**ML – ASSIGNMENT**

**CAP – 1**

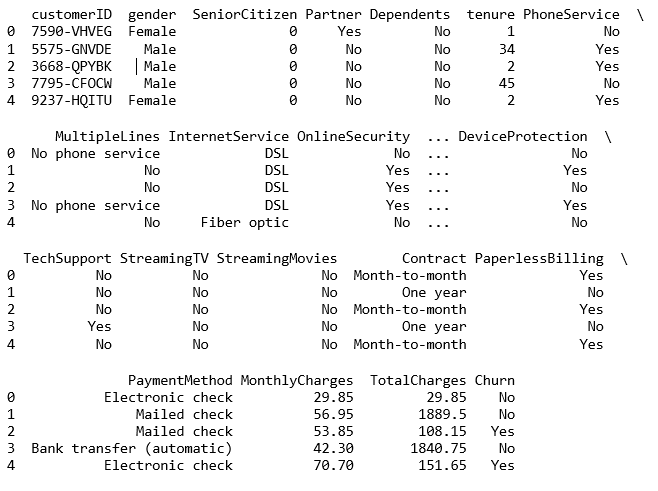
**Dataset used for the assignment :**

*Telco-Customer Dataset ( From Kaggle)*

This dataset is available on Kaggle. It contains 19 columns in total. The column considered as response variable in the dataset is the ‘Churn’ column. The churn column defines if the customer has churned out of the company or not.

1. **Data Reading :**

*Successfully read the data and described the heads of the dataset.*

****

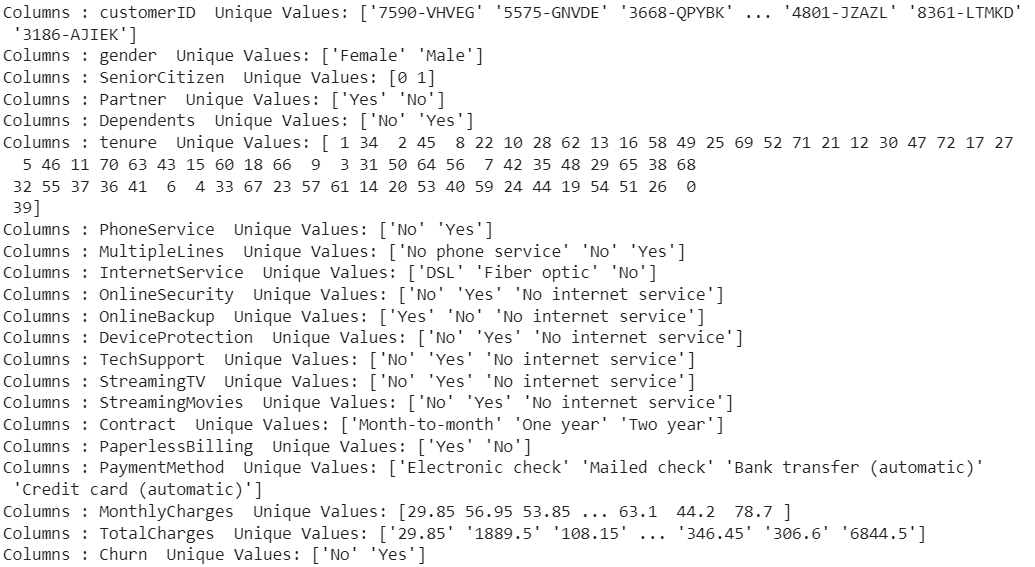
1. **Exploratory Data Analysis :**

*The first step towards EDA, is to analyze the type of variables are present in the column. To do that, we need to find the unique values of all the columns present in the dataset.*

*We find out that, there are* ***13 categorical variables*** *and* ***6 continuous variables.***

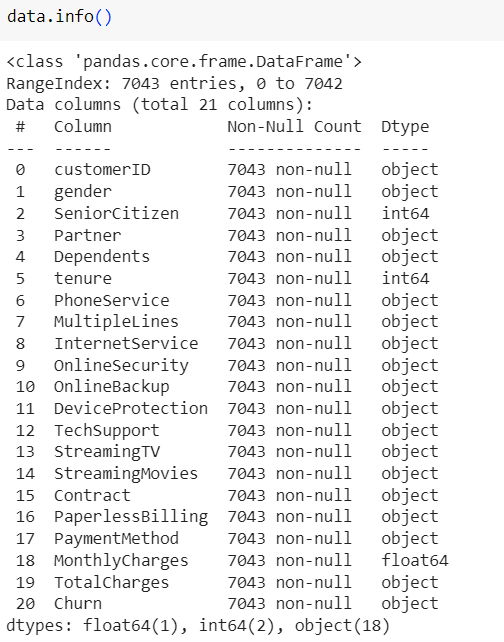
***Categorical Variables :*** *gender, SeniorCitizen, Partner, Dependents, PhoneService, MultipleLines, InternetServices, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies.*

***Continuous Variables :*** *tenure, contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges*

**

*We can further divide the categorical variables into Demographic and Non-Demographic Variables. So now, there are* ***4 Demographic Variables*** *(gender, SeniorCitizen, Partner, Dependents) and* ***9 other are Non-Demographic.***

***Now, we try to find the missing values in the dataset.***

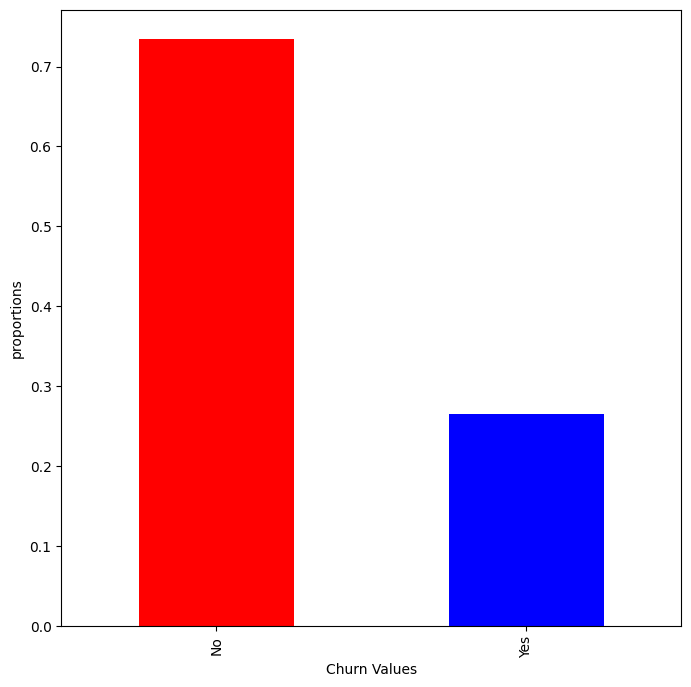
******

* *We conclude that there are 7043 observations and 21 columns in the dataset.*
* *There are no NULL values in the dataset.*
* *We find that TotalCharges has been misclassified as an object when it is numeric variable.*
* *We tried to convert the TotalCharges column to a numeric type but it was affecting the data and making it dirty, so we decide to drop it.*
* *Also we drop the CustomerID since, it is of no use in the analysis.*

*Now we move towards Data Visualisation.*

1. **Data Visualisation :**

*We try to analyse the Response Variable at first i.e. ‘Churn’ column and see its proportions*

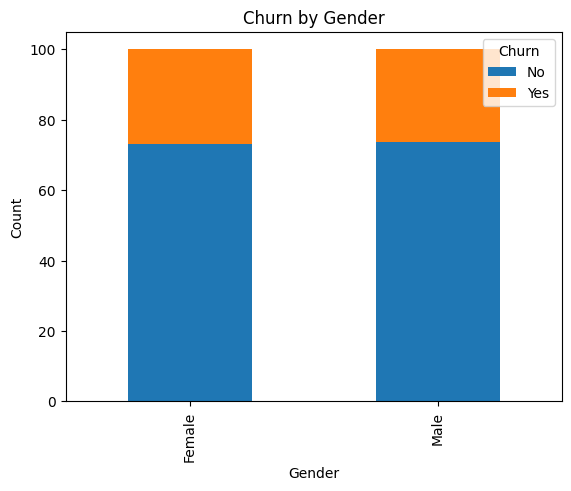
**

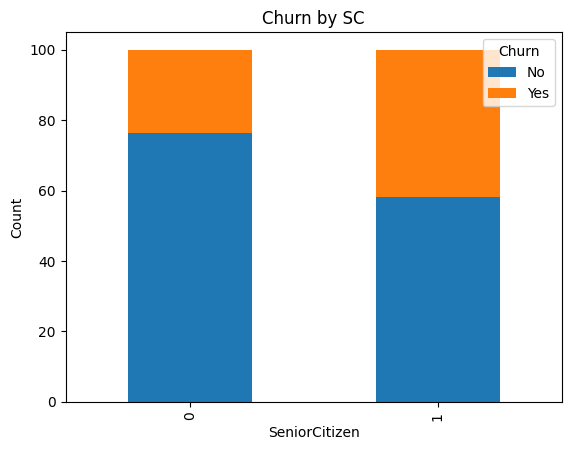
*We can see that the around 75% of the data is not churned and the remaining is churned. This also shows that the data is imbalanced and both the classes i.e. yes and no are not distributed equally in the dataset.*

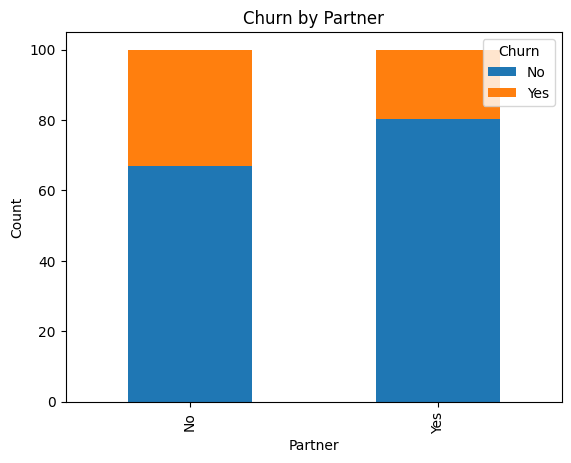
*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

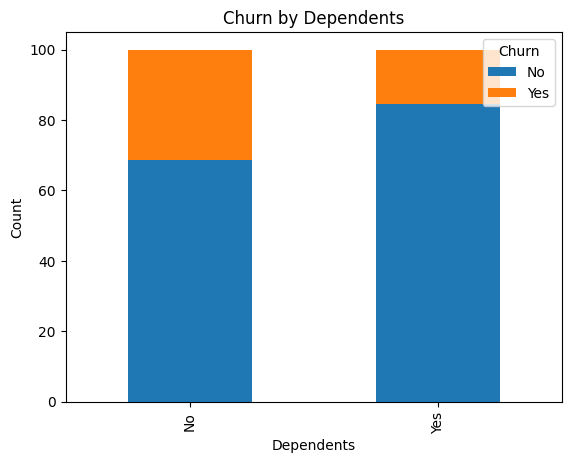
*We now try to analyze the categorical variables, most precisely the demographic ones first.*

*To analyse the categorical variables vs Churn Variable, we can use the Stacked Bar Chart Plot.*









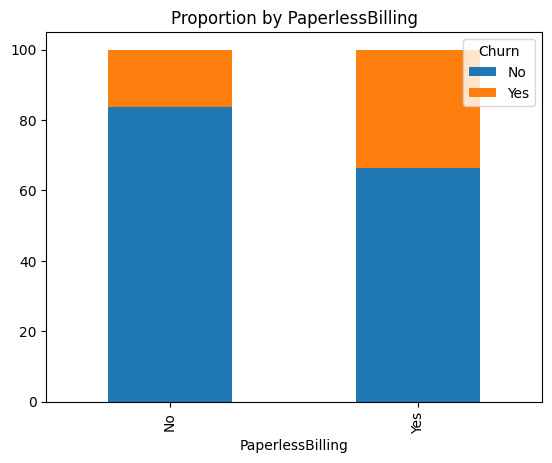
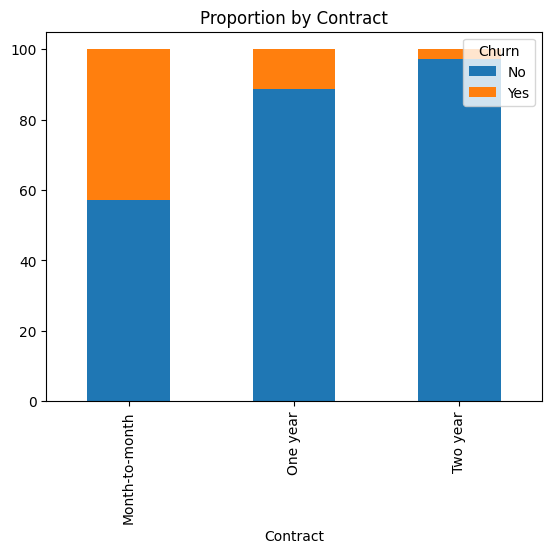
*Analyzing the plots, we can conclude that :*

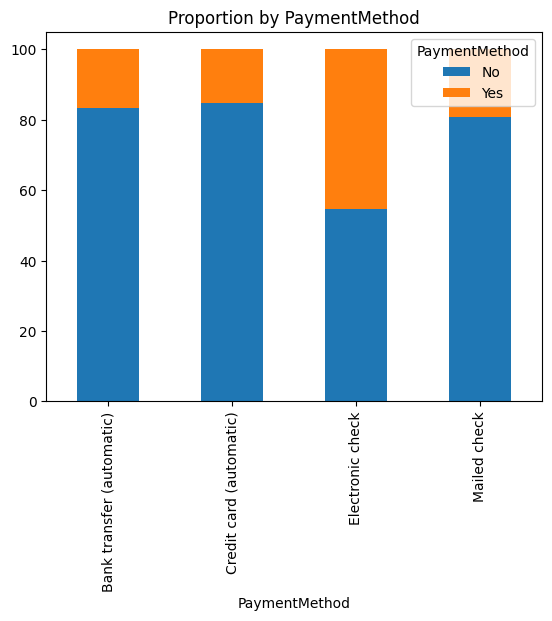
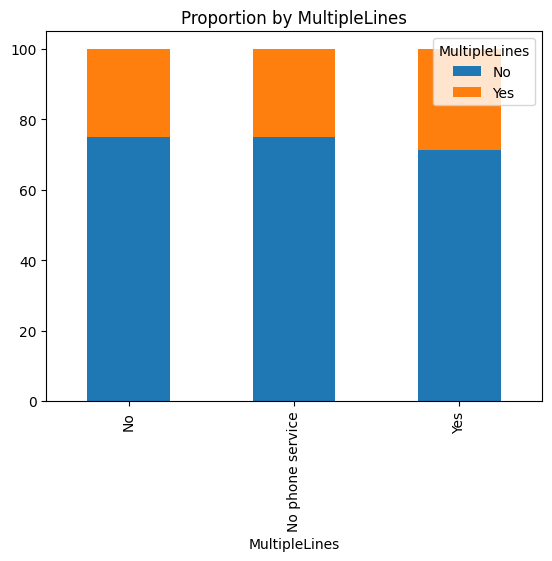
* *We should not use gender column, as it does not predict anything. It shows almost the same level of Churns for both the genders.*
* *For the SeniorCitizen, the churn rate is almost double as compared to the other population.*
* *Churn Rate for Customers with partners is less than those with no partners*
* *Customers with Dependents have a lower Churn Rate.*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Now , we analyze the dataset for the Categorical variables :*

*We take into consideration Contract, PaperlessBilling, MultipleLines and PaymentMethod. Since, all the other categorical variables seem to be more of a service, so we’ll analyze it separately.*

**

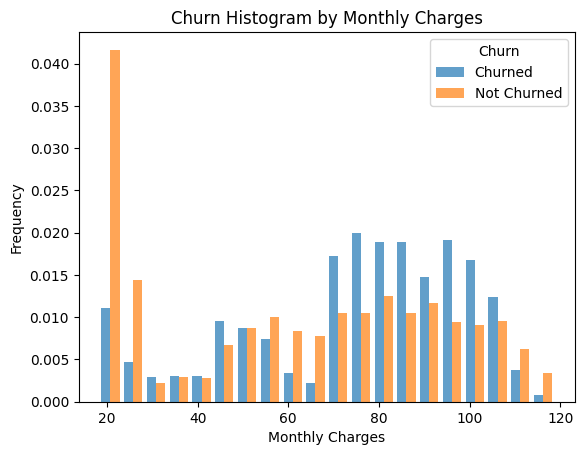
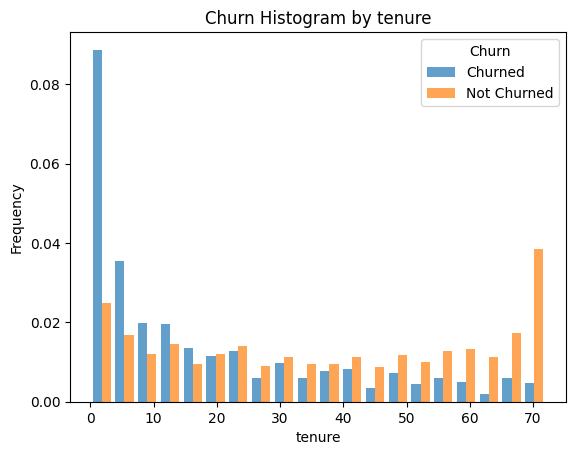
******

*Analyzing the plots we can conclude that :*

* *MultipleLines cannot be used as a parameter to analyze any output.*
* *Longer the period of Contract, Lower will be the Churn Rate.*
* *Customers with PaperlessBillings have a higher Churn Rate.*
* *Customer who have ElectronicCheck PaymentMethod churn more than than any other. All the other have comparatively equal Churn Rates.*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Now we try to analyze the Continuous variables. We usually use Histograms to present Numerical vs Categorical EDA. We analyze tenure, MonthlyCharges and Churn variables and plot its distribution.*

**

*From the above plots, we can conclude that :*

* *Customers with a lower tenure value, have a higher Churn Rate.*
* *Customers with a lower MonthlyCharges values, have a lower Churn Rate value. Customers with MonthlyCharges of 70-105, have a higher Churn Rate value.*

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

1. **Encoding :**

Encoding is an important part of Feature Engineering where we transform the data that can be fit into a model.

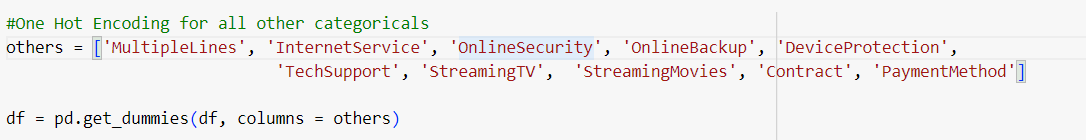
1. ***Label Encoding:***

*Used to replace categorical values with numerical values. Here we replace the values of binary variables into numerical values. The variables, [ gender, Partner, Dependents, PaperlessBilling, PhoneService, Churn ] can be converted to 0s and 1s since they are binary.*

**

1. ***One Hot Encoding :***

*Used to convert categorical variables to binary vectors. Now here when the category is present, it is given as value = 1 and any other scenario would be given the value = 0.*

**

*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

1. **Normalization:**

*We use normalization to bring all the numeric values to a common level. We use the min-max method of Normalization here to rescale the numeric columns [tenure, MonthlyCharges] to a common scale.*