NIKHIL MALVI B 50

EXPERIMENT 1

Titanic Dataset Preprocessing

This code performs preprocessing steps on the Titanic dataset, including:

1. **Data Loading:** Loads the Titanic dataset from a remote URL using pandas.
2. **Exploratory Data Analysis (EDA):**
   - Checks for missing values in each column.
   - Calculates descriptive statistics.
   - Examines data types and the dataset's dimensions.
3. **Data Type Conversion:**
   - Converts relevant columns ('Survived', 'Pclass', 'Sex', 'Embarked') to categorical data types for better analysis.
4. **Data Cleaning and Normalization:**
   - Fills missing values in the 'Age' column with the median age.
   - Applies Z-score normalization to the 'Fare' column, scaling it to have zero mean and unit variance.
5. **Output:**
   - Prints the results of the missing data check, descriptive statistics, data types, and dataset dimensions.
   - Prints the final data types after conversion and cleaning.

```
import pandas as pd

url = 'https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv'
titanic_df = pd.read_csv(url) # Corrected line: removed extra 'lad this' and closed parenthesis
titanic_df.head(5)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
				Futrelle, Mrs. Jacques Heath (Lily								

Next steps: Generate code with titanic_df | View recommended plots | New interactive sheet

```
#Check for missing values
missing_data = titanic_df.isnull().sum()
# Get initial statistics for the dataset
data_description = titanic_df.describe()
# Check the data types and dimensions
data_types = titanic_df.dtypes
dimensions = titanic_df.shape
# Output
print("Missing Data:\n", missing_data)
print("\nData Description:\n", data_description)
print("\nData Types:\n", data_types)
print("\nDataset Dimensions:", dimensions)
```

```
Missing Data:
 PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
```

```
Cabin          687
Embarked         2
dtype: int64

Data Description:
        PassengerId    Survived      Pclass         Age       SibSp  \
count   891.000000  891.000000  891.000000  714.000000  891.000000   
mean    446.000000    0.383838    2.308642   29.699118    0.523008   
std     257.353842    0.486592    0.836071   14.526497    1.102743   
min       1.000000    0.000000    1.000000    0.420000    0.000000   
25%     223.500000    0.000000    2.000000   20.125000    0.000000   
50%     446.000000    0.000000    3.000000   28.000000    0.000000   
75%     668.500000    1.000000    3.000000   38.000000    1.000000   
max     891.000000    1.000000    3.000000   80.000000    8.000000   

            Parch        Fare  
count  891.000000  891.000000  
mean     0.381594   32.204208  
std      0.806057   49.693429  
min      0.000000    0.000000  
25%      0.000000    7.910400  
50%      0.000000   14.454200  
75%      0.000000   31.000000  
max      6.000000  512.329200  

Data Types:
 PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object

Dataset Dimensions: (891, 12)
```

```
# Check for categorical variables and convert them if needed
titanic_df['Survived'] = titanic_df['Survived'].astype('category') # Corrected column name to 'Survived'
titanic_df['Pclass'] = titanic_df['Pclass'].astype('category')
titanic_df['Sex'] = titanic_df['Sex'].astype('category')
titanic_df['Embarked'] = titanic_df['Embarked'].astype('category')
# Normalize continuous numerical columns (Age, Fare, etc.)
titanic_df['Age'] = titanic_df['Age'].fillna(titanic_df['Age'].median())  # Replace missing values with the median
titanic_df['Fare'] = (titanic_df['Fare'] - titanic_df['Fare'].mean()) / titanic_df['Fare'].std()  # Z-score normalization
# Output the final data types after conversion
print("\nFinal Data Types after Conversion:\n", titanic_df.dtypes)
```

```
Final Data Types after Conversion:
 PassengerId       int64
Survived       category
Pclass         category
Name             object
Sex            category
Age             float64
SibSp             int64
Parch             int64
Ticket           object
Fare            float64
Cabin            object
Embarked       category
dtype: object
```

```
titanic_df_encoded = pd.get_dummies(titanic_df, columns=['Sex',
'Embarked'], drop_first=True)
# Display the first few rows
print(titanic_df_encoded) # Removed the extra indent before this line.
```

```
     PassengerId Survived Pclass  \
0              1        0      3   
1              2        1      1   
2              3        1      3   
3              4        1      1   
4              5        0      3   
..           ...      ...    ...   
```

```
886          887         0       2
887          888         1       1
888          889         0       3
889          890         1       1
890          891         0       3

                                                   Name   Age  SibSp  Parch  \
0                               Braund, Mr. Owen Harris  22.0      1      0   
1     Cumings, Mrs. John Bradley (Florence Briggs Th...  38.0      1      0   
2                                Heikkinen, Miss. Laina  26.0      0      0   
3          Futrelle, Mrs. Jacques Heath (Lily May Peel)  35.0      1      0   
4                              Allen, Mr. William Henry  35.0      0      0   
..                                                  ...   ...    ...    ...   
886                               Montvila, Rev. Juozas  27.0      0      0   
887                        Graham, Miss. Margaret Edith  19.0      0      0   
888            Johnston, Miss. Catherine Helen "Carrie"  28.0      1      2   
889                               Behr, Mr. Karl Howell  26.0      0      0   
890                                 Dooley, Mr. Patrick  32.0      0      0   

               Ticket      Fare Cabin  Sex_male  Embarked_Q  Embarked_S  
0           A/5 21171 -0.502163   NaN      True       False        True  
1            PC 17599  0.786404   C85     False       False       False  
2     STON/O2. 3101282 -0.488580   NaN     False       False        True  
3              113803  0.420494  C123     False       False        True  
4              373450 -0.486064   NaN      True       False        True  
..                ...       ...   ...       ...         ...         ...  
886            211536 -0.386454   NaN      True       False        True  
887            112053 -0.044356   B42     False       False        True  
888        W./C. 6607 -0.176164   NaN     False       False        True  
889            111369 -0.044356  C148      True       False       False  
890            370376 -0.492101   NaN      True        True       False  

[891 rows x 13 columns]
```