Documentation of Exploratory Data Analysis (EDA) - Titanic Dataset

Objective:

The goal of this analysis was to explore and understand the Titanic dataset, focusing on key patterns and insights regarding passenger demographics, survival rates, and relationships between features.

Steps Taken:

1. **Data Preprocessing**:
    - **Missing Value Imputation**: Identified and imputed missing values in the dataset to ensure consistency and completeness.
        - For features like **Age** and **Embarked**, imputation was done using the **mean** (for numerical features) and the **mode** (for categorical features).
    - **Feature Engineering**:
        - Created a new feature, AgeGroup, by categorizing passengers into groups like **Child**, **Teen**, **Adult**, **Middle-Aged**, and **Senior** based on their age.
2. **Data Cleaning**:
    - **Duplicates**: Checked for and removed any duplicate rows to avoid data redundancy.
    - **Outliers**: Used boxplots to detect and remove outliers, particularly in numerical features like **Fare**.
3. **Data Transformation**:
    - **Sex Encoding**: Converted the **Sex** column into binary values (0 for male, 1 for female) for better analysis and compatibility with machine learning models.
    - **Correlation Analysis**:
        - Examined the correlation between numerical variables using a correlation matrix.
        - Found strong correlations between features like **Survived** and **Sex**, which is expected given that women had a higher survival rate than men.
4. **Visualizations**:
    - **Barplot for Survival by Gender**: Showed survival counts based on gender, revealing that women had a higher survival rate.

- **Survival by Passenger Class**: Plotted survival rates across different passenger classes, showing that the survival rate was higher for 1st-class passengers.
- **Age Group Distribution by Passenger Class**:
  - Created age groups (Child, Teen, Adult, Middle-Aged, Senior) and plotted the distribution of these groups across the three passenger classes. This revealed that younger passengers (children and teens) were more likely to be found in the 3rd class, while adults and middle-aged passengers dominated 1st and 2nd classes.
- **Survival Rate by Pclass**: Showed survival rates by passenger class, indicating that survival rates were significantly higher in the 1st class compared to 2nd and 3rd classes.
- **Fare Distribution**: Used boxplots to show the relationship between **Fare** and survival. Passengers who survived generally had higher fares.
- **Embarked Port Distribution**: Visualized how many passengers boarded from each embarkation port, showing that the majority embarked from **Southampton**.

5. **Outlier Detection and Removal**:
   - Detected outliers in the **Fare** column using boxplots and removed them to ensure that extreme values didn't skew the analysis.
6. **Categorical Data Analysis**:
   - Plotted the count of passengers from each **Embarked** port, revealing the largest group of passengers boarded from **Southampton (S)**.
   - Analyzed the distribution of **Survived** by **Embarked** port, and found that survival rates were higher for passengers who boarded at **Cherbourg (C)**.
7. **Survival Analysis by Features**:
   - Examined survival rates across different features such as **Pclass**, **Sex**, **Embarked**, and **AgeGroup**.
   - Found that the **Sex** feature had the strongest influence on survival rates, with females having a significantly higher survival rate than males.
   - Survival rates were higher for passengers in **1st class** and those who boarded at **Cherbourg**.
8. **Correlation Matrix**:

- A correlation matrix was used to identify relationships between numerical features. The most significant correlations were:
    - **Survived** and **Sex**: Strong positive correlation (females had a higher chance of survival).
    - **Survived** and **Pclass**: 1st class passengers had a higher survival rate.
    - **Fare** and **Pclass**: Passengers in higher classes generally paid higher fares.

9. **Missing Data Analysis**:
    - **Embarked** and **Age** had some missing values, which were imputed using the mode and mean, respectively.
10. **Class Distribution and Survival**:
    - Plotted survival rates across different **Pclass** and **Embarked** combinations, showing that **1st class** passengers had the highest survival rates, and those who boarded at **Cherbourg** had the highest survival rate.

Key Findings and Observations:

1. **Gender and Survival**: There is a strong correlation between **survival** and **Sex**. Females had a higher survival rate than males, which is consistent with historical accounts of the Titanic disaster, where women and children were prioritized for lifeboats.
2. **Pclass and Survival**: **1st class** passengers had a significantly higher survival rate compared to **2nd** and **3rd class** passengers. This reflects social and economic disparities, where wealthier passengers had better access to lifeboats.
3. **Age Distribution**: The age distribution across passenger classes showed that children and teenagers were primarily in the third class, while adults and middle-aged passengers were more evenly distributed across the first and second **classes**.
4. **Embarked Port**: Most passengers boarded from **Southampton**, but passengers who boarded from **Cherbourg** had a higher survival rate.

Conclusion:

The **Titanic dataset** provides a rich source of information about passenger demographics and survival rates. The analysis shows how factors such as **gender**, **passenger class**, **age**, and **fare** were significant predictors of survival. It highlights the stark class and gender-based disparities in survival, with wealthier and female passengers having higher chances of survival.

This EDA forms the foundation for further modeling, such as building classification models to predict survival based on these features.