****

**PROBLEM STATEMENT:**

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.

You will examine customer data from IBM Sample Data Sets with the aim of building and comparing several customer churn prediction models.

**PROBLEM DESCRIPTION:**

In this article I will be going to analyse the customer churn to achieve customer retention.

**What is customer churn?**

Customer churn is the percentage of customers that stopped using your company's product or service during a certain time frame.

**How is customer churn important?**

**Customer churn** is important because it costs more to acquire new customers than it does to retain existing customers.

**For Example: -** In fact, an increase in the customer retention of just 5% can create at least a 25% increase in profit.

**CUSTOMER RETENTION:**

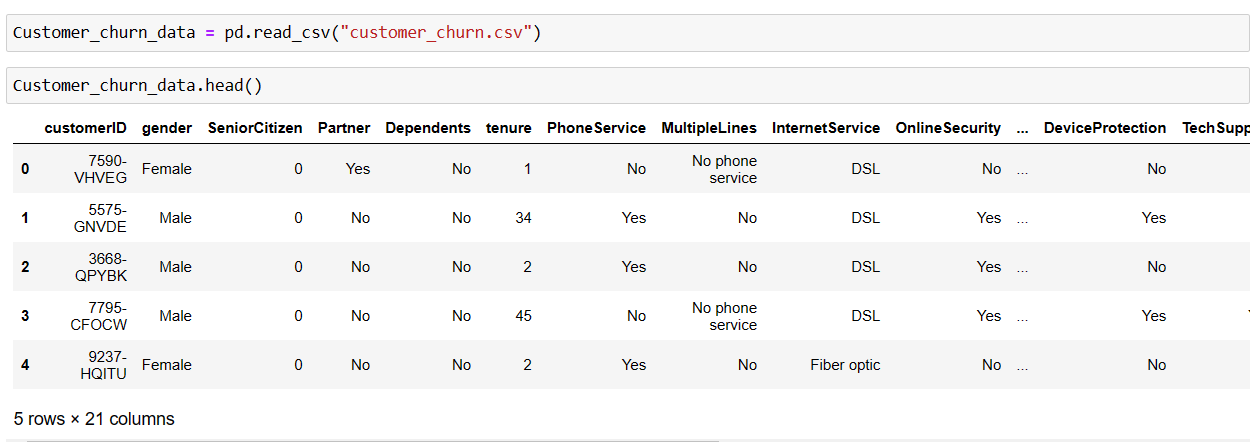
****

This model will take through each and every step present, in detail and helps you to understand the whole machine learning model building process.

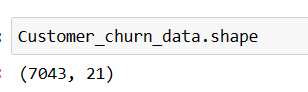
Now-a-days customer churn has become most important issue for many of the companies. Because of this price of the products are affected very badly, which is creating bad impact on the companies or industries in turn ultimately people are also affected.

* **LIBRARIES USED:**
* import NumPy as np
* import pandas as pd
* import matplotlib. pyplot as plt
* import seaborn as sns
* **After importing the libraries, we have to collect the data for analysis.**

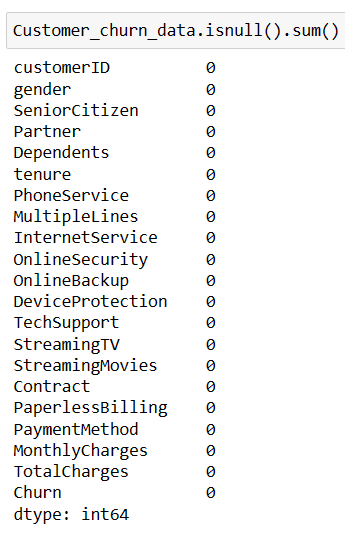
**DATA COLLECTION:**

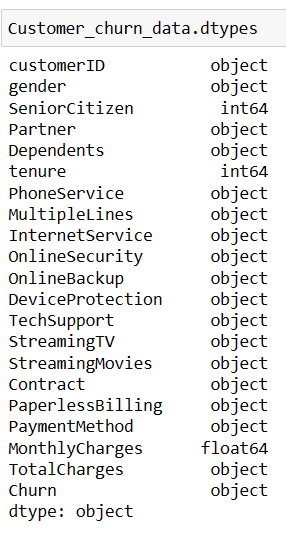
****

* Here we have collected the data and predicting customer churn rate which is converted into csv format and this csv file is converted into dataframe to read the file. In the above collected data we can see that we have 21 columns and also the data is the combination of the numerical and categorical values that means further “Data Cleaning” have to be done for the good accuracy of the model.
* **SIZE OF THE DATA:**

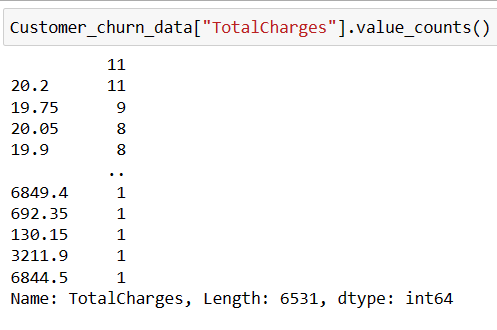
****

* As we can see that we have 7043 rows and 21 columns. Further we have to check the null values, datatypes and also statistical analysis of the data and then we have to analyse to drop any unnecessary columns present in the dataset.



****

* Now here from the above information we can analyse that our dataset doesn’t have any null values and also our dataset contains the data which is the combination of “object”, “float” and “int” datatypes but also the column “Total Charges” is seems to be with "int" or "float" datatypes but it is "Object datatype" so we have to convert it into "float" data type or "int" datatype.
* Now we will check the value\_counts of the column and proceeding with processing of the column proceeding with some changes.



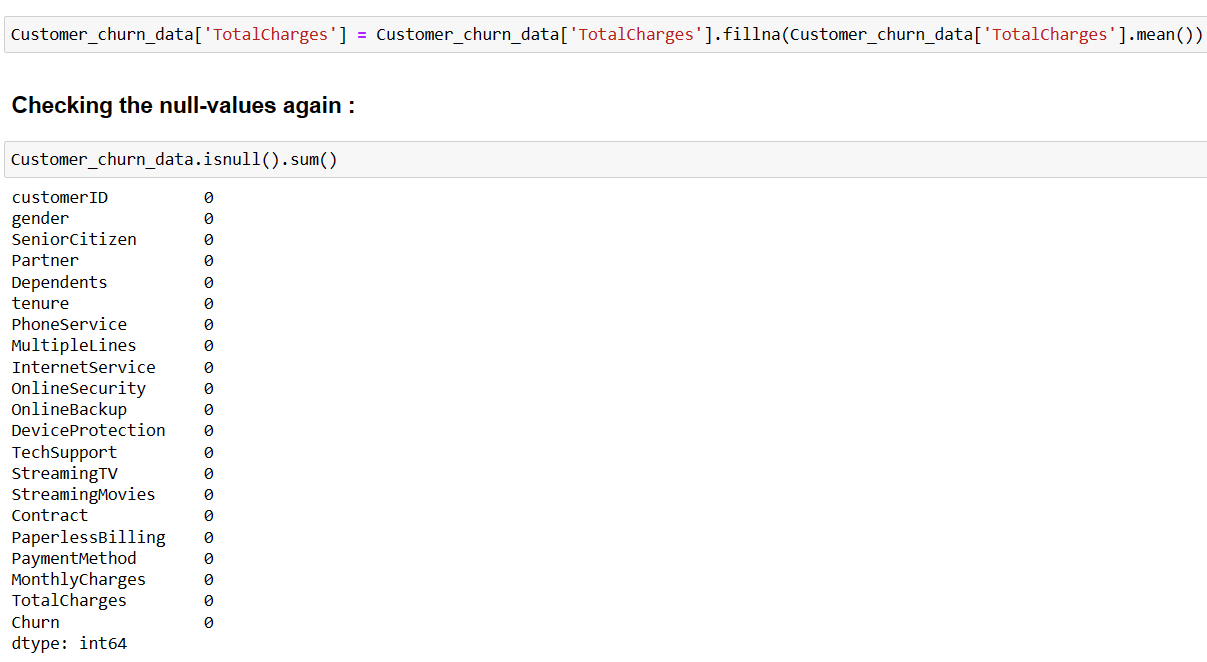
* I can say that there is the data in the columns which is "" and so because of this we are not able to convert into float and so we will replace it with the "nan" data and then we will drop the null values again.



* Here we have converted or replaced the gaps in the column with “nan” values and then after we have checked the null values which we can see that in the column “Totalcharges” we can see that there are 11 nan values which have to be filled further and convert to float datatype to proceed for further preprocessing.



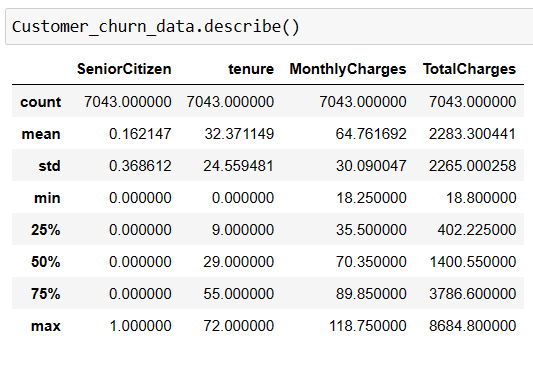
* Here we have successfully converted the column “Totalcharges” from object datatype to float datatype and so now we can proceed filling with the mean values and then check for the null values again.



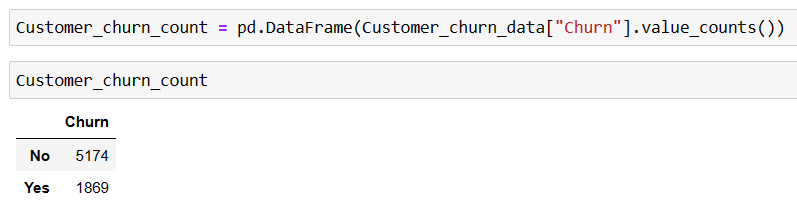
* Here we have filled the null values with “. fillna” method while filling the spaces with “mean” values.
* **Dropping the unnecessary columns:**

****

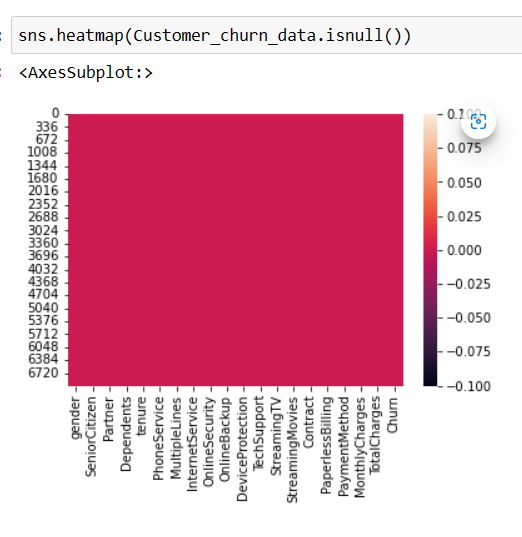
* **T**he column “Customer\_id” has no relationship with our label and thus we have dropped the column and proceed further for our model building.
* **Statistical analysis of the dataset:**



* We can observe that in most of the above numerical columns mean is little greater than standard deviation except for the column “SeniorCitizen”.
* By the observations and the statistical analysed data, it seems that the dataset seems to be perfect and also there are no negative/invalid values present.
* Also, we can observe that “mean” value is greater than “median”ie., 50% quantile in the columns “Tenure” and “Totalcharges” which indicates that these columns are skewed towards right ie., these columns have positive skewness.
* Also, we can observe that the column “Monthlycharges” has the “mean value” smaller than “median value “ie., 50% quantile, which indicates that the column is skewed towards left which means this column has negative skewness.
* Also, we can observe that there is large difference between 75% quantile and mx quantile which indicates that the data has outliers within it which have to be treated in further processing, missing it may affect our model accuracy.
* **Checking the count for our label column “Churn”:**

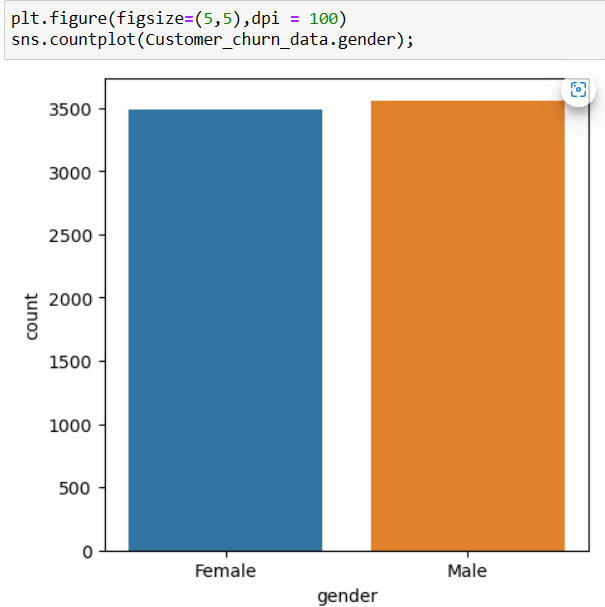
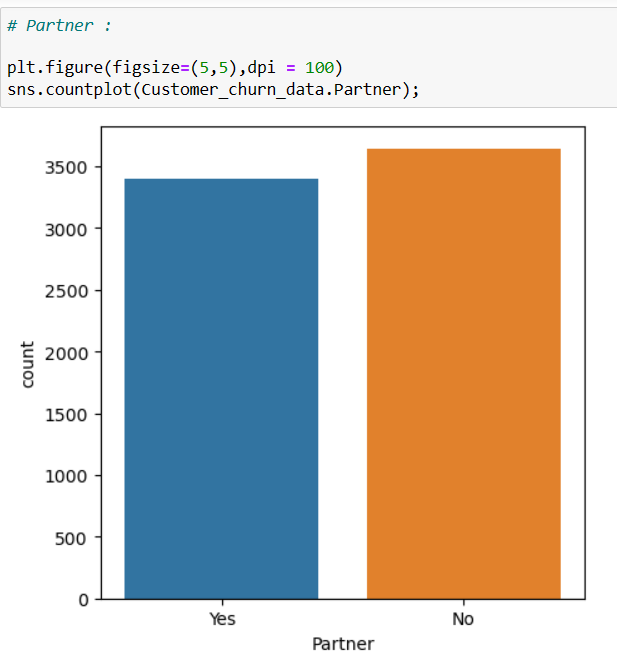
****

* Here we can see that we have taken the “**value\_counts**” for the column “Churn” and also we have converted it into dataframe and also which clearly indicates that the number of customers who have churned is indicated by the category “**Yes**” and the number of customers who have not churned is indicated by the column “**No**”.
* **Now we will plot heatmap for null-values:**

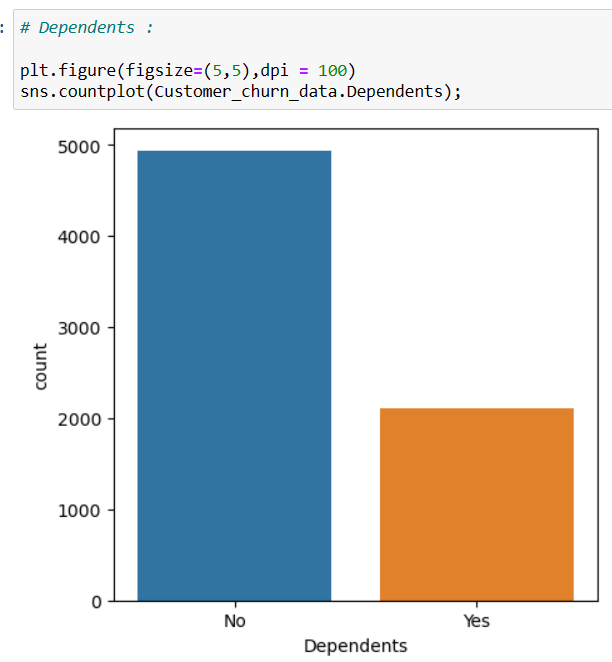


* Here we can see that there no null-values can be seen in heatmap.
* Now let’s visualize our columns individually **(“Univariate analysis”)** and also visualizations of the columns with our label **(“Bivariate analysis”)** and even **“correlation”**.
* **VISUALIZATION:**
* **Univariate analysis:**

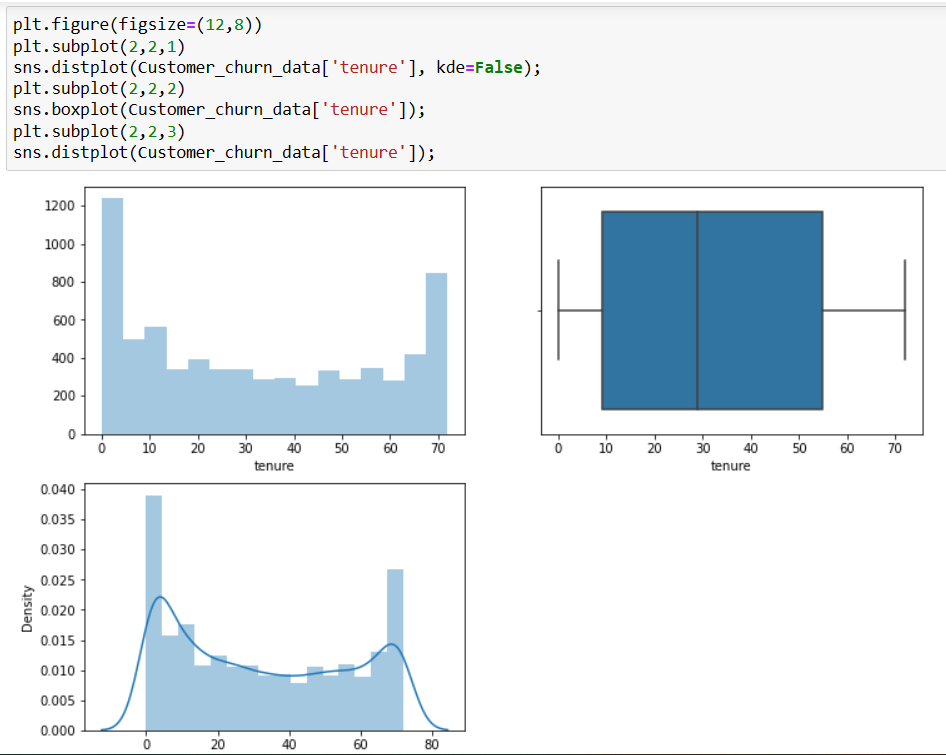
**Gender Partner**

** **

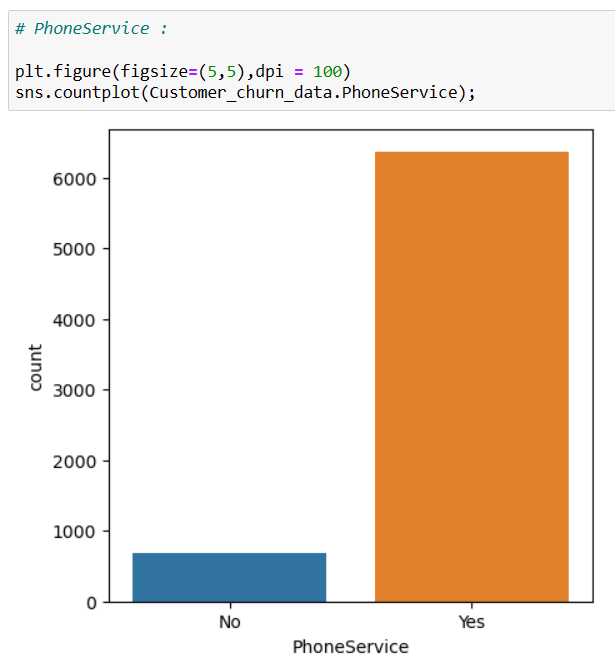
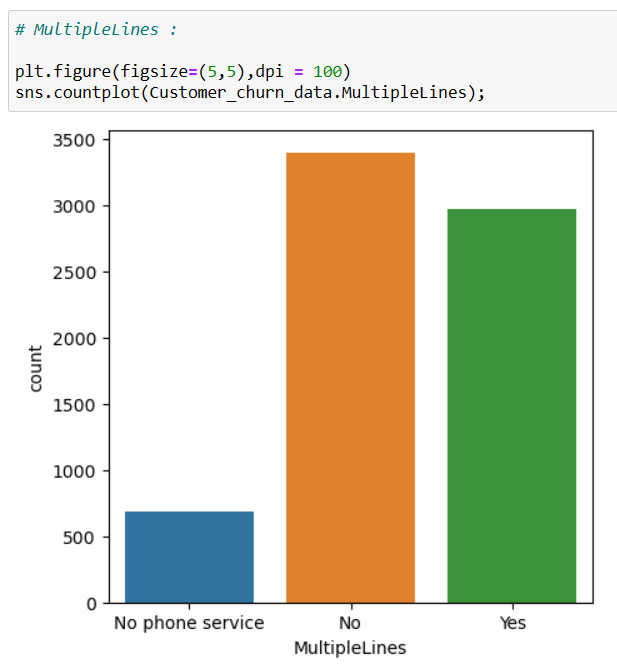
**Dependents**

****

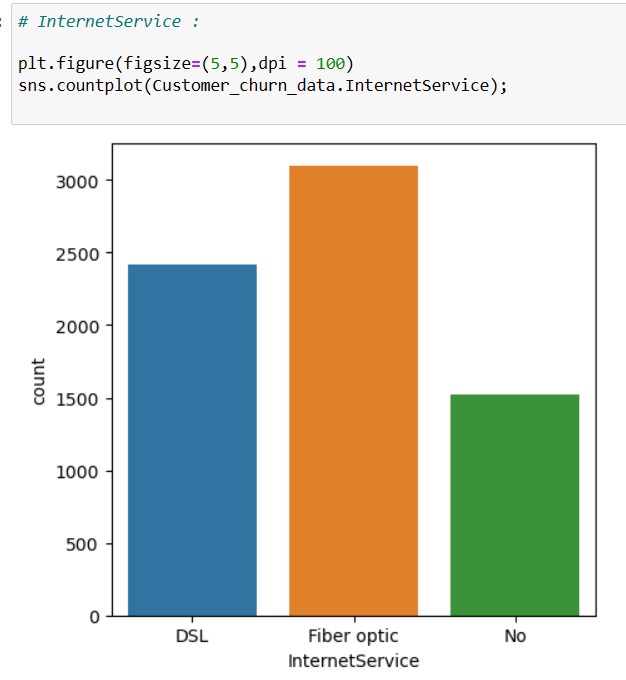
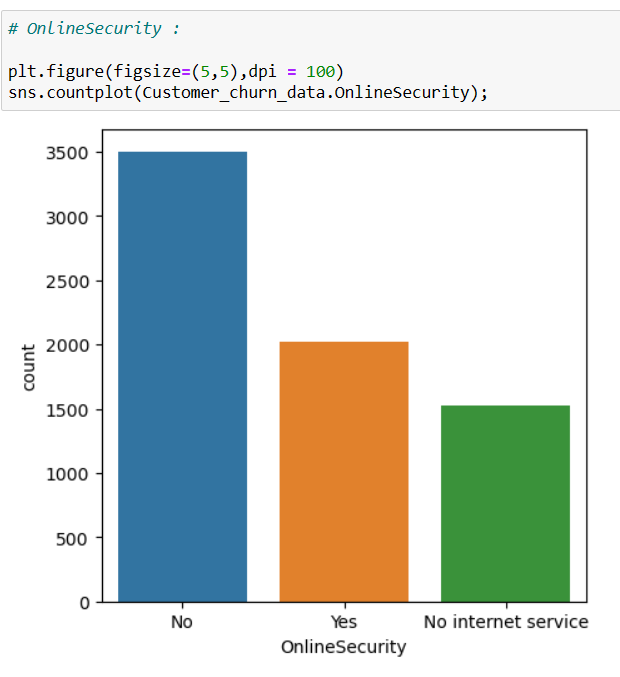
**Tenure**

****

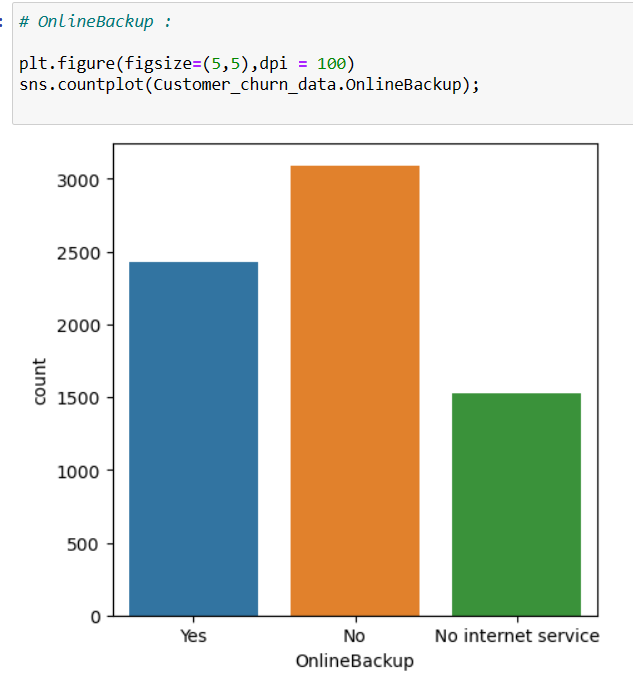
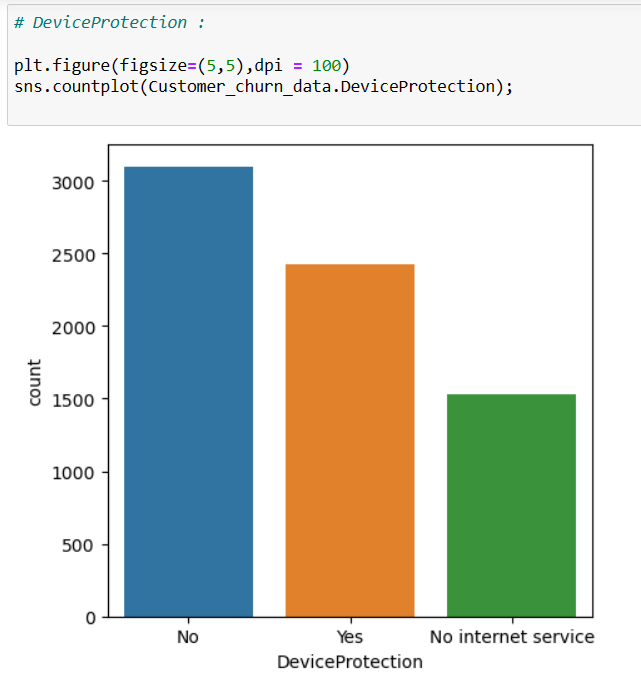
**Phone service Multiple lines**

** **

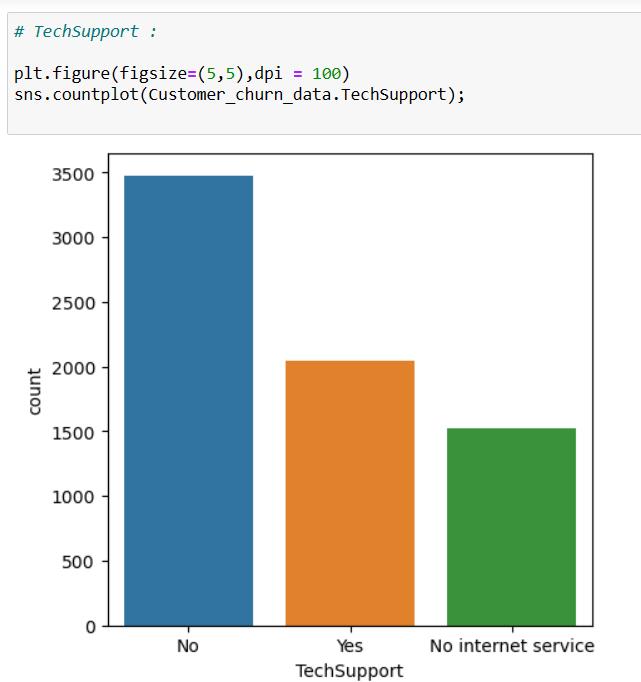
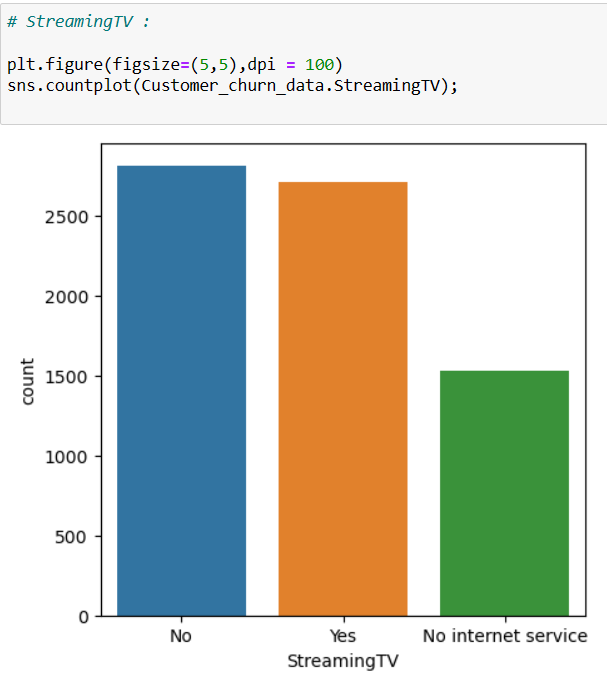
**InternetService OnlineSecurity**

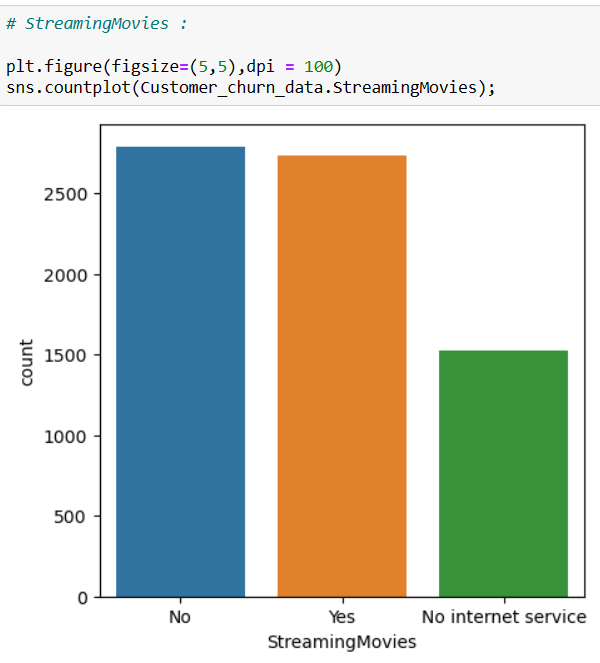
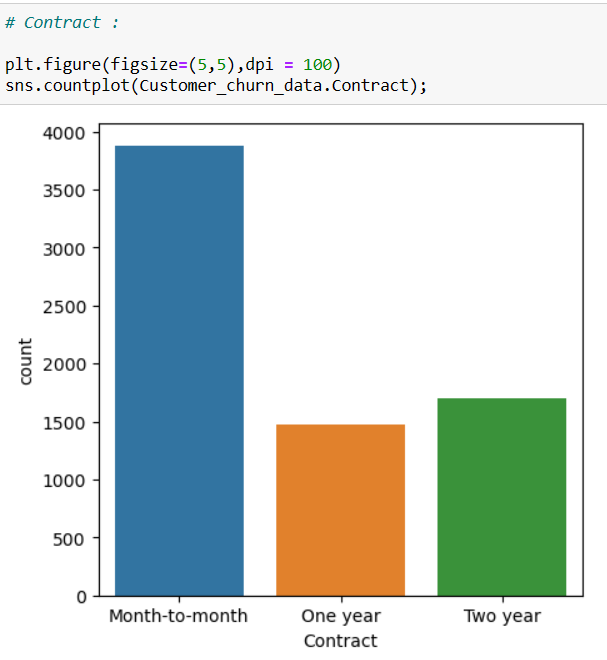
**OnlineBackup** **DeviceProtection**

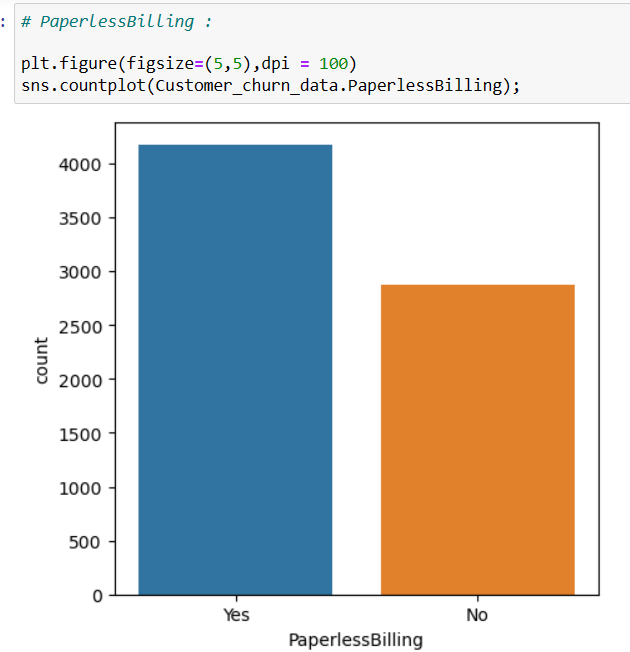
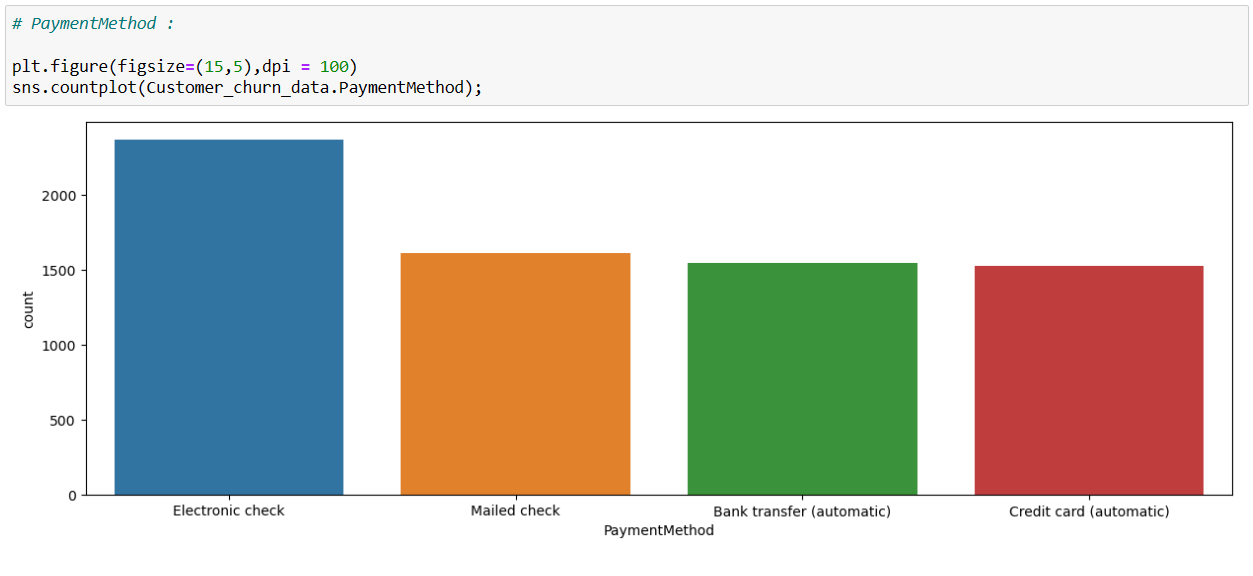
**Techsupport StreamingTv**

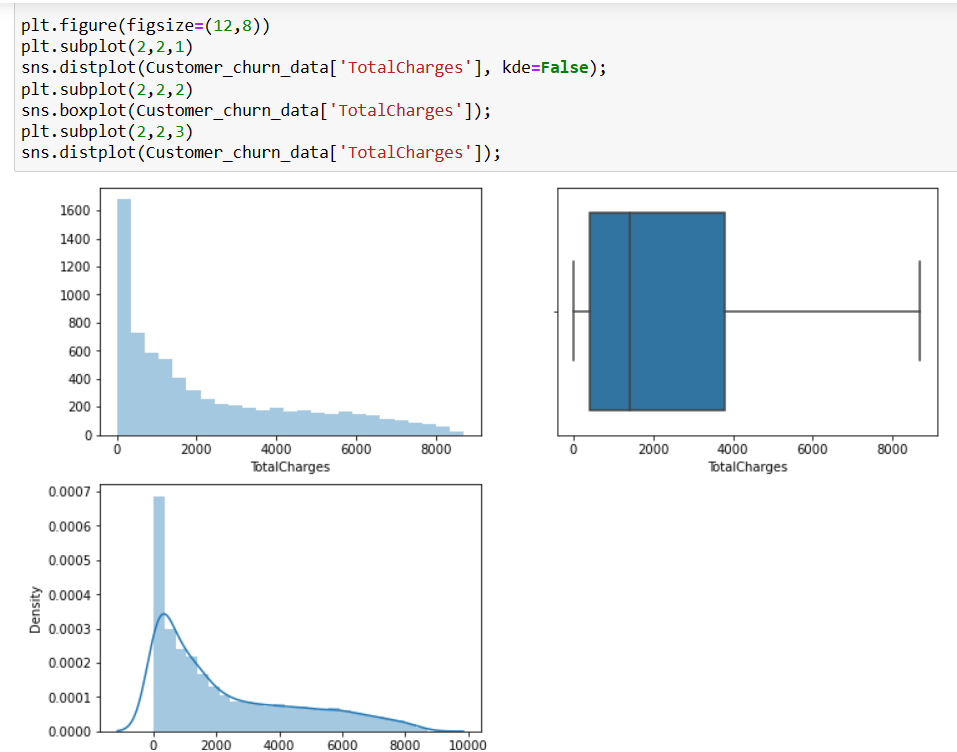
**Streaming movies** **Contract**

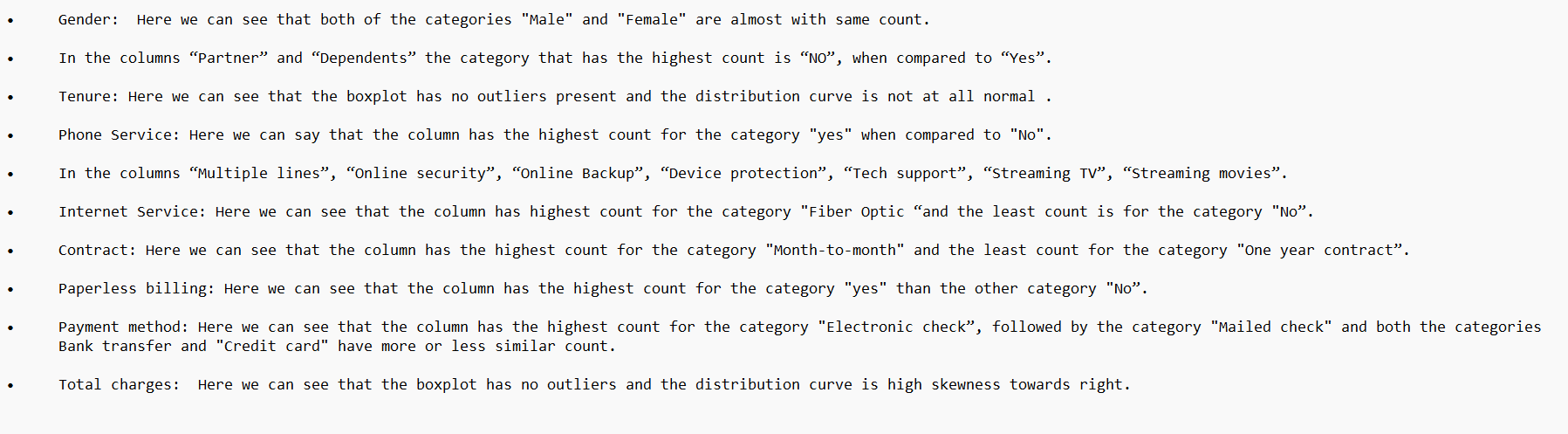
**Paperlessbilling**  **Payment method**

**Totalcharges**



* **OBSERVATIONS:**



* **Bivariate analysis: In this analysis I have used crosstab and generated the churn%.**
* **Gender with Churn**



* **Senior citizen with churn**

****

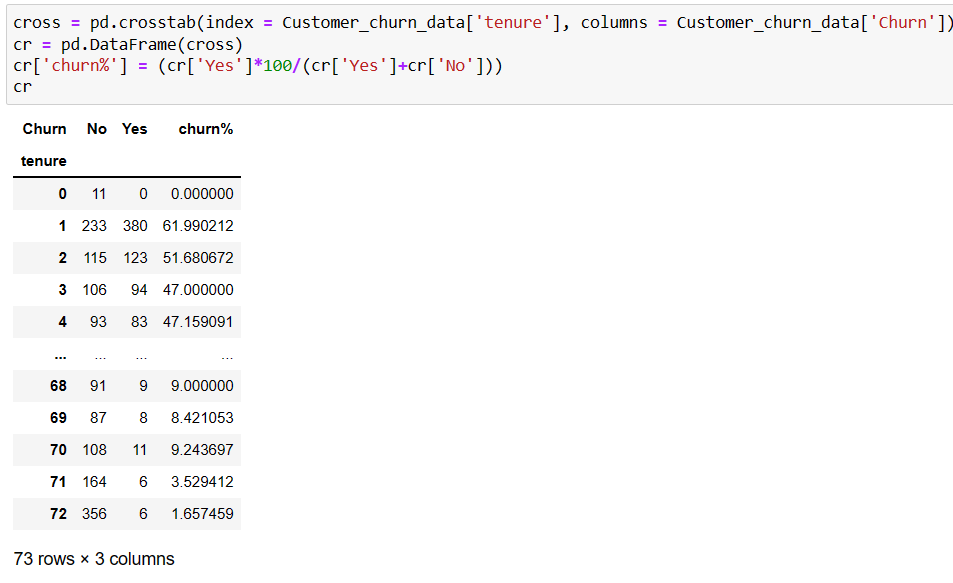
* **Partner with Churn**

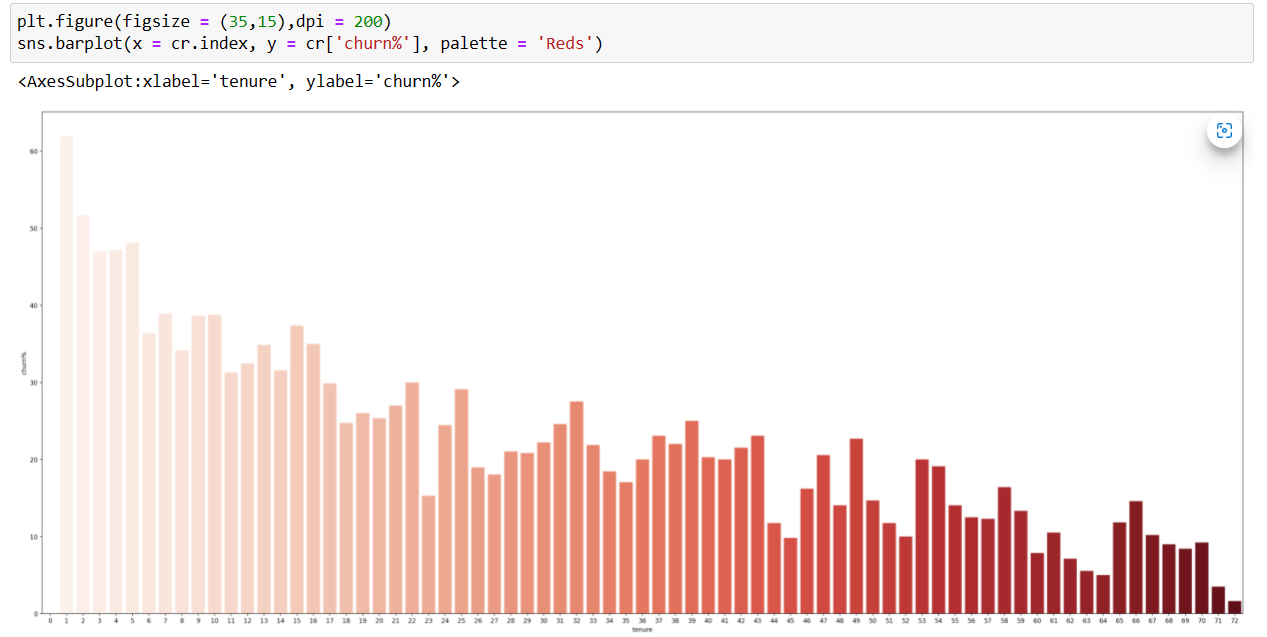
****

* **Dependents with churn**

****

* **Tenure with churn**

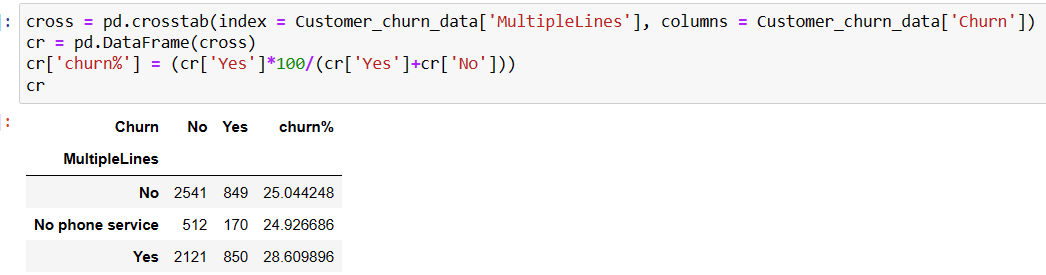
****

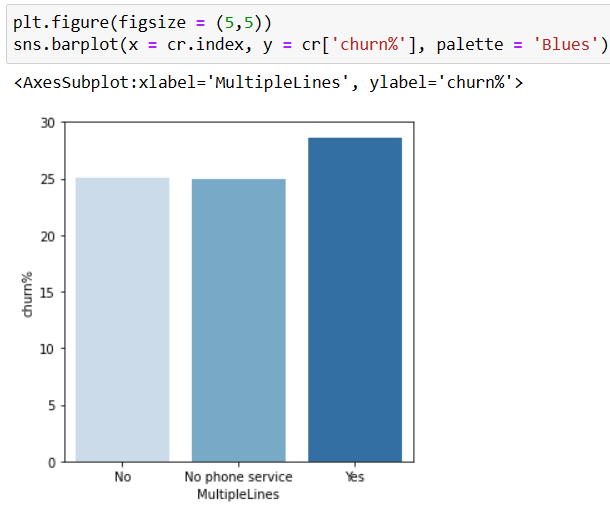
****

* **Phone service with churn**

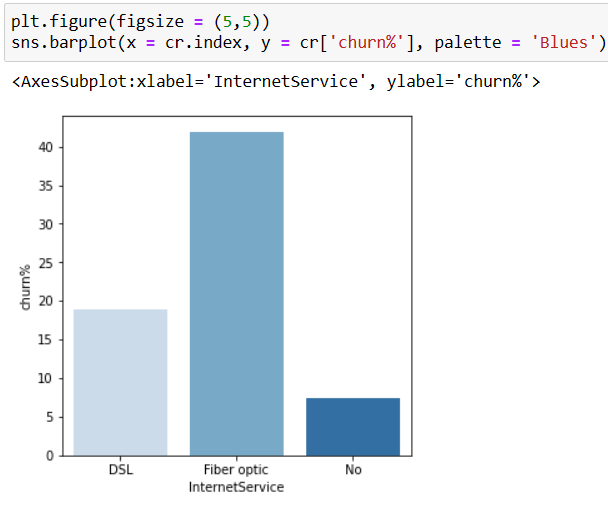
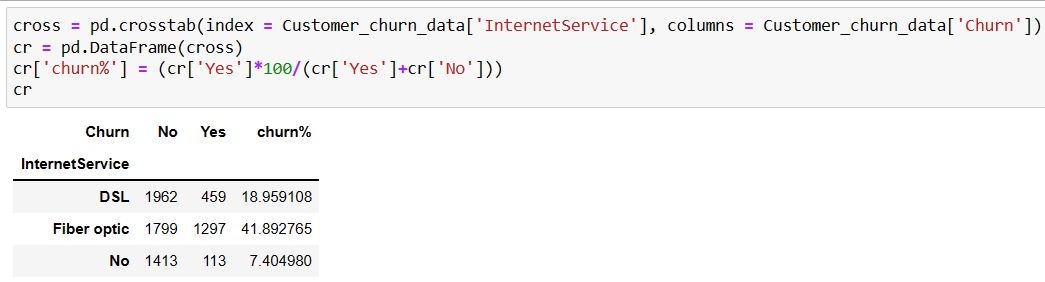
****

* **Multiple lines with churn**

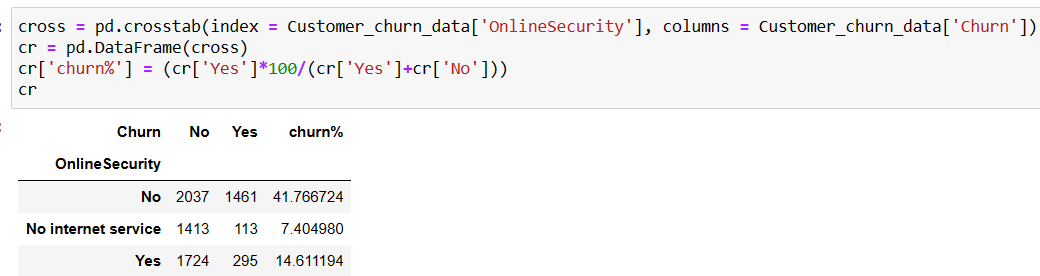
****

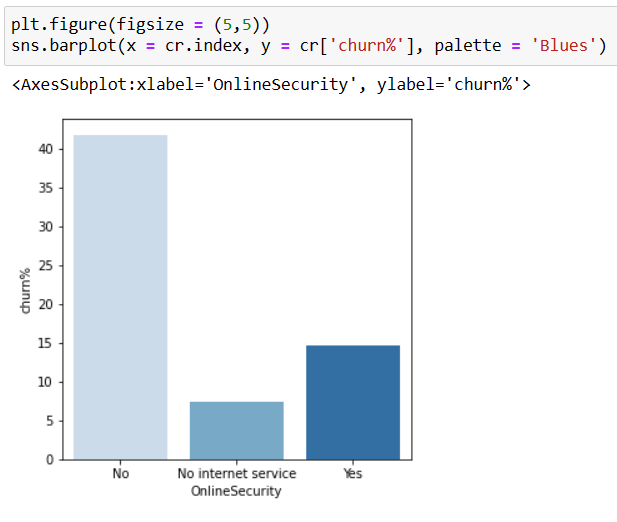
****

* **Internet service with churn**

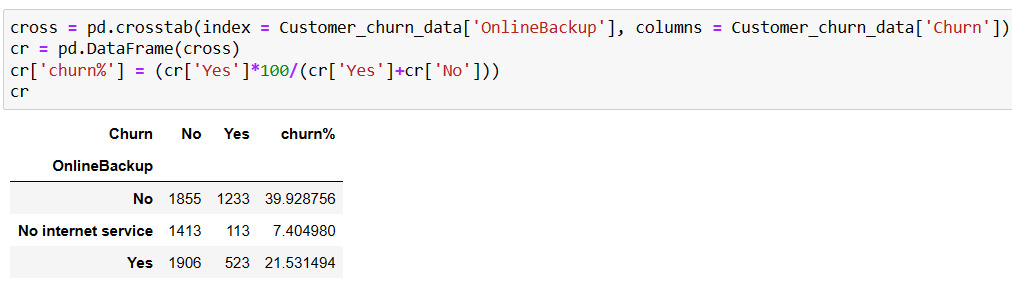
****

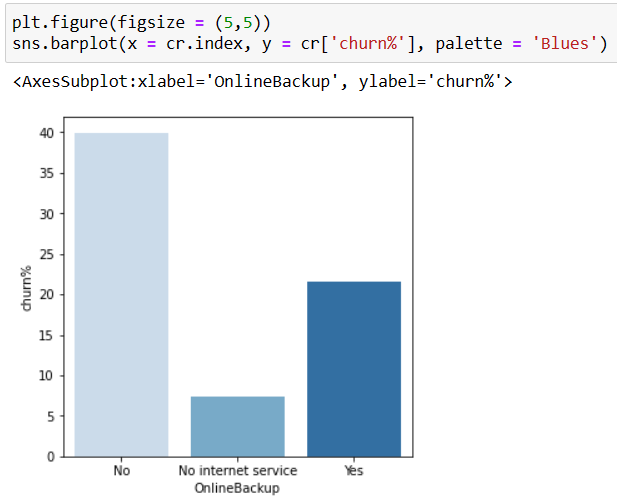
* **Online security with churn**

****

****

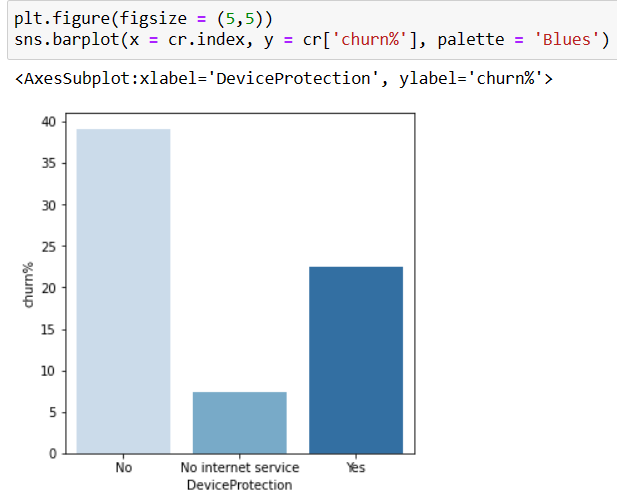
* **Online Backup with churn**

****

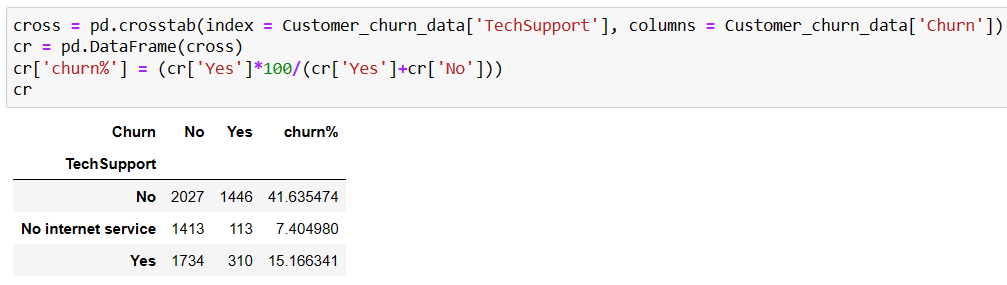
****

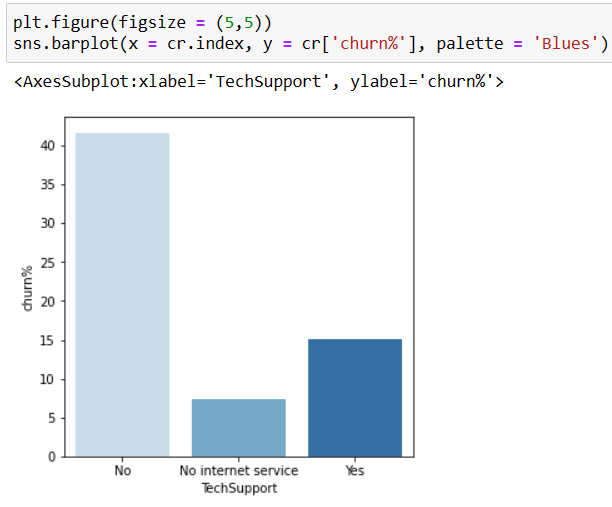
* **Device protection with churn**

****

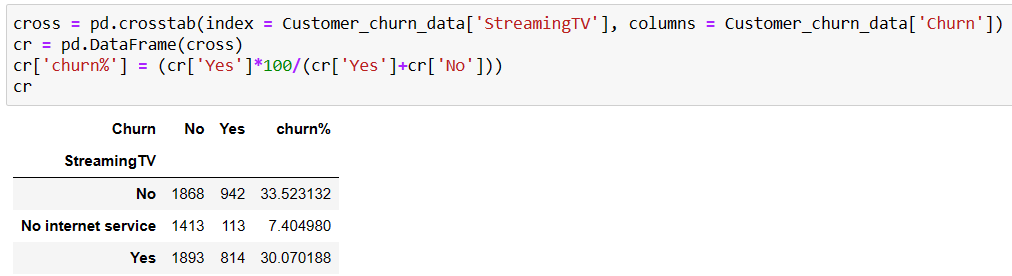
****

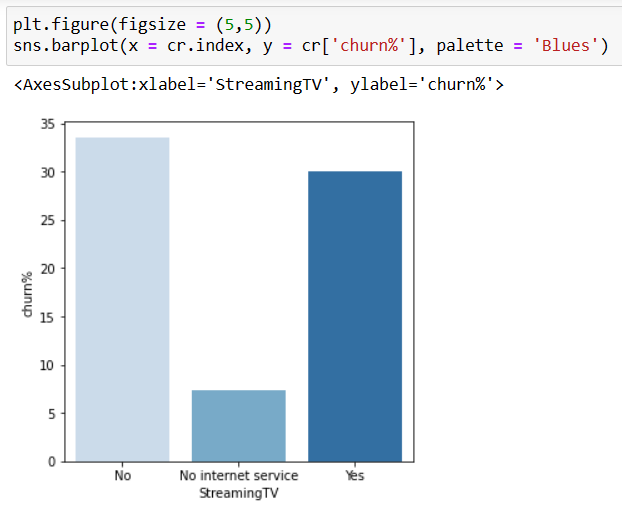
* **Techsupport with churn**

****

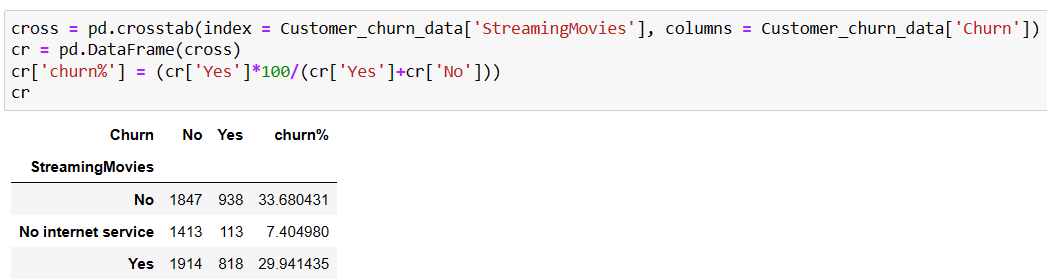
****

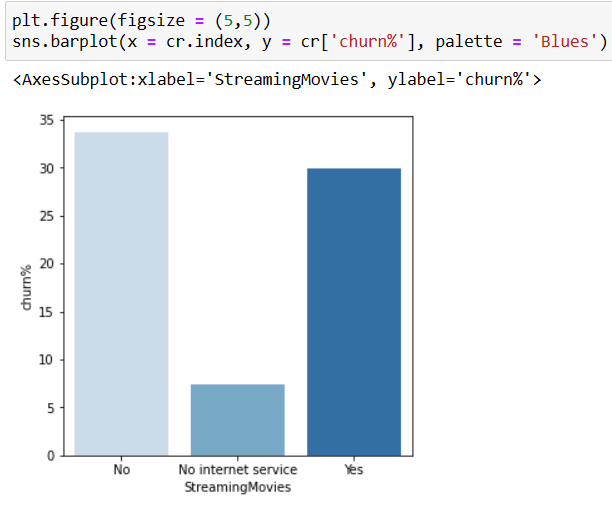
* **Streaming TV with churn**

****

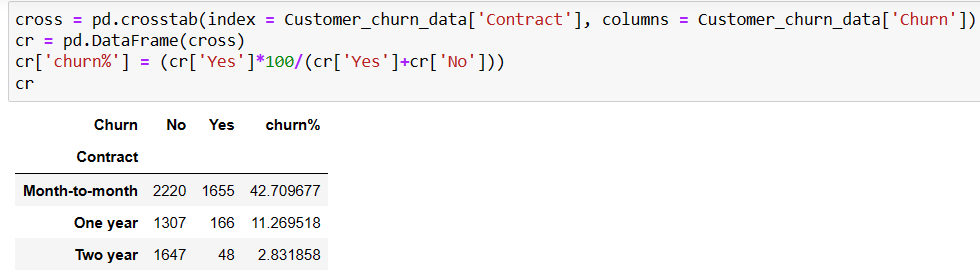
****

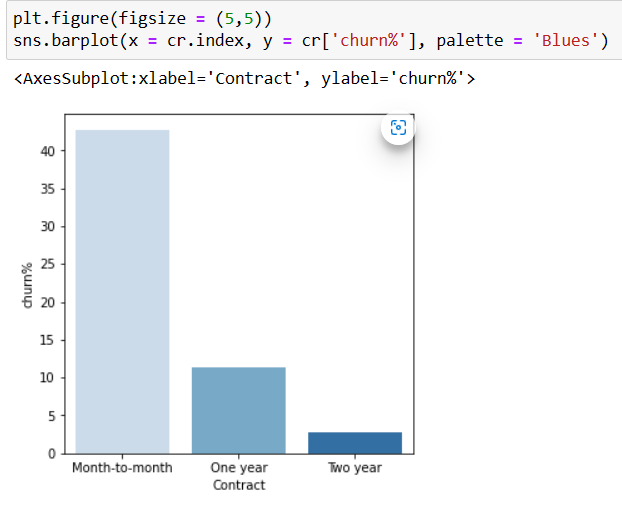
* **Streaming movies with churn**

****

****

* **Contract with churn**

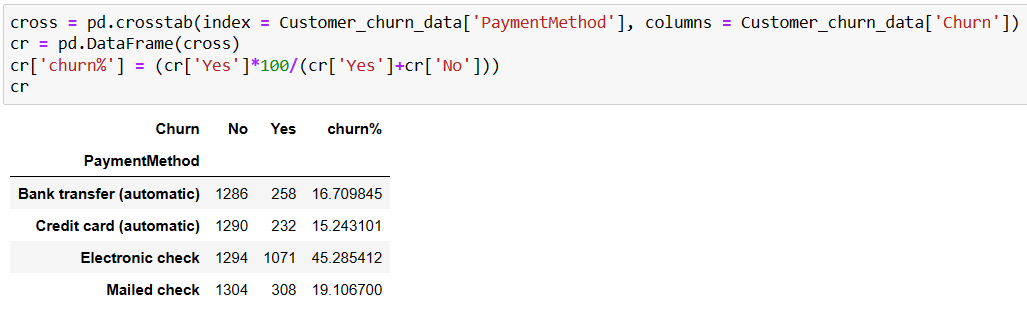
****

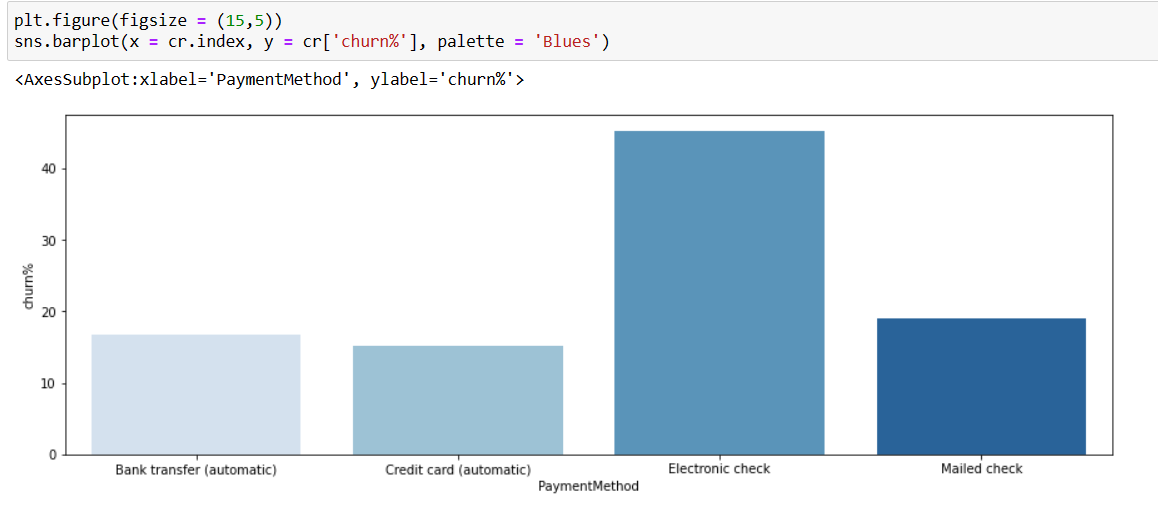
****

* **Paperless billing with churn**

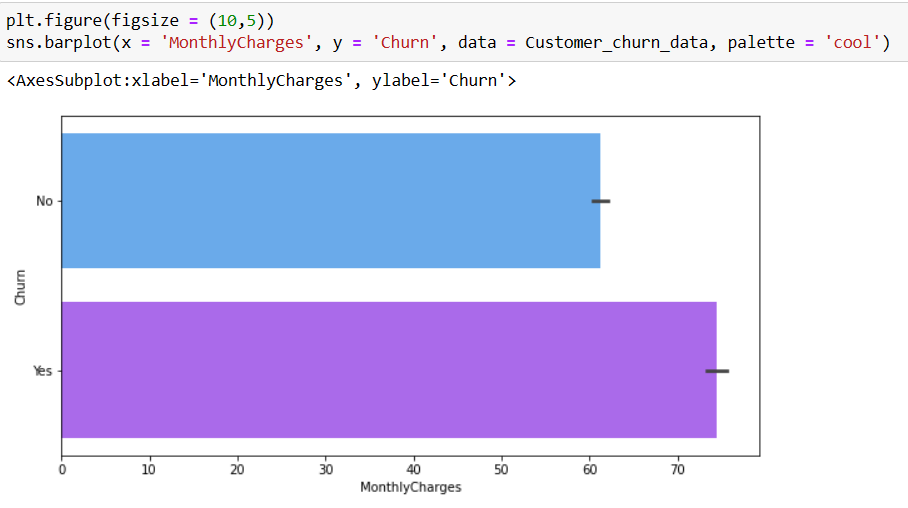
****

* **Payment method with churn**

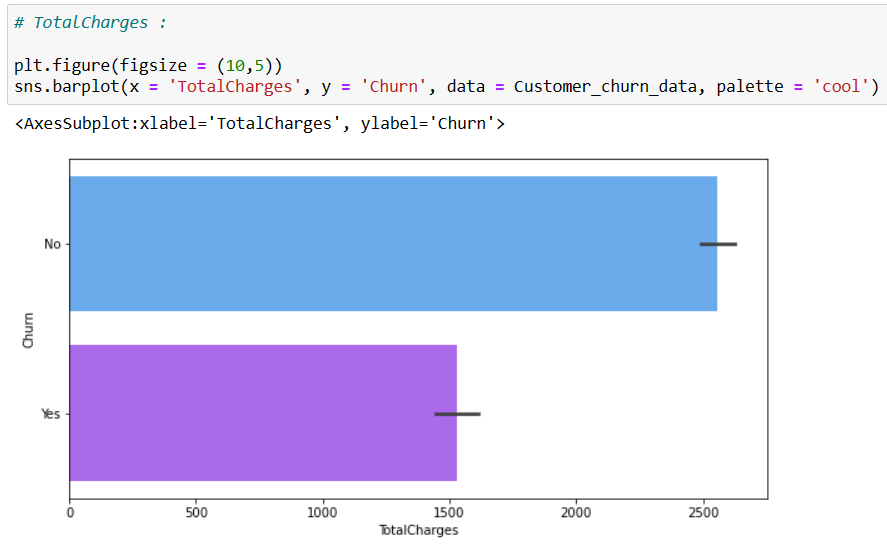
****

****

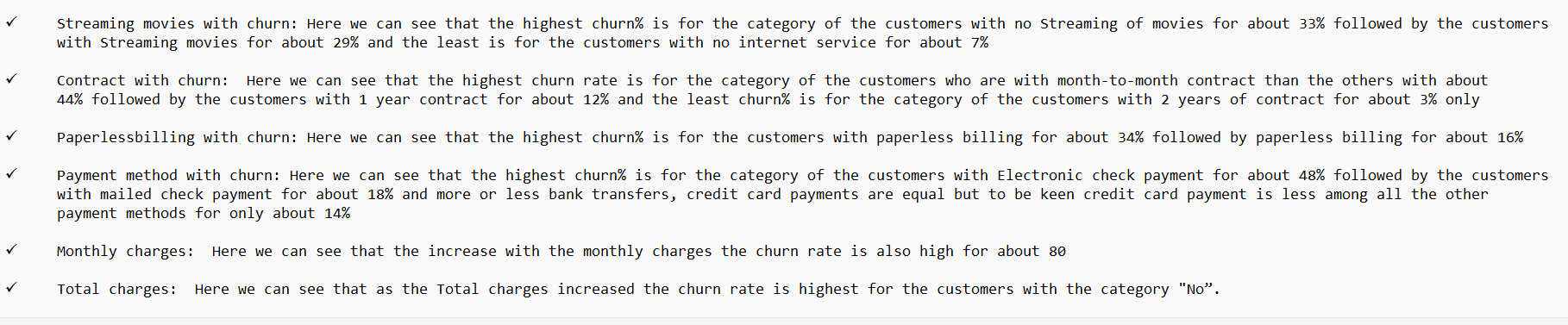
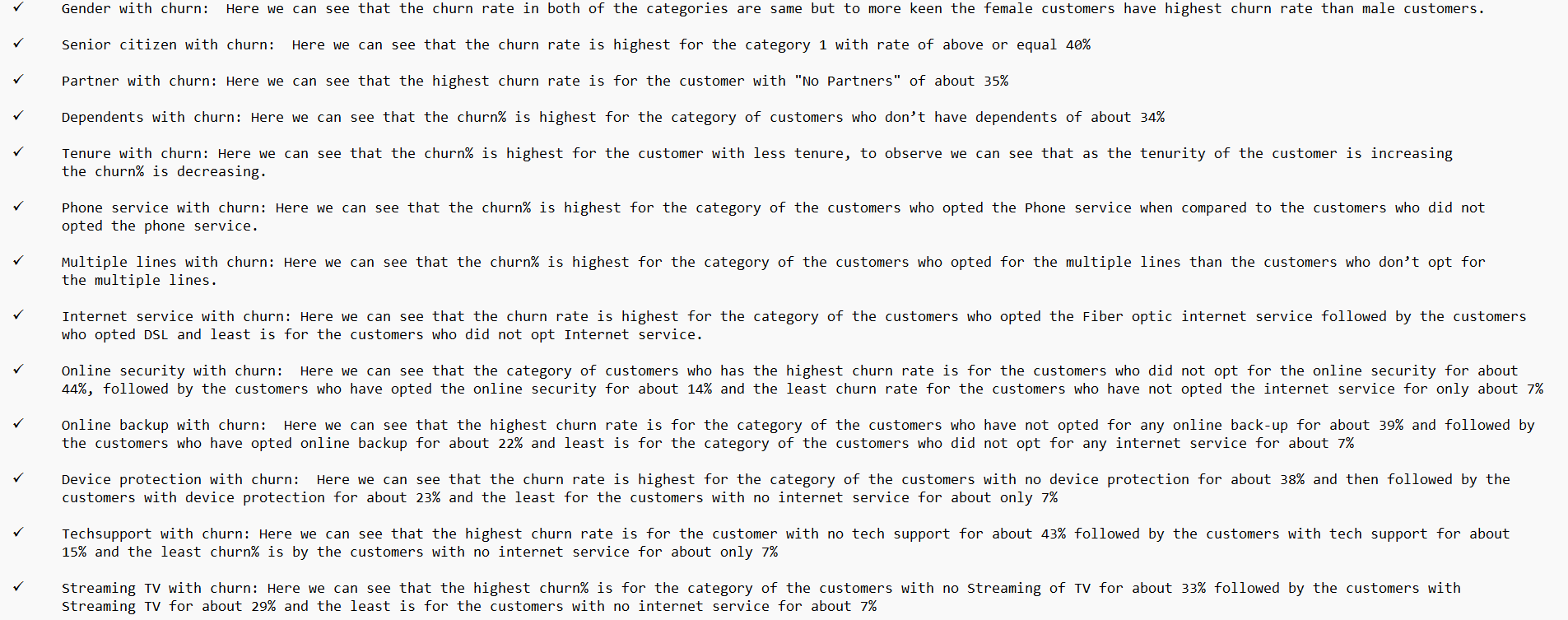
* **Monthly charges with churn**

****

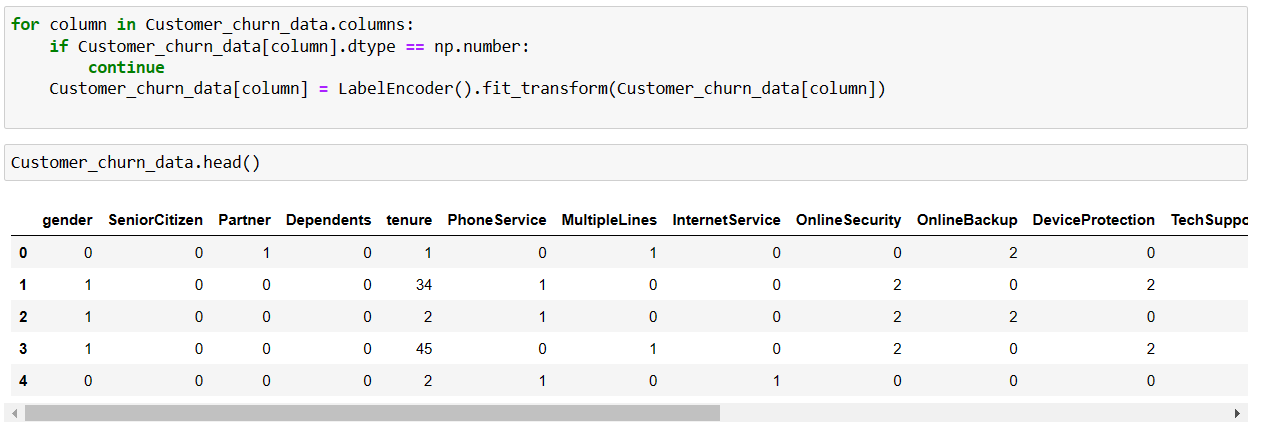
* **Total charges with churn**

****

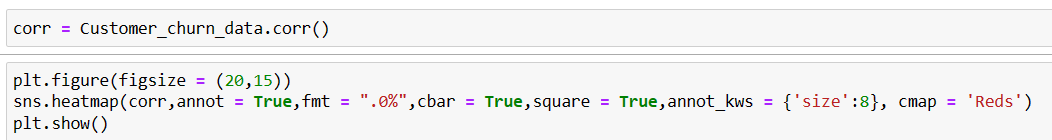
* **OBSERVATIONS:**

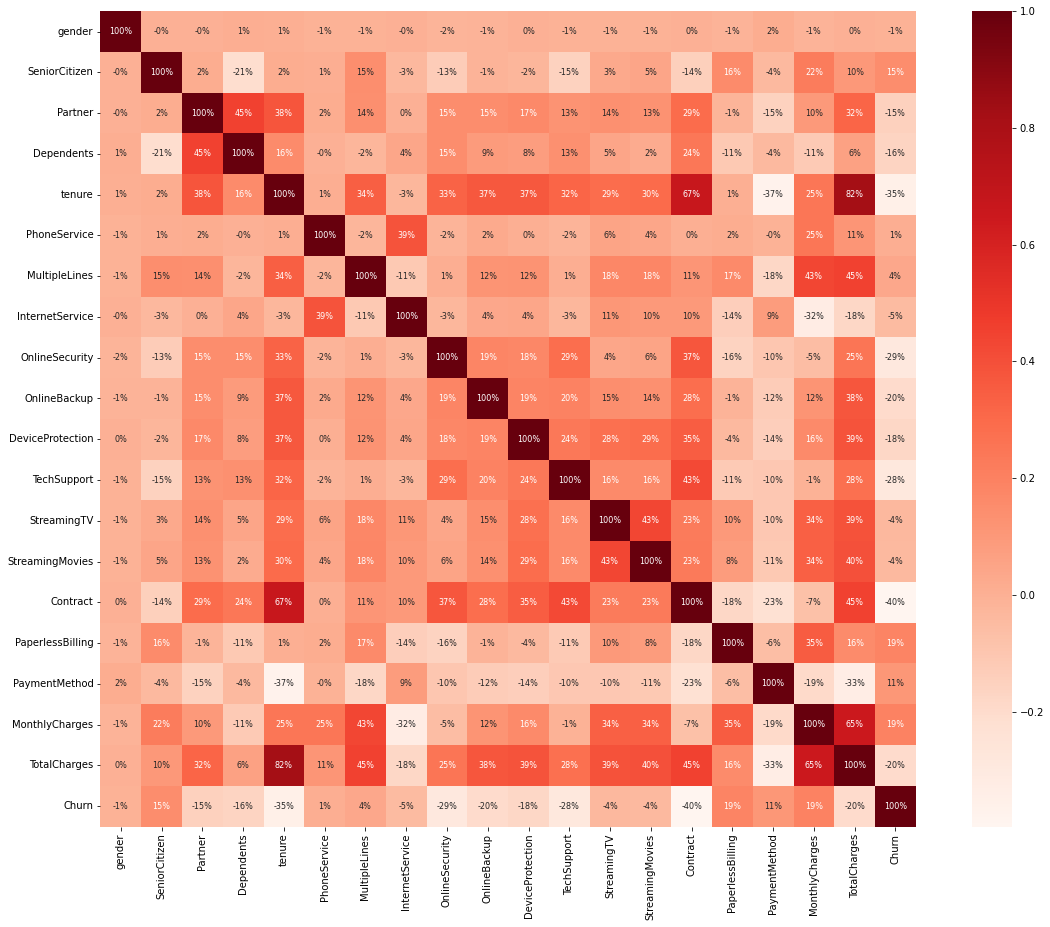


* **Before moving to correlation as our dataset contains the object values, we have to convert all those categorical values into numerical values through “Label Encoder”.**



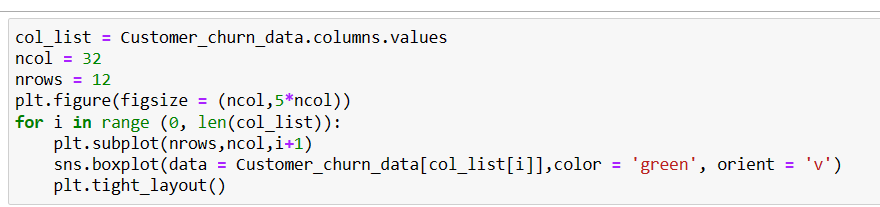
* Here we have converted all the categorical variables into numerical and we can proceed with our heatmap plotting for correlation.
* **CORRELATION:**
* Here we will be going to check the relationship between the variables i.e.., how all these attributes are related with each other which means how much strong is their bond with eachother and also here we will be going to check how all these attributes correlate with our label column.
* Also, with the help of this correlation we can also know the multicollinearity of the attributes if present which helps us to check through heatmap plotting.

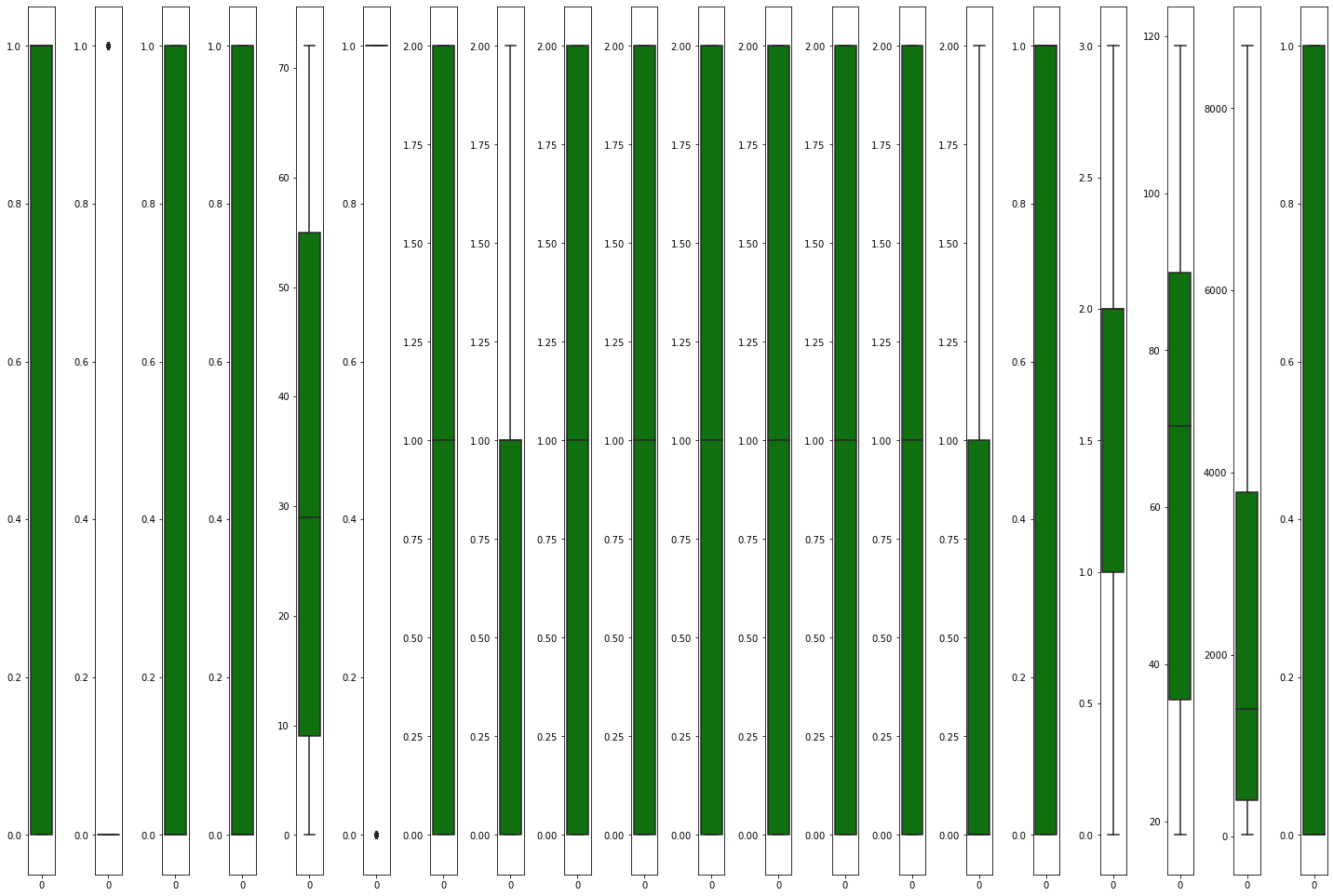




* **OBSERVATIONS:**
* Here we can note that heatmap shows us the positive and negative relationships with eachother and also with the label.
* Here we have few features which are in positive relationship with eachother and with the target variable but there are also few features which with negative relation with eachother and also with the target variable.
* Here we can observe that the column “tenure” has the highest correlation with the column “total charges”.
* The columns like “Senior citizen”, “Monthly charges”, “paperless billing” and “payment method” have the positive correlation with our label column “Churn”.
* The columns like “contract”, “tenure”, “onlinesecurity”, “techsupport”, “totalcharges”, “device protection”, “online backup”, “partner” and “dependents” have negative correlation with our label column “churn”.
* The feature having very least positive correlation with the label is “phoneservice” and also the feature with least negative correlation with the label column is “gender”.
* Also, we can note that no columns are multicollinear to eachother i.e.., multicollinearity is not present in the data.
* **ESTIMATORY DATA ANALYSIS CONCLUSION:**
* Through the analysis above we have analysed that there were few columns which we have to convert their datatype for our better model building.
* Majority of the customers are “Male” and most of the customers have opted “Fiber optic internet service” but also there are customers with “no internet service”.
* Many of the customers are with month-to-month contract and also prefer paperlessbilling with electronic check as their payment method.
* **CHECKING FOR THE OUTLIERS:**

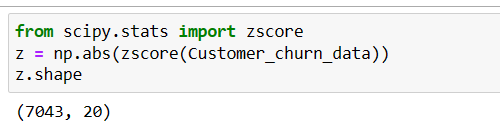
Here we will know the columns with outliers.

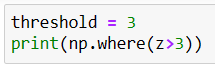


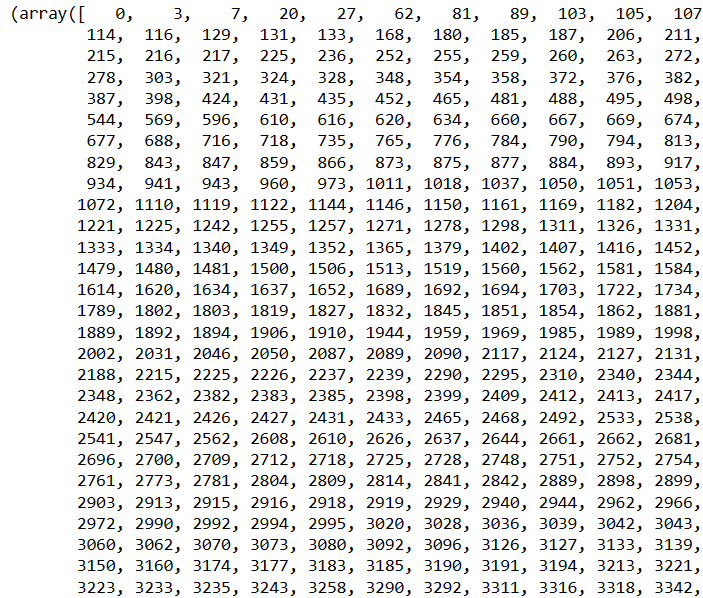


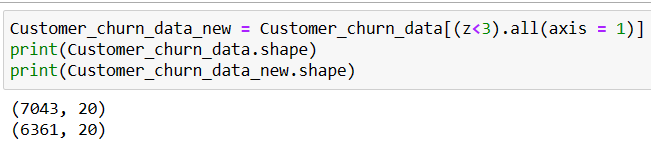
* In the above there are outliers seen in the column “Seniorcitizen” and probably there may be outliers further so its better to treat them.
* **REMOVING THE OUTLIERS:**

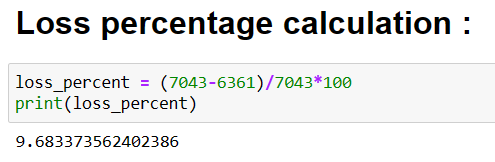
Here for removing the outliers we use “**Z-score method**”.



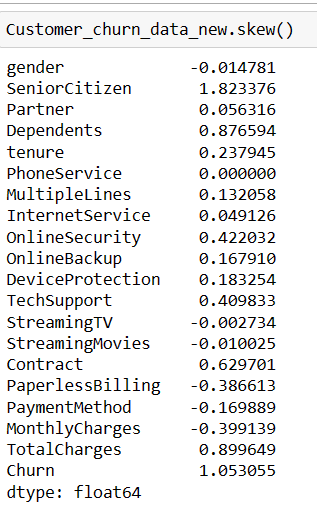


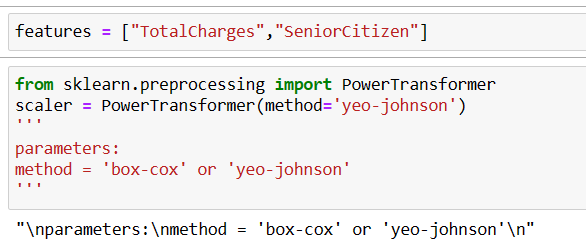


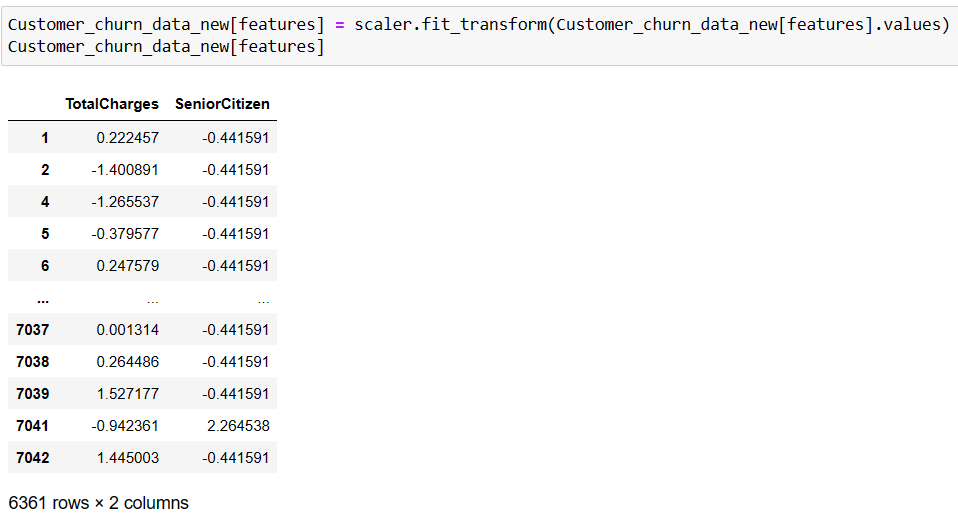


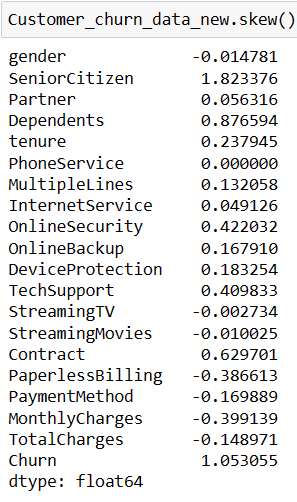


* Here through Z-Score method we have tried to treat the outliers present in the dataset and the threshold number here used is 3 and also, we have assigned a new variable to the data after using threshold along with which we have printed the shapes of both the datasets, the previous dataset and the new one, out of which we can see that the new dataset is with less number of records than the previous dataset and the loss percentage which is calculated is 9.68%.
* **CHECKING THE SKEWNESS:**

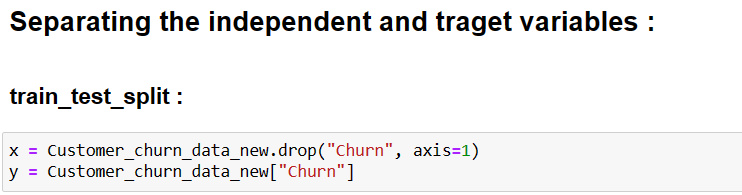
****

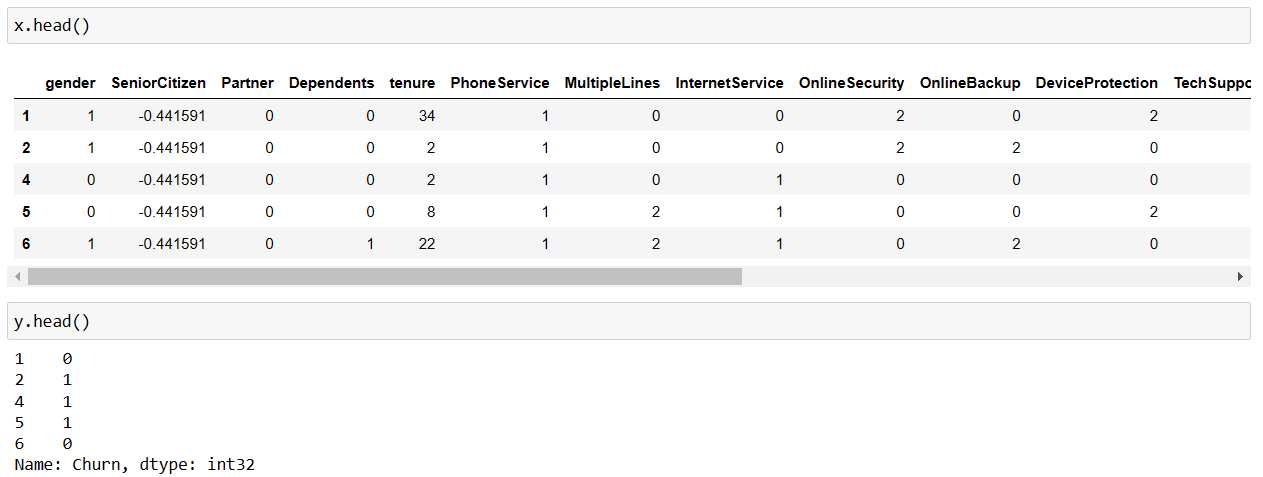
****

****

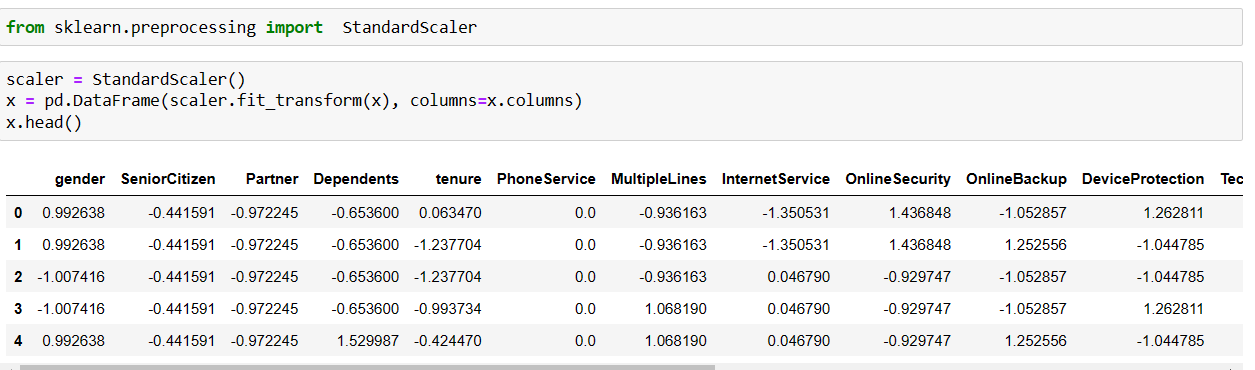
****

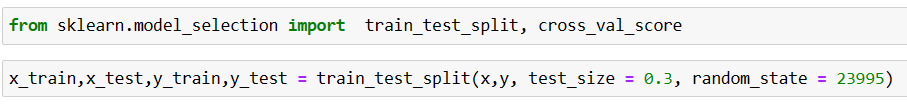
* Here we can see that our column “**Total charges**” reduced its skewness.
* **DATA PRE-PROCESSING:**

****

****

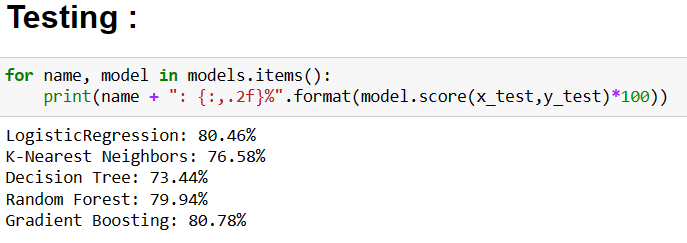
* Here we have divided the dataset into two variables: x and y, in which “x” is variable to which all the features are assigned except the label column “Churn” which is assigned to the variable “y”.
* **Scaling the x\_data using the “Standard Scalar”:**

****

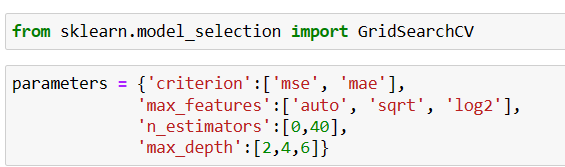
****

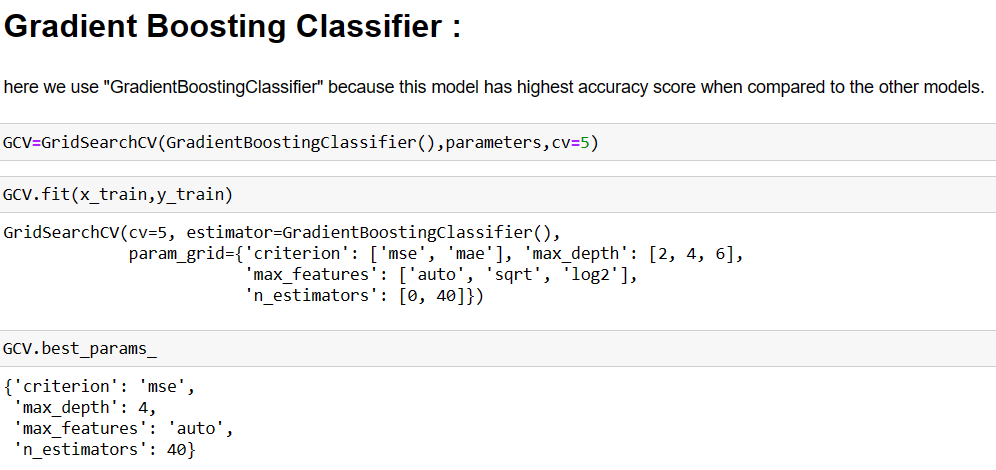
* Here we have scaled the x\_data ie., feature data through “StandardScaler” and then we have imported train\_test\_split and divided the dataset into x\_train, y\_train, x\_test, y\_test with test size = 0.3 which means 70% of the train data is taken and 30% of the test data is considered with random state (it’s a random guess of any number).
* **Building the Machine Learning Models:**



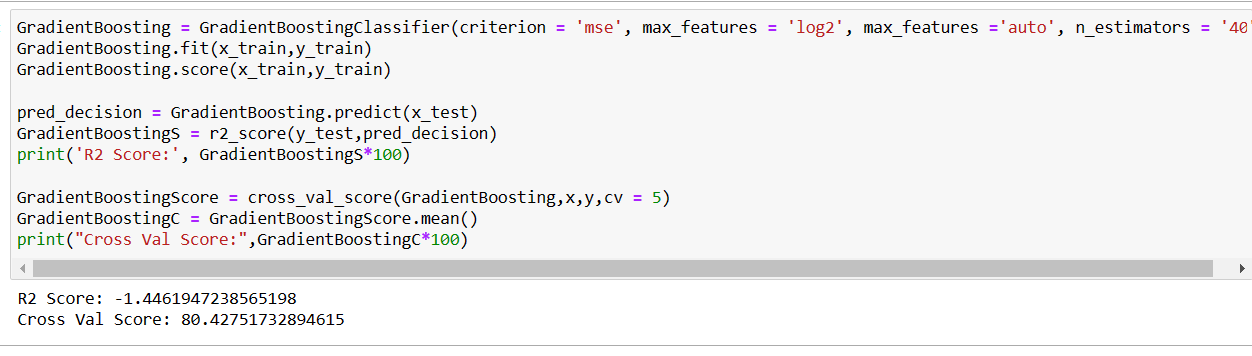


* Here we have imported the necessary libraries required for model building and I have used forloop for training and testing the models.
* In testing all the models “Gradient Boosting” model has the highest test score compared to all the other models.
* **HYPER PARAMETER TUNING:**

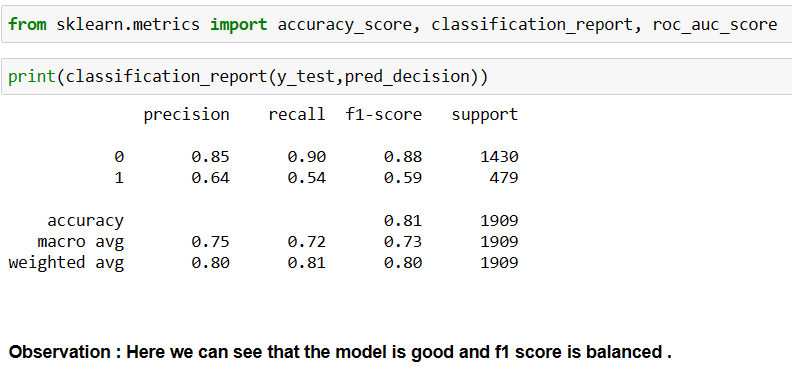


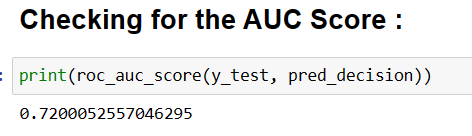


* + Here I have selected the model “Gradient Boosting model” which has the highest accuracy percent among all the other models for hyper parameter tuning.
* For this I have selected few parameters for tuning our model which are: Criterion, max\_depth, max\_features, n\_estimators and got the best parameters for tuning.

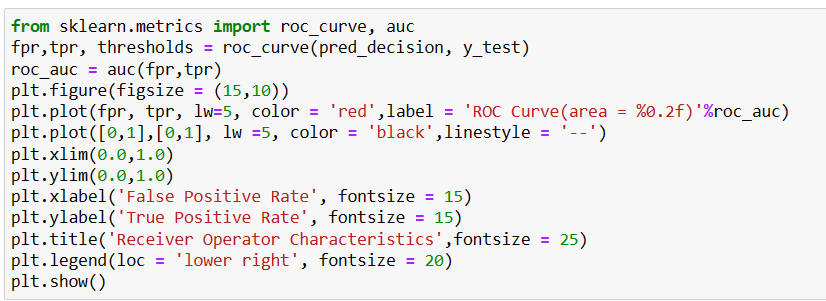
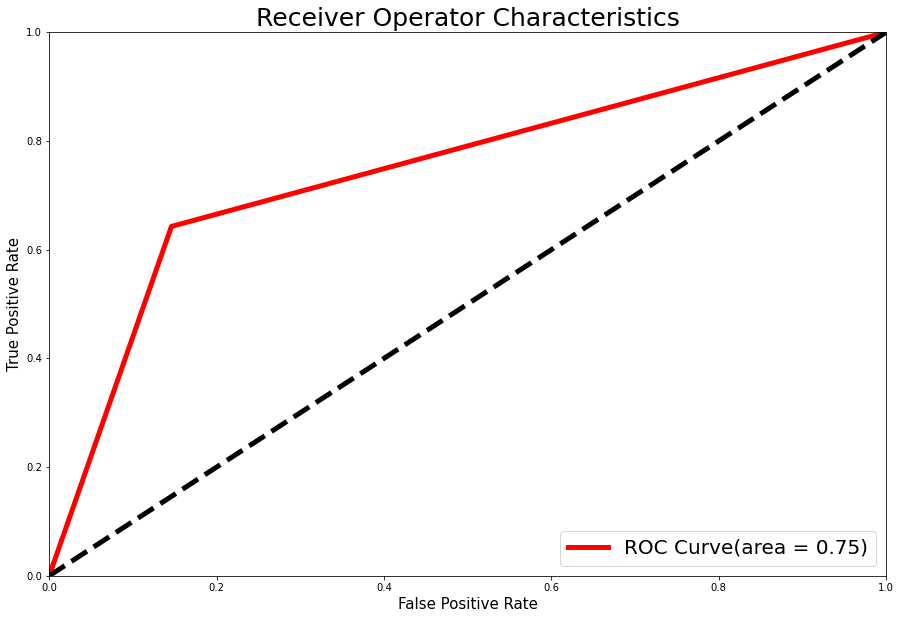


* Here I can see that the cross-validation score is 80.42%.
* After tuning the model accuracy decreased to certain extent.
* **CLASSIFICATION REPORT:**

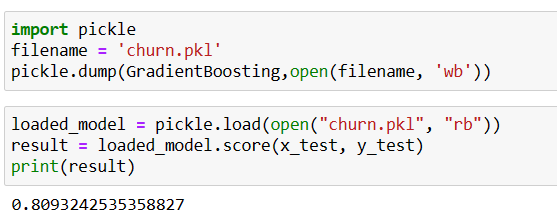
****



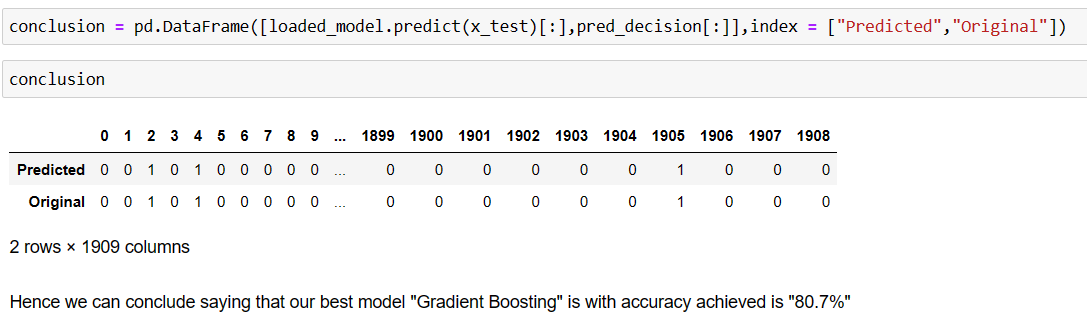
* Here I have checked the **auc\_score** which is **72%**.
* **PLOTTING THE ROC CURVE:**

 ****

* Here I have plotted the ROC Curve for our best model is 75%
* **SAVING THE MODEL:**



* Here I have saved the model and also checked the test score which is 80.9% ie., almost 81%.
* **PREDICTING THE CHURNED VALUES OF THE CUSTOMERS:**

****

* **CONCLUDING REMARKS:**
* This particular problem needs a good vision on data, and in this problem “**Feature Engineering**” is the most crucial thing.
* You can see how we handled numerical and categorical data and also how we build different machine learning models on the same dataset.
* I Think this process gives you easy to understand the “CustomerChurn” for going further step.

**THANKYOU**

HAINDAVI CHAKRAVARTHY

(BATCH NO. 1836)