Titanic Dataset Analysis with feature Engineering

This analysis focuses on the Titanic dataset, which contains information about passengers, their demographics, ticket class, survival status, and other attributes. The notebook systematically processes the data by importing necessary libraries, cleaning and transforming the data, and performing exploratory data analysis (EDA) to derive meaningful insights.

1. Importing Required Libraries

The first step in the notebook is to load the essential Python libraries that will be used throughout the analysis. These libraries help in data manipulation, numerical operations, and visualizations:

- **Pandas**: This library is used for data handling and processing, including reading CSV files and performing various operations on data frames.
- **NumPy**: A numerical computing library that helps with mathematical operations such as calculating the mean of a column.
- **Seaborn & Matplotlib**: These are visualization libraries that allow for data plotting to understand distributions, trends, and relationships in the dataset.
- **%matplotlib inline**: A Jupyter Notebook-specific command that ensures plots are displayed within the notebook itself rather than in a separate window.

2. Loading the Titanic Dataset

The dataset is read from a CSV (Comma-Separated Values) file. This file contains all the information about passengers aboard the Titanic. Once the dataset is loaded, the first few rows are displayed using .head(). This provides an initial glance at the structure and contents of the data, allowing us to see key columns such as:

- **Passenger ID**: A unique identifier for each passenger.
- **Survived**: Indicates whether a passenger survived (1) or not (0).
- **Pclass**: Passenger class (1st, 2nd, or 3rd class).
- **Name, Sex, Age**: Personal details of passengers.
- **SibSp & Parch**: Number of siblings/spouses and parents/children aboard.

- **Ticket, Fare, Cabin**: Ticket number, ticket fare, and assigned cabin.
- **Embarked**: The port where the passenger boarded the Titanic.

To further understand the dataset, the last few rows are displayed using .tail(), which helps verify if there are inconsistencies at the end of the dataset.

3. Understanding Data Structure and Identifying Missing Values

The .info() function provides an overview of the dataset, revealing details such as:

- The number of non-null (available) values in each column.
- The data type of each column (e.g., integer, float, or object/string).
- The presence of missing values.

To determine how many missing values exist, .isna().sum() is used. This function counts the number of missing entries in each column. Common issues in the Titanic dataset include:

- **Cabin**: Many missing values, as not all passengers had an assigned cabin.
- **Age**: Some missing values, possibly because birth dates were not recorded for every passenger.
- **Embarked**: A few missing values, indicating some passengers' boarding locations were not documented.

4. Data Cleaning

To ensure accurate analysis, the dataset must be cleaned by handling missing values and irrelevant columns. The cleaning process includes:

1. **Dropping the Cabin Column**: Since a large portion of this column is missing, it is removed from the dataset as it may not provide reliable insights.
2. **Filling Missing Age Values**: Instead of removing passengers with missing ages, the average (mean) age of all passengers is used to fill in these gaps. This method ensures that the dataset remains as complete as possible without distorting the overall statistics.

After performing these operations, .info() is used again to verify that missing values have been appropriately handled.

5. Data Transformation

To make the dataset more readable and user-friendly:

- The Survived column originally contains numerical values (0 for non-survivors and 1 for survivors). To improve readability, these values are replaced with 'No' (for non-survivors) and 'Yes' (for survivors).
- The dataset is then displayed again to confirm that the transformation was successful.

---

6. Exploratory Data Analysis (EDA)

EDA is performed to uncover trends and patterns in the data through visualization. The following key analyses are conducted:

6.1 Age Distribution Analysis

A histogram with a smooth density curve is plotted to visualize the distribution of passenger ages. This helps in understanding:

- The most common age groups on board.
- Whether the distribution is skewed (more younger or older passengers).
- The presence of any unusual age gaps.

A smooth curve (Kernel Density Estimate, or KDE) is added to the histogram to visualize the overall distribution more clearly.

6.2 Passenger Embarkation Analysis

The "Embarked" column represents the port where each passenger boarded the Titanic. The dataset uses abbreviations:

- **C (Cherbourg)**
- **Q (Queenstown)**
- **S (Southampton)**

To better understand passenger distribution:

- The number of passengers who boarded from each port is counted.
- A bar chart is created to visualize the proportion of passengers from each location.

- Labels are added to make the chart more intuitive.

This analysis helps determine which port contributed the most passengers to the voyage.

6.3 Survival Rate by Gender

One of the most important analyses in the Titanic dataset is survival rates based on gender. To visualize this:

- The number of male and female passengers who survived is counted.
- A bar chart is plotted to compare survival rates between genders.

Since historical accounts suggest that women and children were given priority in lifeboat evacuations, this analysis helps confirm that assumption using real data.

Conclusion

This Titanic dataset analysis follows a structured approach:

1. **Data Loading**: Reading and inspecting the dataset.
2. **Data Cleaning**: Handling missing values and dropping unnecessary columns.
3. **Data Transformation**: Improving readability of key columns.
4. **Exploratory Data Analysis**: Visualizing data to extract meaningful insights.

By examining survival rates by age, gender, and embarkation point, we gain valuable insights into the factors that influenced survival during the Titanic disaster. This structured workflow ensures a clear understanding of the dataset while maintaining data integrity.