Title : Analyzing Customer Churn in a Telecommunications Company

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.

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler, StandardScaler
```

In [2]:
```
#1. Import the "Telecom_Customer_Churn.csv" dataset.
data = pd.read_csv('telecom_customer_churn.csv')
data.head()
```

Out[2]:

	Customer ID	Gender	Age	Married	Number of Dependents	City	Zip Code	Latitude	Longitude
0	0002-ORFBO	Female	37	Yes	0	Frazier Park	93225	34.827662	-118.999073
1	0003-MKNFE	Male	46	No	0	Glendale	91206	34.162515	-118.203869
2	0004-TLHLJ	Male	50	No	0	Costa Mesa	92627	33.645672	-117.922613
3	0011-IGKFF	Male	78	Yes	0	Martinez	94553	38.014457	-122.115432
4	0013-EXCHZ	Female	75	Yes	0	Camarillo	93010	34.227846	-119.079903

5 rows × 38 columns

In [3]:
```
# Step 3: Explore dataset
print("\nDataset info:")
print(data.info())
```

```
Dataset info:
<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                     6361 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 
dtypes: float64(9), int64(6), object(23)
memory usage: 2.0+ MB
None
```

In [4]:
```
print("\nSummary statistics:\n", data.describe())
```

```
Summary statistics:
                Age  Number of Dependents      Zip Code     Latitude  \
count  7043.000000           7043.000000   7043.000000  7043.000000   
mean     46.509726              0.468692  93486.070567    36.197455   
std      16.750352              0.962802   1856.767505     2.468929   
min      19.000000              0.000000  90001.000000    32.555828   
25%      32.000000              0.000000  92101.000000    33.990646   
50%      46.000000              0.000000  93518.000000    36.205465   
75%      60.000000              0.000000  95329.000000    38.161321   
max      80.000000              9.000000  96150.000000    41.962127   

         Longitude  Number of Referrals  Tenure in Months  \
count  7043.000000          7043.000000       7043.000000   
mean   -119.756684             1.951867         32.386767   
std       2.154425             3.001199         24.542061   
min    -124.301372             0.000000          1.000000   
25%    -121.788090             0.000000          9.000000   
50%    -119.595293             0.000000         29.000000   
75%    -117.969795             3.000000         55.000000   
max    -114.192901            11.000000         72.000000   

       Avg Monthly Long Distance Charges  Avg Monthly GB Download  \
count                        6361.000000              5517.000000   
mean                           25.420517                26.189958   
std                            14.200374                19.586585   
min                             1.010000                 2.000000   
25%                            13.050000                13.000000   
50%                            25.690000                21.000000   
75%                            37.680000                30.000000   
max                            49.990000                85.000000   

       Monthly Charge  Total Charges  Total Refunds  Total Extra Data Charges  \
count     7043.000000    7043.000000    7043.000000               7043.000000   
mean        63.596131    2280.381264       1.962182                  6.860713   
std         31.204743    2266.220462       7.902614                 25.104978   
min        -10.000000      18.800000       0.000000                  0.000000   
25%         30.400000     400.150000       0.000000                  0.000000   
50%         70.050000    1394.550000       0.000000                  0.000000   
75%         89.750000    3786.600000       0.000000                  0.000000   
max        118.750000    8684.800000      49.790000                150.000000   

       Total Long Distance Charges  Total Revenue  
count                  7043.000000    7043.000000  
mean                    749.099262    3034.379056  
std                     846.660055    2865.204542  
min                       0.000000      21.360000  
25%                      70.545000     605.610000  
50%                     401.440000    2108.640000  
75%                    1191.100000    4801.145000  
max                    3564.720000   11979.340000  
```

In [5]:
```
print("\nMissing values:\n", data.isnull().sum())
```

```
Missing values:
 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
```

In [6]:

```python
# Step 4: Handle missing values
# We have several columns with missing data
# Strategy:
# - For categorical columns (Offer, Internet Type, Online Security, etc.), fill wit
# - For numerical columns (Avg Monthly Long Distance Charges, Avg Monthly GB Downlo
categorical_cols = ['Offer', 'Multiple Lines', 'Internet Type', 'Online Security',
                    'Online Backup', 'Device Protection Plan', 'Premium Tech Suppor
                    'Streaming TV', 'Streaming Movies', 'Streaming Music', 'Unlimit

numerical_cols = ['Avg Monthly Long Distance Charges', 'Avg Monthly GB Download']

# Fill missing categorical data
for col in categorical_cols:
    data[col].fillna('Unknown', inplace=True)

# Fill missing numerical data
for col in numerical_cols:
    data[col].fillna(data[col].median(), inplace=True)

# Check if there are any missing values left
print("\nMissing values after imputation:\n", data.isnull().sum())
```

```
Missing values after imputation:
 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                                    0
Phone Service                            0
Avg Monthly Long Distance Charges        0
Multiple Lines                           0
Internet Service                         0
Internet Type                            0
Avg Monthly GB Download                  0
Online Security                          0
Online Backup                            0
Device Protection Plan                   0
Premium Tech Support                     0
Streaming TV                             0
Streaming Movies                         0
Streaming Music                          0
Unlimited Data                           0
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
```

In [7]:
```python
# Step 5: Remove duplicates
data.drop_duplicates(inplace=True)
```

In [8]:
```python
# Step 6: Standardize inconsistent data
# Example: Ensure 'Yes'/'No' are consistent in columns like 'Phone Service', 'Paper
yes_no_cols = ['Phone Service', 'Paperless Billing', 'Unlimited Data', 'Multiple Li
for col in yes_no_cols:
    data[col] = data[col].str.strip().str.title()  # remove whitespace and capitali
```

In [9]:
```python
# Step 7: Convert columns to correct data types
# 'Total Extra Data Charges' should be float for consistency
data['Total Extra Data Charges'] = data['Total Extra Data Charges'].astype(float)
```

In [10]:
```python
# Step 8: Handle outliers
# For simplicity, we cap numerical columns at 1st and 99th percentile
num_cols = ['Age', 'Number of Dependents', 'Tenure in Months', 'Avg Monthly Long Di
            'Avg Monthly GB Download', 'Monthly Charge', 'Total Charges', 'Total Re
            'Total Extra Data Charges', 'Total Long Distance Charges', 'Total Reven

for col in num_cols:
    lower = data[col].quantile(0.01)
    upper = data[col].quantile(0.99)
    data[col] = np.clip(data[col], lower, upper)
```

In [11]:
```python
# Step 9: Feature Engineering
# Example features:
# - 'Tenure Group': categorize customers based on tenure
data['Tenure Group'] = pd.cut(data['Tenure in Months'], bins=[0,12,24,48,60,120],
                              labels=['0-1yr','1-2yr','2-4yr','4-5yr','5-10yr'])
```

```
# - 'Avg Charges per Month': ratio of Total Charges to Tenure (avoiding division by
data['Avg Charges per Month'] = data['Total Charges'] / data['Tenure in Months'].re
```

In [12]:
```
# Step 10: Normalize/Scale numerical features
scaler = MinMaxScaler()
scaled_cols = ['Age', 'Number of Dependents', 'Tenure in Months', 'Avg Monthly Long
               'Avg Monthly GB Download', 'Monthly Charge', 'Total Charges', 'Total
               'Total Extra Data Charges', 'Total Long Distance Charges', 'Total Re

data[scaled_cols] = scaler.fit_transform(data[scaled_cols])
```

In [13]:
```
# Step 11: Split dataset into training and testing sets
# Target column: binary churn (1 = churned, 0 = not churned)
data['Churn Binary'] = data['Customer Status'].apply(lambda x: 1 if x.strip().lower

X = data.drop(['Customer ID', 'Customer Status', 'Churn Category', 'Churn Reason'],
y = data['Churn Binary']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_sta
```

In [14]:
```
# Step 12: Export cleaned dataset
data.to_csv("telecom_customer_churn_cleaned.csv", index=False)

print("\nData cleaning and preparation completed. Cleaned dataset exported.")
```

```
Data cleaning and preparation completed. Cleaned dataset exported.
```