ASSIGNMENT 2

20BCG10024 – SATYAM SONI

****

1. Use the dataset: The dataset can be downloaded from the link provided.

2. Load the dataset: The dataset can be loaded into a Pandas Data Frame using the following code:

3. I) Perform Univariate Analysis

df.describe()

II)Perform Bi-Variate Analysis

df.plot.scatter(x='Age', y='Survived')

III) Perform Multi-Variate Analysis

df.plot.scatter(x='Age', y='Survived', c='Pclass')

4. Perform Descriptive Statistics

df.describe()

5. Code for handling missing values, finding outliers, and replacing outliers in the Titanic dataset

```
import pandas as pd
# Load the dataset
df = pd.read_csv('titanic.csv')
# Check for missing values
print(df.isnull().sum())

# There are 177 missing values in the Age column.
# Replace the missing values with the mean of the column
df['Age'].fillna(df['Age'].mean(), inplace=True)
```

```
# Check for outliers
Q1 = df['Age'].quantile(0.25)
Q3 = df['Age'].quantile(0.75)
IQR = Q3 - Q1

# Outliers are any data points that are outside of the range of (Q1 - 1.5 * IQR) to (Q3 + 1.5 * IQR)
outliers = df[(df['Age'] < Q1 - 1.5 * IQR) | (df['Age'] > Q3 + 1.5 * IQR)]

# There are 14 outliers in the Age column.
# Replace the outliers with the mean of the column
df.loc[outliers.index, 'Age'] = df['Age'].mean()

# Check for categorical columns
categorical_columns = df.select_dtypes(include='object').columns

# There are 3 categorical columns in the dataset: Sex, Cabin, and Embarked.
# Encode the categorical columns
df = pd.get_dummies(df, columns=categorical_columns)
```

6. Check for Categorical Columns and Perform Encoding

```
import pandas as pd

# Load the dataset
df = pd.read_csv('titanic.csv')

# Check for categorical columns
```

```
categorical_columns = df.select_dtypes(include='object').columns

# There are 3 categorical columns in the dataset: Sex, Cabin, and Embarked.
# Encode the categorical columns
df = pd.get_dummies(df, columns=categorical_columns)
```

7. Split the Data into Dependent and Independent Variables

```
X = df[['Pclass', 'Age', 'Sex', 'Fare', 'SibSp']]
y = df['Survived']
```

8. Scale the Independent Variables

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
X = scaler.fit_transform(X)
```

9. Split the Data into Training and Testing

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25)
```