CUSTOMER CHURN ANALYSIS.

MADE BY: - SAMREEN KHAN

PROBLEM STATEMENTS: -

**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.**

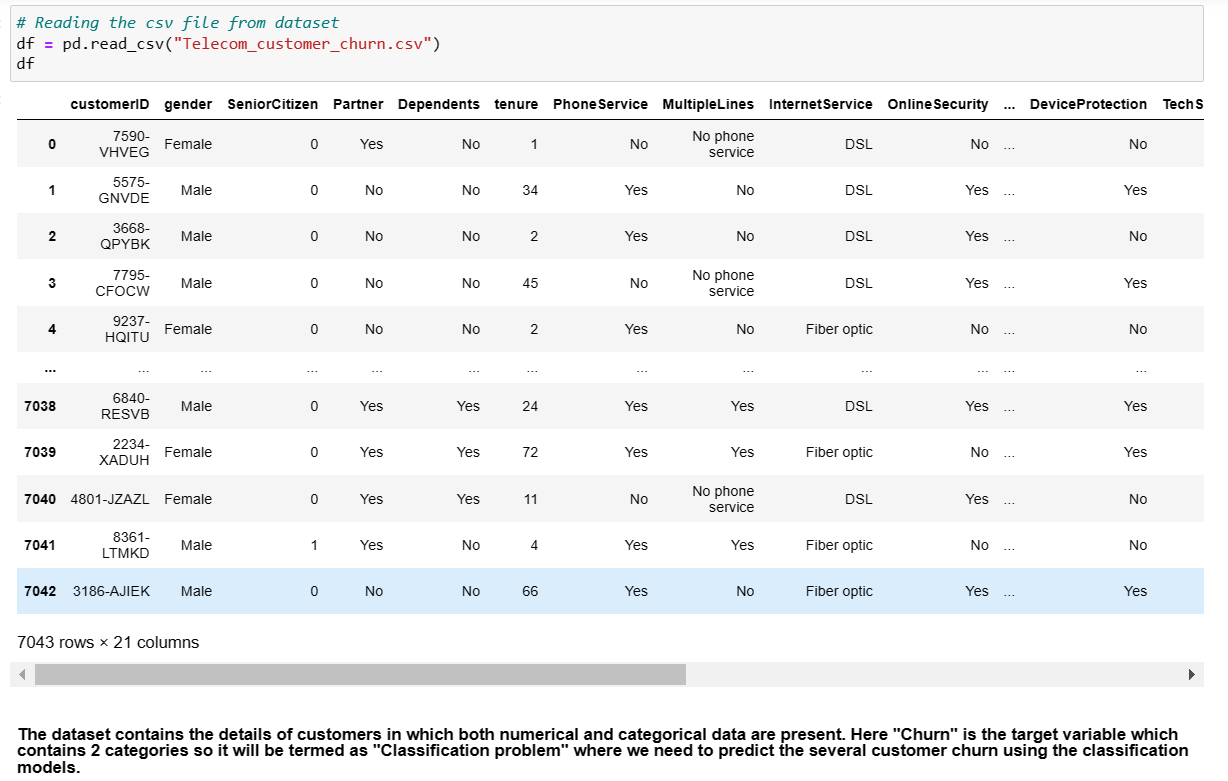
**Now let’s see what’s inside the project**

**LET’S IMPORT SOME IMPORTANT LIBRARIES.**

****

IMPORTED HERE SOME OF THE BASIC AND IMPORTANT LIBRARIES.

NOW LET’S IMPORT THE DATASET TO EXAMINE.

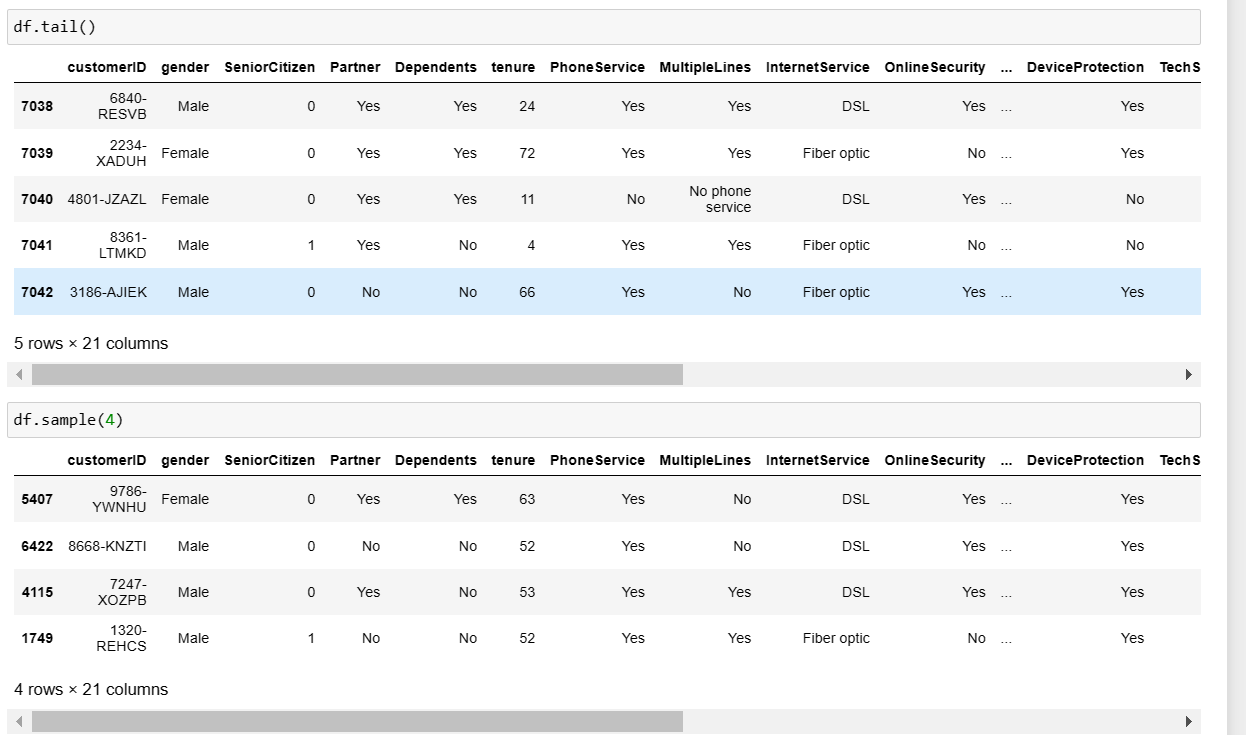


THE DATASET CONTAIN BOTH NUMERICAL AND CATAGORIAL COLUMNS.

HERE “CHURN” IS A TARGET VARIABLE WHICH CONTAINS 2 CATAGORIES SO IT WILL BE TERMED AS CLASSIFICATION PROBLEM WHERE WE NEED TO PREDICIT SERVAL CUSTOMER CHURN USING CLASSIFICATION MODELS.

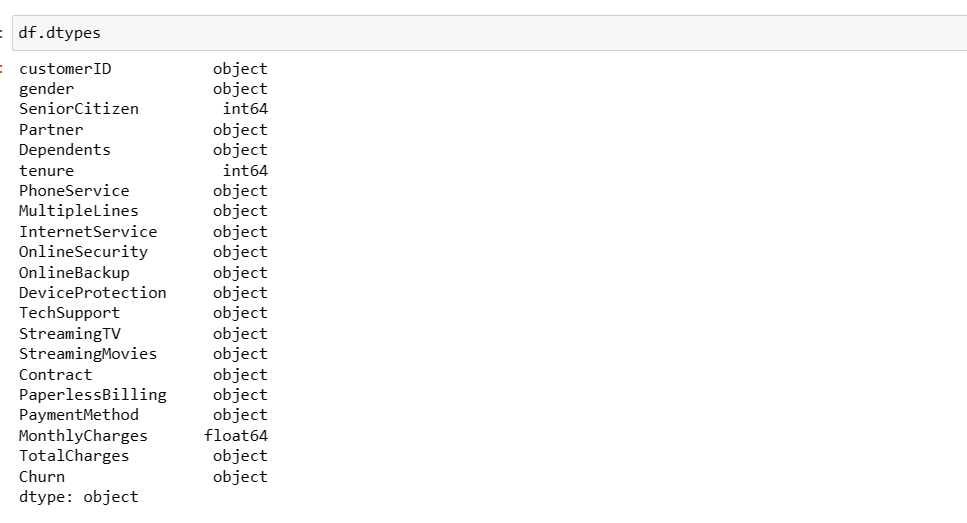
THE DATASET HAVE 7043 ROWS AND 21 COLUMNS IN THE DATASET.

NOW LET’S SHUFFLE THE DATASET IN ORDER TO SEE ANY KIND OF ABNORMAL DATA’S PRESENT.



AFTER SHUFFLING THROUGH THE DATASET, I WAS NOT ABLE TO FIND OUT ANY ABNORMAL VALUES PRESENT IN THE DATASET.

NOW LET’S SEE THE DATATYPE OF THE ATTRIBUTES PRESENT IN THE DATASET.

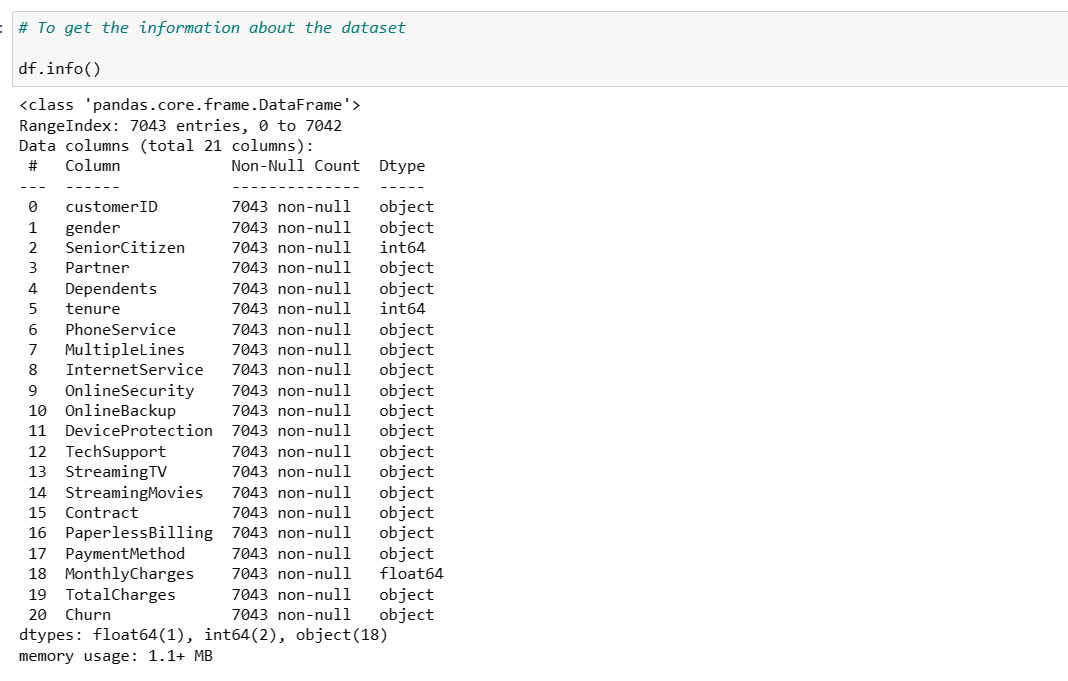


WE CAN SEE THAT BOTH CATAGORICAL AND NUMERICAL COLUMNS ARE PRESENT IN THE DATASET.

OUR TARGET VARIABLE IS IN ‘OBJECT’ FORMAT WHICH MAKES IT A CLASSIFICATION PROBLEM.

EDA

LET’S CHECK INFORMATION ABOUT THE DATASET.

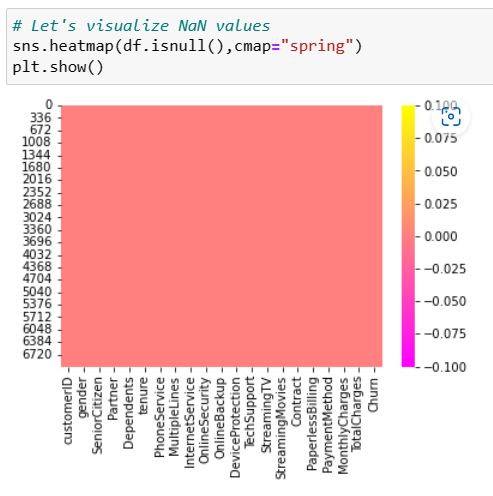


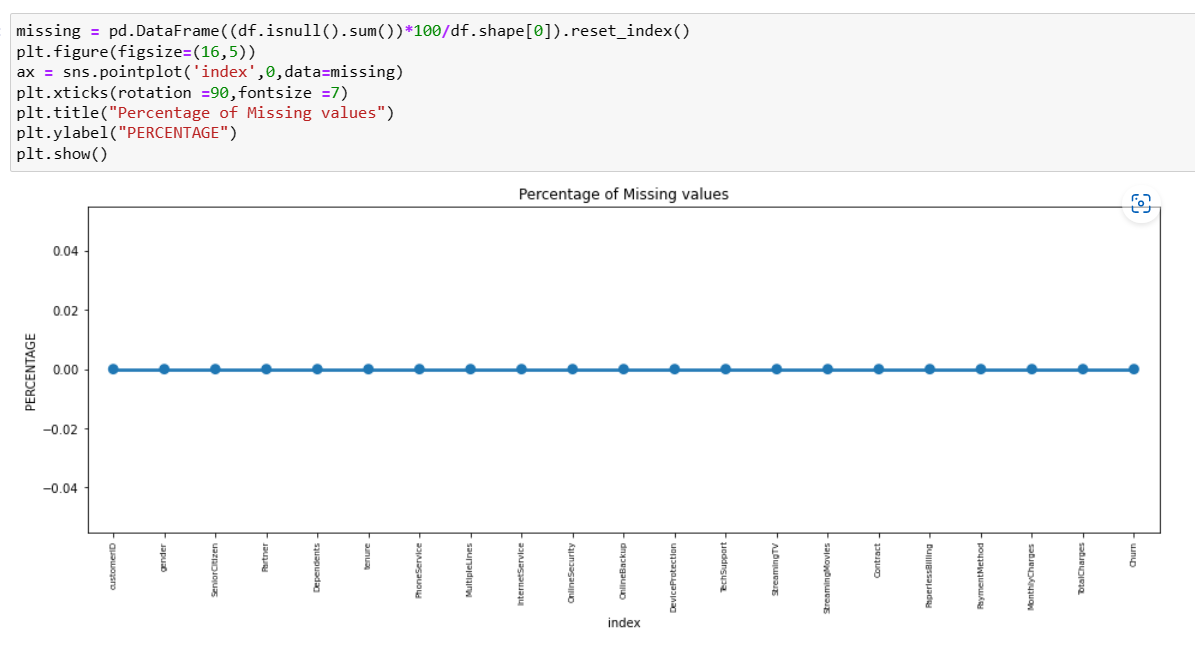
HERE WE CAN SEE COLUMNS NAME, NON-NULL COUNT AND DTYPE OF THE DATASET.

HERE THE NON-NULL COUNT OF ALL THE COLUMNS IS 7043 WHICH MEANS NO NULL VALUE IS PRESENT IN THE DATASET.

AND 3 TYPES OF DATA TYPES ARE PRESENT IN THE DATASET WHICH ARE FLAOT64, INT64 AND OBJECT.

AS WE KNOW NULL VALUES ARE NOT PRESENT IN THE DATASET SO LET’S CHECK IT AGAIN.



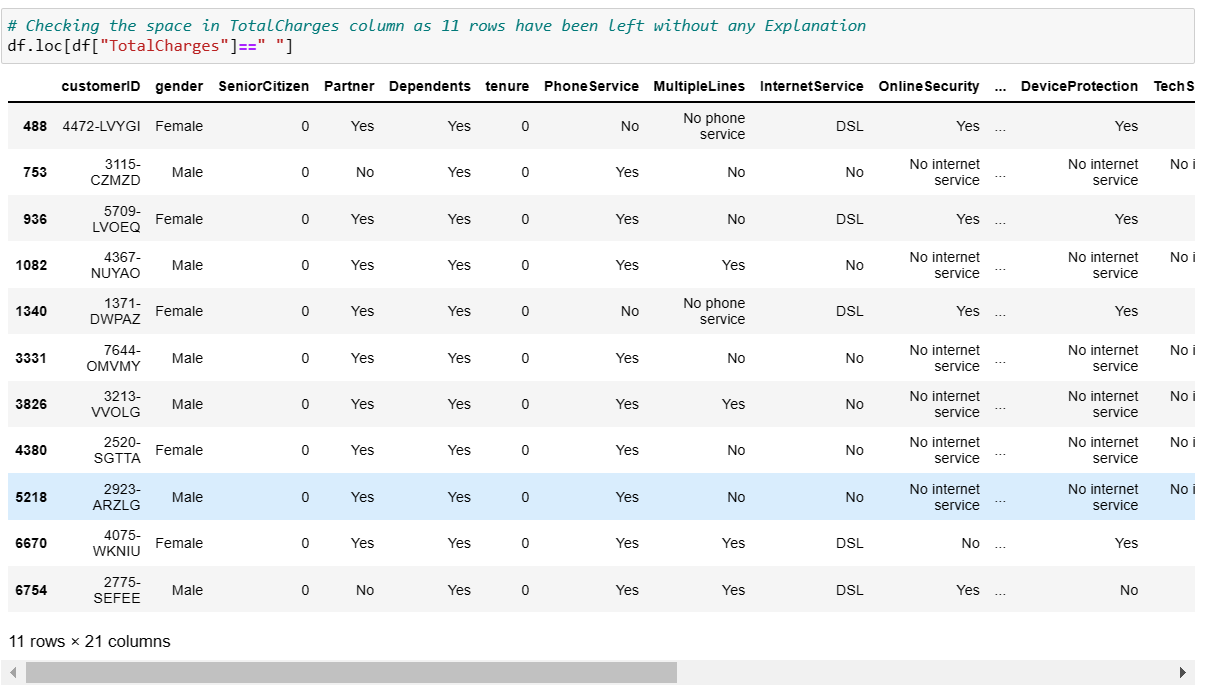


SO HERE WE CAN CLEARLY SEE FROM ABOVE ALL PLOTTING THAT NO NULL VALUES ARE PRESENT IN THE DATASET.

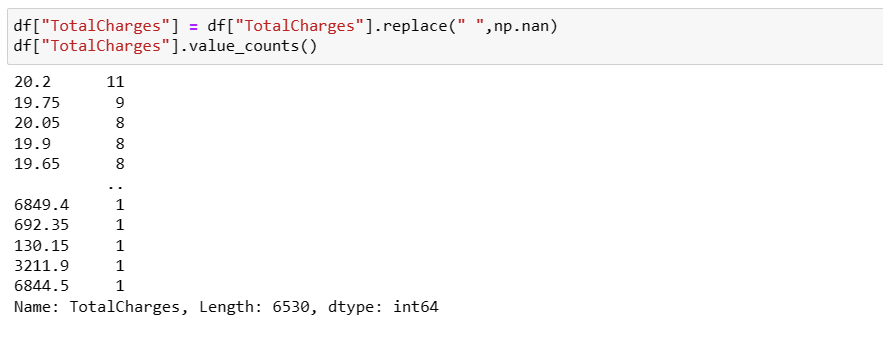
NOW LET’S CHECK ANY ABNORMAL VALUES PRESENT IN THE DATASET SUCH AS SPACES OR ANYOTHER SYMBOLS.



WHILE WE CHECKING UNIQUE VALUES OF ALL THE ATTRIBUTES IN THE DATASET, WE FOUND OUT THAT 11 ROWS IN TOTALCHARGES HAVE BEEN LEFT BLANK, SO WE WILL INITIALLY FILL IT UP AS ‘NAN’ AND AFTER THAT WE WILL FILL IT UP ACCORDING TO THE DATA TYPE.



THESE ARE THE 11 ROWS WHICH ARE LEFT BLANK.



#### Now we have replaced all the missing value with np.nan.

#### Now let’s fill the nan value but before filling the nan value lets convert the data type of TotalCharge which is in object to float time as it is format of continuous data.

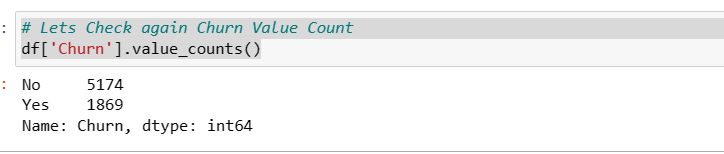


NOW LET’S FILL NAN VALUES WITH MEDIAN.



CLEARLY, WE CAN SEE NO NAN VALUES ARE PRESENT IN THE DATASET NOW.

NOW LET’S SEE THAT VALUE COUNTS OF TARGET VARIABLE.



There are two categories in the column Churn namely No and Yes. We can assume that "No" stands for the customers who have not churned and "Yes" stands for the customers who have got churned from the company.

DROP ALERT



DROPPING CUSTOMERID AS IT WAS OF NO USE.

VISUALIZATIONS



So count of both gender are pretty same. but still male is little bit more than female.



Here 0 represents the non senoir citizens and 1 represents the senior citizens. The count of 0 is high in data compared to 1 which means the number non seniorcitizens are quite high compared to senior citizens data in the given dataset. Around 83% of the customers are non senior citizens and only 16% are senior citizens.



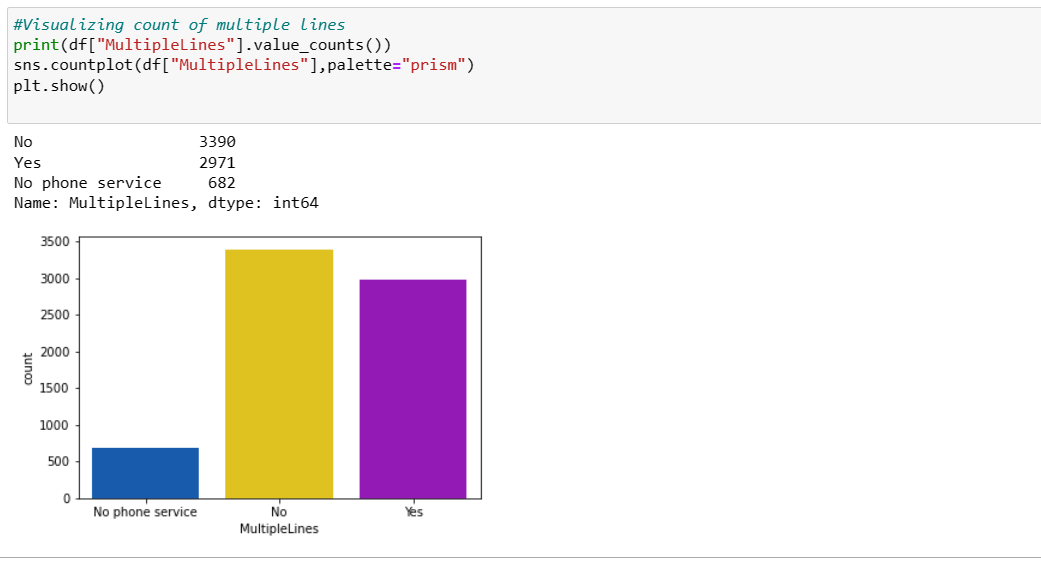
So here 0 means who do not have partner and 1 means who have partner... both are almost equal but no partner people are higher than people having partner.



So here people who have dependents on them are less and people who do not have dependents are more.



So people who have Phone service are higher in number than a people who don’t have phone service.



The customers who have phone services from single line have high counts compared to the customers having phone services from multiple lines, also the customers who do not have phone services have covered very less data compared to others.



Most of the customers have chosen to get Fiber optic internet followed by DSL, but there are many customers who do not get an internet service.



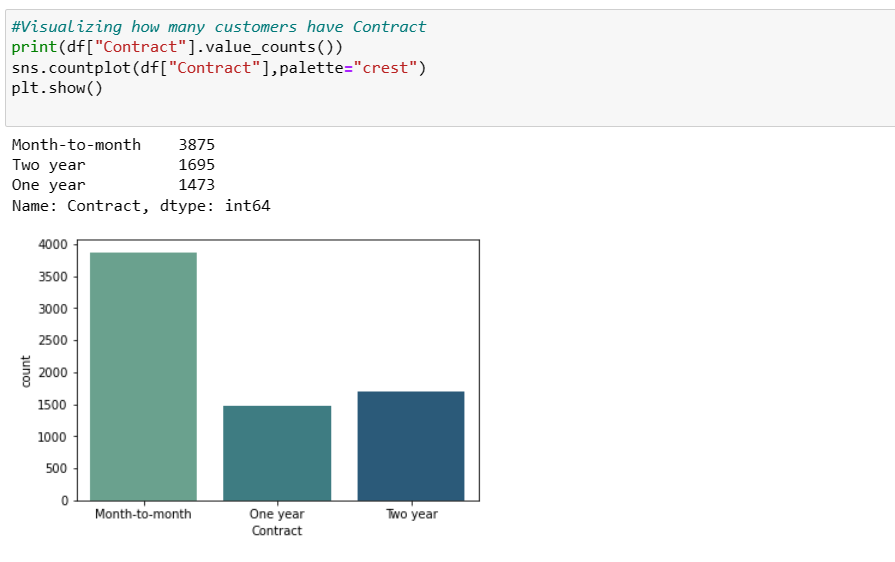
It is obvious that the customers who do not own internet services and online security, they do not need online backup usage. From the plot we can see the majority of customers who own internet services they do not have Online backup and the customers who own internet services have very less online backup. Also, the customers who do not have internet services have very less online backup counts compared to others.



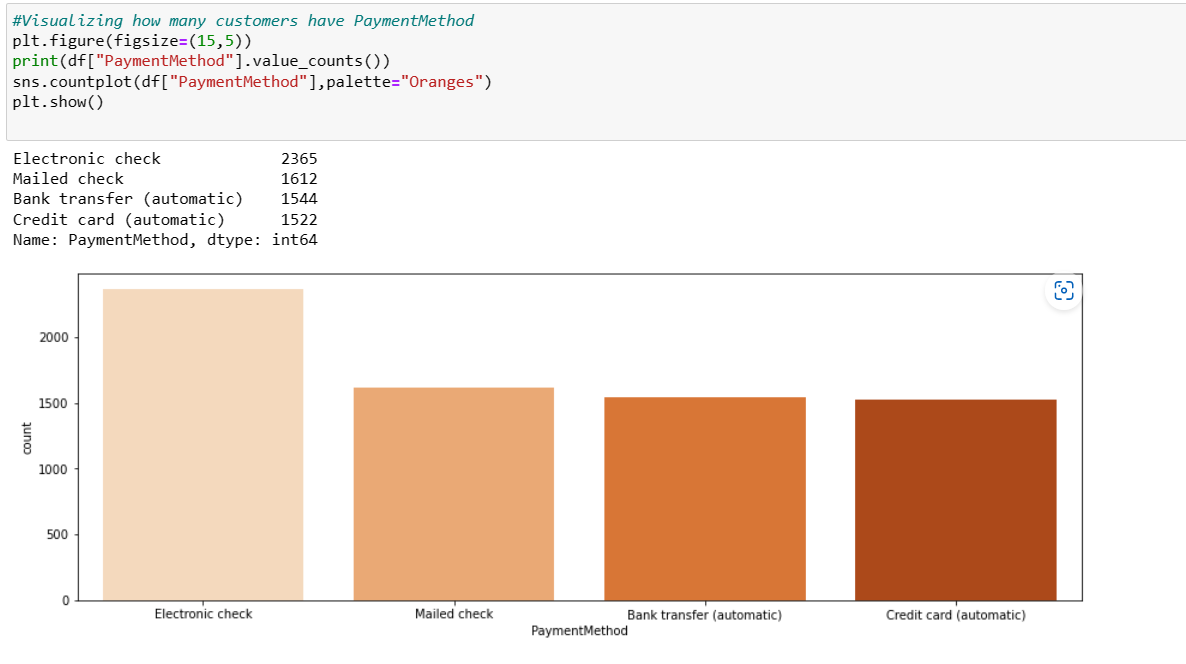
We can see in count plot that people with No device protection are higher in count as compared to people who have device protection. and people who don’t have internet service do not need device protection.



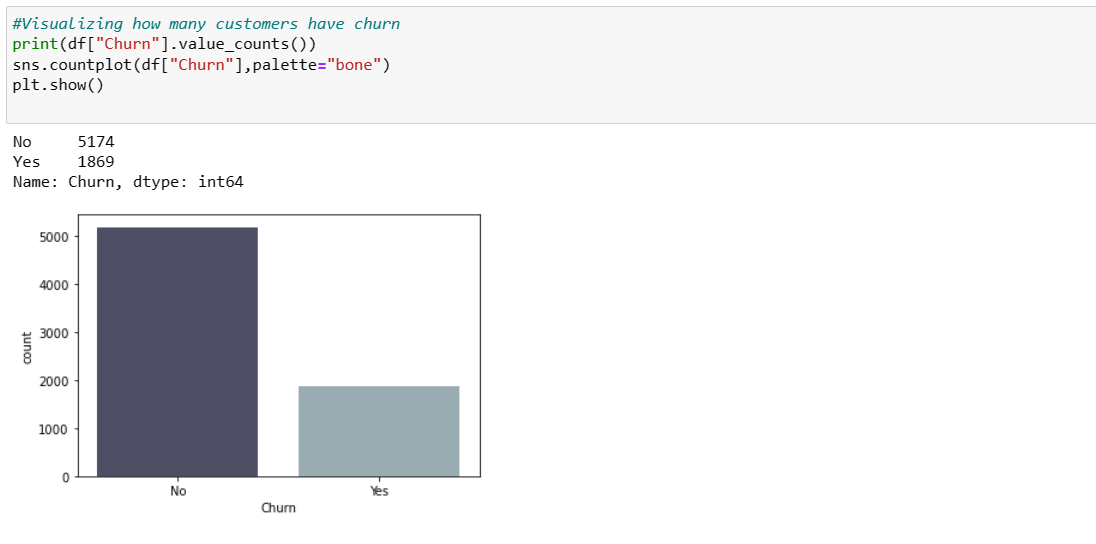
Customer who do not have streaming TV are 39.90% whereas customer who have streaming TV are 38.44%. which means customer who do not have streaming TV are higher in count than customer who have streaming tv.



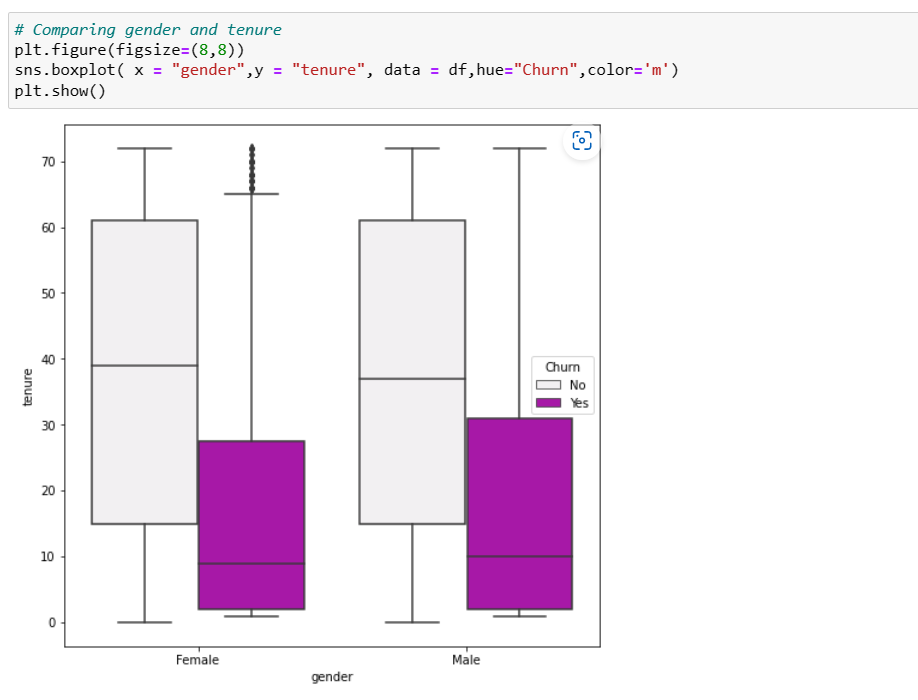
Most of the customer opt for month-to-month contract in comparison to one year or two-year contract.



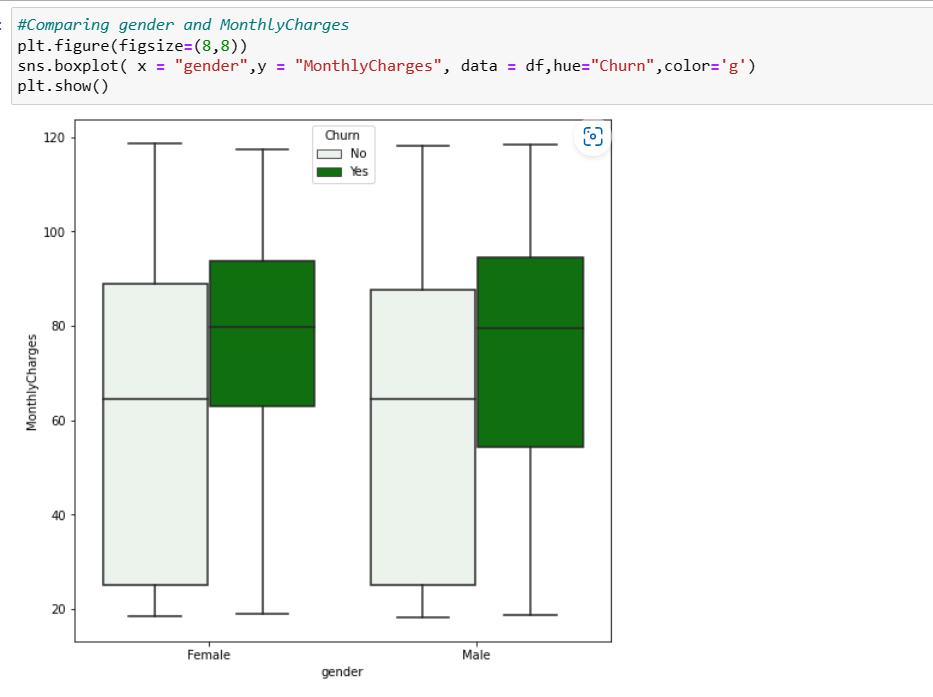
Most of customer prefer electronic check over any payment method like mailed check, bank transfer(automatic), credit card(automatic)... while mailed check, bank transfer(automatic), credit card(automatic) is nearly equally chosen for payment.



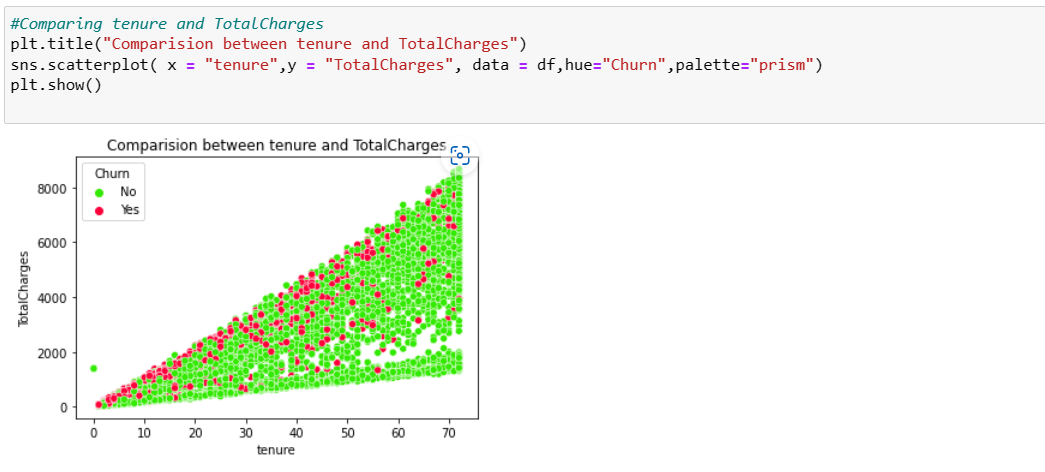
Customers who do not churn are higher in count than a customer who have churn. and this is our target variable which seems to be imbalanced.



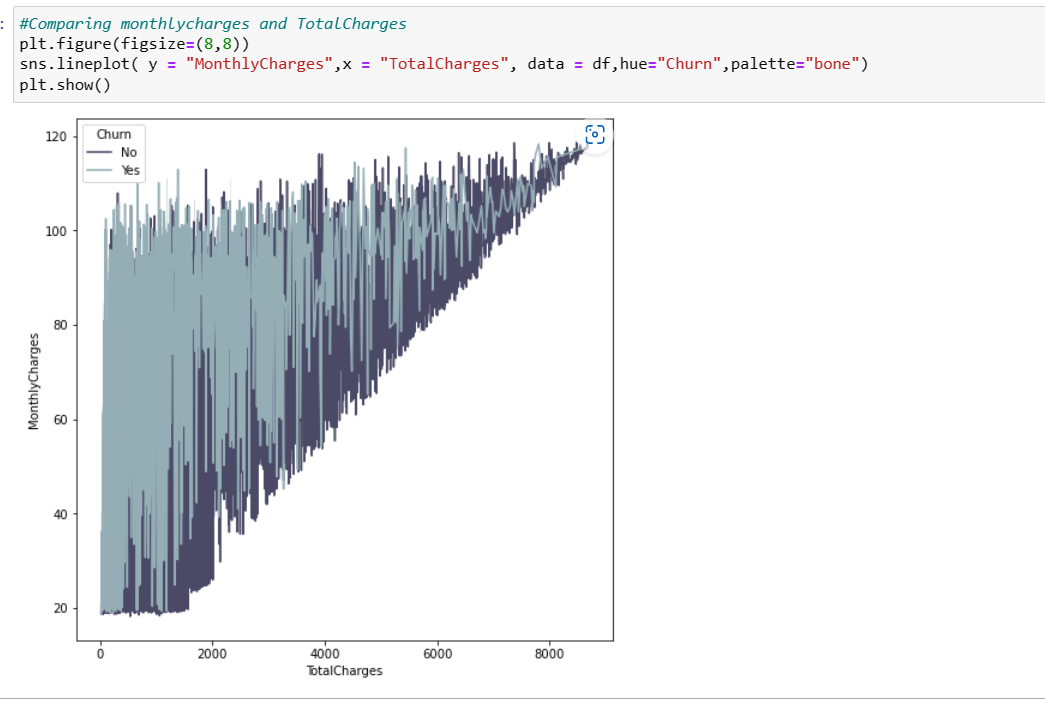
Here male are more prone to churn in comparison to female in tenure of 60 years.



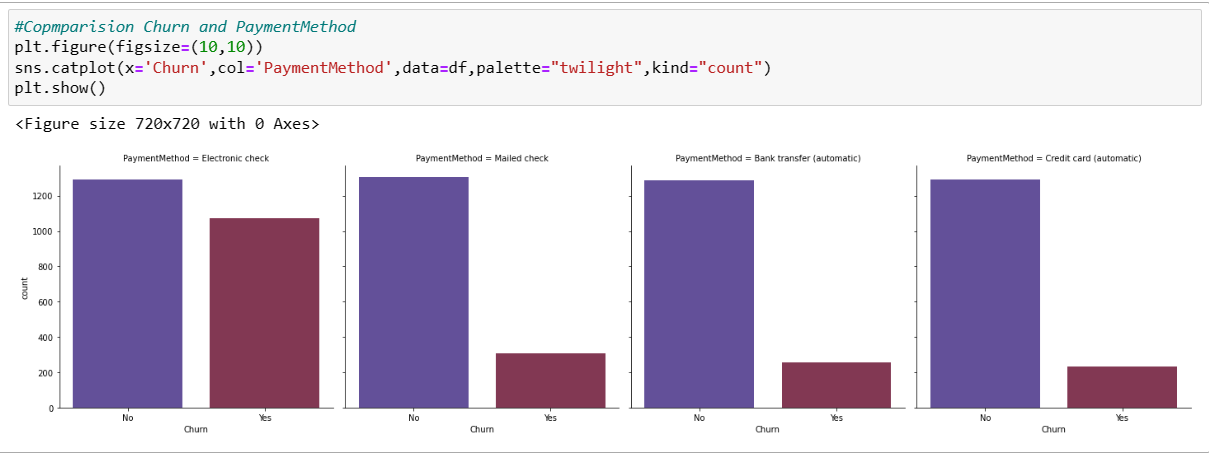
Both male and female with monthly charges above 60 have high chances to churn.



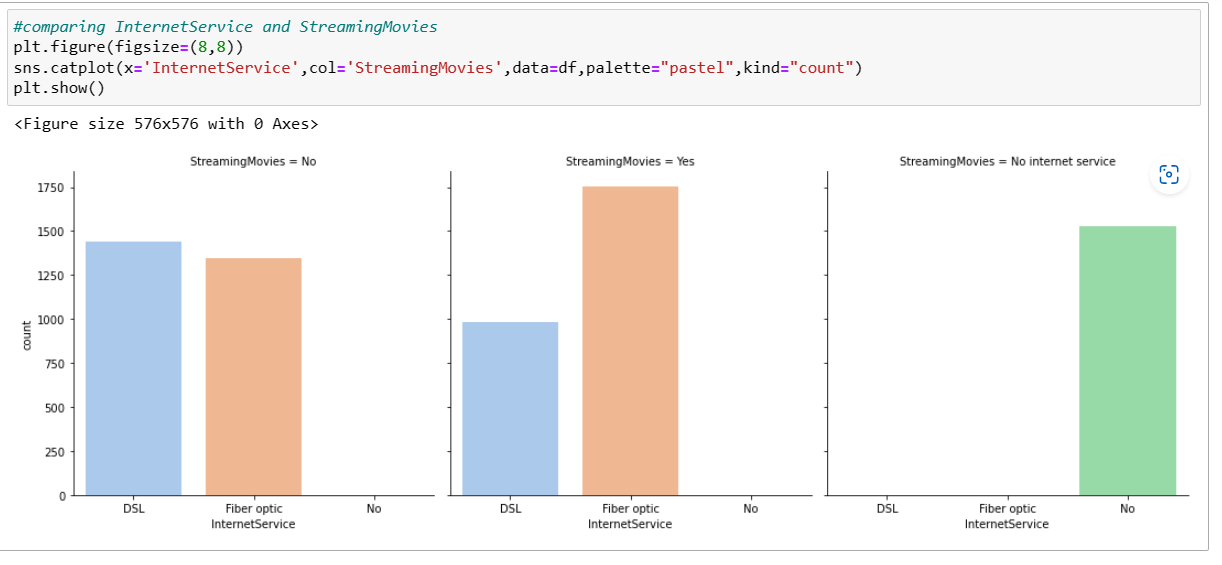
So here tenure and totalcharges have linear relationship...when tenure increases total charges increases. customers with less tenure service are more likely to churn.



monthlycharges and totalcharges have linear relationship... as monthly charges increase total charges also increases. The customers with high monthly charges have high tendency to stop the services since they have high total charges. Also, if the customers ready to contribute with the monthly charges, then there is an increment in the total charges.



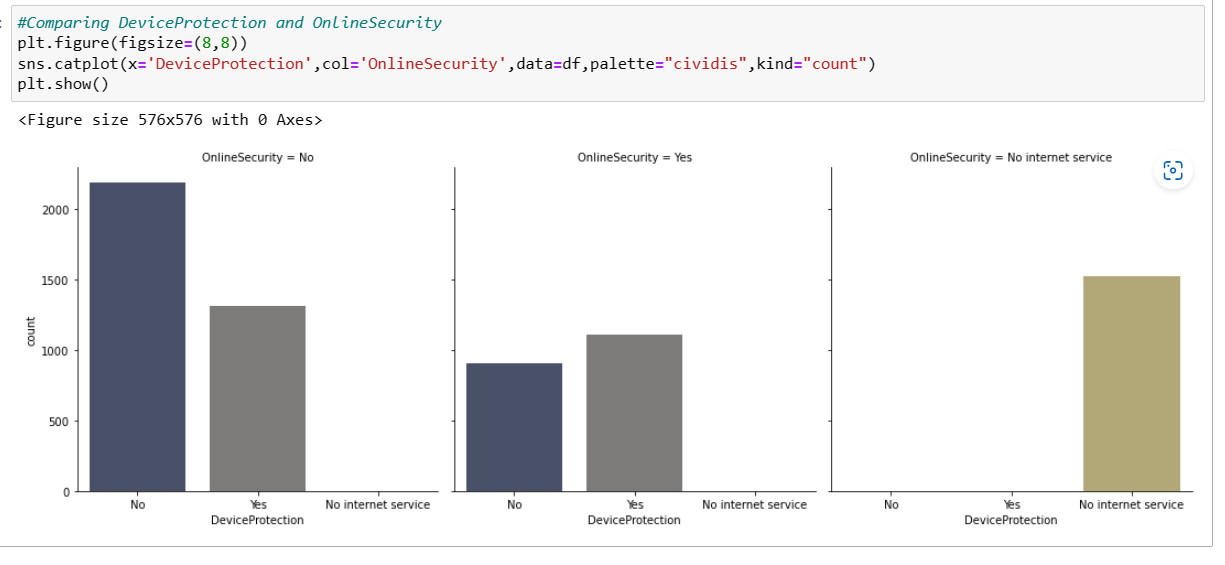
Customer who get paid through Electronic Check have highest churn. Whereas other payment method have nearly equal churn.



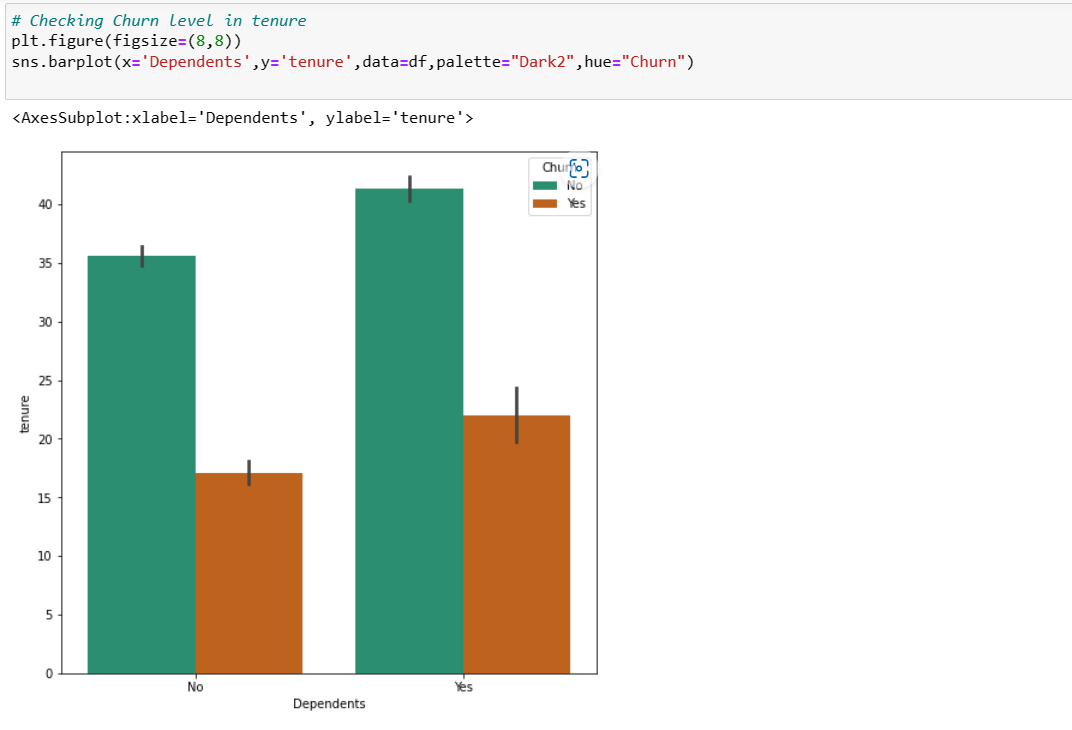
Customers with Fiber optic Internet service are highest in streaming movies followed by DSL. whereas customers with no internet service are not all streaming movies.



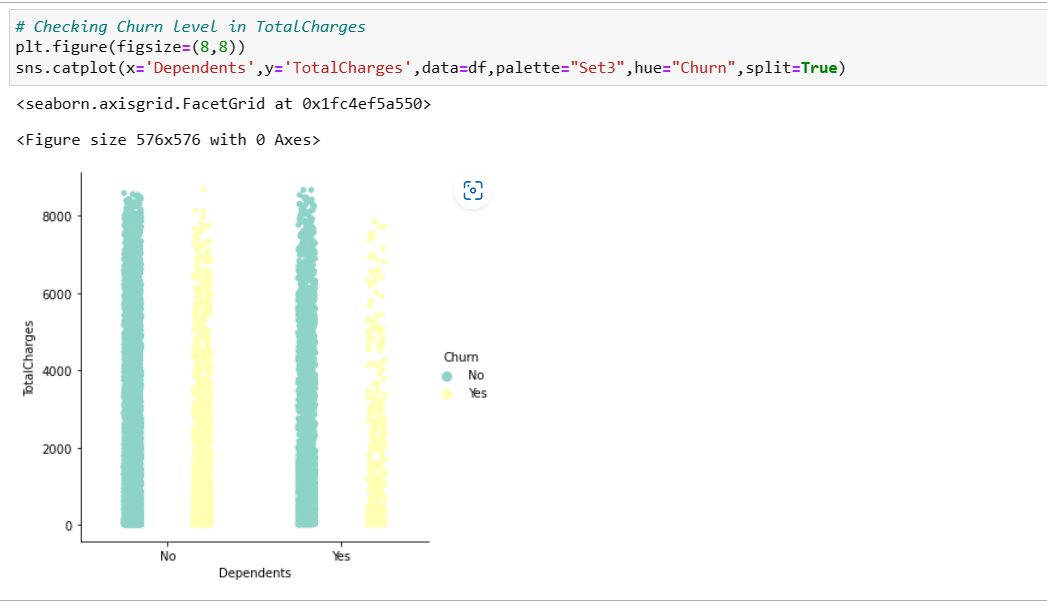
Customers who opted for onlinesecurity is highest from DSL internet service followed by Fiber Optic internet services...and customers with no internet services do no need online security



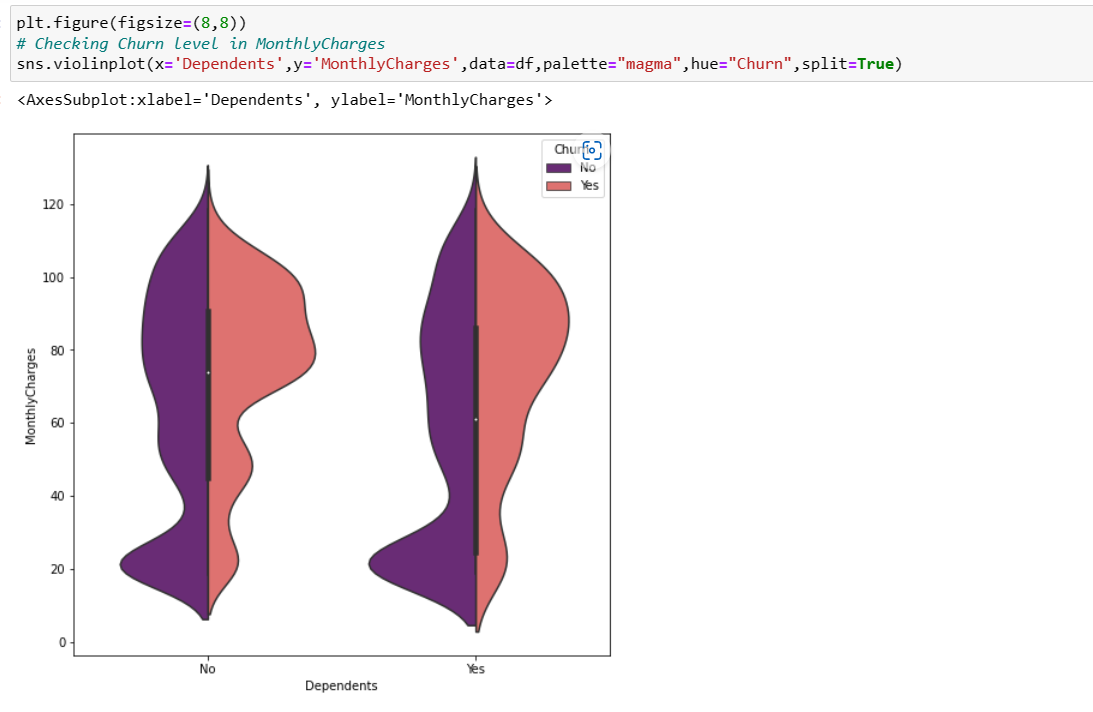
We get to know that it is on the will of customer to opt for either Online Security and Device protection or both at a same time. Here customer who have not taken online security but have taken Device protection above 1250 counts and the one who have opted for online security and have taken Device protection too is little less than the one who have not taken Online Security.



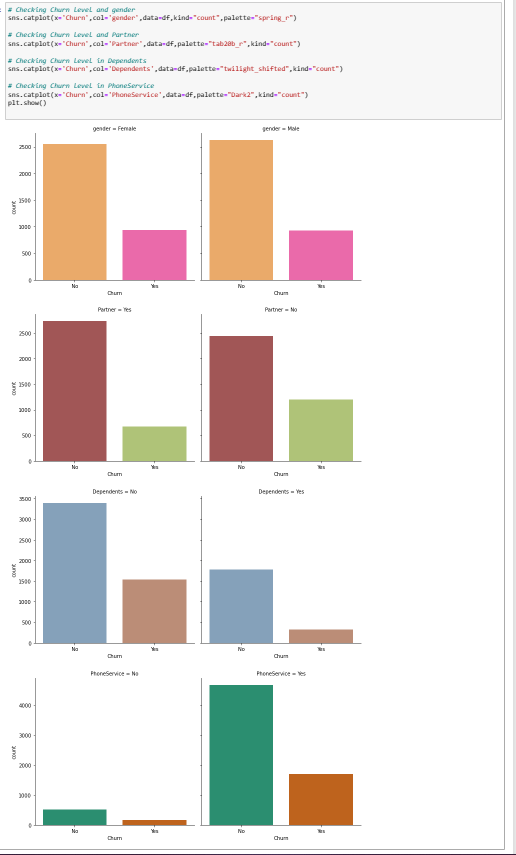
The customers who have dependents with high tenure, then the churned level is high.



Customer who have total charges in range of 0-2000 with dependents have high chance of churn.



The customers having Monthly charges between 70-110 with dependents have high churn rate and when the customers have no dependents and having monthly charges around 20 then the ratio of churn is very high.



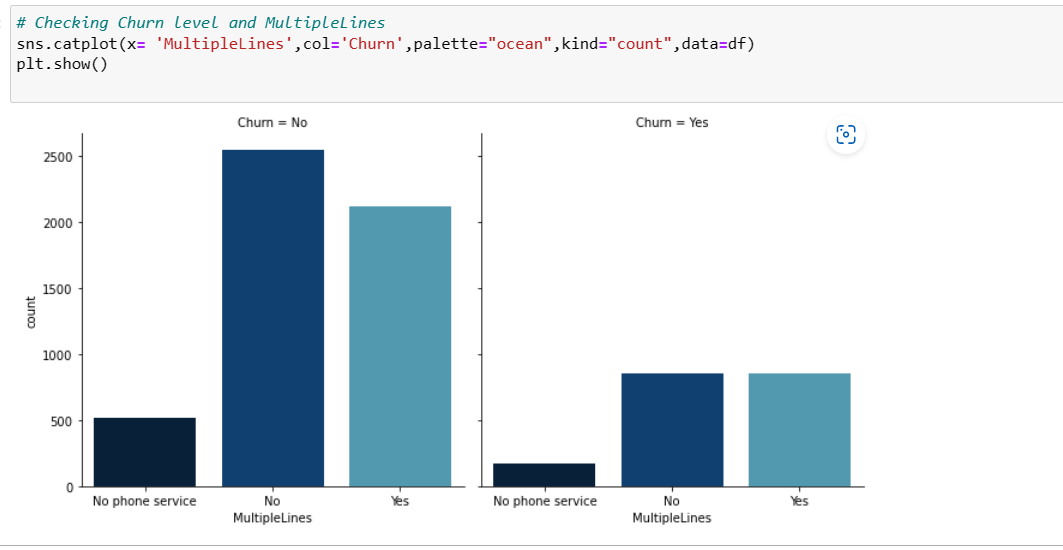
Observation:

- In the first plot we can see there is no significance difference in the genders, both the genders have equal churn level.

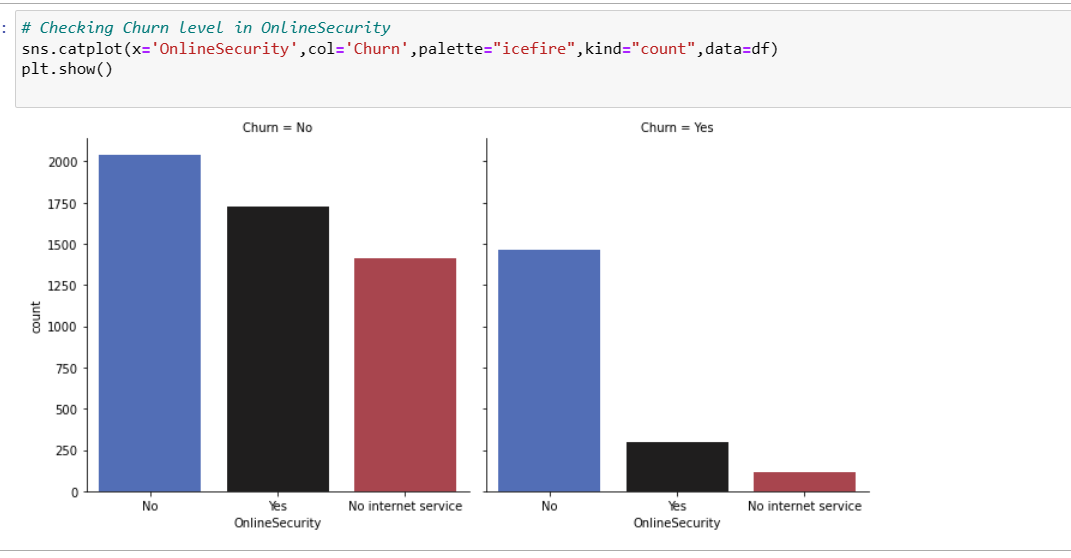
- In the second plot we can see the customers without partners have high churn rate compared to the customers with partners.

- The customers who do not have any dependency have high churn rate compared to the customers who have dependents.

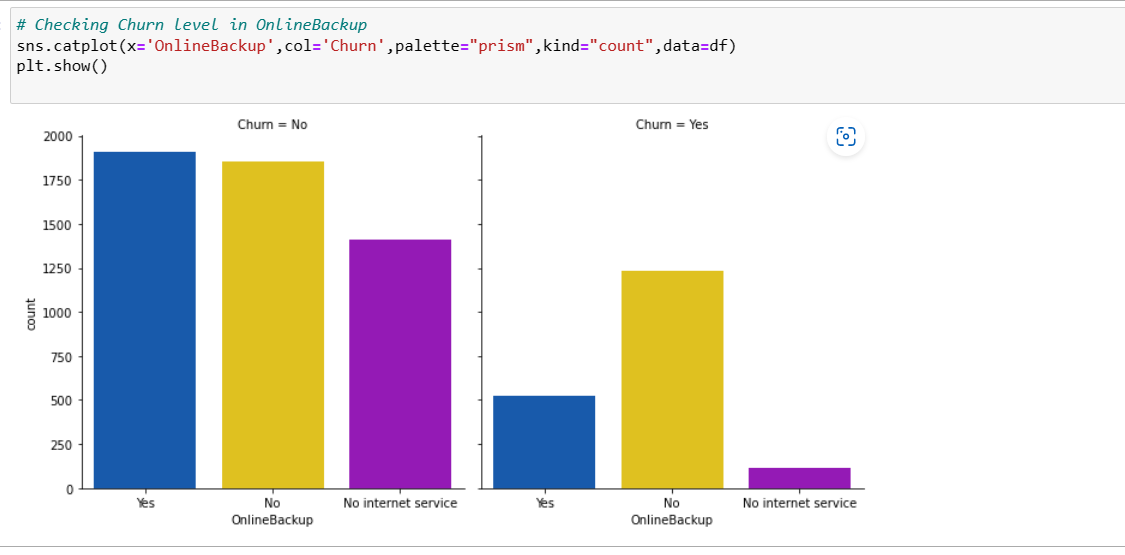
- In the last plot we can notice the customers who have phone service have high tendency of getting churned.



The customers who have phone services from single line have high churn rate compared to the customers having phone services from multiple lines, also there are very less number of customers who do not have phone services.



The customers who have no internet service have very less churn rate and the customers who do not have online security services have high tendency to getting churned.



It is also same as in the case of online security. It is obvious that the customers having who do not have internet services they do not need any online backup. The customers who do not have online backup services they have high churn rate.

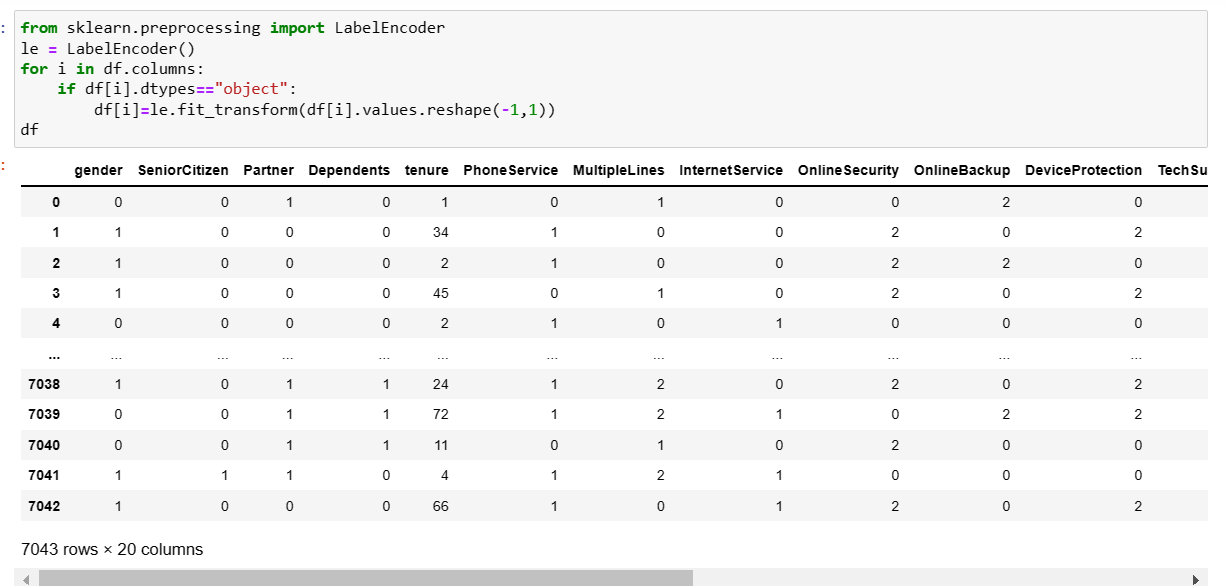


The customers who have churned are mostly having month to month contract.



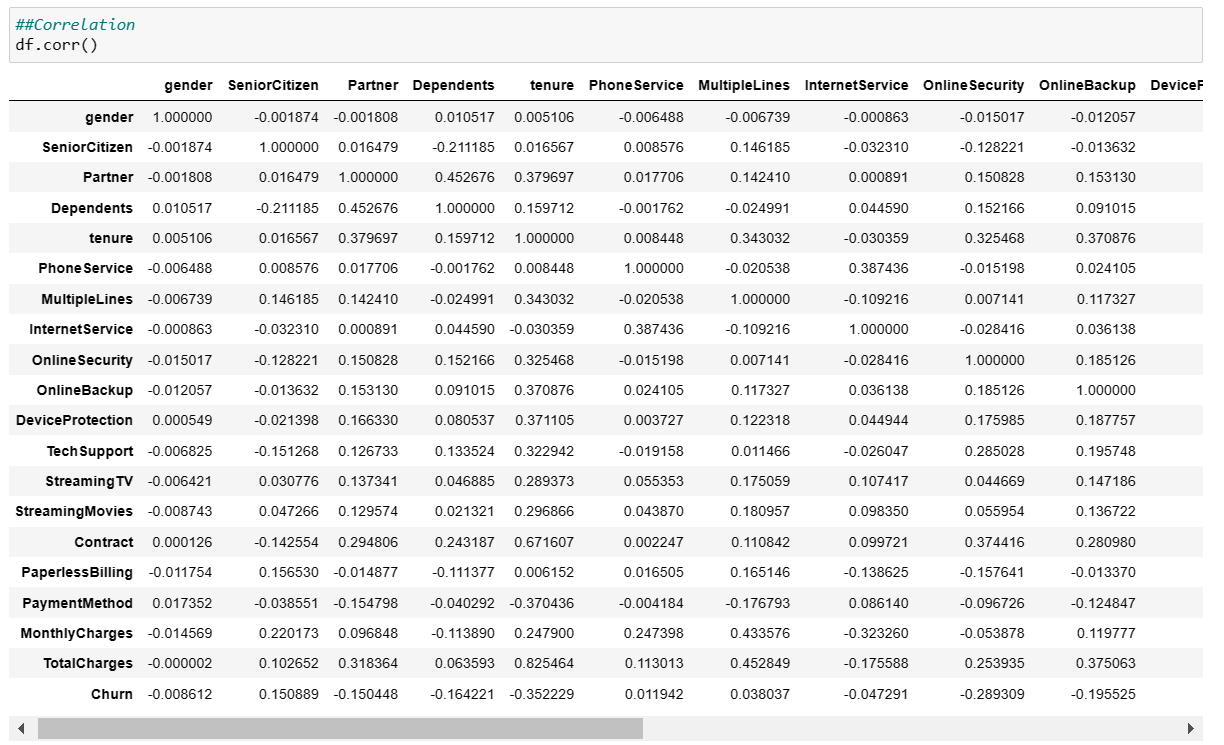
The customers who prefer paperless billing they have high churn rate in comparison with customers who have paper billing.

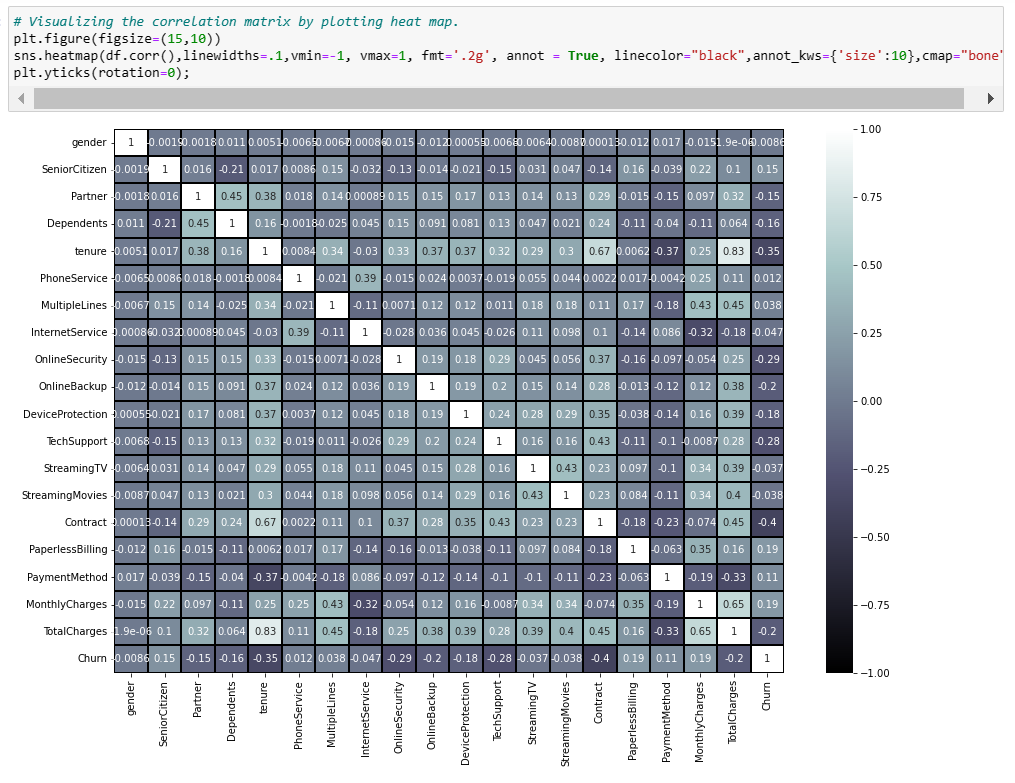
ENCODING TECHNIQUE



I have done encoding of all the object columns from the dataset.

CORRELATION





Observation:

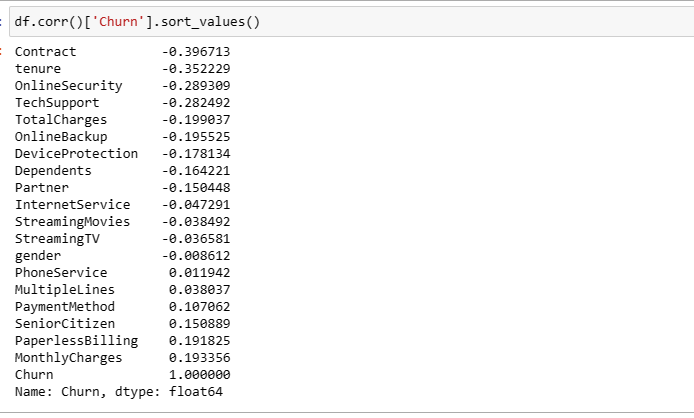
- This heatmap shows the correlation matrix by visualizing the data. we can observe the relation between feature to feature and feature to label. This heat map contains both positive and negative correlation.

- There is no much positive correlation between the target and features.

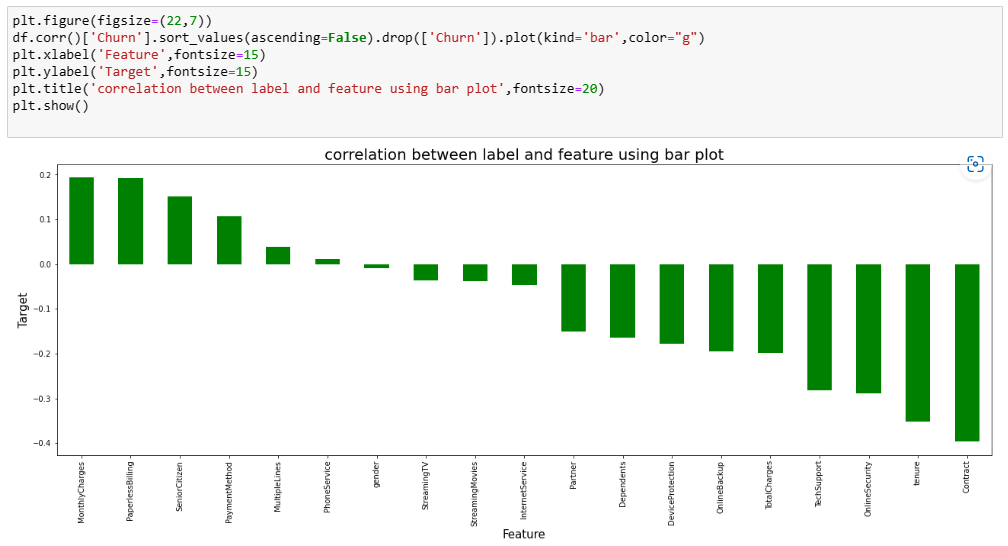
- The columns MonthlyCharges, PaperlessBilling, SeniorCitizen and PaymentMethod have positive corelation with the label Churn.

- The label is negatively correlated with Contract, tenure, OnlineSecurity, TechSupport, TotalCharges, DeviceProtection, OnlineBackup, Partner and Dependents.

- Also, the column gender has very less correlation with the label, we can drop it if necessary.

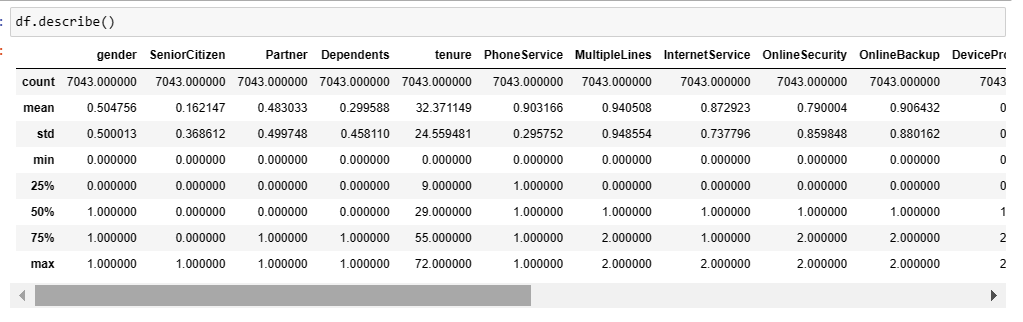


We can observe positive and negative correlation with target.



This shows features relationship with target. Column gender and phone service have least relationship with target.

DESCRIPTIVE STATISTICS: -



Observation:

- This gives the statistical information of the numerical columns. The summary of this dataset looks perfect since there is no negative/ invalid values present.

- From the above description we can observe the following things.

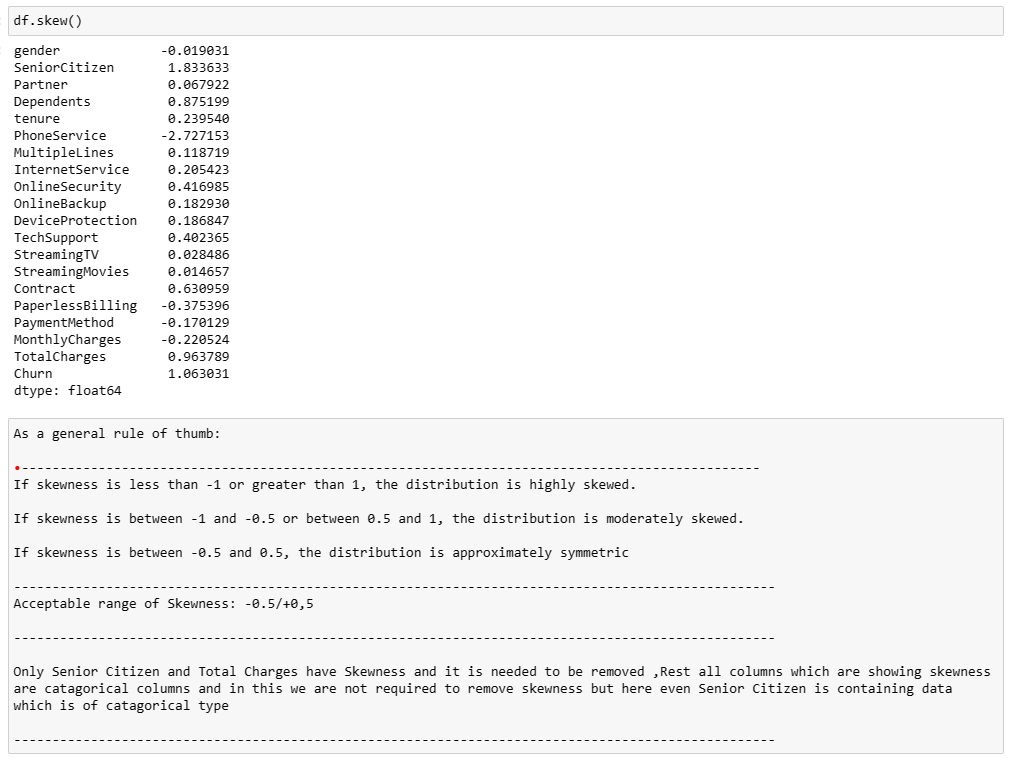
- The counts of all the 4 columns are same which means there are no missing values in the dataset.

- The mean value is greater than the median (50%) in tenure and TotaCharges columns which means the data is skewed to right in these columns.

- The data in the column MonthlyCharges have mean value less than median that means the data is skewed to left.

- We can also notice the Standard deviation, min, 25% percentile values from this describe method.

LET’S CHECK SKEWNESS

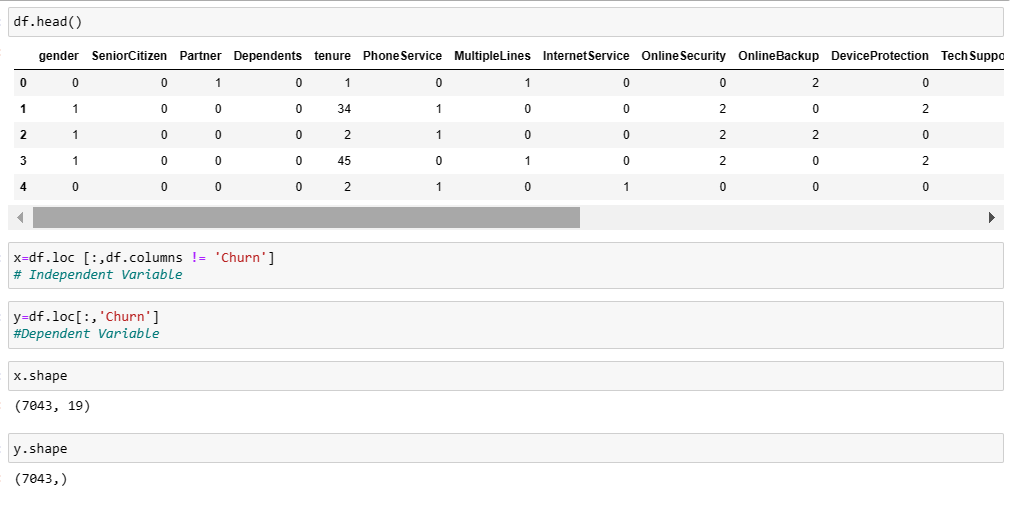


NOW LET’S REMOVE SKEWNESS

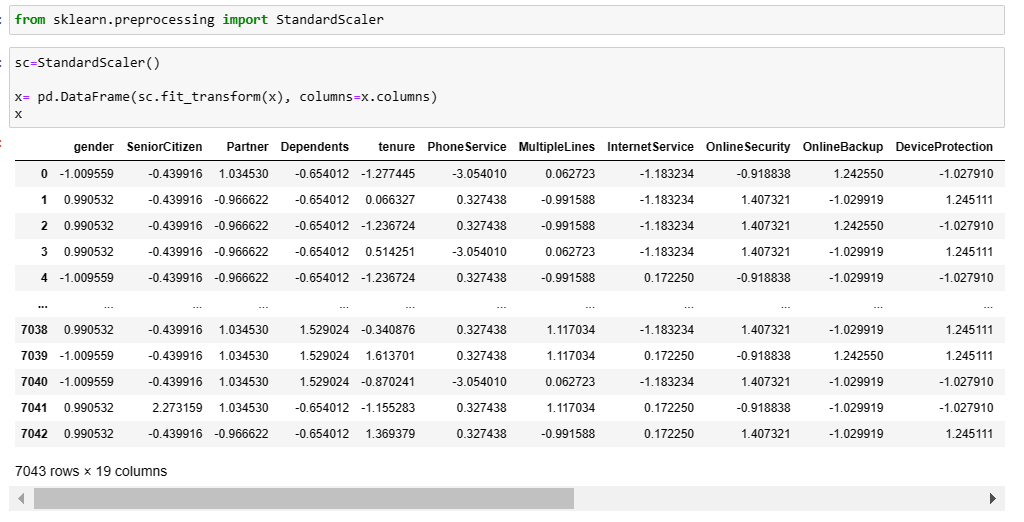


Skewness from Total Charges has been removed

NOW LET’S SPLIT THE DATASET INTO INDEPENDENT AND DEPENDENT VARIABLE: -



NOW LET’S SCALE THE DATASET : -



We have scaled the data using standard scalarization method to overcome with the issue of data biasness.

CHECKING MULTICOLLINEARITY USING VIF METHOD : -

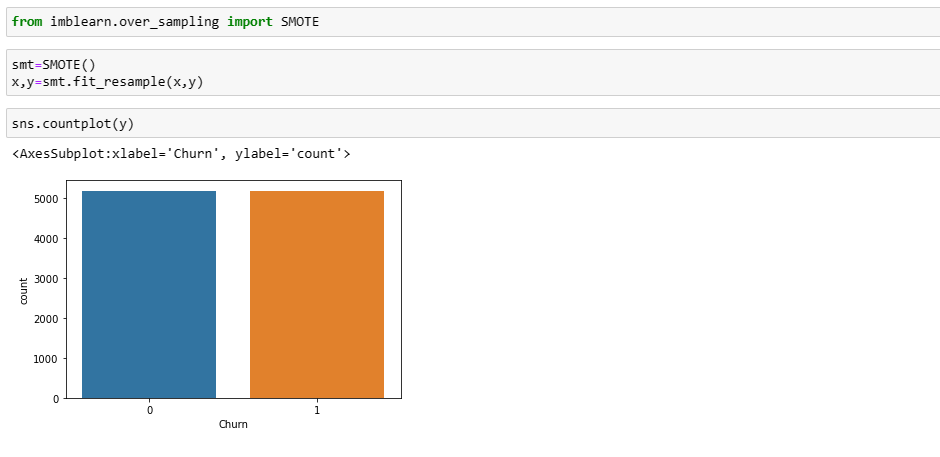


By checking VIF value we can find the features which causing multicollinearity problem. Here we can find the feature TotalCharges and tenure have VIF value greater than 10 which means they have high correlation with the other features. We will drop one of the column first, if the same issue exist then we will try to remove the column having high VIF(above 10).

NOW DROPPING TOTALCHARGES AS IT WAS MULTICOLLINEARITY ABOVE 10 AND CHECKING AGAIN.



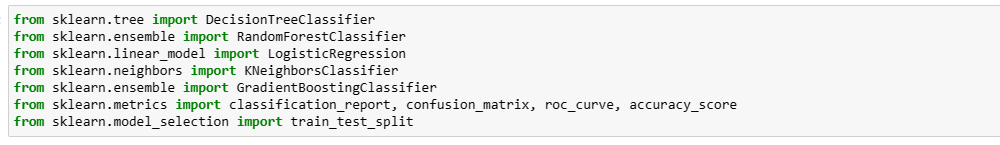
LET’S BALANCE THE IMBALANCED TARGET VARIABLE.



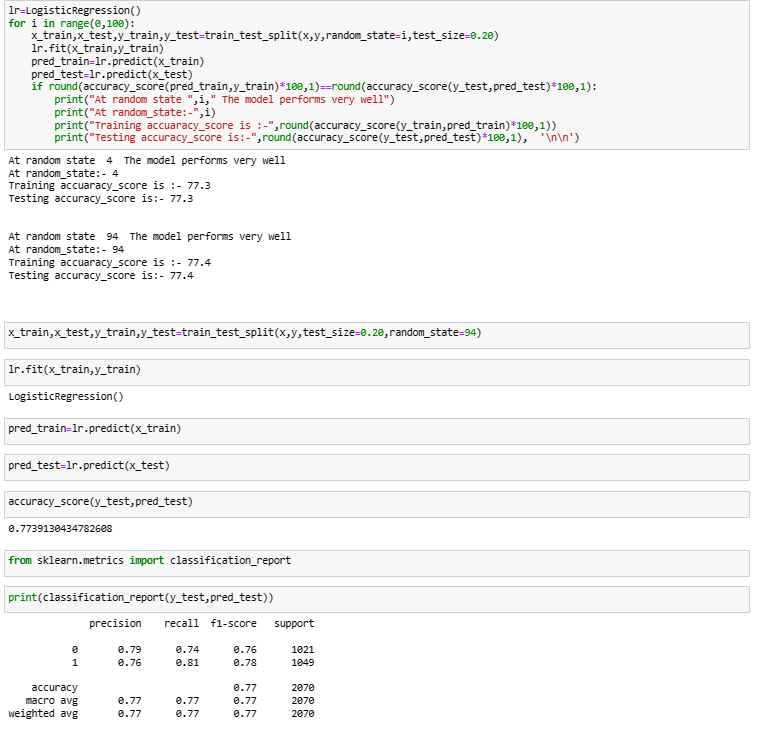
NOW THE DATASET IS BALANCED.

NOW DO MODEL TRAINING AND TESTING.

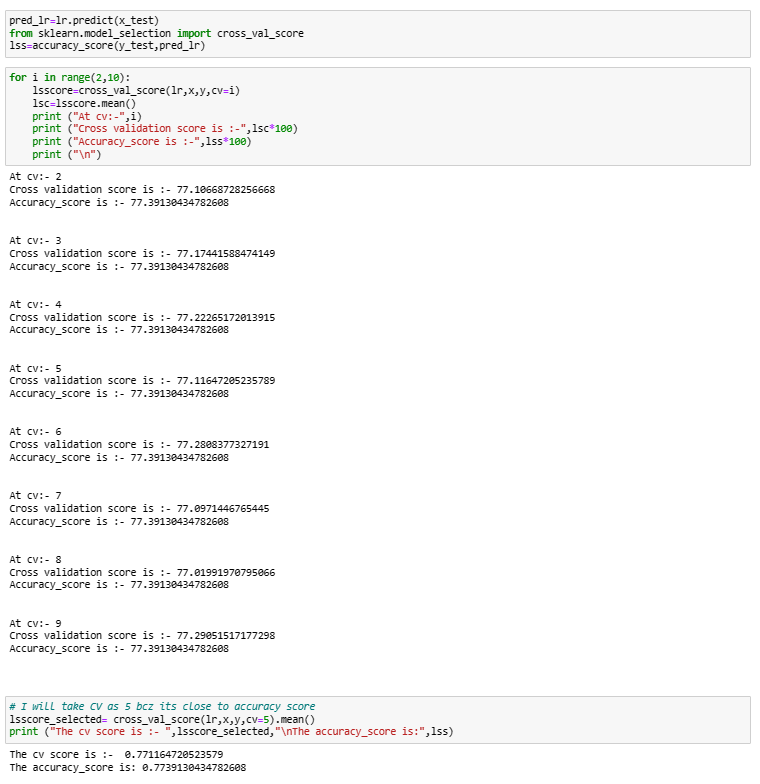
IMPORTING SOME OF THE IMPORTANT DATASET: -



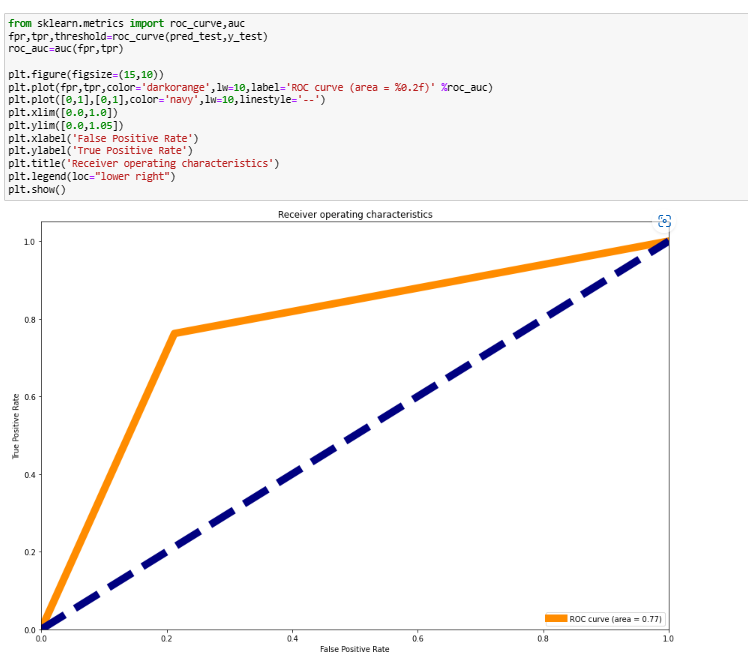
LOGISTIC REGRESSION



CV FOR LOGISTIC REGESSION

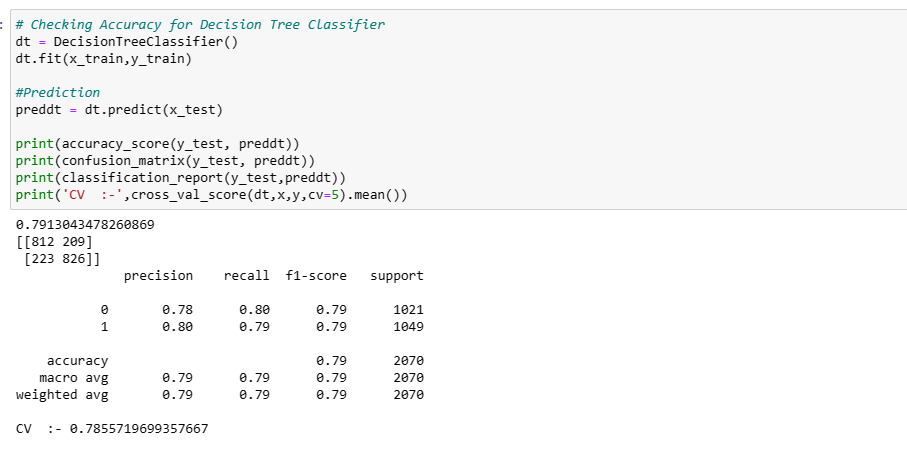


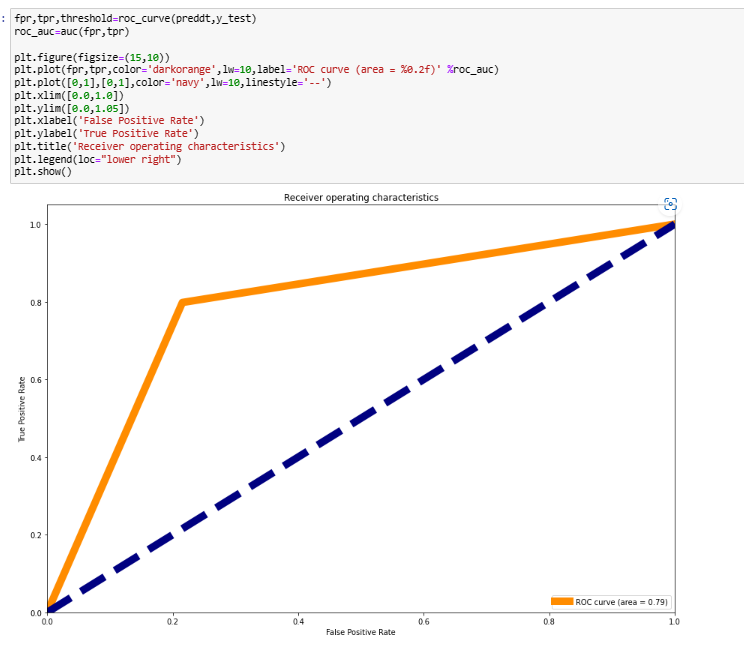
AUC-ROC CURVE FOR LG.



LOGISTIC REGRESSION COVERS 77% UNDER ROC CURVE.

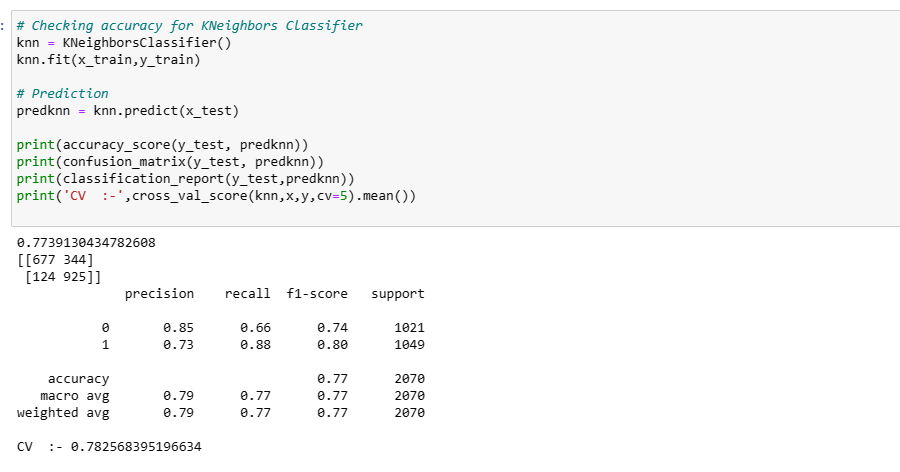
DECISION TREE CLASSIFIER

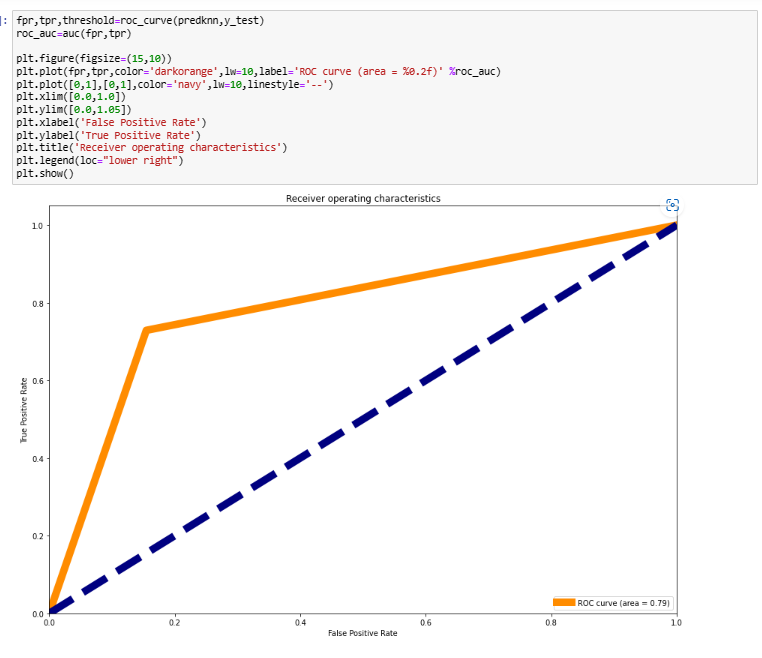




IN DTC 79% AREA IS UNDER THE CURVE.

KNEIGHBORS CLASSIFIER: -

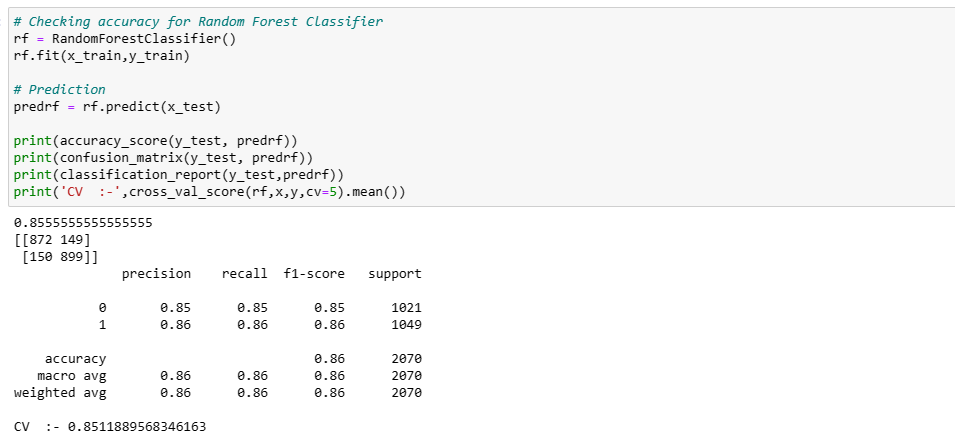


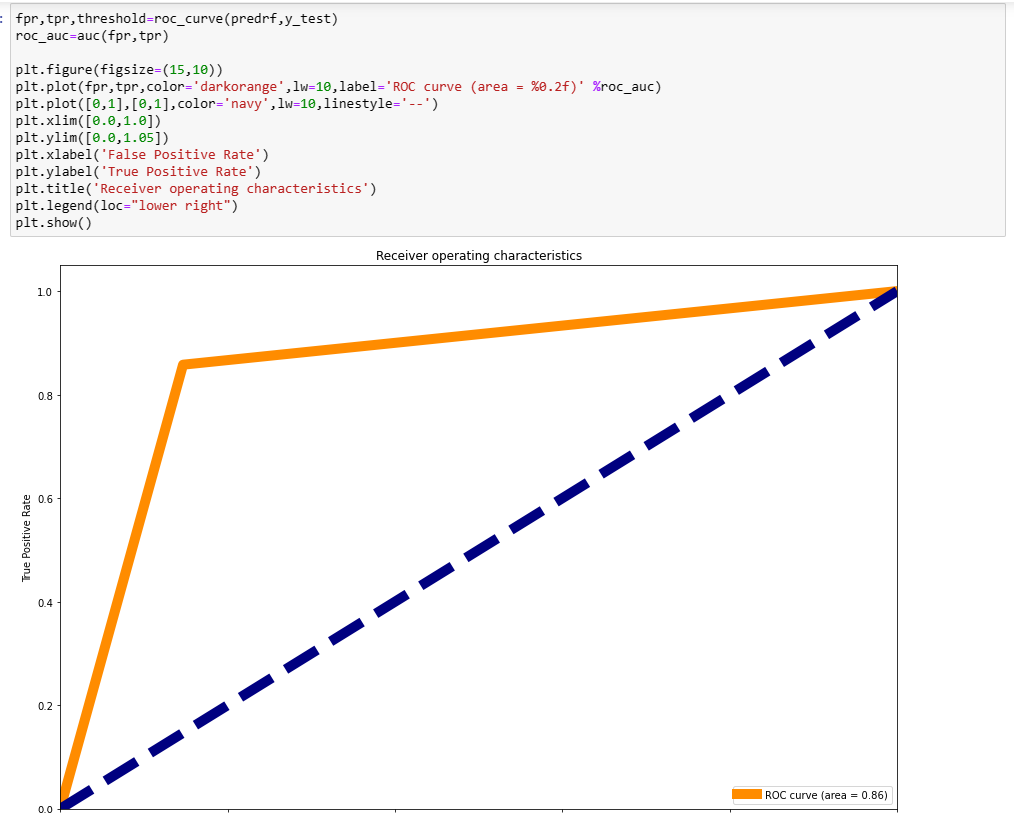


KNN IS COVERING 79% AREA UNDER ROC CURVE.

ENSEMBLE TECHNIQUE: -

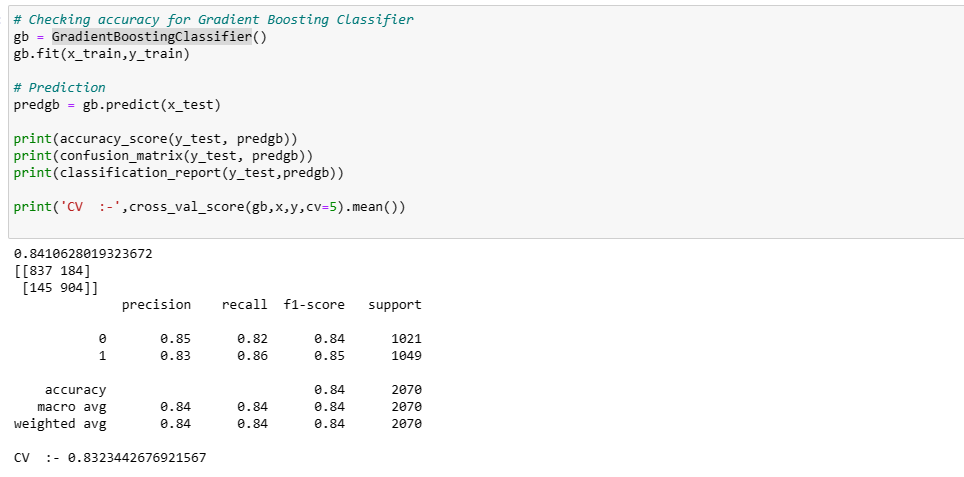
RANDOM FOREST CLASSIFIER

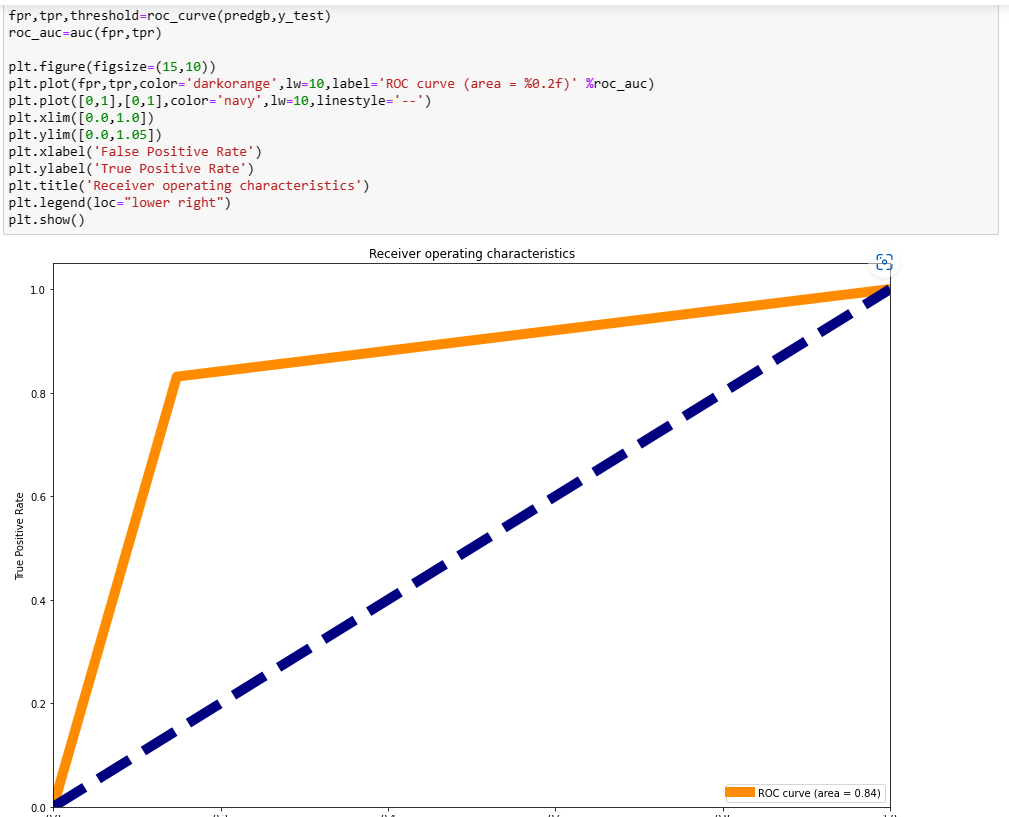




RF Classifier is covering 86% under ROC Curve.

Gradient Boosting Classifier: -

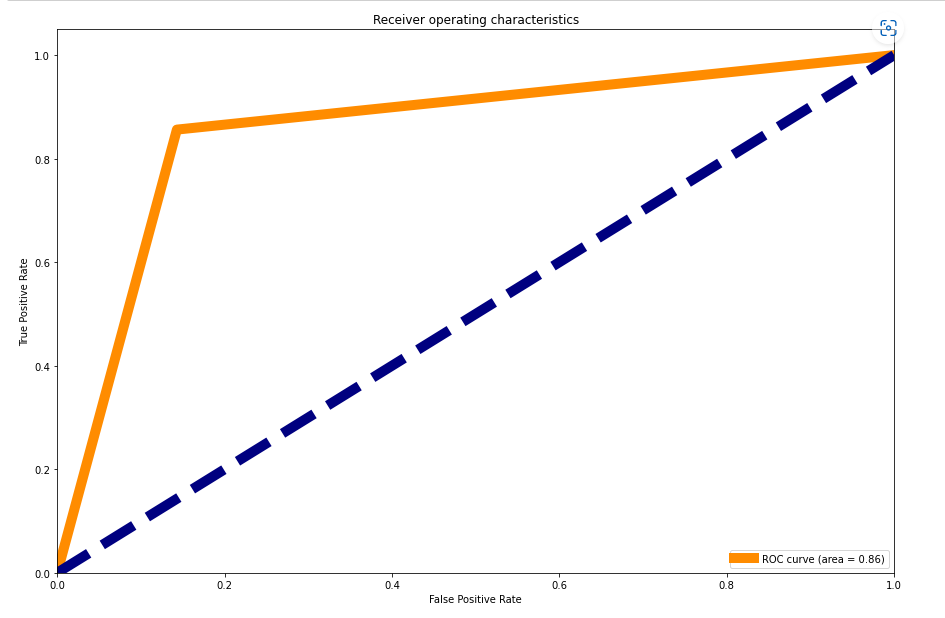




GB Classifier covering 84% under ROC Curve.

EXTRA TREE CLASSIFIER: -





Extra tree Classifier is covering 86% under ROC Curve.

LET’S SEE WHICH MODEL PERFORMED THE BEST: -

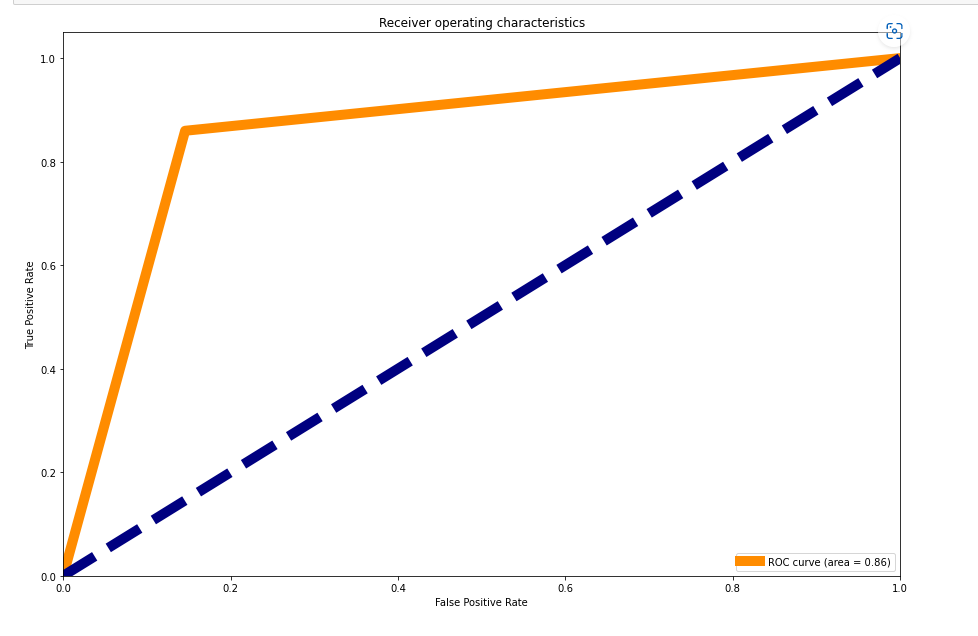


From the difference between the accuracy score and the cross validation score we can conclude that Extra Trees Classifier as our best fitting model which is giving very less difference compare to other models.

HYPER PARAMETER TUNING : -







Our final model is covering 86% under ROC Curve in Hyperparameter tuning.

NOW SAVING THE MODEL: -



OUR PREDICTION AND ORIGINAL DATA ARE ALMOST EQUAL.

CONCLUSION: -

This blog talks about what Customer Churn is, and how Customer Churn Analysis is carried out. It also highlights several of the finer points regarding this topic like the importance of Churn Analysis, the process involved, methods to reduce Customer Churn, and the challenges one faces while carrying out Churn Analysis.