Task 5: Exploratory Data Analysis (EDA)

The objective of this task was to explore the Titanic dataset and uncover the factors that influenced passenger survival. Using Python libraries such as Pandas, Matplotlib, and Seaborn, I aimed to analyze the data, clean it, and visualize key patterns that could help explain how social, demographic, and economic variables affected outcomes.

The Titanic dataset contains information about each passenger, including their age, gender, ticket class, fare, and port of embarkation. It also includes the survival status of each individual, making it a rich dataset for identifying relationships between different variables.
I began by examining the basic structure of the dataset using functions like .info(), .describe(), and .isnull().sum(). This initial inspection helped me understand the number of rows and columns, data types, and the presence of missing values that needed attention before analysis.

There were some missing values, particularly in the Age and Embarked columns.
To make the dataset more complete, I filled the missing Age values with the mean age and replaced missing entries in the Embarked column with the most frequent value, which was S .
After cleaning, I added two new columns one to identify whether a passenger was an adult (Age greater than 18), and another to categorize passengers into different fare ranges.
These transformations made it easier to visualize and interpret trends across different groups.

When plotting the age distribution, it became evident that most passengers were between twenty and forty years old, showing that young adults formed the largest age group on board. A quick look at the gender breakdown revealed that around two-thirds of the passengers were male, while about one-third were female. This imbalance later helped explain differences in survival rates.

The gender-based survival analysis showed a very clear pattern women had a much higher chance of survival compared to men. This directly reflected the women and children first evacuation principle that was followed during the disaster. When I looked at survival rates across passenger classes, a strong class divide appeared. First-class passengers had the highest survival rates, while third-class passengers had the lowest. This suggested that access to safety was not equal and was likely influenced by social and economic position.

The age factor showed a similar trend. Children below eighteen years of age had slightly higher chances of survival compared to adults, reinforcing the idea that younger passengers were given priority during rescue operations. When exploring the relationship between fare and survival, it became apparent that passengers who paid higher fares generally had better survival rates. This aligned with the earlier observation that wealthier passengers, who mostly travelled in first class, had better access to lifeboats and safety measures.

Through this exploration, several clear insights emerged. Most passengers were young adults, but their survival was heavily dependent on their gender and class. Women and children were more likely to survive than men, and those who travelled in higher classes had better chances of making it out alive. Fare, age, and class were closely interlinked in determining survival outcomes.

Overall, this analysis helped me understand how to approach real-world datasets from cleaning and transforming raw information to drawing meaningful insights through visualization. Beyond the technical process, it was fascinating to see how numbers could tell a human story of inequality and chance. Working on this task gave me a practical understanding of how exploratory data analysis bridges the gap between data and narrative, turning raw figures into patterns that reveal something deeper about human experiences.