About the data

Churn is one of the biggest problem in the telecom industry. Research has shown that the average monthly churn rate among the top 4 wireless carriers in the US is 1.9% - 2%

Data shape
7043 rows
21 Columns
We have a row for each customer and the features' categories are described below:

Demographic customer information

- gender
- SeniorCitizen
- Partner
- Dependents

Services that each customer has signed up for

- PhoneService
- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies

Customer account information

- tenure
- Contract
- PaperlessBilling
- PaymentMethod
- MonthlyCharges
- TotalCharges

Customers who left within the last month

- Churn

EDA

1. UNIQUE VALUES

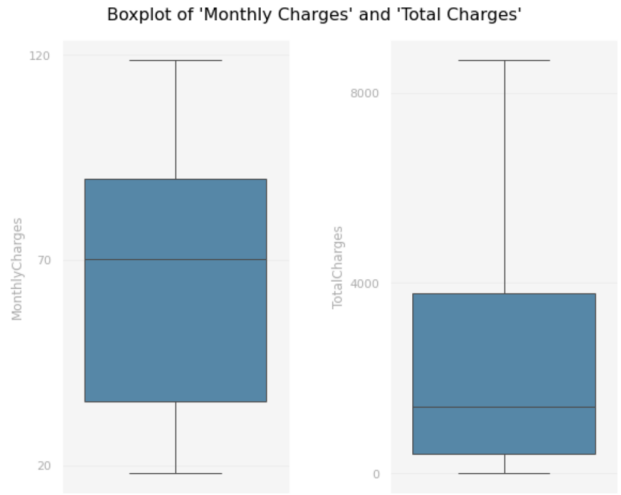
```
# number of unique observations per column
df.nunique()
```

```
customerID          7043
gender                 2
SeniorCitizen          2
Partner                2
Dependents             2
tenure                73
PhoneService           2
MultipleLines          3
InternetService        3
OnlineSecurity         3
OnlineBackup           3
DeviceProtection       3
TechSupport            3
StreamingTV            3
StreamingMovies        3
Contract               3
PaperlessBilling       2
PaymentMethod          4
MonthlyCharges      1585
TotalCharges        6531
Churn                  2
dtype: int64
```

Observe that we have 21 features for 7043 clients. Among all features, 3 of them are numerical, 10 are categorical and the remaining can be considered binary. The feature 'customerID' represents a unique value for each row.

2.CHECKING THE CONSISTENCY OF DATA:

At first glance, everything looks ok with our numerical features. We didn't spot any outlier in MonthlyCharges nor TotalCharges.



Before starting to apply some feature engineering, let's check the value distribution for our target variable Churn.

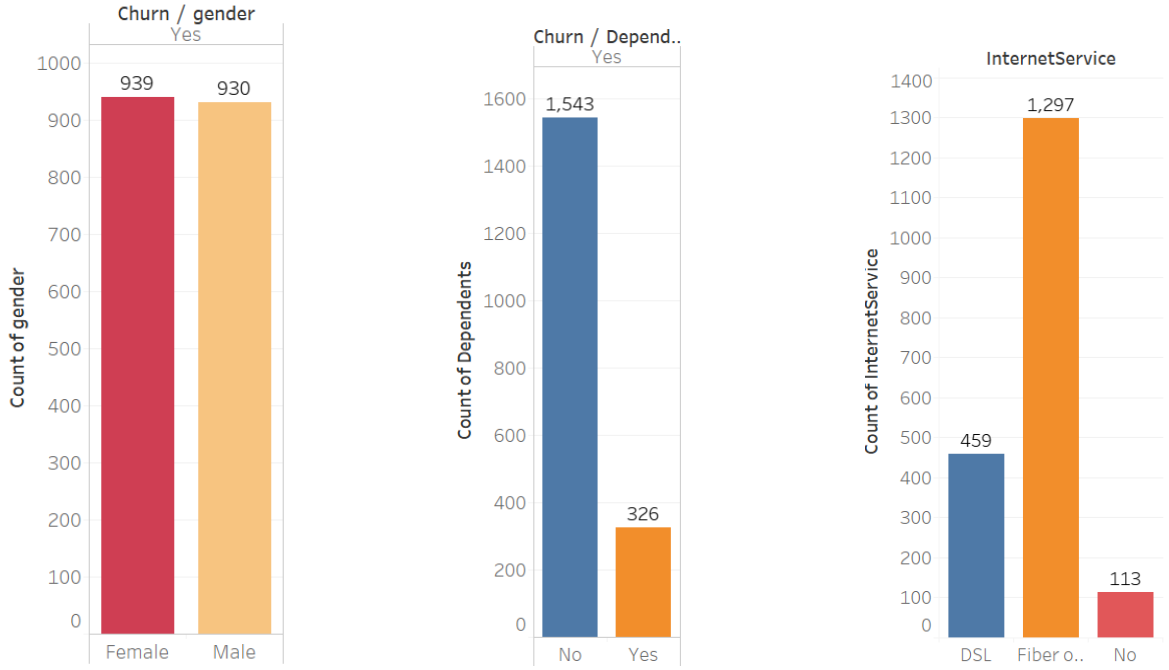
We are dealing with an unbalanced dataset as shown below:

```
No      5174
Yes     1869
Name: Churn, dtype: int64

Total Churn Rate: 26.54%
```

Observe that the churn rate is 26.5%, meaning that the quantity of "No" values is substantially higher than that of "Yes" values.

3. CHURN RATIO WITH RESPECT TO OTHER COLUMNS:



Looking at the example above, we can interpret that Gender probably won't be a meaningful variable to the model, as the churn rate is quite similar for both male and female customers. On the other hand, clients with dependents are less prone to stop doing business with the company.

As for internet service, customers with fiber optic plans are more likely to quit. Their churn rate is more than double that of DSL and no internet users.

FEATURE SELECTION

We did churn modelling using algorithms like logistic regression and XGBoost and found out what are the columns which are the most important while finding the churn.

Below are the importance score for all the column names:

importance score	Column names
627340.305176	TotalCharges
16278.923685	tenure
3680.787699	MonthlyCharges
519.895311	Contract_Month-to-month
488.578090	Contract_Two year
426.422767	PaymentMethod_Electronic check
374.476216	InternetService_Fiber optic
286.520193	InternetService_No
176.123171	Contract_One year
147.295858	OnlineSecurity
135.559783	TechSupport
134.351545	SeniorCitizen
133.036443	Dependents
105.680863	PaperlessBilling
99.582057	PaymentMethod_Credit card (automatic)
82.412083	Partner
76.485913	PaymentMethod_Bank transfer (automatic)
71.313180	InternetService_DSL
45.651590	PaymentMethod_Mailed check
31.217694	OnlineBackup
20.226662	DeviceProtection
17.334235	StreamingTV
16.242531	StreamingMovies
6.548512	MultipleLines_Yes
3.874782	MultipleLines_No
0.907148	MultipleLines_No phone service
0.258699	gender
0.097261	PhoneService

Out of these, first ten are the primary factors on which 'çhurn' depends and based on it we have done our churn visualization.