used directly in this EDA.

Data Summary and Missing Values

Dataset Info

- **Total entries**: 891
- **Features**: 12
- **Data types**: Mixture of numerical and categorical
- **Missing values** are primarily found in:

Feature	Missing Count
Age	177
Cabin	687
Embarked	2

Statistical Summary (Selected Features)

Feature	Mean	Std Dev	Min	25%	50%	75%	Max
Age	29.70	14.52	0.42	20.12	28.00	38.00	80.00
Fare	32.20	49.69	0.00	7.91	14.45	31.00	512.33
SibSp	0.52	1.10	0	0	0	1	8
Parch	0.38	0.81	0	0	0	0	6

Univariate Visual Analysis

Distribution of Age

- The Age distribution is **right-skewed**, with the majority of passengers between 20 and 40 years old.
- Children and infants also form a noticeable group.

Passenger Class Count

- Most passengers traveled in **3rd class**, indicating that the dataset is skewed toward lower socioeconomic passengers.

Survival Count

- Survival was low: **~38% of passengers survived**, while the majority perished.

Gender Count

- About **65% were male**, and **35% were female**, a significant skew which is important for modeling.

Bivariate & Multivariate Visual Analysis

Survival by Passenger Class

- Clear trend: **higher class → higher survival**.
    - **1st class** passengers had the best survival outcomes.
    - **3rd class** had the lowest, despite having the most passengers.

Correlation Heatmap

A heatmap of correlations among selected numerical features reveals:

Feature Pair	Correlation
Survived vs Pclass	-0.34
Survived vs Fare	+0.26
Age vs Pclass	-0.37

- Negative correlation between Pclass and Survived confirms the importance of socioeconomic status.
- Fare has a mild positive correlation with survival.
- Age has no strong linear relationship with survival.

Pairplot of Key Features

A **Seaborn pairplot** was created using the following numerical variables:
['Survived', 'Pclass', 'Age', 'SibSp', 'Parch', 'Fare']

Insights:

- Survivors cluster more in **lower Pclass** (1st), **higher Fare**, and **younger ages**.
- Overlapping clusters are present, indicating the challenge of linearly separating survivors from non-survivors.

Key Insights & Observations

- **Survival was influenced strongly by class and gender.**
    - 1st class passengers and females were more likely to survive.
- **Age was not a strong predictor on its own**, but combining it with other features could help.

- **Fare correlated with both class and survival**, suggesting it could be a useful engineered feature.
- **Missing values**, particularly in Age and Cabin, must be addressed in preprocessing.

Conclusion

This exploratory analysis provides a foundational understanding of the Titanic dataset, especially the features most associated with survival. It highlights the importance of passenger class, gender, and fare, and sets the stage for further modeling through feature engineering and machine learning.