

Assignment No-3

CODE:-

```
# Imports
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from google.colab import files

# Upload dataset
uploaded = files.upload()
filename = next(iter(uploaded))
df = pd.read_csv(filename)

# Explore dataset
df.info()
print(df.isnull().sum())
print(df.describe())

# Handle missing values
df.fillna(df.median(numeric_only=True), inplace=True)
for col in df.select_dtypes(include=['object']).columns:
    df[col].fillna(df[col].mode()[0], inplace=True)

print(df.isnull().sum())

# Remove duplicates
df.drop_duplicates(inplace=True)
print(f'Duplicates removed, remaining: {df.duplicated().sum()}')

# Standardize categorical data example
if 'Gender' in df.columns:
    df['Gender'] = df['Gender'].str.strip().str.lower()
if 'SeniorCitizen' in df.columns:
    df['SeniorCitizen'] = df['SeniorCitizen'].replace({1: 'Yes', 0: 'No'})
```

```

# Convert data types
if 'Churn' in df.columns:
    df['Churn'] = df['Churn'].astype('category')
if 'CustomerID' in df.columns:
    df['CustomerID'] = df['CustomerID'].astype(str)

print(df.dtypes)

# Handle outliers using IQR for numeric columns only
Q1 = df.select_dtypes(include=np.number).quantile(0.25)
Q3 = df.select_dtypes(include=np.number).quantile(0.75)
IQR = Q3 - Q1

condition = ~((df.select_dtypes(include=np.number) < (Q1 - 1.5 * IQR)) |
              (df.select_dtypes(include=np.number) > (Q3 + 1.5 *
IQR))).any(axis=1)
df_no_outliers = df[condition]
print(f"Shape after removing outliers: {df_no_outliers.shape}")

# Feature engineering: create 'Tenure' if date column exists
if 'ContractStartDate' in df.columns:
    df['Tenure'] = pd.to_datetime(df['ContractStartDate'],
errors='coerce').apply(lambda x: (pd.to_datetime('today') - x).days // 30
if pd.notnull(x) else np.nan)
    df['Tenure'].fillna(df['Tenure'].median(), inplace=True)

# Scale numerical columns if they exist
num_cols = ['MonthlyCharges', 'TotalCharges']
for col in num_cols:
    if col in df.columns:
        scaler = MinMaxScaler()
        df[col] = scaler.fit_transform(df[[col]])

```

```
# Prepare features and target
if 'Churn' in df.columns:
    X = df.drop(columns=['Churn', 'CustomerID'], errors='ignore')
    y = df['Churn']

    # Split into train/test
    X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

    print(f"Train shape: {X_train.shape}, Test shape: {X_test.shape}")

# Export cleaned data
df.to_csv('Cleaned_Telecom_Customer_Churn.csv', index=False)

# Download the cleaned file
files.download('Cleaned_Telecom_Customer_Churn.csv')
```

OUTPUT :-

The screenshot shows a Google Colab notebook with the following content:

```
Choose File Telecom_C_hum.csv
Saving Telecom_Customer_Churn.csv(text/csv) - 1435200 bytes, last modified: 9/17/2025 - 100% done
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 38 columns):
 #   Column                                Non-Null Count  Dtype
---  -
 0   Customer ID                          7043 non-null   object
 1   Gender                                7043 non-null   object
 2   Age                                    7043 non-null   int64
 3   Married                               7043 non-null   object
 4   Number of Dependents                 7043 non-null   int64
 5   City                                  7043 non-null   object
 6   Zip Code                             7043 non-null   int64
 7   Latitude                             7043 non-null   float64
 8   Longitude                            7043 non-null   float64
 9   Number of Referrals                  7043 non-null   int64
10   Tenure in Months                     7043 non-null   int64
11   Offer                                3166 non-null   object
12   Phone Service                        7043 non-null   object
13   Avg Monthly Long Distance Charges    6361 non-null   float64
14   Multiple Lines                       7043 non-null   object
15   Internet Service                     7043 non-null   object
16   Internet Type                         5517 non-null   object
17   Avg Monthly GB Download              5517 non-null   float64
18   Online Security                      5517 non-null   object
19   Online Backup                        5517 non-null   object
20   Device Protection Plan               5517 non-null   object
21   Premium Tech Support                 5517 non-null   object
22   Streaming TV                         5517 non-null   object
23   Streaming Movies                     5517 non-null   object
24   Streaming Music                      5517 non-null   object
25   Unlimited Data                       5517 non-null   object
26   Contract                             7043 non-null   object
27   Paperless Billing                     7043 non-null   object
28   Payment Method                       7043 non-null   object
29   Monthly Charge                       7043 non-null   float64
30   Total Charges                        7043 non-null   float64
31   Total Refunds                        7043 non-null   float64
32   Total Extra Data Charges             7043 non-null   int64
33   Total Long Distance Charges          7043 non-null   float64
34   Total Revenue                        7043 non-null   float64
35   Customer Status                      7043 non-null   object
36   Churn Category                       1869 non-null   object
37   Churn Reason                         1869 non-null   object
```

The screenshot shows the same Google Colab notebook with the following content:

```
36 churn category                                1869 non-null   object
37 Churn Reason                                1869 non-null   object
dtypes: float64(9), int64(6), object(23)
memory usage: 2.0+ MB

Customer ID                                0
Gender                                      0
Age                                          0
Married                                     0
Number of Dependents                       0
City                                         0
Zip Code                                    0
Latitude                                    0
Longitude                                    0
Number of Referrals                         0
Tenure in Months                           0
Offer                                        3877
Phone Service                              0
Avg Monthly Long Distance Charges           682
Multiple Lines                             682
Internet Service                            0
Internet Type                              1526
Avg Monthly GB Download                     1526
Online Security                            1526
Online Backup                              1526
Device Protection Plan                     1526
Premium Tech Support                       1526
Streaming TV                               1526
Streaming Movies                           1526
Streaming Music                            1526
Unlimited Data                              1526
Contract                                    0
Paperless Billing                           0
Payment Method                             0
Monthly Charge                             0
Total Charges                              0
Total Refunds                              0
Total Extra Data Charges                    0
Total Long Distance Charges                 0
Total Revenue                              0
Customer Status                            0
Churn Category                             5174
Churn Reason                               5174
dtype: int64

Age      Number of Dependents      Zip Code      Latitude \
count    7043.000000              7043.000000    7043.000000    7043.000000
mean      46.589726                0.468692    93486.078567    36.197455
std       14.761813                0.967881     1454.767346      7.468078
```

Search results - shikant...Search results - Google...Classwork for DMV BE...PR - Google Drive...PR 3 - Google Docs...Data cleaning in Colab...Untitled3.ipynb - Colab...Telecom Custom...

colab.research.google.com/drive/1bOZdPg1g1eQgMn267TBYR7kd1Nj?authuser=0

Untitled3.ipynb

FileEditViewInsertRuntimeToolsHelp

CommandsCodeTextRun all

countAgeNumber of DependentsZip CodeLatitude \

dtype: int64

count7043.0000007043.0000007043.0000007043.000000

mean46.5897260.46889293486.07856736.197435

std16.7593520.9620821856.7675052.468929

min19.0000000.00000090001.00000032.555828

25%32.0000000.00000092101.00000033.990646

50%46.0000000.00000093518.00000036.205465

75%60.0000000.00000095329.00000038.161321

max80.0000009.00000096150.00000041.962127

countLongitudeNumber of ReferralsTenure in Months \

count7043.0000007043.0000007043.000000

mean-119.7566841.95186732.386767

std2.1544253.00119924.542061

min-124.3813720.0000001.000000

25%-121.7880900.0000009.000000

50%-119.5952930.00000029.000000

75%-117.5607953.00000055.000000

max-114.19290111.00000072.000000

countAvg Monthly Long Distance ChargesAvg Monthly GB Download \

count6361.0000005517.000000

mean25.42851726.189958

std14.28037419.586585

min1.0100002.000000

25%13.05000013.000000

50%25.69000021.000000

75%37.60000038.000000

max49.90000085.000000

countMonthly ChargeTotal ChargesTotal RefundsTotal Extra Data Charges \

count7043.0000007043.0000007043.0000007043.000000

mean63.5961312280.3812641.9521826.860713

std31.2847432266.2204627.98261425.104078

min-18.00000018.0000000.0000000.000000

25%38.400000400.1500000.0000000.000000

50%70.0500001394.5500000.0000000.000000

75%89.7500003786.0000000.0000000.000000

max118.7500008684.80000049.790000150.000000

VariablesTerminal

10:48 AMPython 3

20°CLight rain

Search

ENGIN13:30:1317-09-2025