Comprehensive Data Cleaning and Summary Statistics Analysis 📊

In my recent project on **Comprehensive Data Cleaning and Summary Statistics Analysis**, I aimed to meticulously clean and summarise a dataset from a CSV file to ensure its reliability for further analysis. Here's a detailed breakdown of the steps I undertook, the methodologies employed, and the significant results achieved.

Task 1: Data Cleaning and Preparation 🧹

Step 1: Load and Read the Data 📥

I started by importing the necessary libraries: pandas for data manipulation and numpy for numerical operations. I loaded the dataset using pd.read_csv(), which allowed me to read the data from a specified file path.

Step 2: Initial Data Exploration 🔍

To understand the dataset, I performed an initial exploration by using .info() to retrieve metadata about the dataset. This included checking the number of entries, data types, and any null values present.

- **Check for Duplicates**: I identified and removed any duplicate rows using .duplicated().sum() to ensure that my analysis was based on unique data.
- **Data Type Correction**: I checked and corrected the data types for relevant columns, converting Survived and Embarked to categorical types for better analysis.

Step 3: Display Initial Statistical Summary 📈

Using .describe(), I generated a statistical summary, providing insights into the distribution and characteristics of numerical columns. This helped in understanding the dataset's structure.

Step 4: Check for Missing Values 🚫

I examined the dataset for missing values using .isnull().sum(). This helped identify any necessary data cleaning actions needed for specific columns.

Step 5: Data Cleaning - Remove Missing 'Age' Values 🧹

Since Age was crucial for analysis, I opted to drop rows with missing Age values to retain only complete cases.

Step 6: Fill Missing 'Fare' Values with the Median 💰

For Fare, I filled in missing values using the median fare, which is robust to outliers and thus provided a more reliable estimate.

Step 7: Handle Missing Values for 'Embarked' and 'Cabin' 🚢

I addressed missing values for Embarked by filling with the mode (most frequent value). For Cabin, I opted to fill missing entries with 'Unknown', ensuring no null values remained.

Step 8: Check for Missing Values After Cleaning ✔️

After cleaning, I re-checked for missing values to confirm that all actions taken were effective.

Step 9: Check for Outliers Before Handling (Age) 📏

I examined outliers in the Age column using the Z-score method, identifying potential outliers that might skew my analysis.

Step 10: Check for Outliers Before Handling (Fare) 💸

Similarly, I assessed the Fare column for outliers using the Interquartile Range (IQR) method, enabling me to determine extreme values.

Step 11: Remove Outliers Using Z-Score Method (Age) ❌

I employed the Z-score method to filter outliers in the Age column, retaining only those within three standard deviations from the mean.

Step 12: Remove Outliers Using IQR Method (Fare) 📊

For the Fare column, I applied the IQR method to remove outliers, maintaining a clean dataset for analysis.

Step 13: Final Data Information 📝

Once cleaning was complete, I provided a final overview of the cleaned dataset using .info() to confirm the absence of duplicates, missing values, and the expected data types.

Step 14: Save Cleaned Data to a New CSV File 💾

I saved the cleaned dataset to a new CSV file, facilitating easy access for future analysis or sharing.

Task 2: Calculate Summary Statistics 🧮

In the second task, I computed summary statistics such as mean, median, and mode for Age and Fare. I also calculated the count of survivors, providing critical insights into the dataset's outcomes.

Step 16: Display Summary Statistics 📊

I presented the calculated summary statistics, which included:

- Mean Age: **29.70**
- Median Age: **28.00**
- Mode Age: **24.00**
- Mean Fare: **32.20**
- Count of Survived: **549 Not Survived, 342 Survived**

Step 17: Detailed Statistical Summary for Numerical Columns 📋

Finally, I generated a detailed statistical summary for all numerical columns, which included measures of central tendency and dispersion, providing a comprehensive understanding of the dataset.

Skills Acquired

Through this project, I gained experience in: Data Cleaning, Statistical Analysis, Data Visualization, Outlier Detection, Data Wrangling, and Data Management.

Hashtags

#DataScience #DataAnalysis #Pandas #Numpy #DataCleaning #DataWrangling #Statistics #OutlierDetection #DataVisualization #DataPreparation #DataInsights #MachineLearning #Python #EDA #BigData #Analytics #Dataset #DataQuality #DataExploration #DataManagement

This project has not only enhanced my analytical skills but also deepened my understanding of data preparation processes, which are critical for any data-driven decision-making.