Project Description: Exploratory Data Analysis (EDA) on Titanic

"Project 3: Data Analytics - Exploratory Data Analysis (EDA) on Titanic."

The primary objective of this project is to perform EDA on the Titanic dataset, with a focus on understanding the data's characteristics and relationships between features.

Key aspects and skills demonstrated in this notebook include:

- **Data Wrangling:** This involves cleaning and preparing the data for analysis.
    - **Handling Missing Values:** The notebook addresses missing values in columns like 'age' (imputing with median based on title) and 'embarked' (filling with the mode).
    - **Dropping Irrelevant/High-Null Columns:** Columns such as 'body', 'cabin', 'boat', and 'home. Dest' were dropped due to irrelevance or a high number of null values.
    - **Feature Engineering:** A new 'title' column is created by extracting titles from passenger names, which are then categorized into common groups like 'Mr.', 'Mrs.', 'Miss', 'Master', 'Officer', and 'Noble'.
    - **Handling Zero Values:** Zero values in the 'fare' column are replaced with Nan and then filled with the median fare.
    - **Removing Duplicates:** Duplicate rows in the dataset are dropped.
- **Data Visualization:** Utilizes libraries like Seaborn and Matplotlib to create basic visualizations, such as histograms and correlation heatmaps, to explore data distribution and relationships between features.
- **Basic Statistics:** Demonstrates a fundamental understanding of statistical concepts during the EDA process.

The project uses Python libraries including pandas, NumPy, matplotlib, and seaborn for data manipulation and visualization.

Dataset Description: titanic3.csv

The titanic3.csv dataset contains information about passengers on the Titanic, commonly used for predictive modeling and exploratory data analysis to determine factors associated with survival.

- **Number of Entries:** 1309 passengers

- **Number of Columns:** 14

Columns and their descriptions:

- **pclass** (int64): Passenger Class (1st, 2nd, 3rd). A proxy for socio-economic status.
- **survived** (int64): Survival status (0 = No, 1 = Yes). This is often the target variable in predictive models.
- **name** (object): Full name of the passenger.
- **sex** (object): Sex of the passenger (male/female).
- **age** (float64): Age of the passenger. Contains missing values (1046 non-null out of 1309).
- **sibsp** (int64): Number of siblings/spouses aboard the Titanic.
- **parch** (int64): Number of parents/children aboard the Titanic.
- **ticket** (object): Ticket number.
- **fare** (float64): Passenger fare. Contains one missing value.
- **cabin** (object): Cabin number. Contains a significant number of missing values (only 295 non-null).
- **embarked** (object): Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton). Contains two missing values.
- **boat** (object): Lifeboat number, if the passenger boarded one. Contains many missing values.
- **body** (float64): Body identification number for victims. Contains a very high number of missing values (only 121 non-null).
- **home.dest** (object): Home/destination of the passenger. Contains a substantial number of missing values.

This dataset is well-suited for tasks such as:

- **Exploratory Data Analysis (EDA):** To understand the distribution of variables, identify patterns, and visualize relationships between features (e.g., how pclass and sex relate to survived).
- **Feature Engineering:** Creating new features from existing ones (e.g., extracting titles from name, creating family size from sibsp and parch).

- **Missing Value Imputation:** Handling the significant number of missing values in age, cabin, embarked, boat, and body.
- **Predictive Modeling:** Building classification models (e.g., Logistic Regression, Decision Trees, Random Forests) to predict survived status.