Name : Devkumar Biswas

Class: BE(AI&DS)

Div: B

Subject : DMV(CL-I Lab)

Roll no. : BEAD21267

**Assignment No. - 2**

**Problem Statement** : Data Cleaning and Preparation

Problem Statement: Analyzing Customer Churn in a Telecommunications Company

Dataset: "Telecom\_Customer\_Churn.csv"

Description: The dataset contains information about customers of a telecommunications company and whether they have churned (i.e., discontinued their services). The dataset includes various attributes of the customers, such as their demographics, usage patterns, and account information. The goal is to perform data cleaning and preparation to gain insights into the factors that contribute to customer churn.

Tasks to Perform: 1. Import the "Telecom\_Customer\_Churn.csv" dataset. 2. Explore the dataset to understand its structure and content. 3. Handle missing values in the dataset, deciding on an appropriate strategy. 4. Remove any duplicate records from the dataset. 5. Check for inconsistent data, such as inconsistent formatting or spelling variations, and standardize it. 6. Convert columns to the correct data types as needed. 7. Identify and handle outliers in the data. 8. Perform feature engineering, creating new features that may be relevant to predicting customer churn. 9. Normalize or scale the data if necessary. 10. Split the dataset into training and testing sets for further analysis. 11. Export the cleaned dataset for future analysis or modeling.

CODE:-

import pandas as pd #data manipulation

import numpy as np #numerical computations

from sklearn.model\_selection import train\_test\_split

from sklearn import metrics #evaluating the performance of machine learning model

data = pd.read\_csv("Telecom\_Customer\_Churn.csv")

print(data.index)

print(data.columns)

data.shape

print(data.head())

data.isna().sum()

data.isnull().sum()

print("Number of rows before removing duplicates:", len(data))

# Remove duplicate records

data\_cleaned = data.drop\_duplicates()

# Check the number of rows after removing duplicates

print("Number of rows after removing duplicates:", len(data\_cleaned))

data.describe()

#Measure of frequency destribution

unique, counts = np.unique(data['Tenure in Months'], return\_counts=True)

print(unique, counts)

unique, counts = np.unique(data['Total Charges'], return\_counts=True)

print(unique, counts)

import seaborn as sns #Seaborn library for data visualization

sns.pairplot(data)

X = data.drop("Total Revenue", axis=1)

y = data["Total Revenue"]

# Split the dataset into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

X\_train.shape

y\_train.shape

X\_test.shape

y\_test.shape # Export the cleaned dataset to a CSV file

data.to\_csv("Cleaned\_Telecom\_Customer\_Churn.csv", index=False)

OUTPUT: 









