Titanic Dataset – Exploratory Data Analysis (EDA) Report

The objective of this task is to perform a comprehensive Exploratory Data Analysis (EDA) on the Titanic dataset to uncover key patterns and relationships that influenced passenger survival. Using Python libraries like Pandas, Matplotlib, and Seaborn, the analysis aims to explore the distribution of variables, detect missing values and outliers, and examine how features such as age, gender, class, and fare relate to survival. This process helps build a deeper understanding of the data and prepares it for further preprocessing and predictive modeling.

1. Dataset Description

The Titanic dataset contains information about 418 passengers aboard the RMS Titanic. Each entry corresponds to a unique passenger and includes demographic, travel-related, and socio-economic information. The dataset aims to help understand which factors contributed to a passenger's likelihood of survival in the maritime disaster of 1912.

The dataset includes both categorical and numerical features:

- **PassengerId, Survived, Pclass**
- **Name, Sex, Age**
- **SibSp, Parch, Ticket, Fare**
- **Cabin, Embarked**

2. Observations from Visual Analysis

- `df.head()`
  ➤ Returns the first five rows of the dataset to get a quick overview.
- `df.info()`
  ➤ Displays column names, data types, and non-null counts—useful for

detecting missing values and data types.

- `df.describe(include='all')`
  ➤ Provides summary statistics for both numerical and categorical columns (like mean, std, unique, top, etc.).

- `df.isnull().sum()`
  ➤ Shows the total number of missing (null) values per column.

- `df.nunique()`
  ➤ Counts the number of unique values in each column—helps identify categorical columns or ID-like fields.

- `df['Embarked'].value_counts()`
  ➤ Shows the frequency of each category in the 'Embarked' column—useful for understanding distribution.

These functions are foundational in understanding the structure and quality of the dataset.

🔹 Age Distribution

- Most passengers were between **20 and 40 years old**.
- There were several children and elderly passengers.
- The distribution is **right-skewed**.
- Missing values (~20%) in Age need to be imputed.

🔹 Fare Distribution

- Most passengers paid fares between **0 and 50**.
- A small number of passengers paid **very high fares** (outliers), likely from 1st class.
- Distribution is heavily **right-skewed**.

🔹 Passenger Class (Pclass)

- **3rd class** had the largest number of passengers.
- **1st class** passengers had the **highest survival rate**.

🔹 Gender (Sex)

- There were more **male** passengers than female.
- However, **females had a much higher survival rate** than males.

🔹 Embarked Port

- Most passengers boarded from **Southampton (S)**.
- Survival rates varied across embarkation ports, with **Cherbourg (C)** having relatively higher survival.

🔹 Boxplot of Age and Fare

- Fare showed several **extreme outliers**.
- Age had more **consistent spread**, but with some older individuals standing out.

🔹 Scatter Plot (Age vs Fare)

- Survivors tend to be clustered around **younger age and higher fare** regions.
- Non-survivors were scattered across low fare and higher age.

🔹 Pairplot (Multivariate)

- Shows **interplay between features** like Age, Fare, Pclass, SibSp, Parch.
- Clear clustering of survivors vs. non-survivors in multiple dimensions.

🔹 Correlation Heatmap

- Strong **positive correlation** between Fare and Survival.
- Strong **negative correlation** between Pclass and Survival (1st class survived more).
- SibSp and Parch were also positively correlated, indicating family travel patterns.

3. Summary of Findings

- **Gender** is a major predictor: females had much higher survival rates.
- **Class** influenced survival: 1st class passengers were more likely to survive.
- **Fare** positively correlated with survival, likely due to its connection with class.
- **Most passengers** boarded from Southampton, but Cherbourg passengers had higher survival.
- **Family connections (SibSp and Parch)** had a subtle effect on survival; solo travelers were at a disadvantage.
- Missing data in **Age** and **Cabin** requires handling before modeling.

4. Conclusion

The Titanic dataset demonstrates clear patterns of inequality in survival outcomes based on class, gender, and economic status. Higher-class passengers and women were more likely to survive, while men in lower classes had the lowest survival rates. These patterns reflect social norms of the early 20th century and provide a rich foundation for predictive modeling and classification tasks.