**Exploratory Data Analysis (EDA) on Customer Churn**

The exploratory data analysis (EDA) of the customer churn dataset reveals insightful patterns and distributions. The initial dataset exploration shows the presence of missing values, which were subsequently handled by filling numerical columns with mean values and categorical columns with mode values. Outlier analysis through boxplots indicated some extreme values, which were addressed using the IQR method to ensure a more robust dataset. Encoding categorical variables and normalizing the data prepared it for deeper analysis. The statistical summary and histograms illustrated the distributions of numerical features, highlighting various trends and anomalies. Scatter plots against the target variable, 'Churn,' were used to visualize potential relationships and identify predictors of churn. The correlation matrix further elucidated the strength of relationships between features, with significant correlations to the 'Churn' variable identified, suggesting key factors influencing customer retention.

**Data Collection:**

Used a telecom churn dataset from Kaggle named as “customer\_churn”.

The dataset consists of 3,333 entries with 11 columns, including features like AccountWeeks, ContractRenewal, DataPlan, and Churn (the target variable indicating whether a customer churned or not). The dataset's initial exploration shows a mix of numerical and categorical features, with some missing values in DayCalls and RoamMins.  
**Data Preprocessing:**

Missing values were handled effectively: numerical missing values were filled with the mean, and categorical missing values with the mode. After addressing these, the dataset was cleaned and ready for analysis. The cleaned dataset now has no missing values.

The dataset was normalized to ensure all features are on a similar scale. This process adjusts the values so they lie between 0 and 1, making comparisons between features more straightforward. The normalized data shows values adjusted for analysis.

**Exploratory Analysis:**

The statistical summary of the cleaned dataset provides insights into the distribution of numerical features:

* **Churn**: The dataset does not contain positive churn values, which may need verification.
* **AccountWeeks**: Ranges from 1 to 205 weeks, with a mean of about 100 weeks.
* **DataUsage**: Varies significantly from 0 to 4.56 GB, with a mean of approximately 0.82 GB.
* **CustServCalls**: Average of 1.32 calls, with a range from 0 to 3 calls.
* **MonthlyCharge**: Ranges from $16 to $93.8, with an average charge of $54.95.
* **RoamMins**: Average roaming minutes are about 10.2, ranging from 3.1 to 17.3 minutes.

The correlation matrix did not show any significant relationships between features and the Churn variable, suggesting that no strong predictors are apparent in the dataset. This indicates that other methods or additional features might be required to better understand the factors leading to customer churn.

**Results:**

First few rows of the dataset:

Churn AccountWeeks ContractRenewal DataPlan DataUsage CustServCalls \

1 0 107 1 1 3.70 1

2 0 137 1 0 0.00 0

11 0 74 1 0 0.34 0

12 0 168 1 0 0.00 1

13 0 95 1 0 0.44 3

DayMins DayCalls MonthlyCharge OverageFee RoamMins

1 161.6 123.0 82.0 9.78 13.7

2 243.4 114.0 52.0 6.06 12.2

11 187.7 127.0 49.4 8.17 9.1

12 128.8 96.0 31.0 5.25 11.2

13 156.6 88.0 52.4 12.38 12.3

Dataset Info:

None

Missing values in each column:

Churn 0

AccountWeeks 0

ContractRenewal 0

DataPlan 0

DataUsage 0

CustServCalls 0

DayMins 0

DayCalls 0

MonthlyCharge 0

OverageFee 0

RoamMins 0

dtype: int64

Missing values after handling:

Churn 0

AccountWeeks 0

ContractRenewal 0

DataPlan 0

DataUsage 0

CustServCalls 0

DayMins 0

DayCalls 0

MonthlyCharge 0

OverageFee 0

RoamMins 0

dtype: int64

Statistical summary of the dataset:

Churn AccountWeeks ContractRenewal DataPlan DataUsage \

count 2402.0 2402.000000 2402.0 2402.000000 2402.000000

mean 0.0 99.915071 1.0 0.281432 0.818555

std 0.0 39.323585 0.0 0.449791 1.247201

min 0.0 1.000000 1.0 0.000000 0.000000

25% 0.0 73.000000 1.0 0.000000 0.000000

50% 0.0 100.000000 1.0 0.000000 0.000000

75% 0.0 127.000000 1.0 1.000000 1.882500

max 0.0 205.000000 1.0 1.000000 4.560000

CustServCalls DayMins DayCalls MonthlyCharge OverageFee \

count 2402.000000 2402.000000 2402.000000 2402.000000 2402.000000

mean 1.320150 173.134763 100.132806 54.949500 9.905271

std 0.969507 47.446997 19.078202 15.313186 2.434603

min 0.000000 41.900000 48.000000 16.000000 3.220000

25% 1.000000 141.625000 87.000000 44.000000 8.220000

50% 1.000000 175.300000 100.000000 52.000000 9.910000

75% 2.000000 207.075000 113.000000 63.450000 11.597500

max 3.000000 301.700000 152.000000 93.800000 16.640000

RoamMins

count 2402.000000

mean 10.194047

std 2.596899

min 3.100000

25% 8.400000

50% 10.200000

75% 12.000000

max 17.300000

Correlation with target variable (Churn):

Churn NaN

AccountWeeks NaN

ContractRenewal NaN

DataPlan NaN

DataUsage NaN

CustServCalls NaN

DayMins NaN

DayCalls NaN

MonthlyCharge NaN

OverageFee NaN

RoamMins NaN

Name: Churn, dtype: float64





