**Customer Churn Analysis**

**1.Problem Definition:**

Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals.

Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to defect means that a company can prioritise focused marketing efforts on that subset of their customer base.

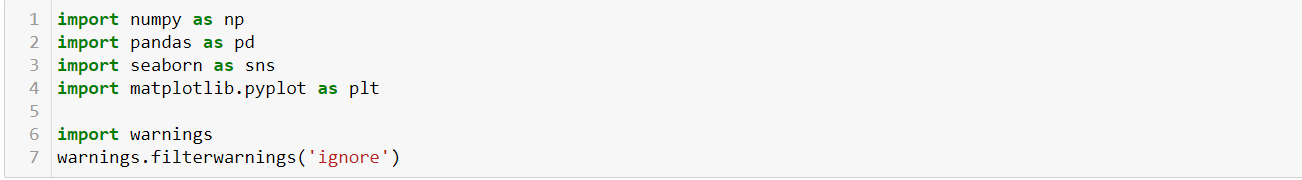
Preventing customer churn is critically important to the telecommunications sector, as the barriers to entry for switching services are so low.

<https://github.com/dsrscientist/DSData/blob/master/Telecom_customer_churn.csv>

From the problem statement. We can predict that this is classification problem. Why because in customer churn columns it has only Yes & No. We need to predict ‘Churn’ column. That why I predict that it is Classification problem. Let us, dive into deep check whether it is True or not

**2. Data Analysis:**

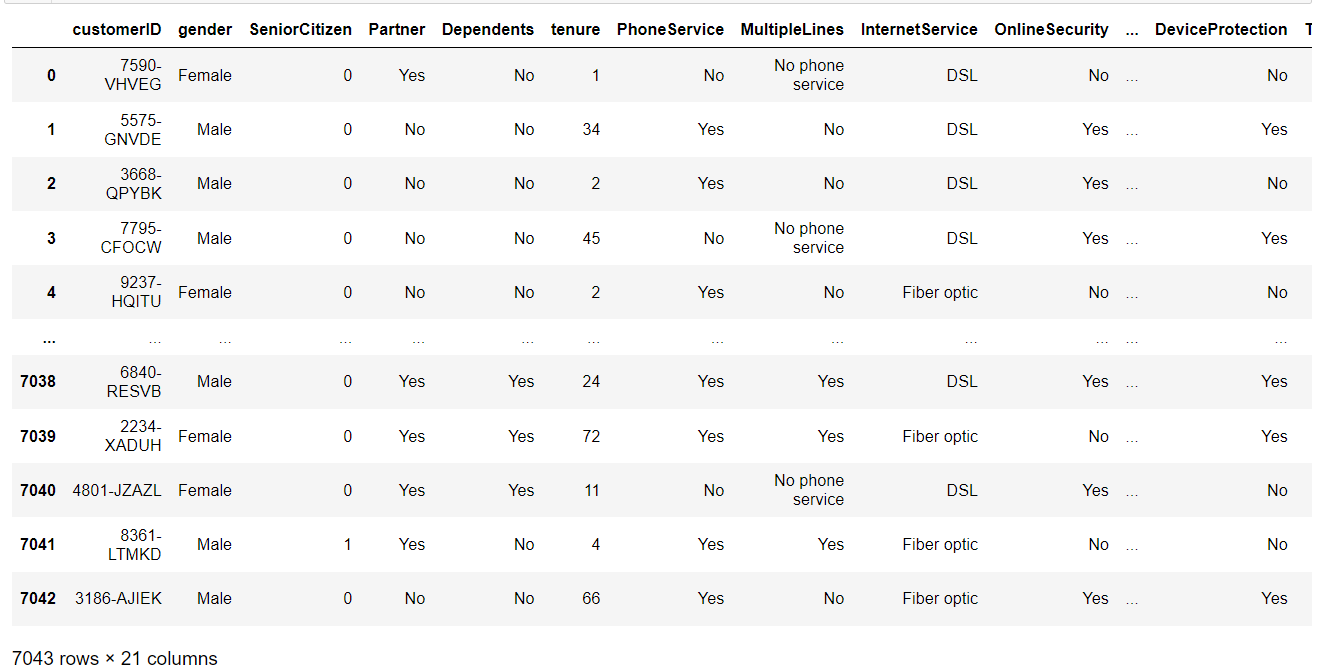
We will start by importing necessary libraries required for performing EDA. like numpy, pandas and seaborn, matplot for visualization purpose. import Warnings to avoid the unnecessary warnings.



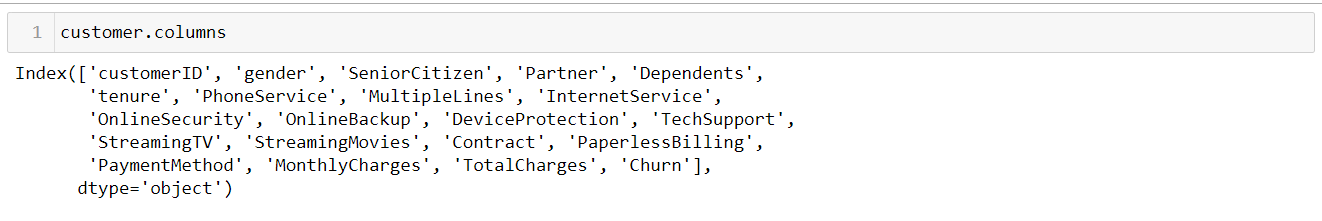
**Reading data**:



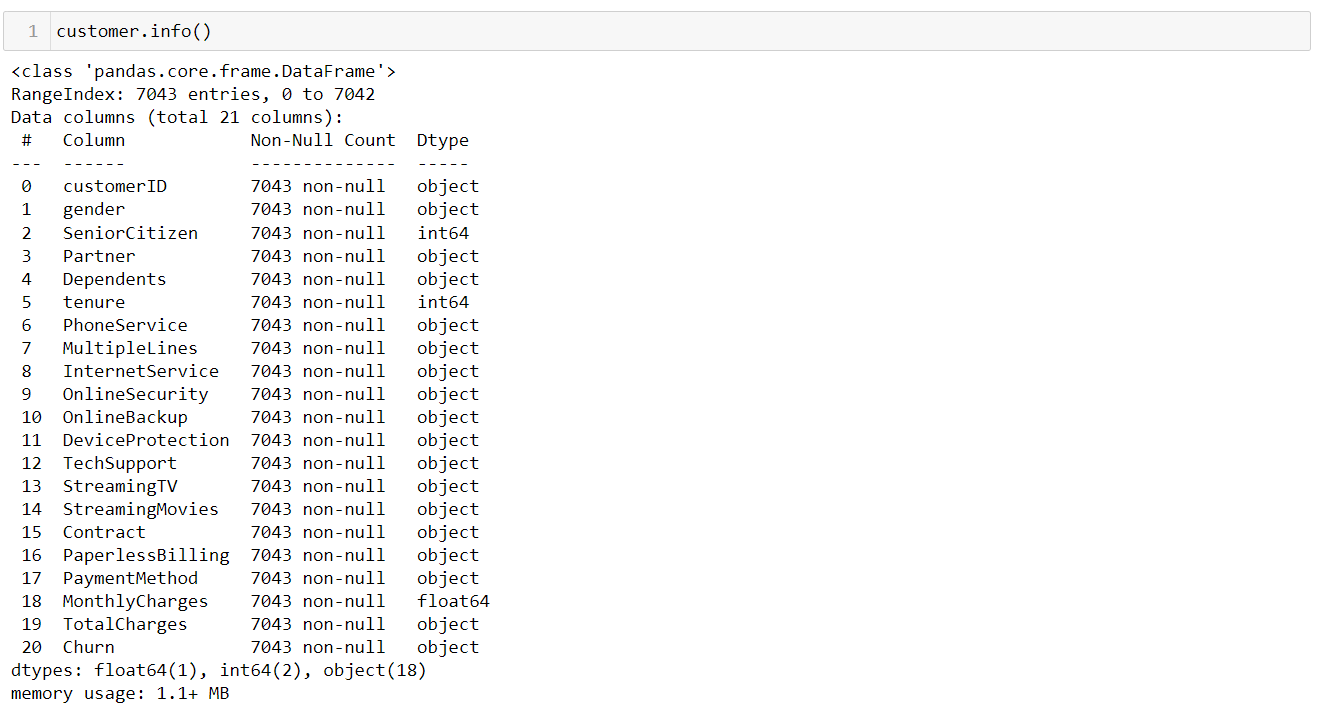
**Resultant:**



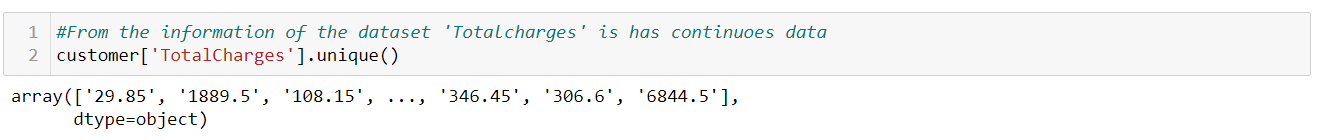
The data set has 7043 rows and 21 columns. Letus check for the names of the columns that are present in the dataset.



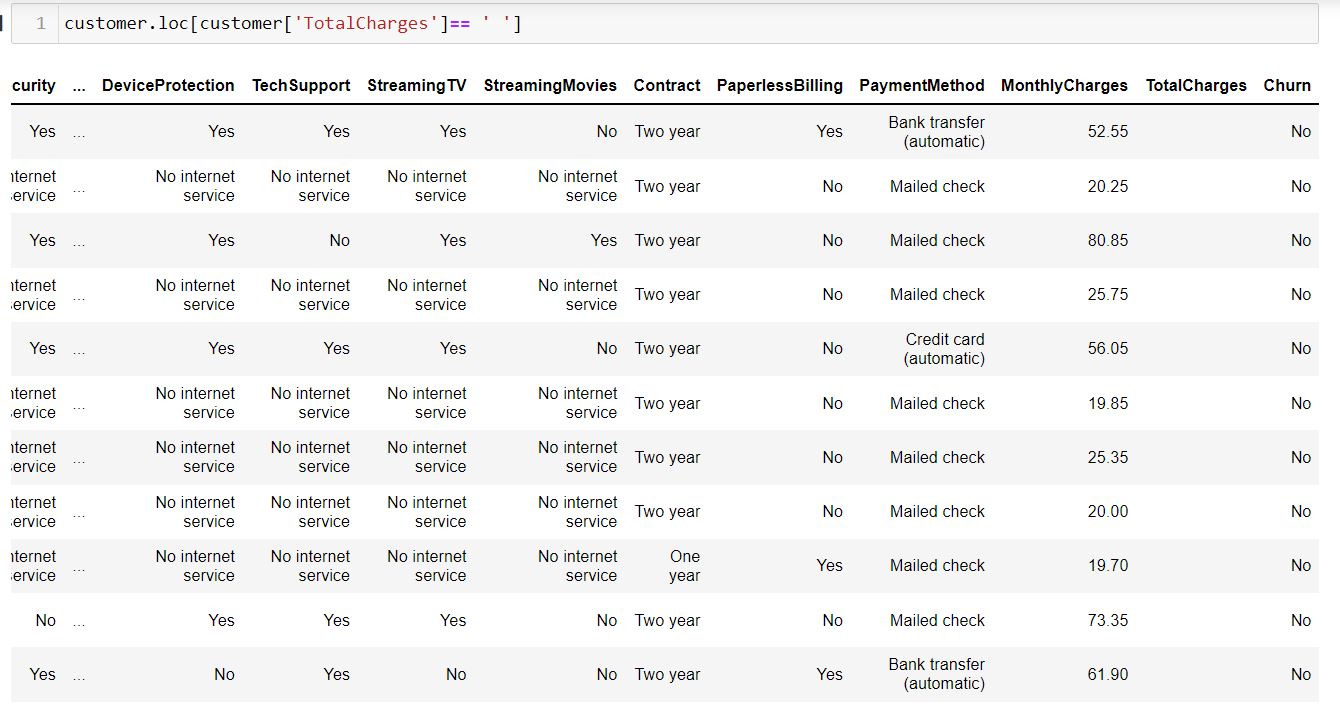
Now,check for the information of the dataset.like datatype of the columns.How many integer,float and object type are present.



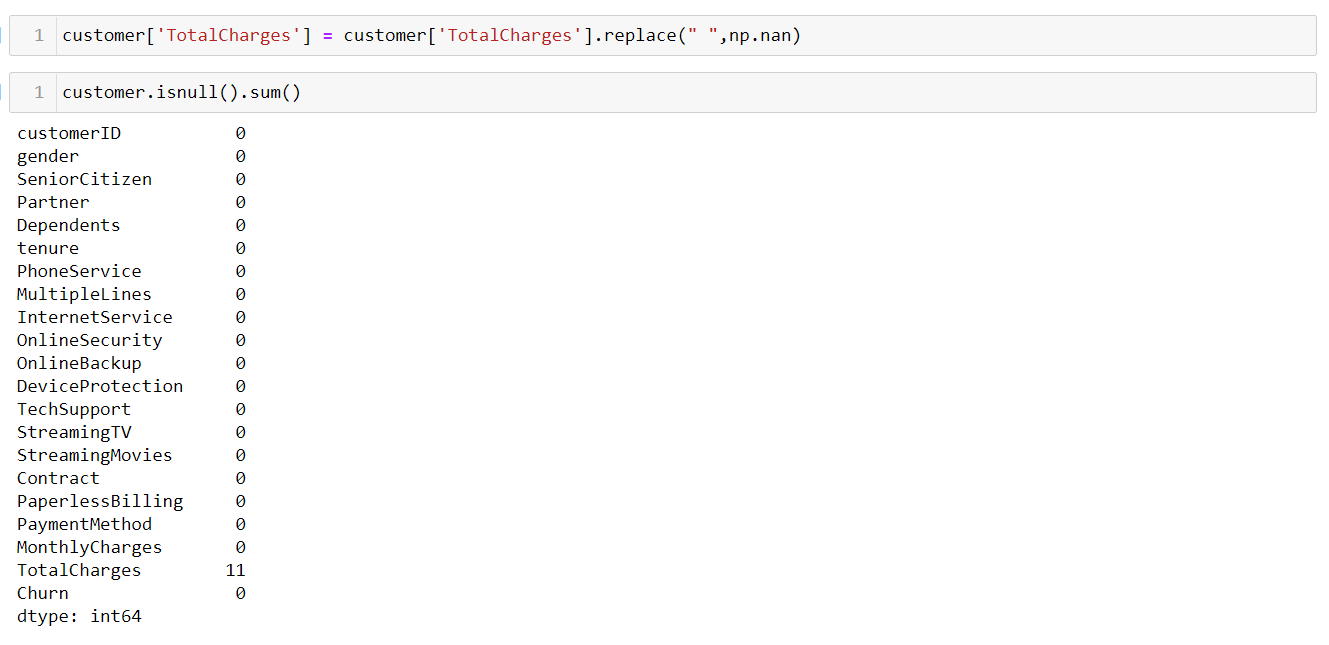
From the information we can observe that 64-int ,1-float and 18-object type of variables are present in our dataset.we can observe another thing here,that is dtype of ‘TotalCharges’ are represented by ‘Object’.But It consists of float value means ‘continuoes data’.Why it represented as object type.Letus exam in ‘TotalCharges’ column.

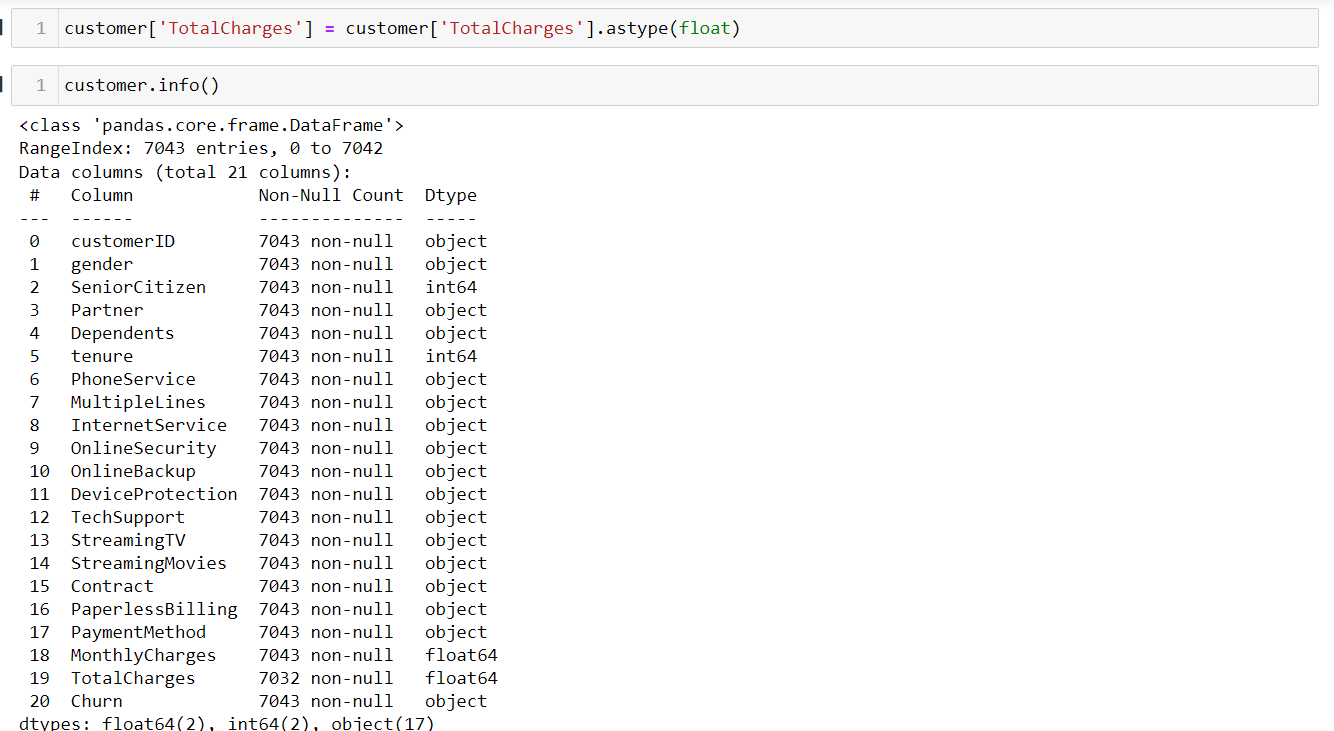


Checking whether their is any space exist in the place of value.



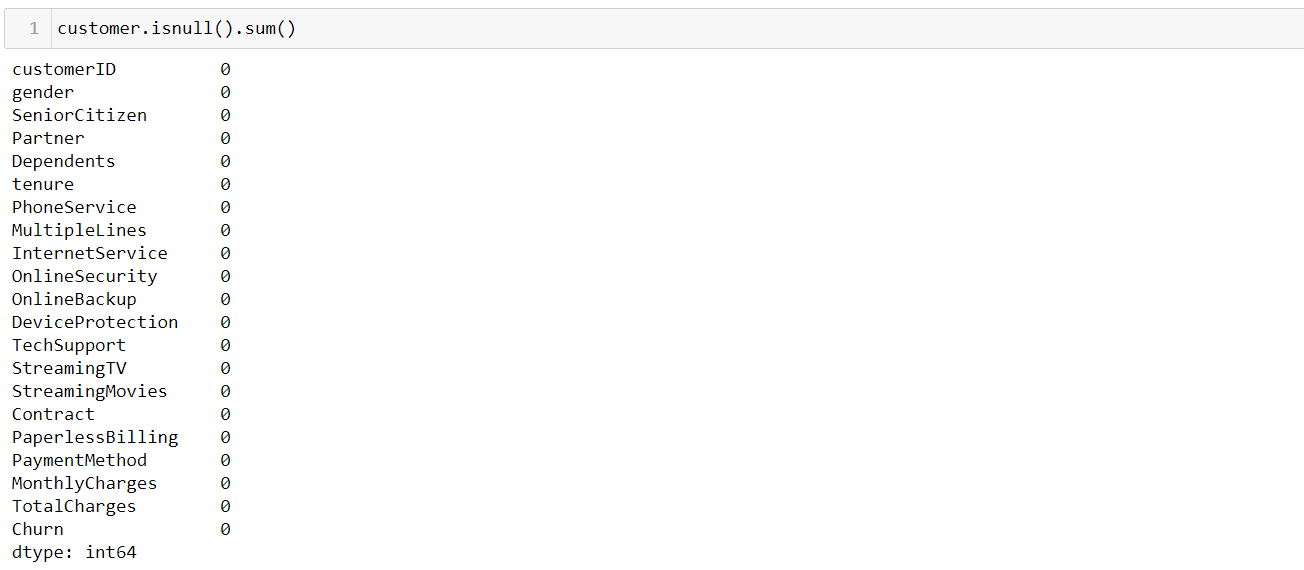
Our examination is true.their exist a space in the place of values.For this we need to fill the space with null value first and then it is continuoes variable. So, we go for ‘mean’ of the values for filling NaN’s.





Previously datatype of the ‘TotalCharges’ is object. But the data consists in the columns is float type.That why after filling the space with NaN.we fill that NaN with mean of the values.and then after change datatype of the variable to float type.Now,finally checking the information.

Now, we going to check for the null values present in the data set or not.if present we need to check datatype of the variable to fill NaN’s if it is categorical we need to fill with most occurred value.if it is continuoes value we can fill NaN’s with mean value.

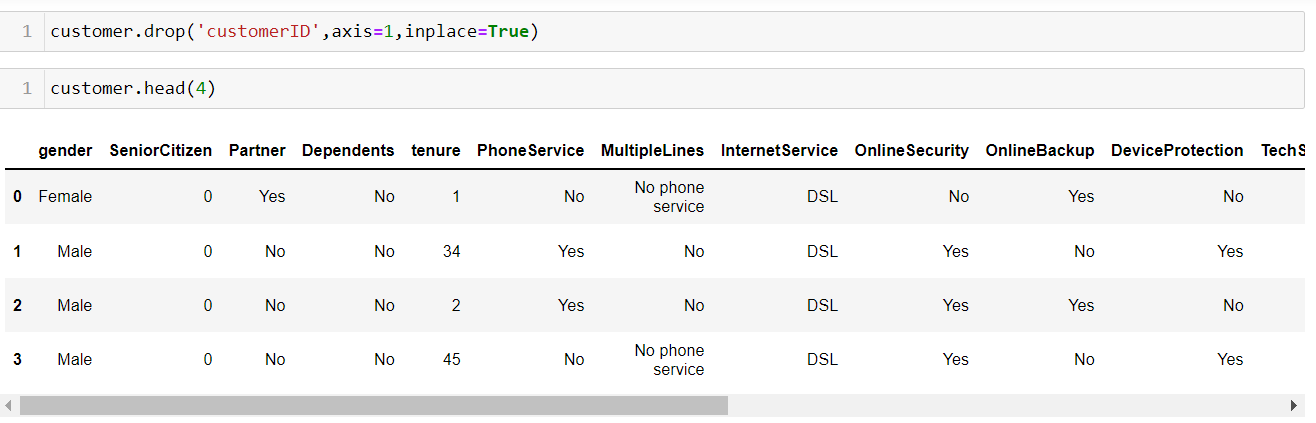


Luckily, our data set has no Nan’s. so, we are jumping into the next step.

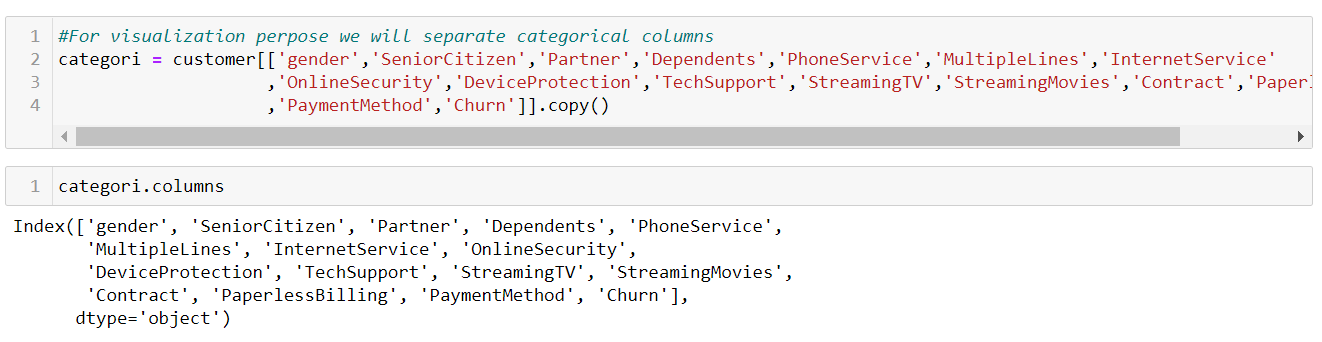
**3. EDA:**

**Visualization**:

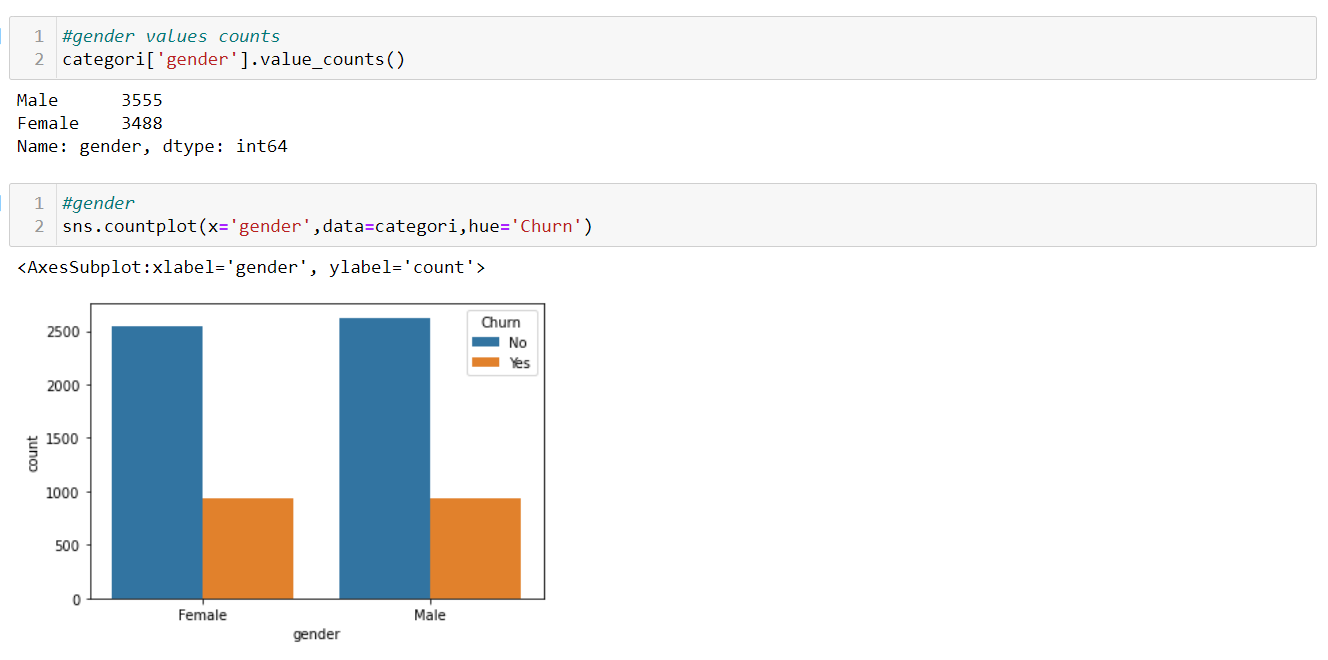
1St of all we need to divide the numerical and categorical data type of the columns for visualization purpose. Before this I’m deleting unnecessary columns present in the dataset. that is ‘Customer ID’.



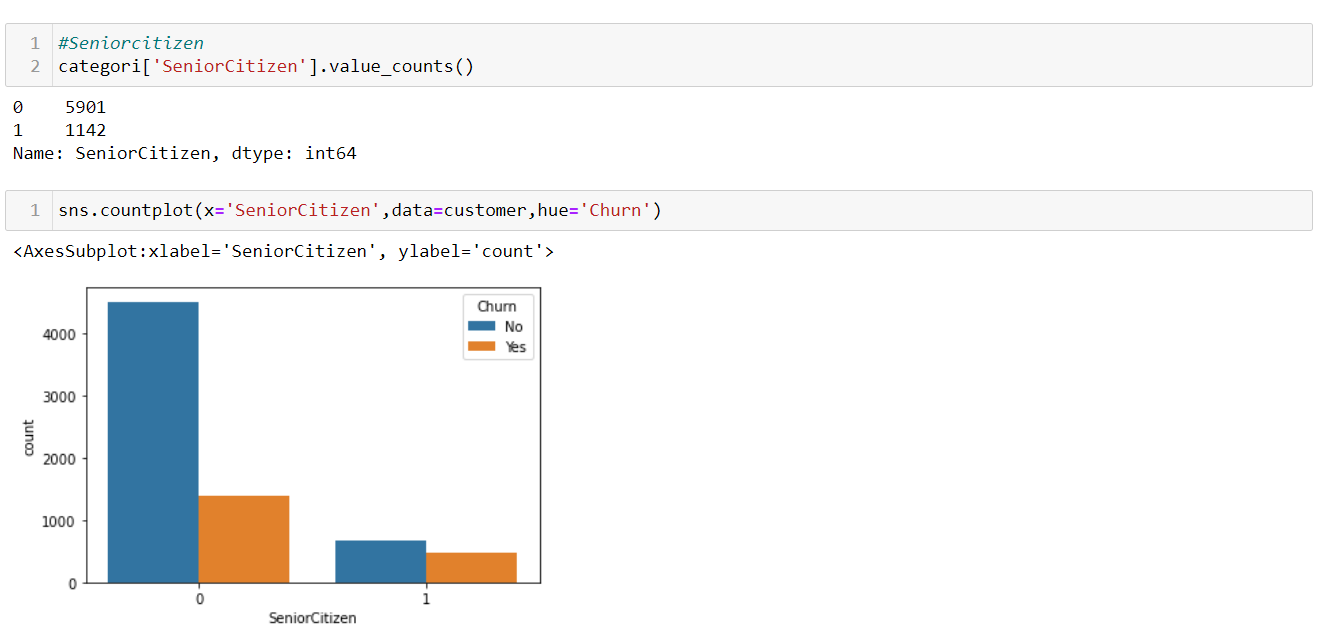
Then, separating the categorical data



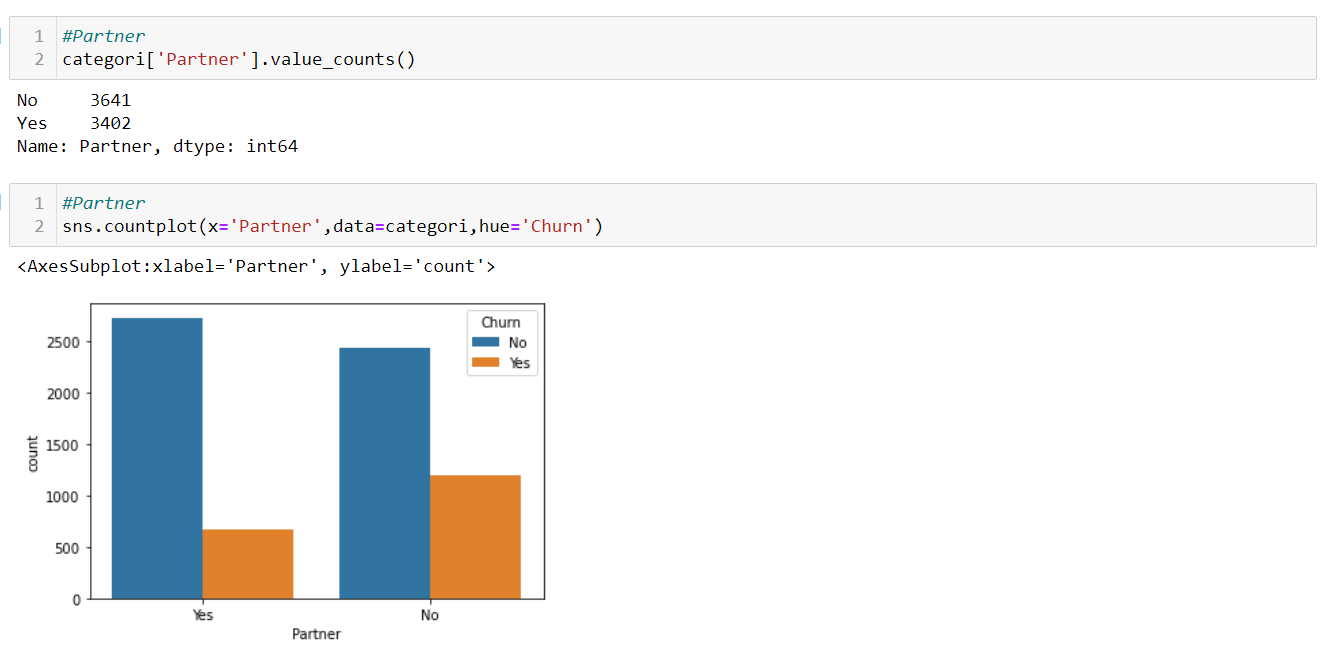
**‘Gender’ Vs ‘Churn’**



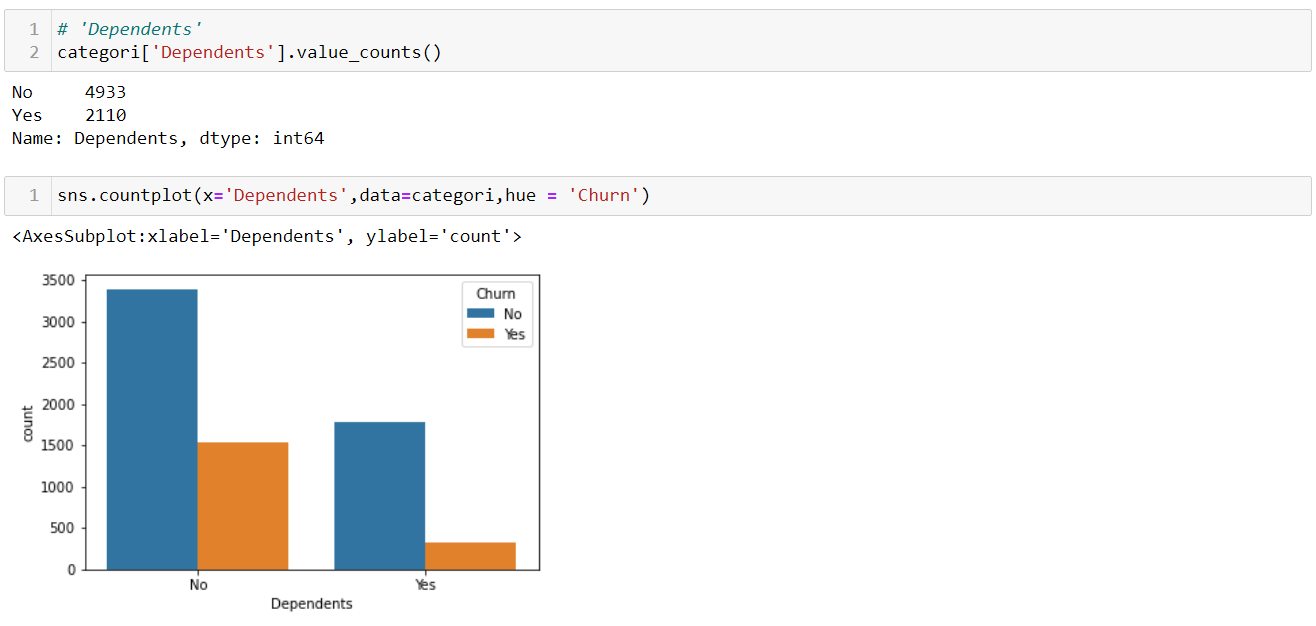
**‘Senior Citizen’ Vs ‘Churn’**



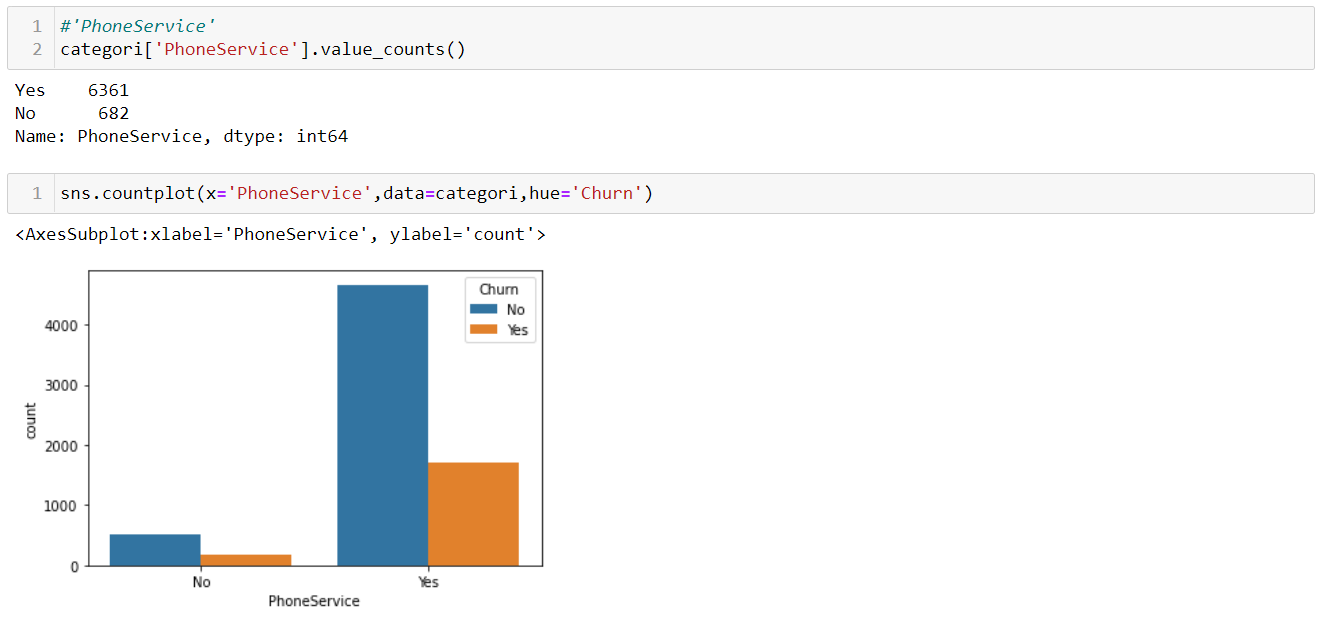
**‘Partner’ Vs ‘Churn’**



‘Dependents’ Vs ‘Churn’



'Phone Service' Vs ‘Churn’



'Multiple Lines' Vs ‘Churn’



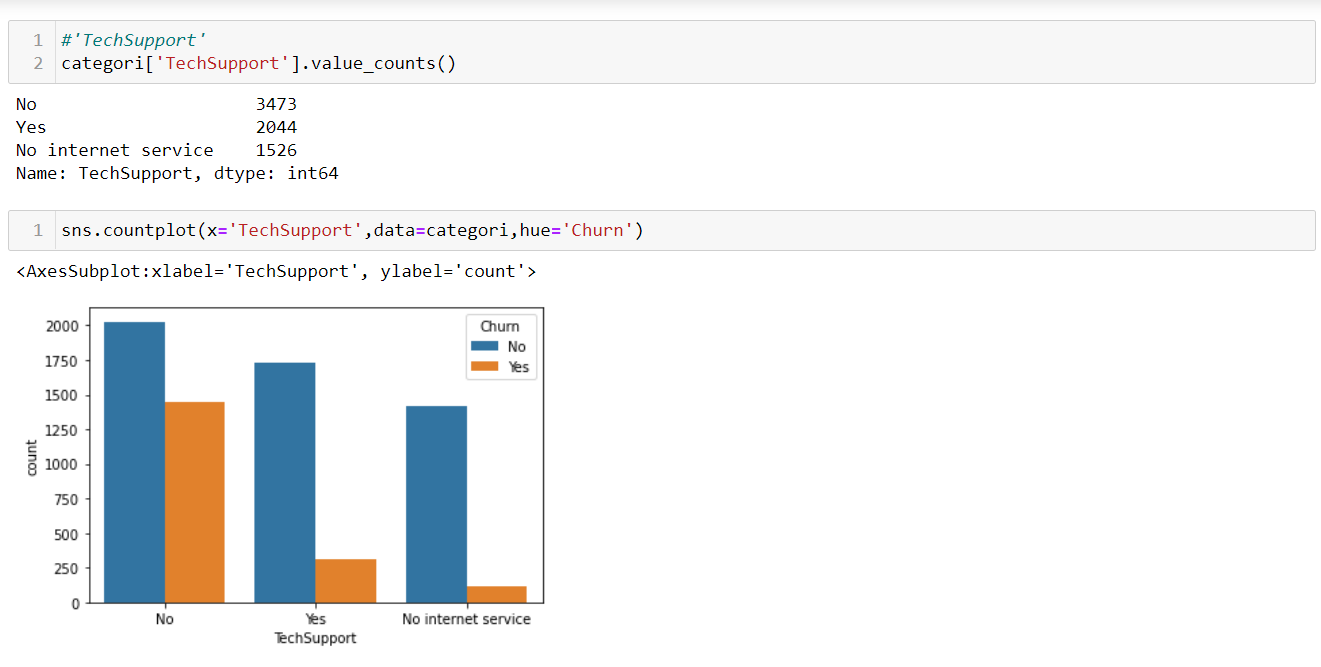
‘Internet Service’ Vs ‘Churn’



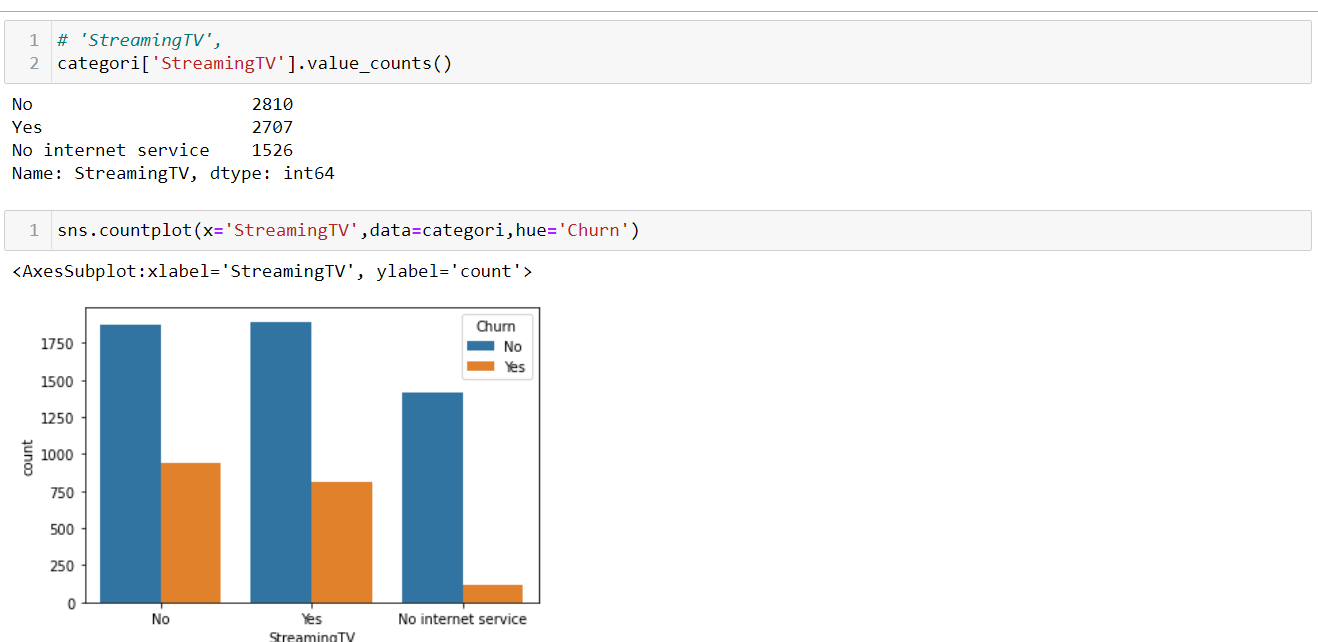
'Online Security' Vs ‘Churn’



'Tech Support' Vs ‘Churn’



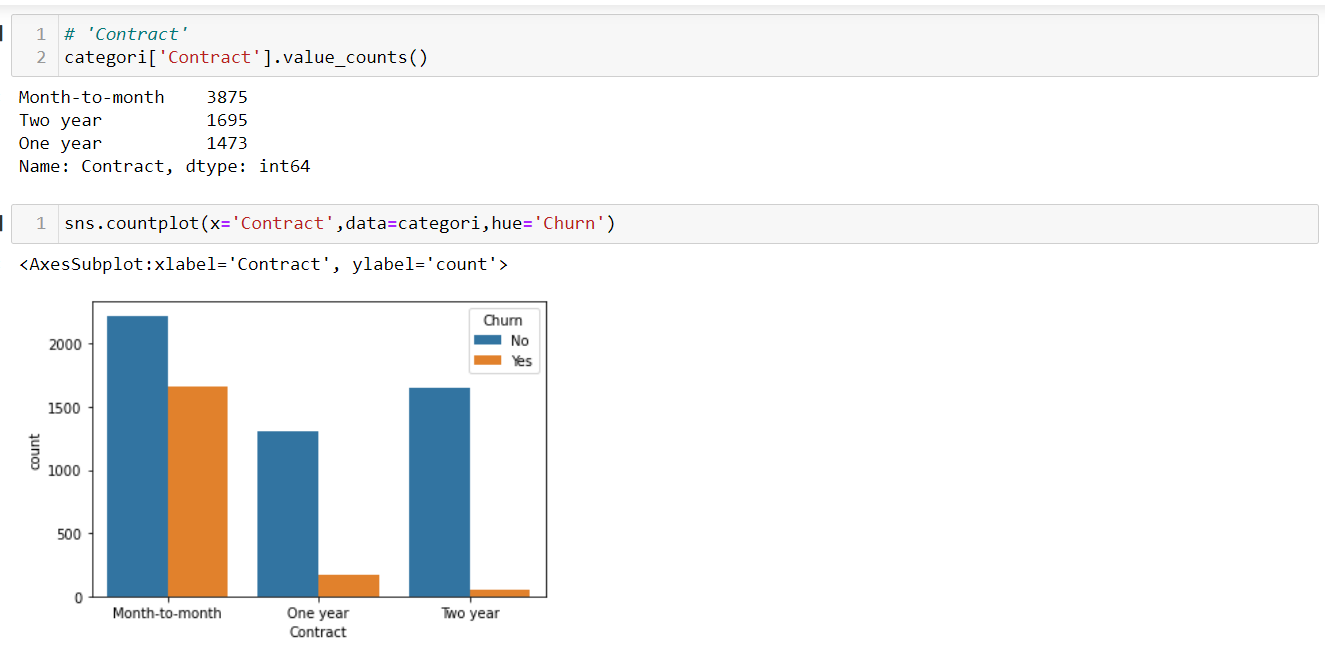
'Streaming TV' Vs ‘Churn’



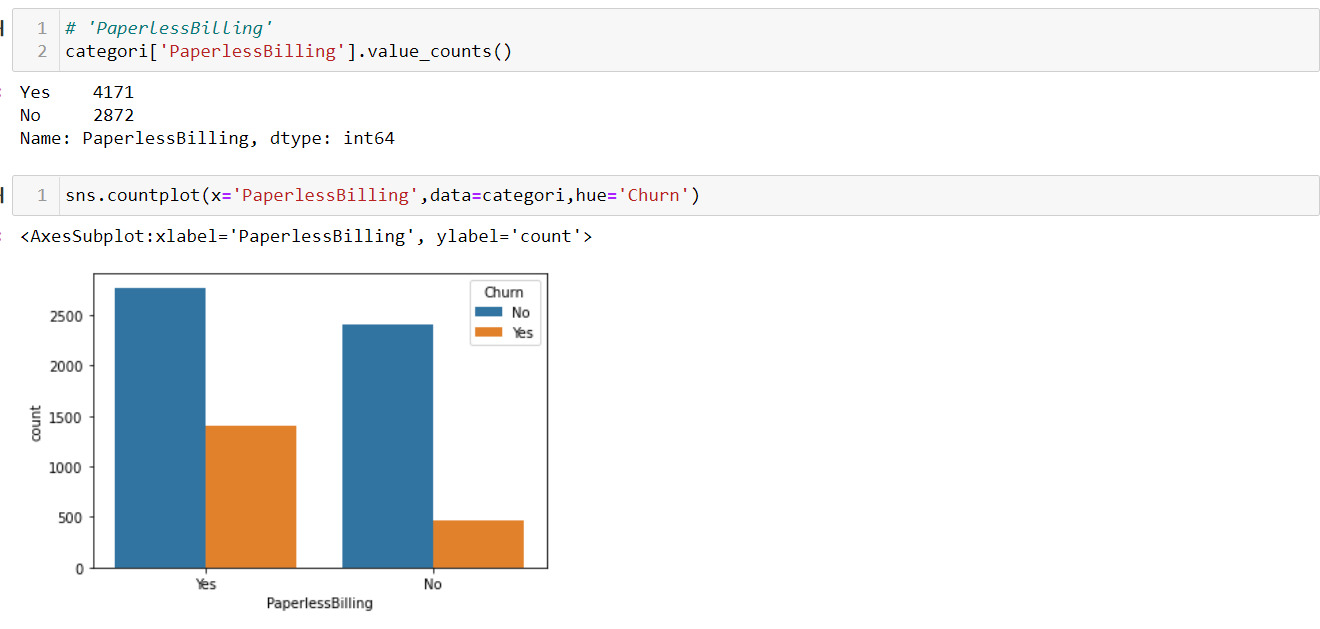
'Streaming Movies' Vs ‘Churn’



'Contract' Vs ‘Churn’



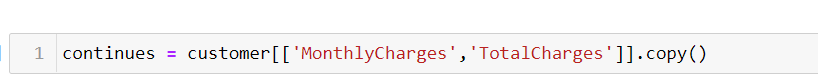
'Paperless Billing' Vs ‘Churn’

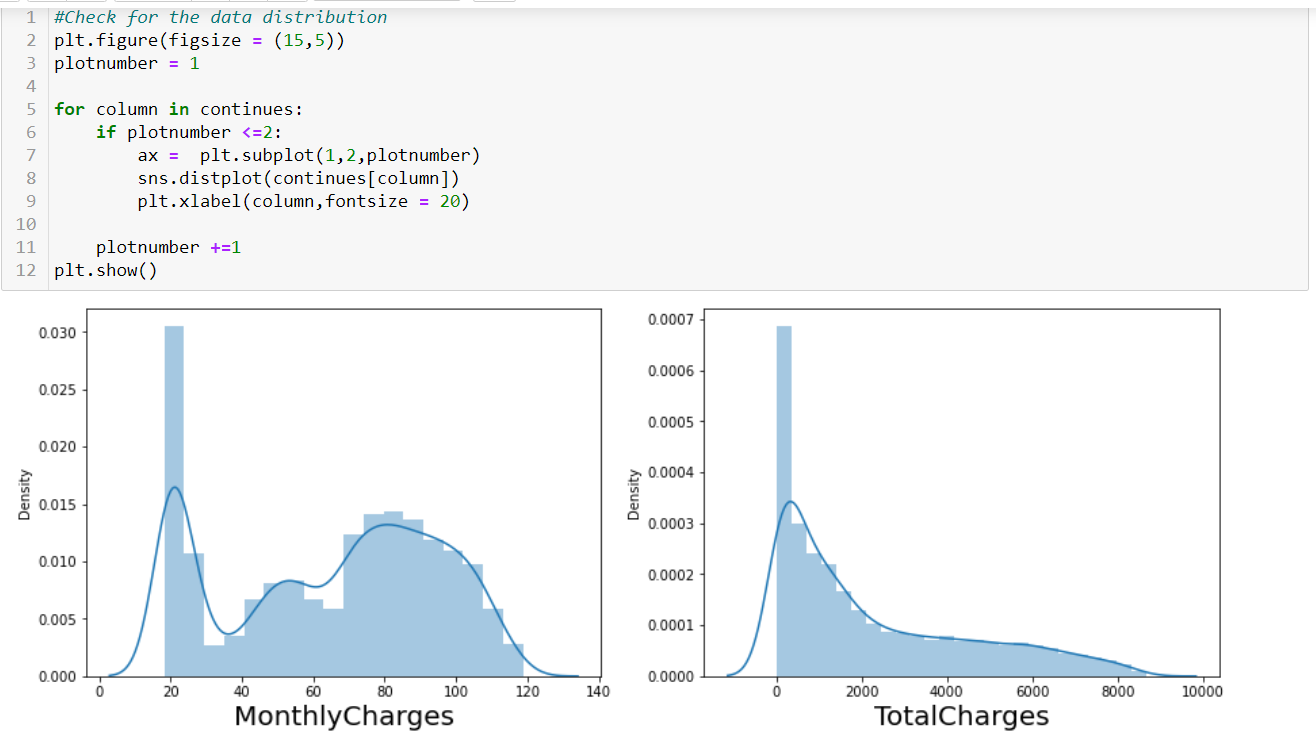


'Payment Method' Vs ‘Churn’



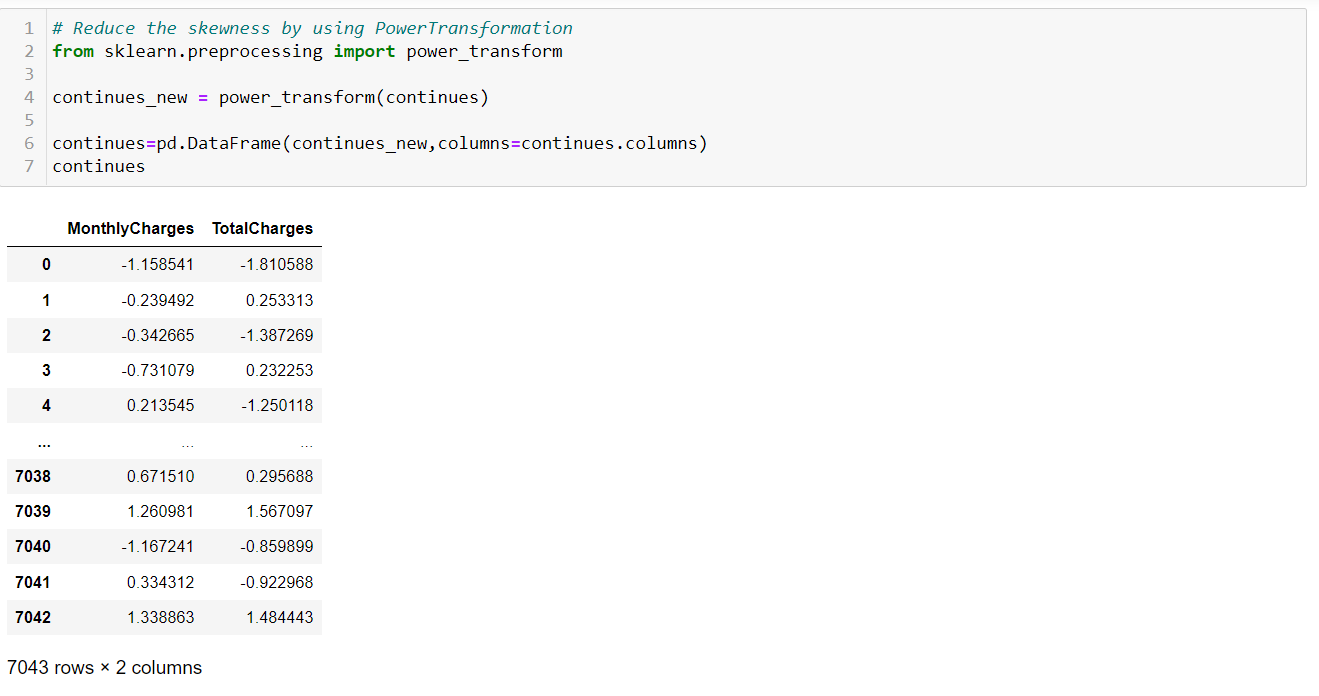
Continuous data:



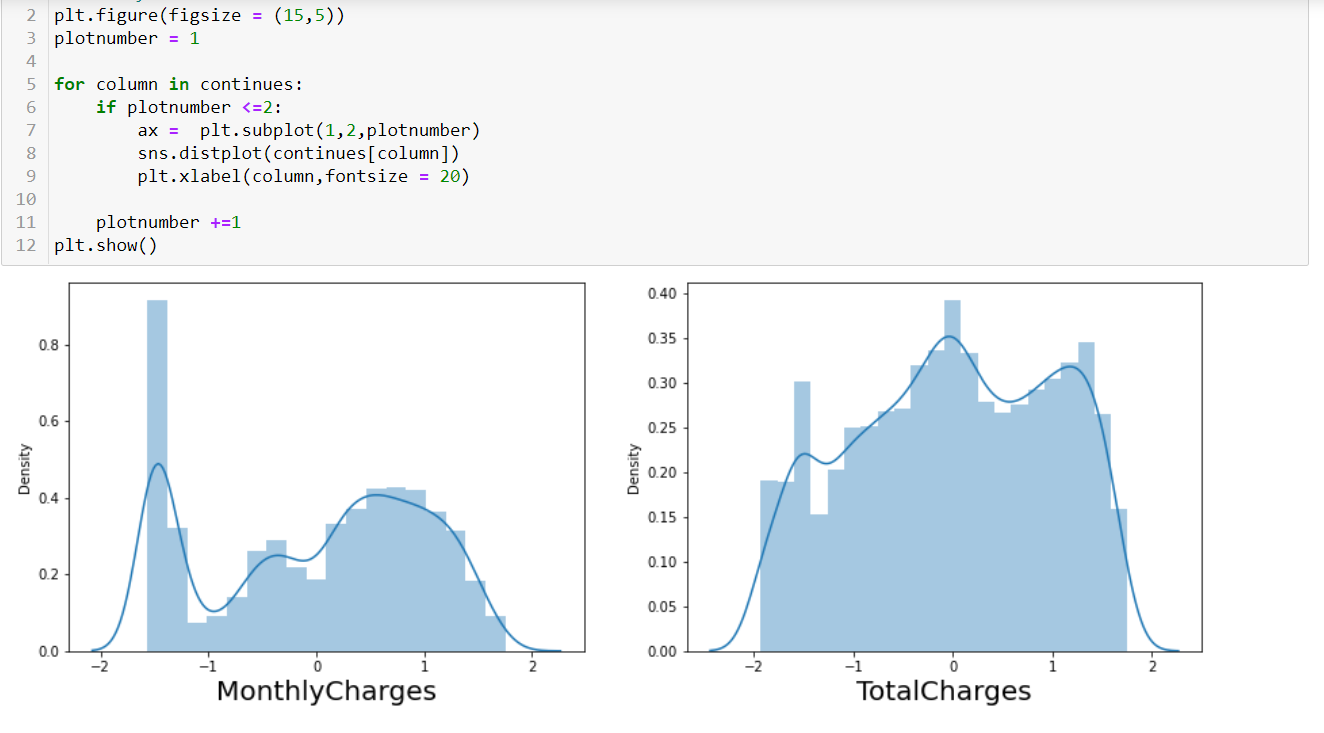


Checking data distribution. ’Monthly charges’ and ‘Total Charges’ are right skewed. We need to reduce the skewness by applying Power transformation methods.

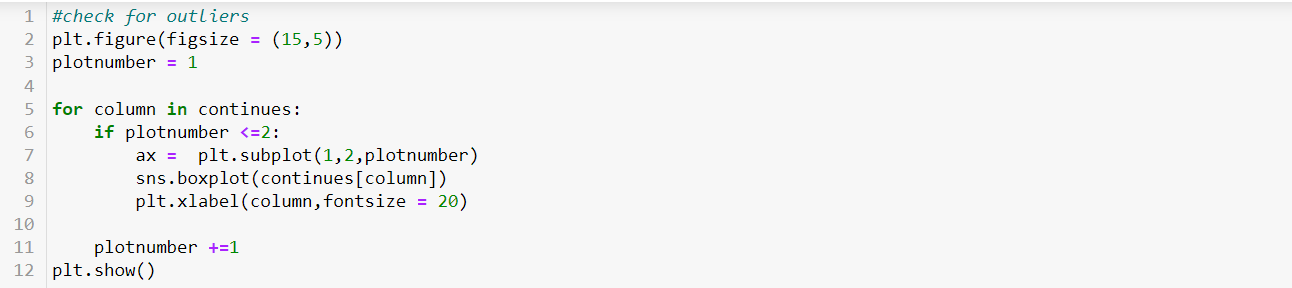


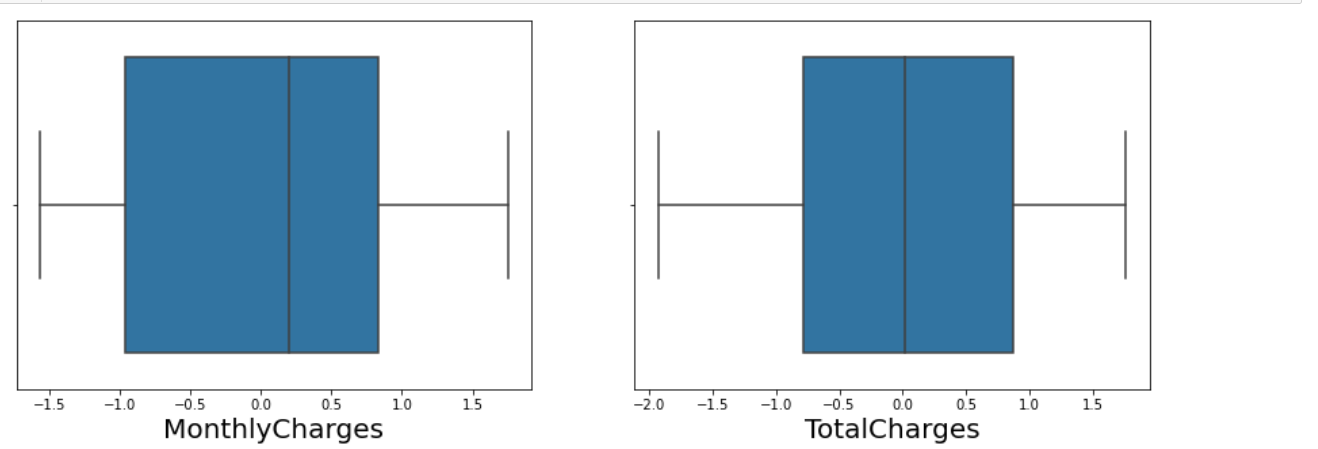






Checking Outlier





There is no outliers present in the ‘Monthly Charges’ and ’Total Charges’ so stepping into next step.

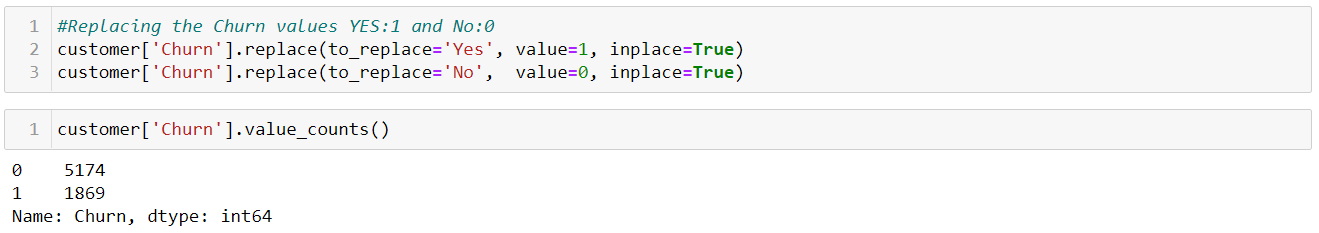
Our dataset almost all has string values. so, we need to encode all of them.1st encode the categorical variable and we go for nominal data.

Categorical data: usually categorical data is represented in ‘String’ or ‘Character’ or finite numbers. For Example: City name, Education degree, Blood groups like that.

Categorical variable is of 2 types: 1. Ordinal and 2. Nominal

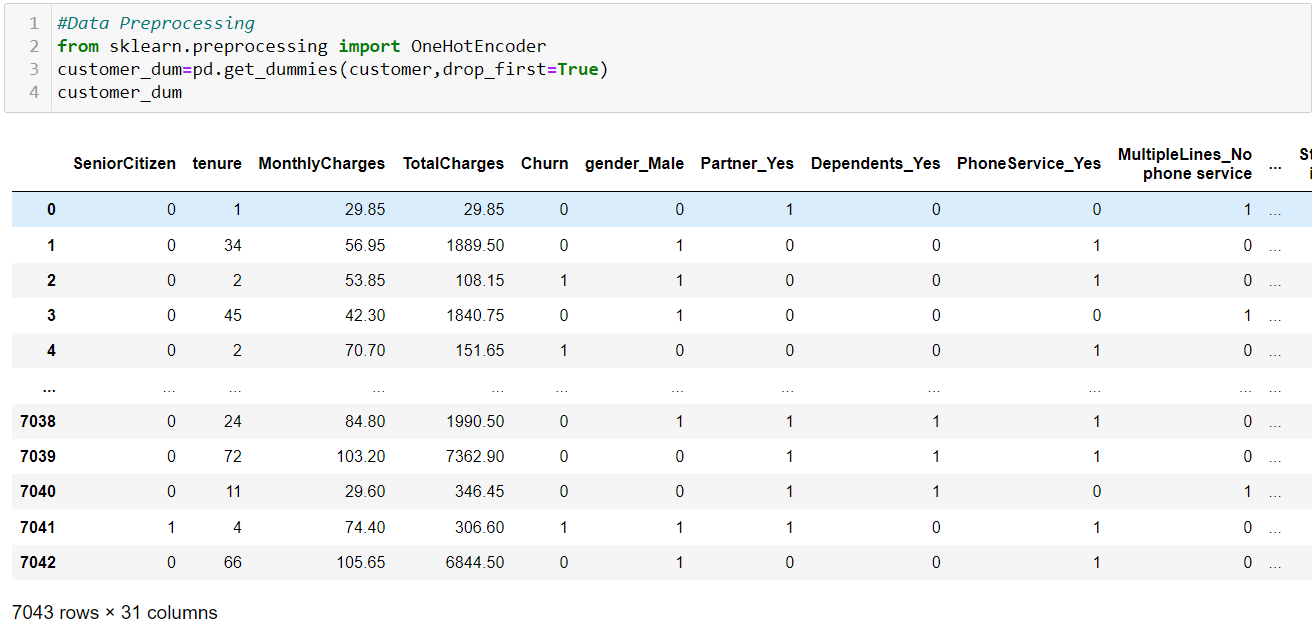
**1.ordinal data**: It have an inherent order. It should retain information regarding the order of the data.

**2.Nomina data**: As name represents no order. It doesn’t retain any order. It just represents the presence or absence category of the data.



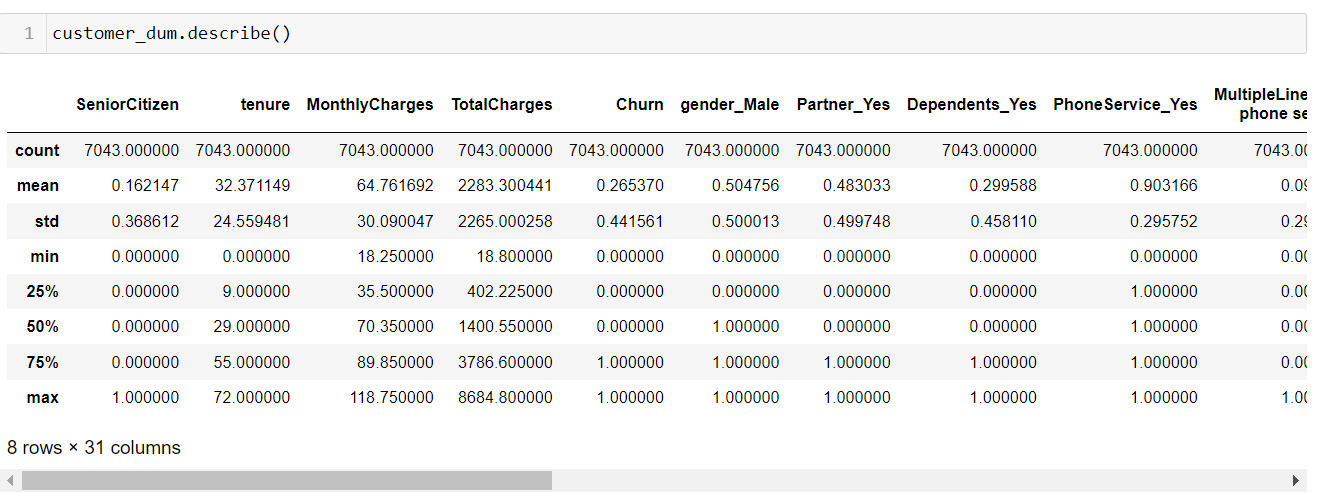
Here, I have only two types of data ‘Yes’ or ‘No’ that why I preferred to convert them manually for I convenience. If we want to convert them automatically then we can go for Label Encoding by importing Label Encoder from sklearn. preprocessing.

For the remaining data. I preferred to go for One hot Encoder. Because they don’t follow any order. Before encoding the data, we need to importing necessary libraries.



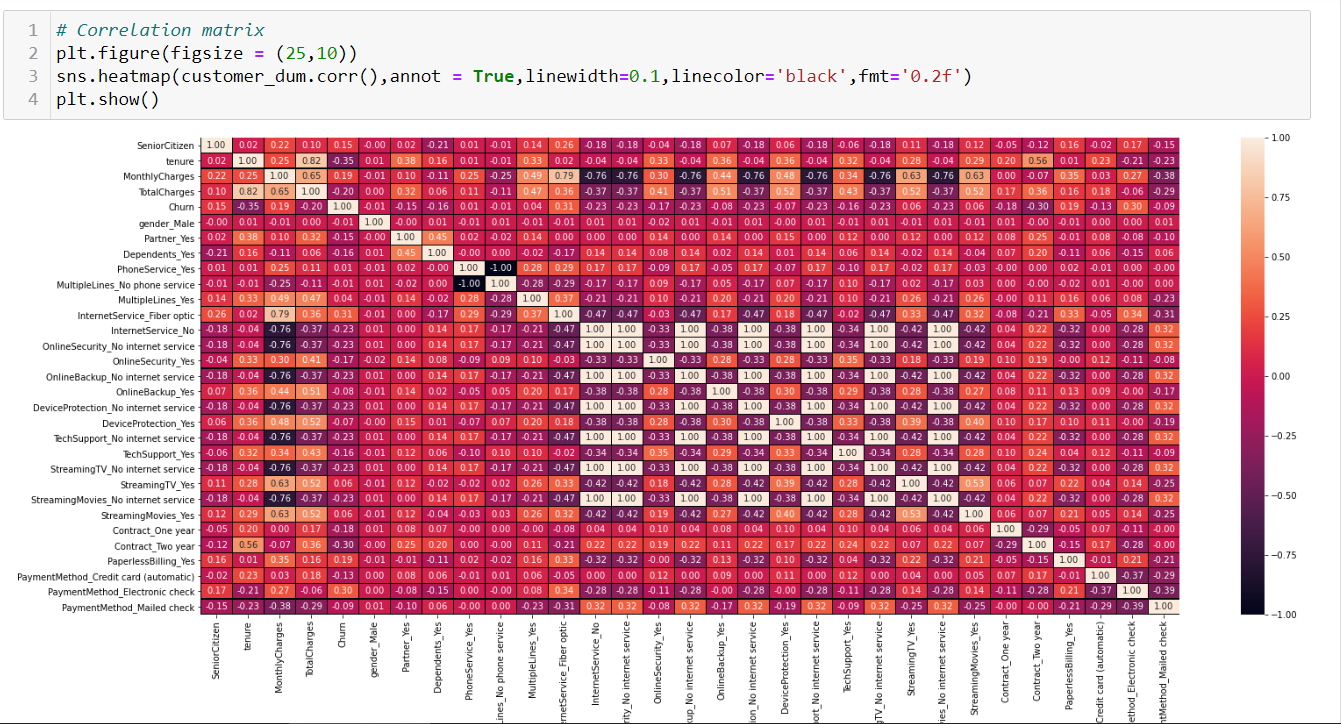
After encoding number of columns and rows in the dataset are 7043 rows and 31 columns. Before encoding 7043 rows and 20 columns.

Statistical Measure:



**Correlation Matric:**

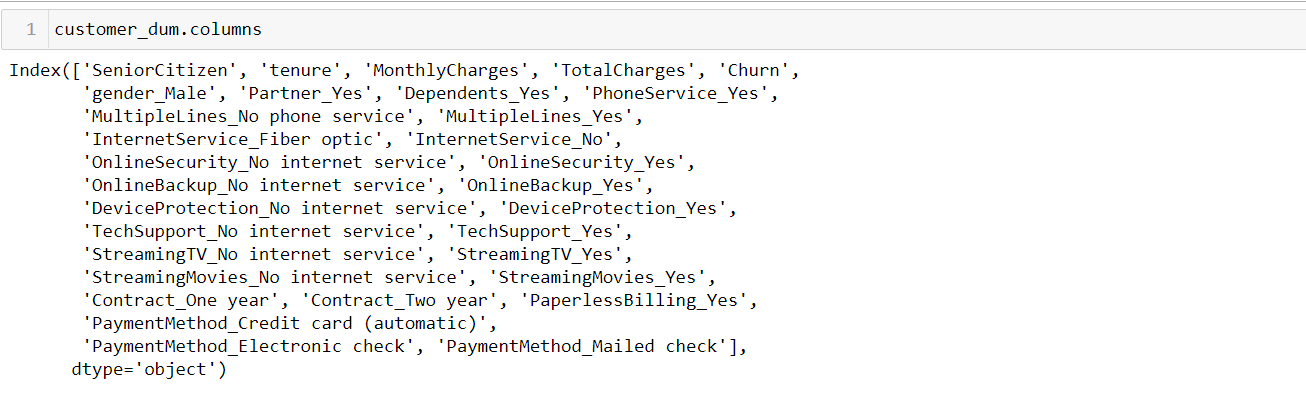
By this correlation matrix we can calculate or measure the relation between each variable and also, we can find if the variable is positively correlated or negatively correlated with each We

an represent or visualize the correlation matrix with heat map.

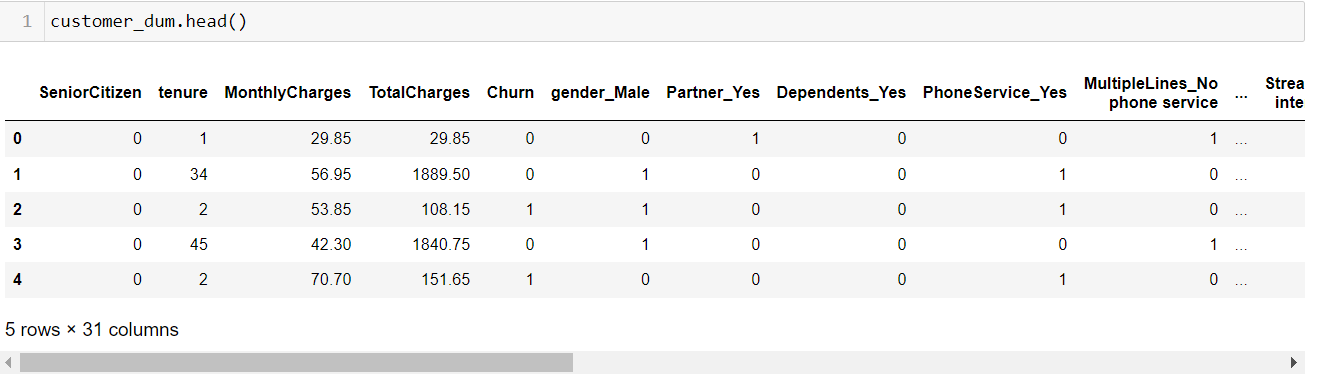
From the heat map, we can observe that most of the columns are internally positively correlated with each other. If we observe the target column ‘Churn’ with ‘Senior Citizen’, ’Monthly Charges’, ’Phone Service’, ‘Multiple Lines’, ‘Internet Service Fibre optic’, ’Streaming TV’, ’Streaming Movies’, ’Paperless Billing’, ’payment method electronic’ are positively correlated with ‘Churn’.

And remaining ‘tenure’, ‘Total charges’, ‘gender’, ‘ Partner’, ‘Dependent’ , ’Multiple lines no phone service’, ’Internet Service No’, ‘Online security No internet’, ’Online backup’, ’device protection‘, ‘ Technical support’, ‘ Contract’, ‘ payment method credit card’ and so on are negatively correlated with ‘Churns’.

Our final dataset columns and data look likes:



**Dataset**:



**4. Pre-Processing Pipeline**:

Before going to modelling the data, we need to pre-process the data. The data preparation process can be complicated by issues.

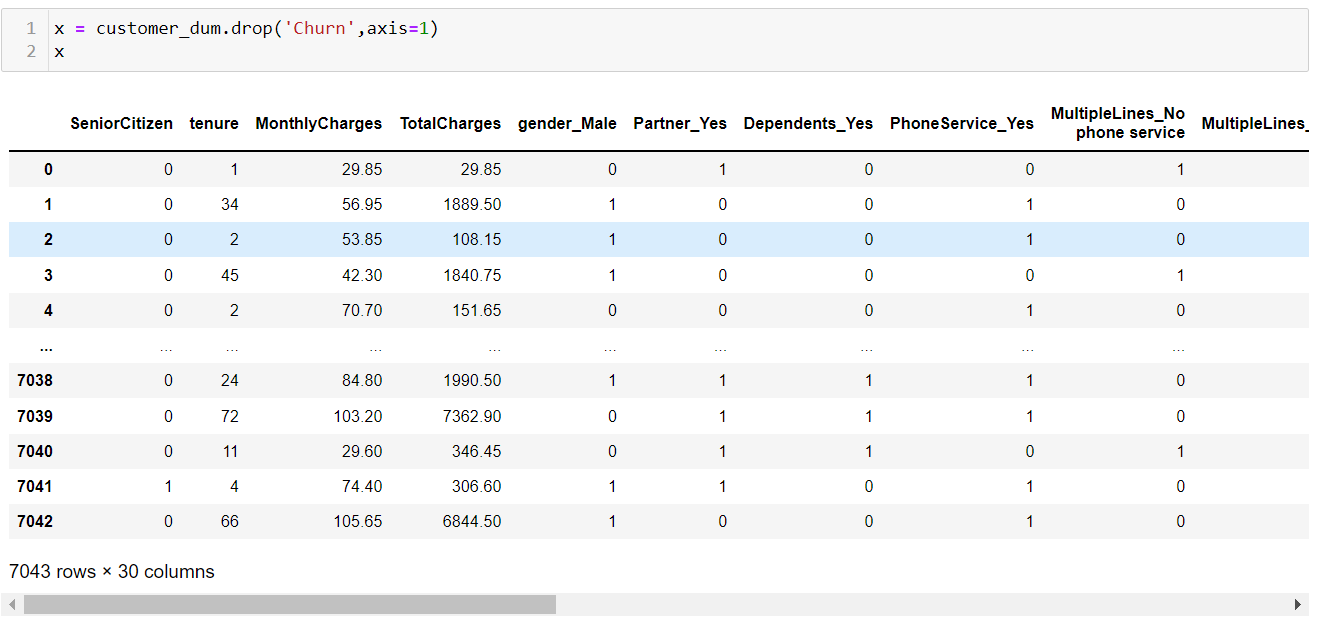
1.Data cleaning: Is nothing but finding the missing values and null values. Replace with them according to type of the data set.

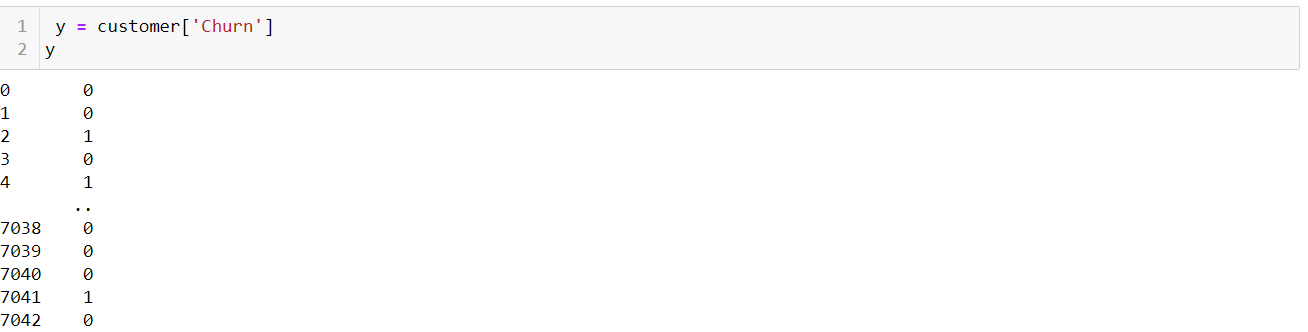
2.Data Transformation: Check for whether the data is normally distributed or not. if not applying some techniques square root, cube root and log based on the type of the data.

3.Feature Engineering**: in simple terms, is the act of converting raw observations into desired features using statistical or machine learning approaches.**

4.Dimentional Reduction: This reduces the dimension of the data set as per requirement

5.Feature selection: Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.



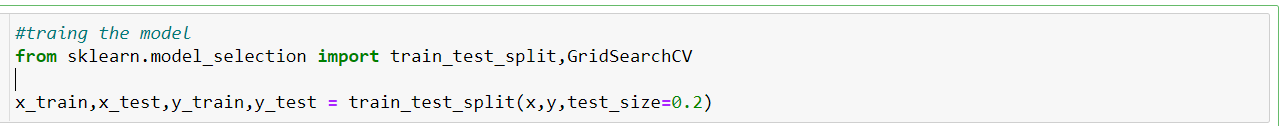


**Data Scaling**:



**5. Building Machine Learning Models**:

Training the data:



Before modelling the data, we need know the definition of What is Classification Accuracy,

Confusion matrix and classification report in that precision, recall, f1\_score and support.

**Classification Accuracy**: Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.



It works well only if there are equal number of samples belonging to each class.

**Confusion matrix:**

|  |  |  |
| --- | --- | --- |
|  | **Positive** | **Negative** |
| **Positive** | **TP** | **FP** |
| **Negative** | **FN** | **TN** |

Where the terms have meaning:

. **True Positive (TP):** A result that was predicted as positive by the classification model and also is positive.

. **True Negative (TN):** A result that was predicted as Negative by classification model but actual is negative.

. **False Positive (FP):** A Result that was predicted as Negative by classification model but positive actual is Negative positive.

. **False Negative (FN):** A Result that was predicted as Negative by classification model but actually is positive

**Recall or Sensitivity**: It is measure of total number of positive results. how many positive we correctly predicted by the model.

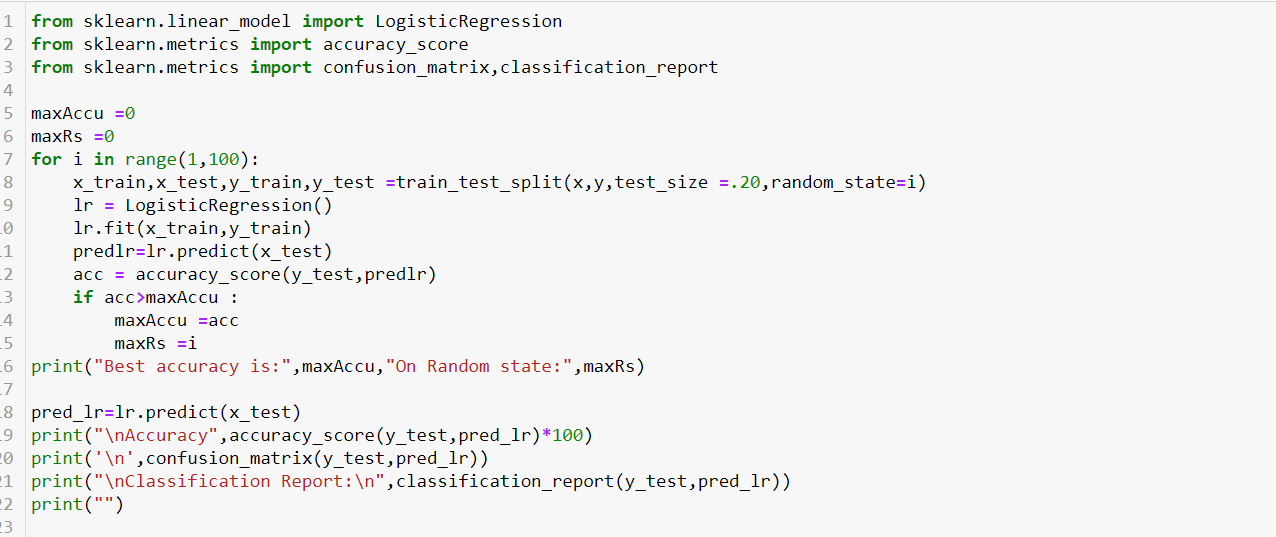
Recall = (TP) / (TP +FN)

**Precision:** It is measured of amongst all of the positive predictions. How many of them were actually positive?

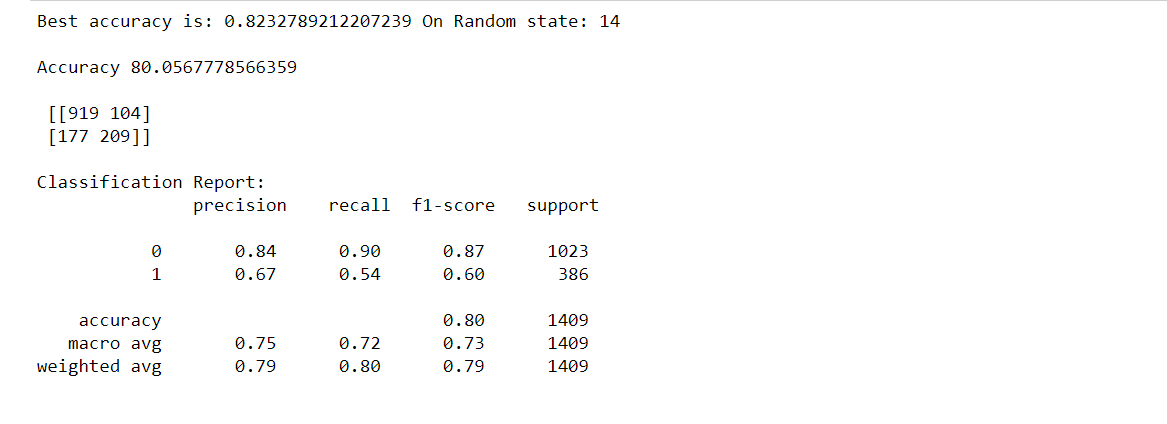
Precision = (TP) / (TP + FP)

**F1 Score:** F1 Score is used to measure a test’s accuracy.

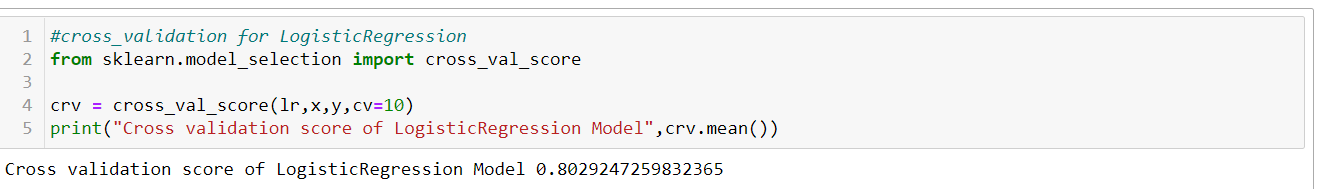
**Logistic Regression:**



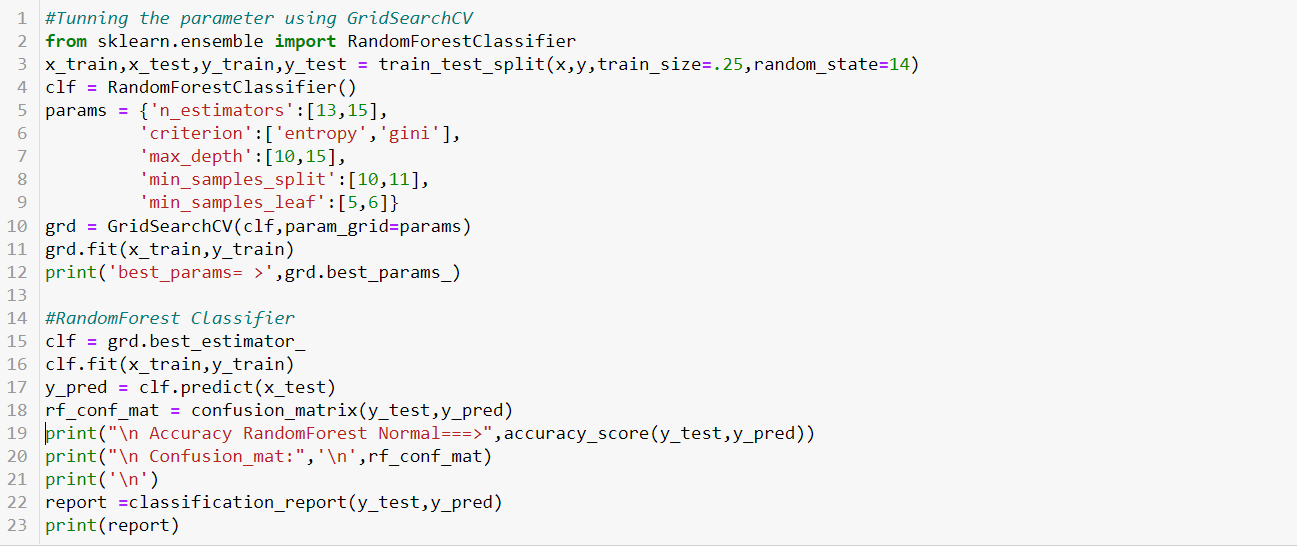
**Resultant Matric:**

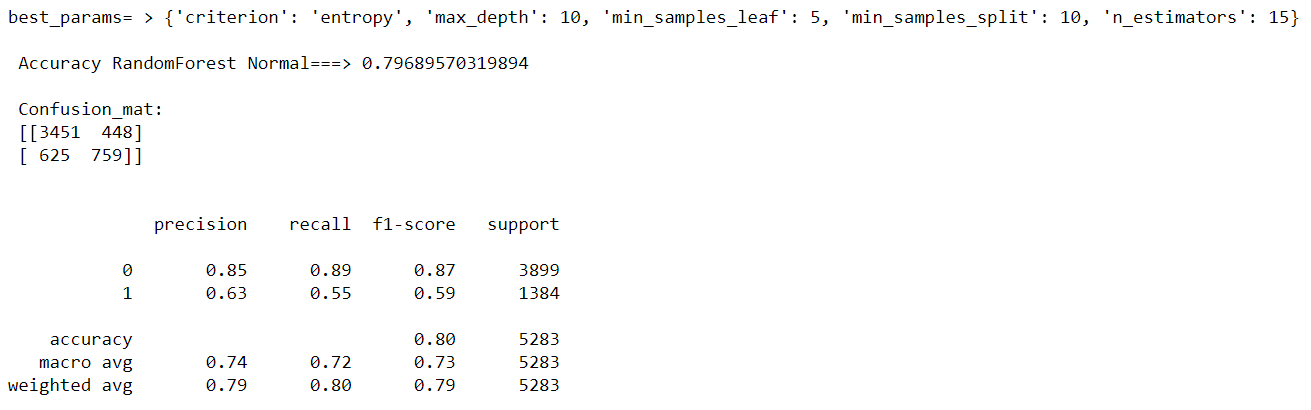


**Cross Validation:**

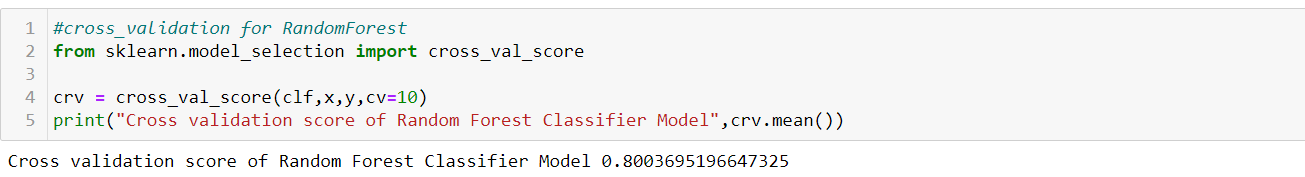


**Random Forest Classifier**

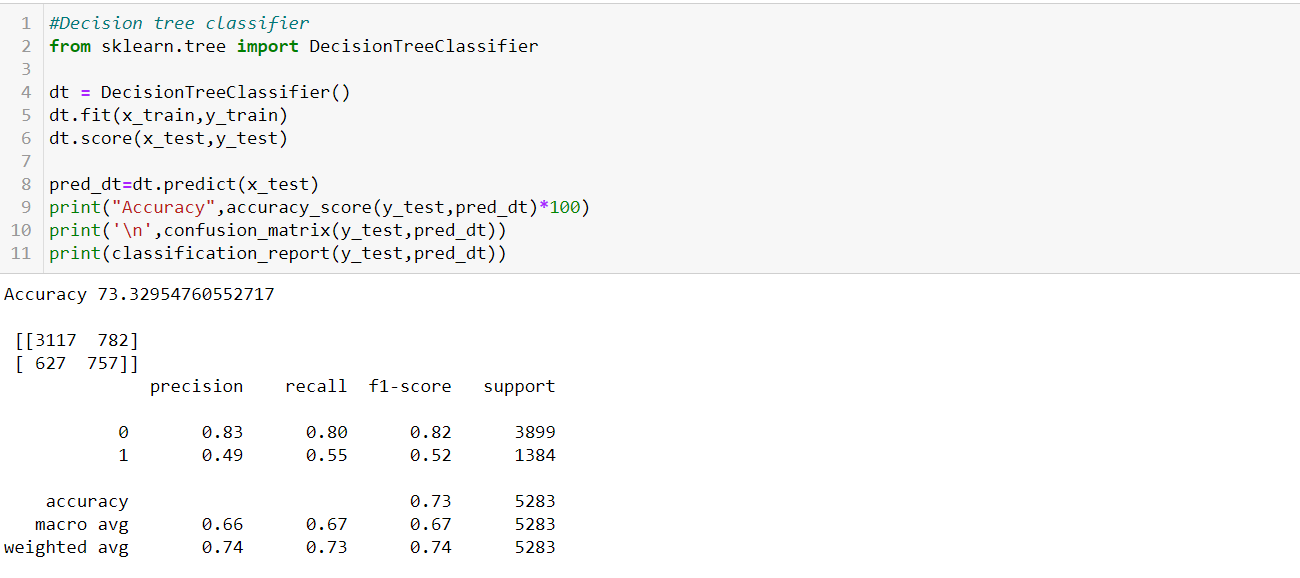




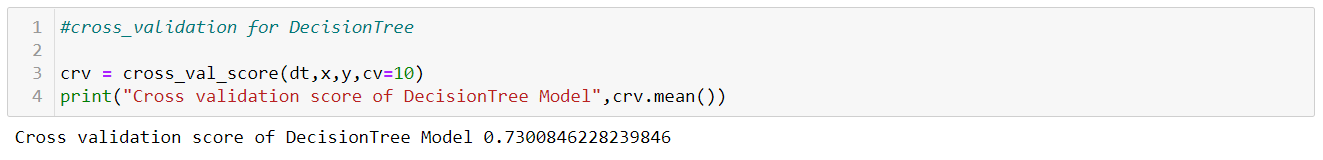
**Cross Validation:**



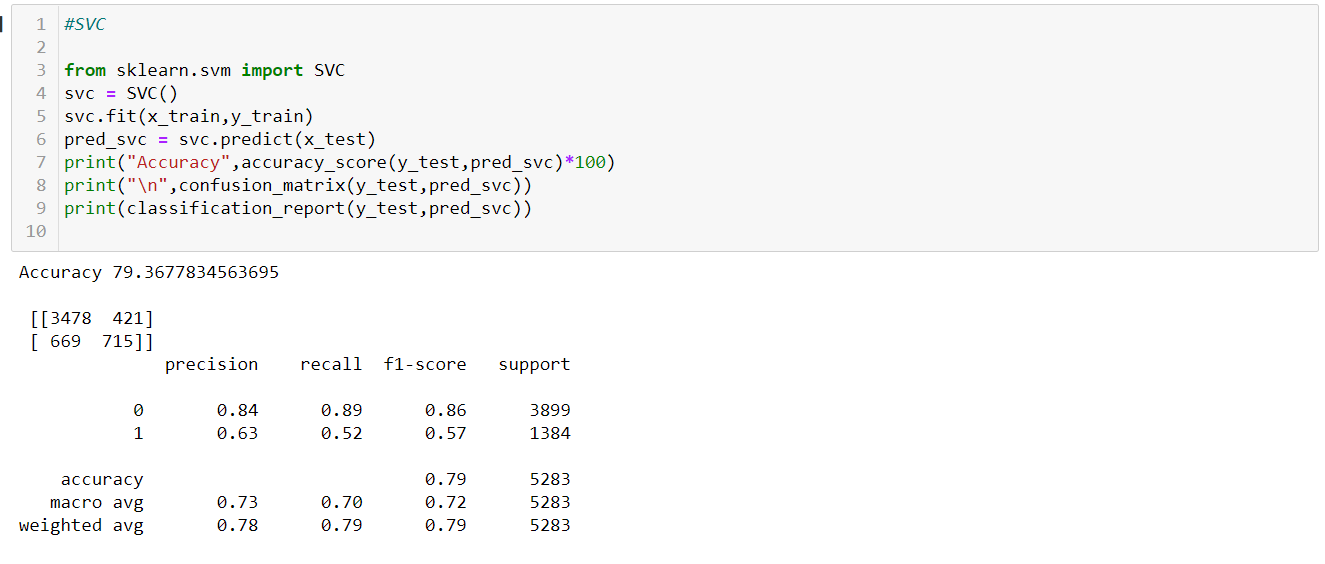
**Decision Tree:**

****

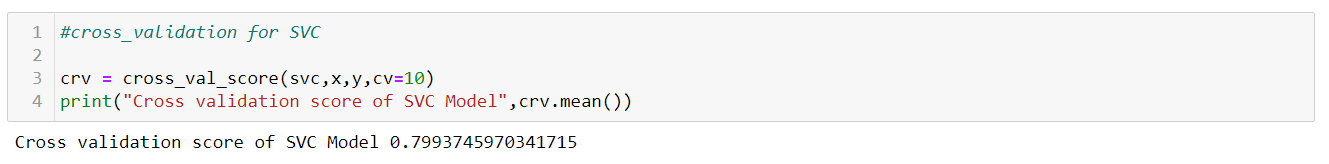
**Cross validation:**

****

**SVC:**

****

**Cross Validation:**

****

**From the model building i observed one thing that Accuracy score and CV score are**

# For Logistic Regression Accuracy score is: 82% and CV Score:80%

# for Random Forest Accuracy score is: 79.9% and CV Score:79.9%

# For Decision Tree Accuracy score is: 73.7% and CV Score:73.9%

# For SVC Accuracy score is: 79% and CV Score:79.9%

# So, from the result Random Forest Classifier and SVC model giving the best results

**AUC\_ROC\_CURVE:**

**Area Under Curve:** Area Under Curve (AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem.

. **True Positive Rate (Sensitivity)**: True Positive Rate is defined as*TP/ (FN+TP)*. True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

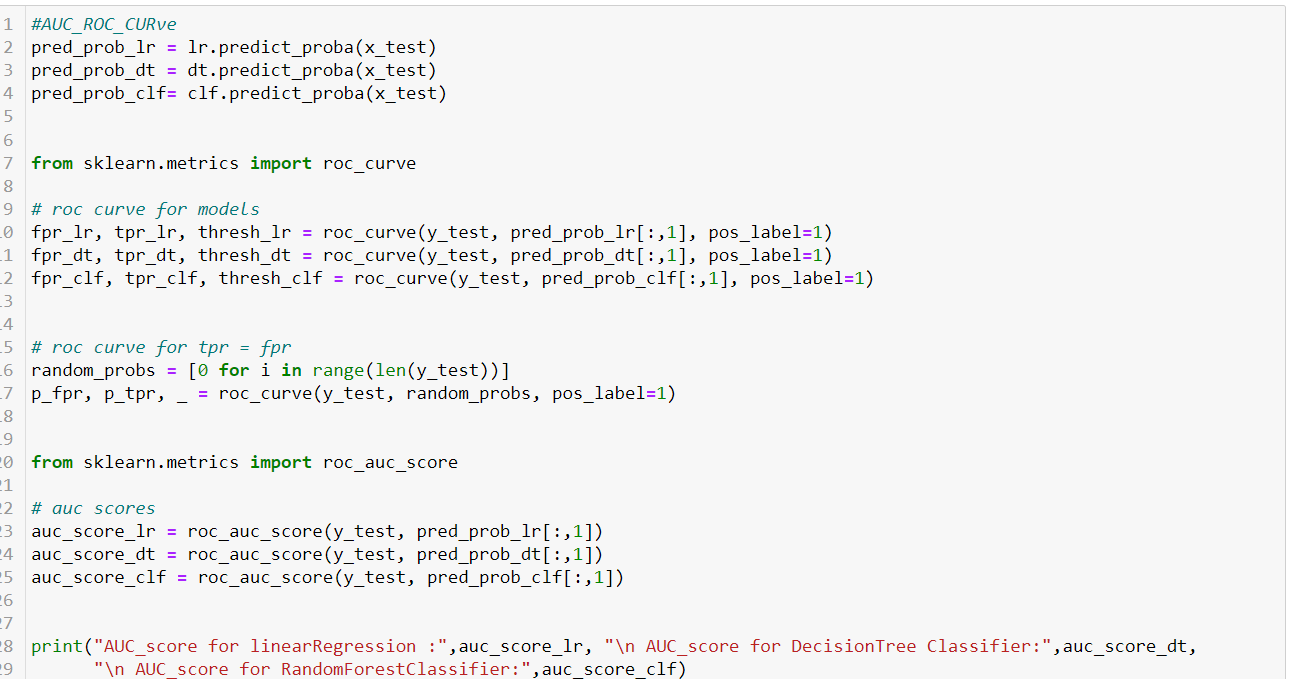


**. True Negative Rate (Specificity)**: True Negative Rate is defined as *TN / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are correctly considered as negative, with respect to all negative data points.

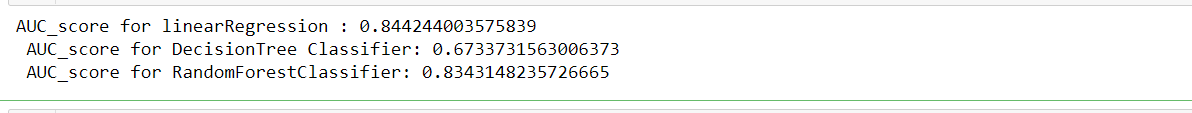


**. False Positive Rate**: False Positive Rate is defined as *FP / (FP+TN)*. False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

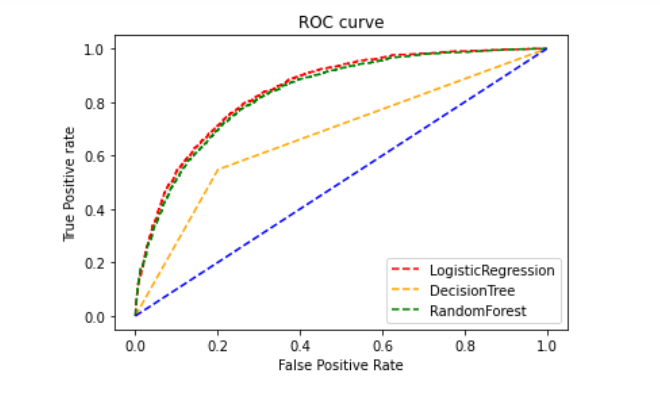




Result:



****

****

**6.Conclusion:**

* We came to the end. From the **EDA** can help us identify which features contribute to customer churn. From the above ROC Curve Logistic Regression and Random Forest gives 83% of result for our model. We can build several **machine learning models** with **recall** approximately equal to **78%**, meaning that they can successfully detect almost 80% of those customers more luckily to churn. Perhaps, adding more features or/and records could help us improve predictive performance.