**Titanic Dataset: Data Cleaning and Preprocessing for Machine Learning**

This repository contains a Python script for cleaning, preprocessing, and preparing the classic Titanic dataset for machine learning tasks. The goal is to transform the raw data into a clean, structured format suitable for model training.

**Project Objective**

The main objective of this project is to demonstrate a standard workflow for data preprocessing, which is a crucial first step in any machine learning pipeline. This includes handling missing data, converting data types, normalizing features, and managing outliers.

**Dataset**

The project uses the "Titanic - Machine Learning from Disaster" dataset, which is widely used for introducing machine learning concepts.

* **Source:** [Kaggle Titanic Dataset](https://www.kaggle.com/c/titanic/data)
* **File:** Titanic-Dataset.csv

The dataset provides information on passengers of the Titanic, including whether they survived or not, along with features like age, class, sex, and fare.

**Tools and Libraries**

* **Python 3.x**
* **Pandas:** For data manipulation and analysis.
* **NumPy:** For numerical operations.
* **Scikit-learn:** For feature scaling.
* **Matplotlib & Seaborn:** For data visualization, specifically for identifying outliers.

**Data Cleaning & Preprocessing Workflow**

The Python script (clean\_titanic.py) performs the following steps:

1. **Load and Explore Data:**

* The Titanic-Dataset.csv is loaded into a Pandas DataFrame.
* Initial exploration is done using .info() and .head() to understand data types, non-null counts, and the overall structure.

1. **Handle Missing Values:**

* **Age:** Missing age values are imputed using the **median** age of all passengers.
* **Embarked:** Missing port of embarkation values are filled with the **mode** (the most frequent port).
* **Cabin:** The 'Cabin' column is dropped from the dataset due to a high number of missing values.

1. **Encode Categorical Features:**

* The Sex and Embarked columns, which are categorical, are converted into numerical format using one-hot encoding with pd.get\_dummies(). This is necessary for most machine learning algorithms.

1. **Standardize Numerical Features:**

* The Age and Fare columns are standardized using StandardScaler from Scikit-learn. This scales the features to have a mean of 0 and a standard deviation of 1, preventing features with larger scales from dominating the model.

1. **Visualize and Remove Outliers:**

* Boxplots are generated for the Age and Fare columns to visually identify outliers.
* Outliers are then programmatically removed using the Interquartile Range (IQR) method. Any data point that falls outside of 1.5 times the IQR below the first quartile or above the third quartile is considered an outlier and removed.

**How to Run This Project**

1. **Clone the repository:**  
   git clone <your-repository-url>  
   cd <repository-name>
2. Install dependencies:  
   Make sure you have Python installed. Then, install the required libraries using pip.  
   pip install pandas numpy scikit-learn matplotlib seaborn
3. Place the dataset:  
   Ensure that the Titanic-Dataset.csv file is in the root directory of the project.
4. Execute the script:  
   Run the Python script from your terminal.  
   python clean\_titanic.py

**Files in This Repository**

* Titanic-Dataset.csv: The original, raw dataset.
* clean\_titanic.py: The Python script containing all the preprocessing logic.
* README.md: This file, providing an overview of the project.

**Output**

After running the script, the following files will be generated in the project directory:

* Titanic-Dataset-Cleaned.csv: The final, cleaned dataset, ready for machine learning.
* outlier\_boxplots.png: An image file containing boxplots of the 'Age' and 'Fare' columns, visualizing the data distribution before outlier removal.