**Customer Churn Analysis in Telecom Industry**

* **Objective:**

In this article, I’m going to apply data analytics skills on the telecom customer churn dataset which contains customer level information and try to find insights about the key attributes which are causative for customer analysis, we try to find the hidden trends from customer behavior through visualization techniques, build several predictive classification models. This article concludes by providing successfully predict potential customer churn, Companies can use such ML pipelines to initiate retention strategies on those customers who are classified as likely targets of churn.

* Abstract:

With rapid development of telecommunication industry, the service providers are more inclined towards expansion of the customer base. Customer acquisition and retention has become a key concern for several industries and is particularly acute in fiercely competitive and growing business. Finding the Key factors which triggers the customer churn plays important role in early initiation of customer retention policies and cut back the churn. We will focus on analyzing the customer data, perform exploratory data analysis, to get insight about which variables are contributing to customer churn, implementing the machine learning algorithms to identify potential churn customers and label them based on usage patterns and visualize the results.

* Problem Statement:

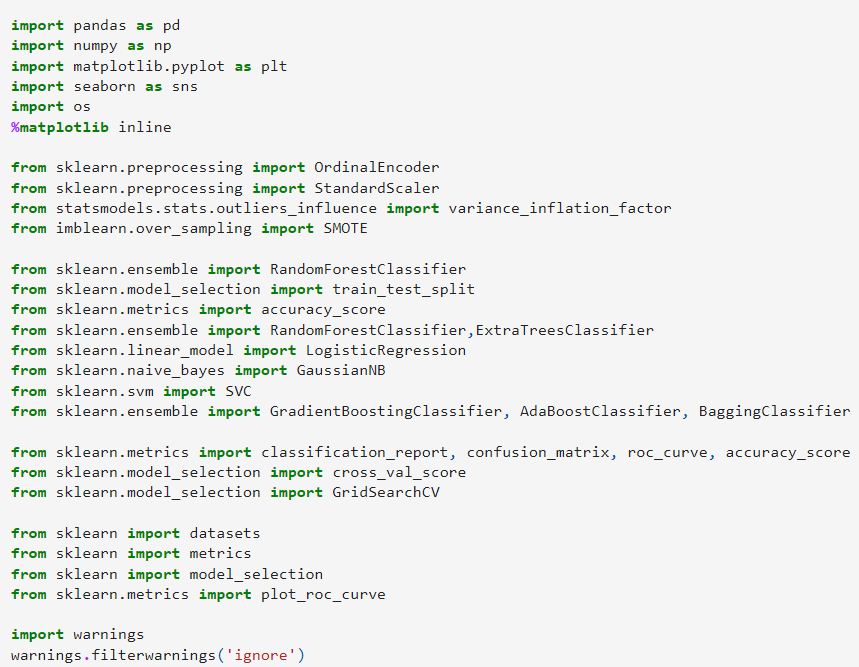
Customer churn is when a company’s customers stop doing business with that company. Businesses are very keen on measuring churn because keeping an existing customer is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales 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 prioritize 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.

* Data Analysis:

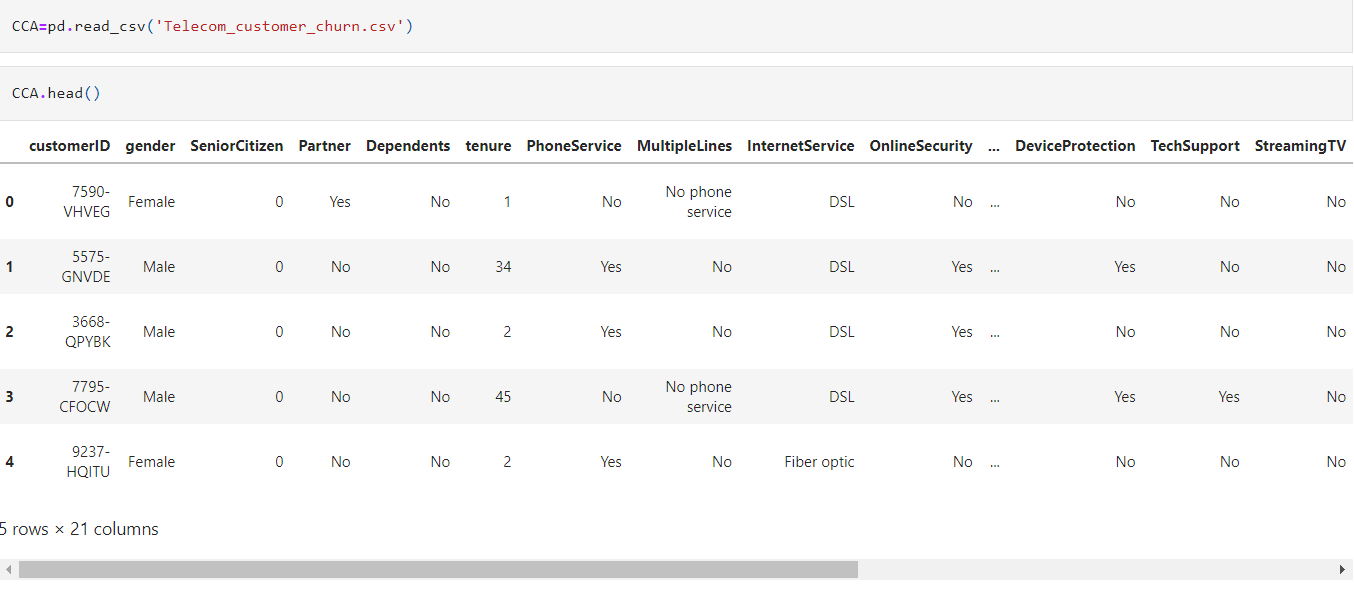
Let us import all the necessary libraries which are required for the data analysis, these inbuilt libraries in the python language.



* Overview of dataset

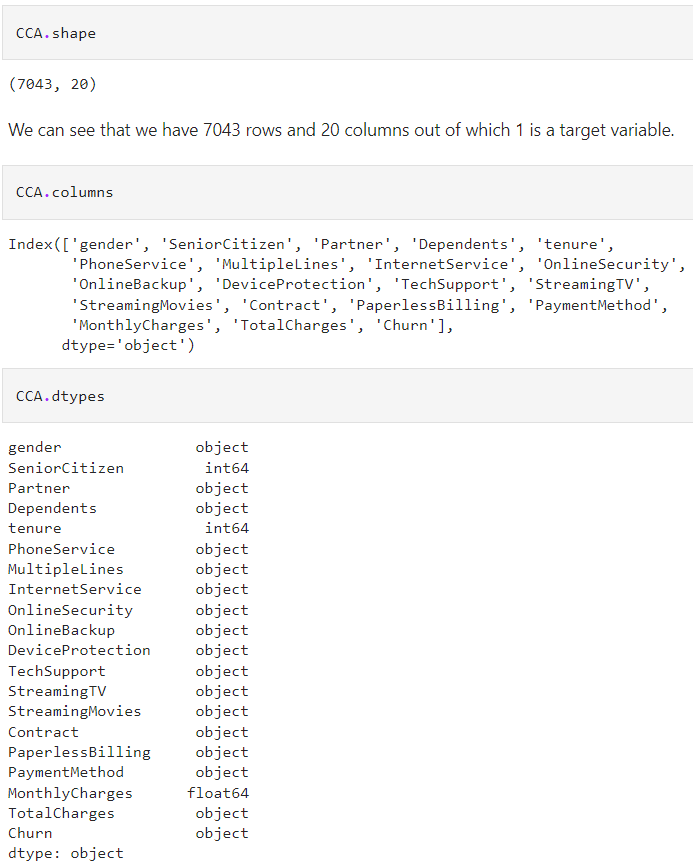
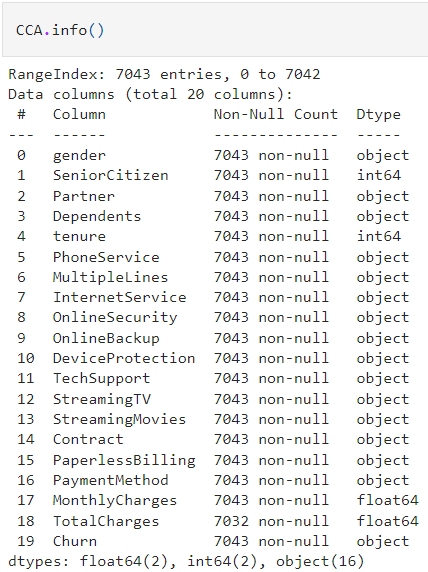
Let us import the dataset, understand the data and what each record in row and column is about. Here each column consists of details of customer information which is recorded by company during the service. There are 4 types of information,

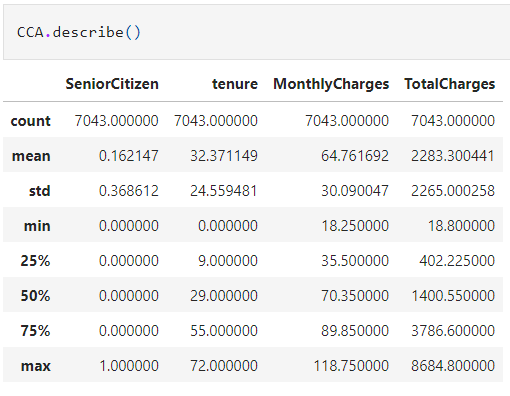
I. Services used by customer. II. Customer Demographic information (basic details) III. Customer churn Details. IV. Customer Account information.



Following are the attributes available from the datasets which contains all features and (Churn)target variable.

* Exploratory Data Analysis (EDA)



* Observations:

1. The dataset consists of 7043 rows and 21 columns,
2. The dataset contains 18 categorical and 3 numerical values represented in right table. Total charges should be numerical data, but it is showing categorical data, let us check what is missing.
3. There are no null values in the dataset.
4. The descriptive statistical overview of the dataset is given in right table

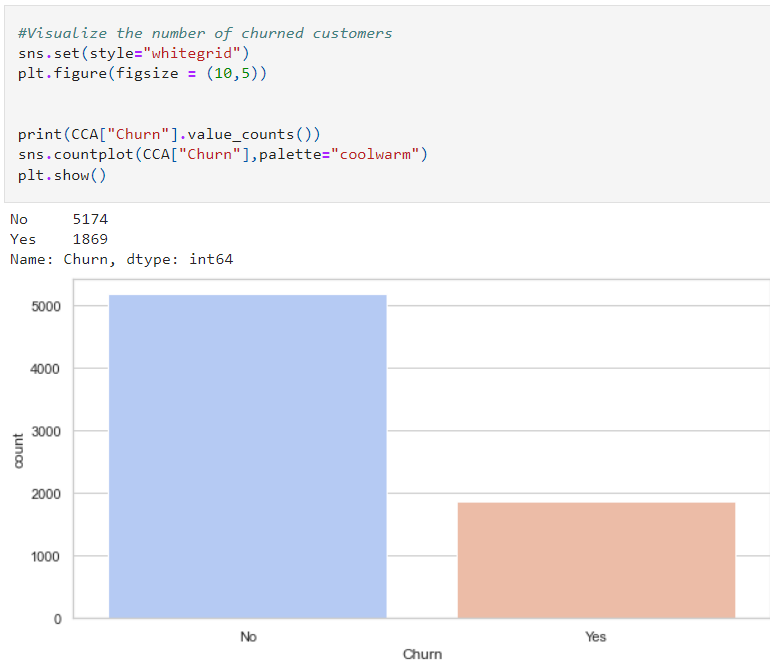
* The below table provides brief understanding of all the columns and types of data present

|  |  |  |
| --- | --- | --- |
| Column | Definition | Comments |
| Customer ID | Unique ID provided by company |  |
| gender | Gender of customer | Male, Female |
| Senior Citizen | If customer is Senior Citizen | 1= Yes, 0=No |
| Partner | If customer have partner | Yes, No |
| Dependents | If customer have any dependents | Yes, No |
| tenure | Since how many years customer is using service of company |  |
| Phone Service | If customer has phone service | Yes, No |
| Multiple Lines | If customer uses multiple line service | Yes, No, No phone service |
| Internet Service | Does customer have internet service | Fiber Optic, DSL, No |
| Online Security | Does customer use online security service | Yes, No, No internet service |
| Online Backup | Does customer use online backup service | Yes, No, No internet service |
| Device Protection | Does customer use device protection service | Yes, No, No internet service |
| Tech Support | Does customer use tech support of company | Yes, No, No internet service |
| Streaming TV | Does customer streams TV? | Yes, No, No internet service |
| Streaming Movies | Does customer Streams movies ? | Yes, No, No internet service |
| Contract | What type of contract does customer use? | Month-to-Month, One year, Two year |
| Paperless Billing | Does customer prefer paperless billing? | Yes, No |
| Payment Method | What is mode of payment customer opt for? | Payment Method |
| Monthly Charges | What is monthly charge of customer |  |
| Total Charges | Total charges since using the service |  |
| 'Churn' | Does customer leaves company or continues with the company service? | Yes, no |

* Data Visualization:

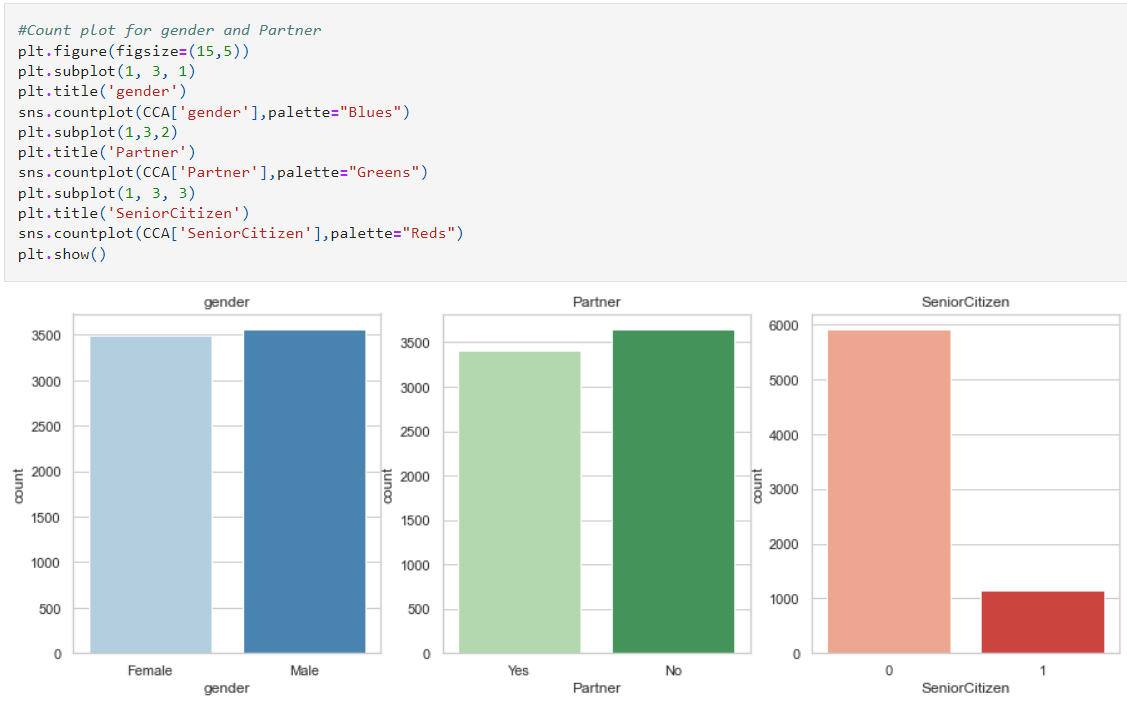
Let us proceed with the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test, check assumptions with the help of summary statistics and graphical representations. And summarize the main characteristics for model building.

* **Univariate Analysis**



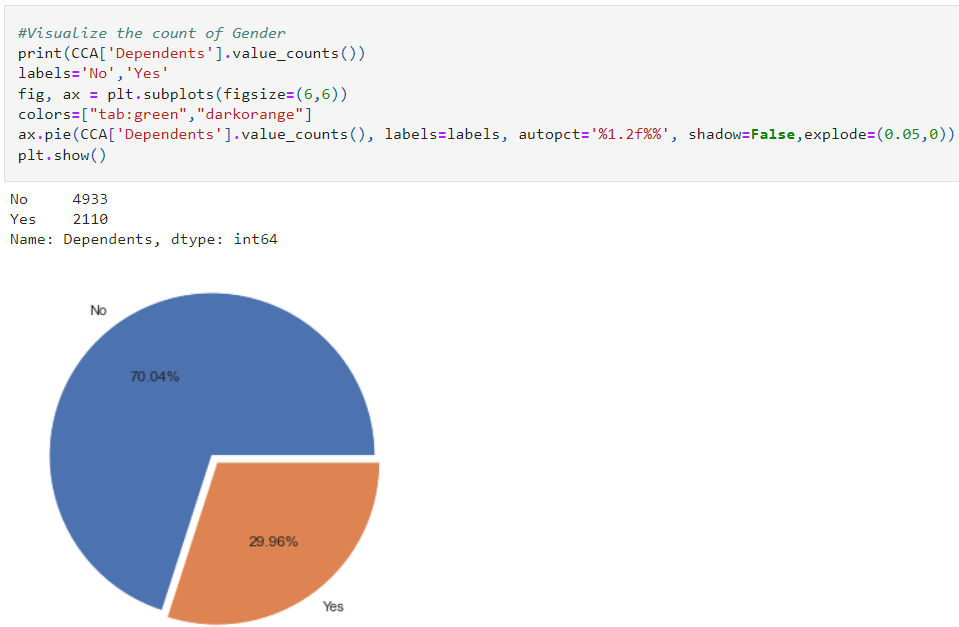
From the count plot we can observe that the count of "No Churn" are high compared to the count of "Yes Churn". That is there are more number of customers who have not churned. This leads to class imbalance issue in the data, we will rectify it by using oversampling method in later part.

* Count plot for gender and partner:

****

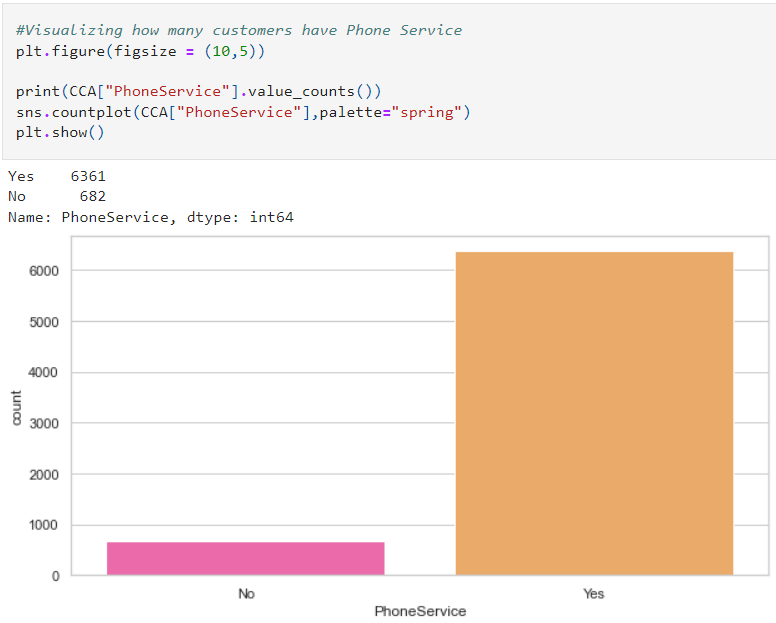
Both the genders has same count almost and having partner or not also has same count. In senior citizen we see low number in senior citizen

* **Count of Gender:**

****

The customers who have dependents are very less in counts that means they do not have anyone dependent on them. Here around 70% of customers have dependents only 29.96% have no dependents.

* **Customers having phone services:**

****

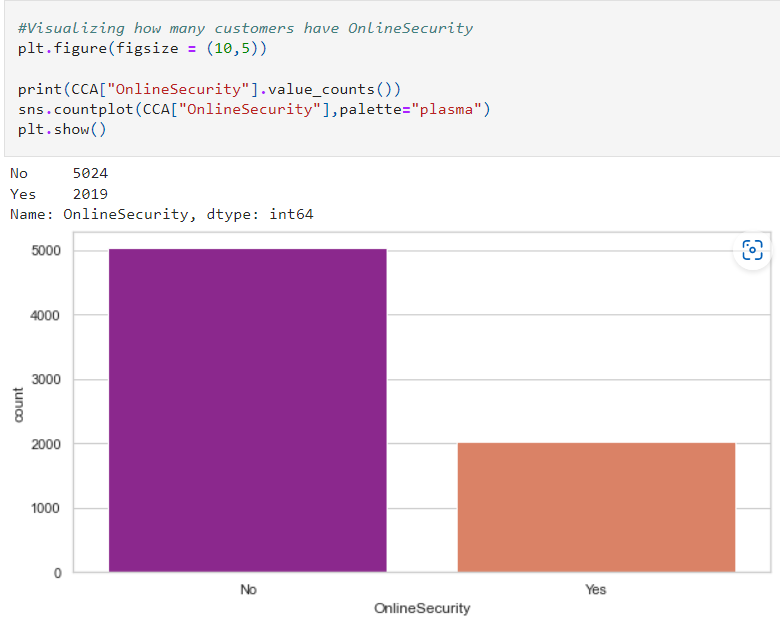
The customers who have phone services are large in numbers and who do not own phone services are very less in number.

* **Customers having internet services:**

****

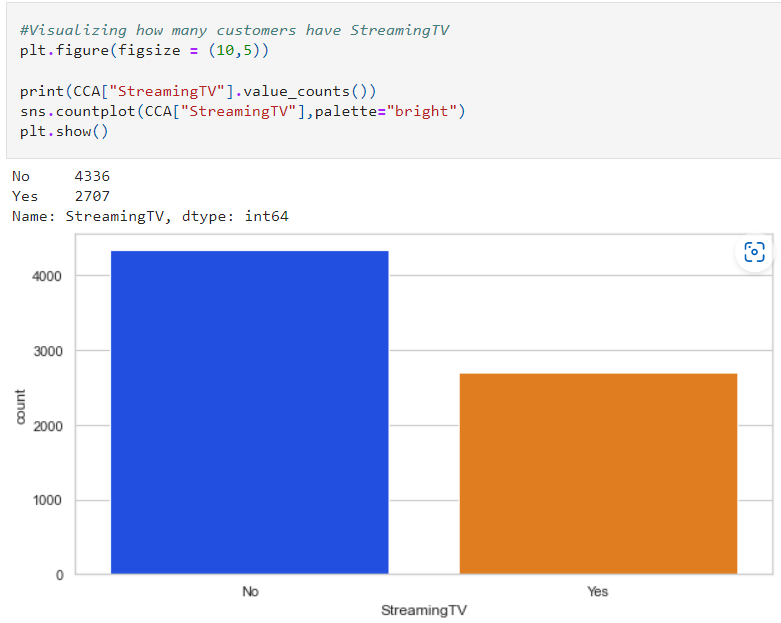
Most of the customers have chosen to get Fiber optic internet followed by DSL, but there are many customers who do not get an internet service.

* **Customers having online security :**



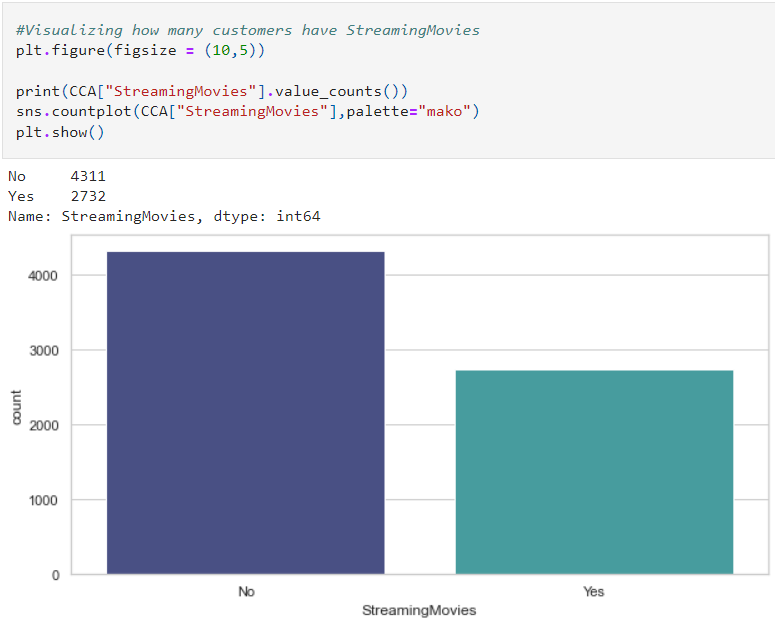
In the count plot we can observe the majority of customers who have internet services but they do not use any online security.

* Customers having streaming TVs:



It is clearly showing that most customers do not have tv streaming

* Customers streaming movies:



It is clearly showing that most customers do not have Streaming Movies.

* Customers having contracts:



Most of the customers prefer Month to Month contract compared to 1 year and 2 years contract.

* Payment method by customers:



Most of the customers prefer paperless billing and average number of customers who do not prefer paper less billing, they may like to receive paper billing.

# Checking the distribution:

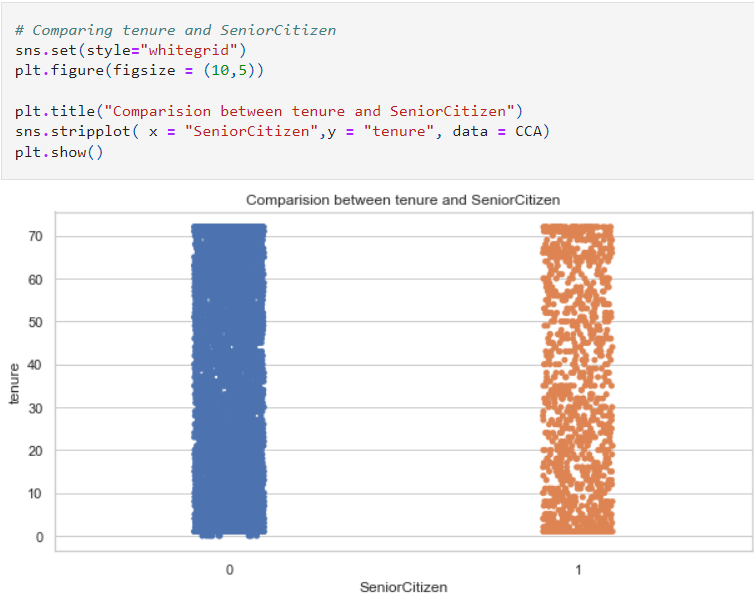
# 

# From the above distribution plots we can notice that the data almost looks normal in all the columns except SeniorCitizen. And the data in the column TotalCharges is skewed to right. Other two columns tenure and MonthlyCharges do not have skewness.

# Bivariate Analysis

Let us visualize the relationship between the attributes and target variables.

* **Tenure and senior citizens:**



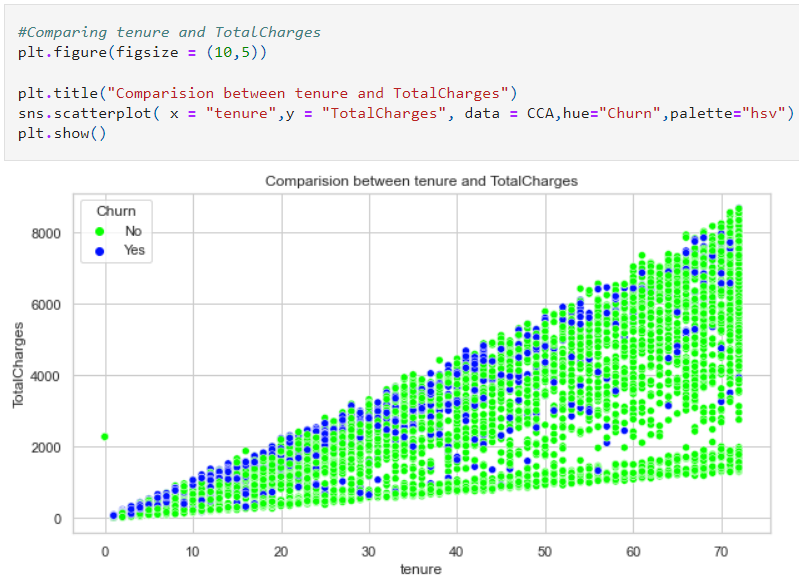
There is no significant difference between the features, here both the feature are in equal length.

* **Monthly charges and gender:**



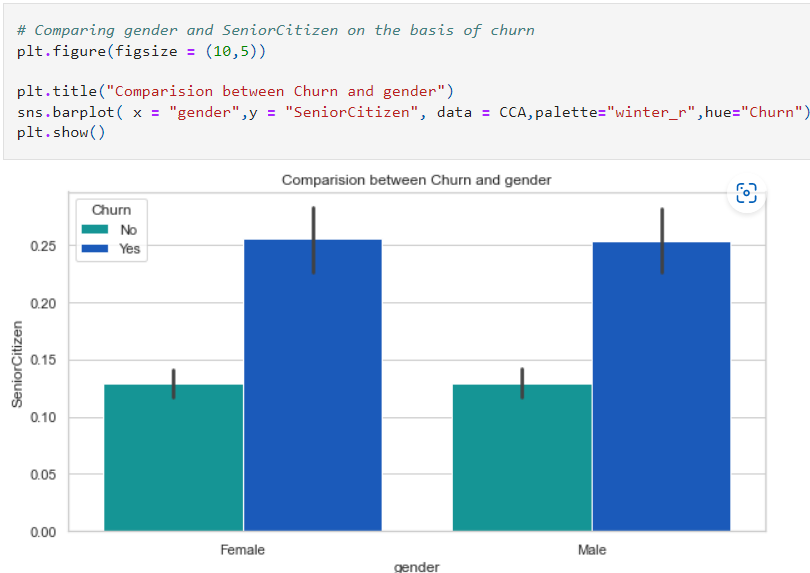
Most both male and female customers with monthly charges above 60 have high chances of getting churned.

* **Tenure and total charges**



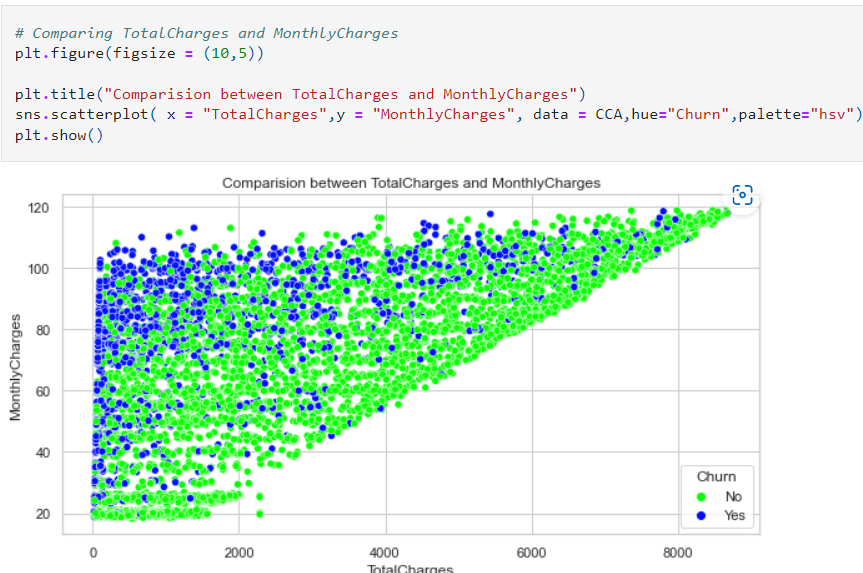
Here we can notice the strong linear relation between the features. As the tenure increases, TotalCharges also increases rapidly. If the customers have low tenure services then there is high chance of churn.

* **Gender and Senior citizen:**

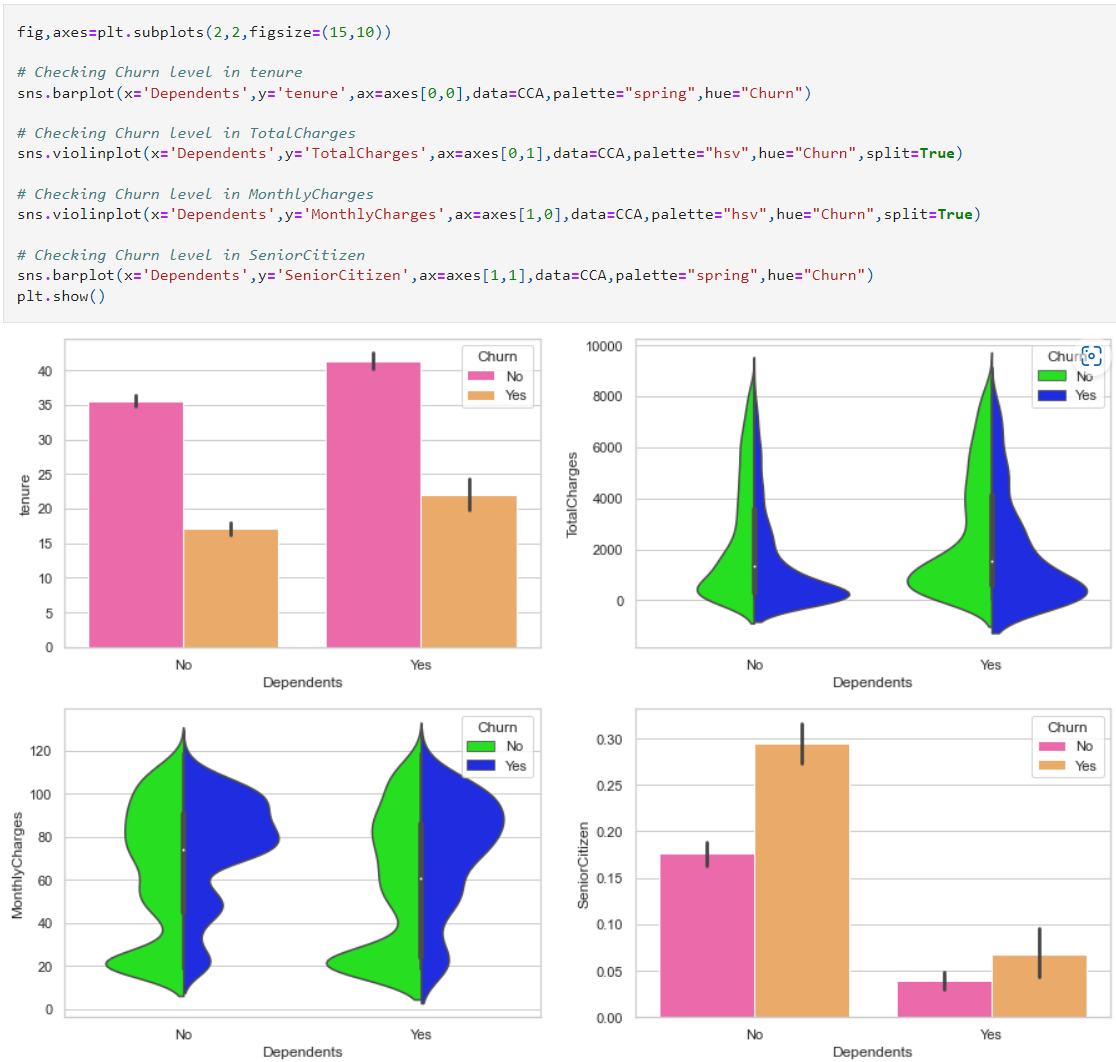


There is no significant difference between the columns. The customer's churns remains unaffected in gender and SeniorCitizen case.

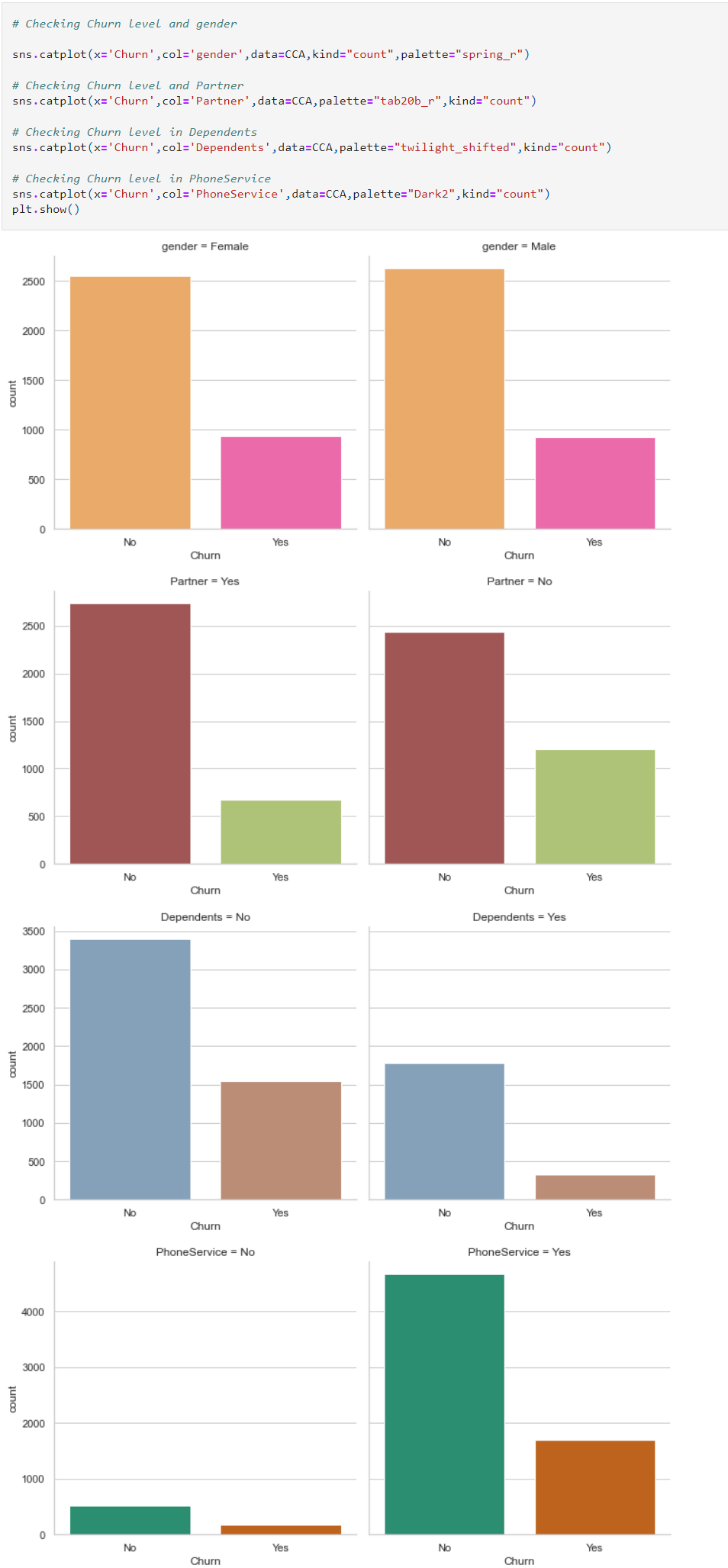
* **Total charges and Monthly charges:**



There is a linear relation between the features. The customers with high monthly charges have high tendancy to stop the services since they have high total charges. Also the if the customers ready to contribute with the monthly charges then there is an increment in the total charges.

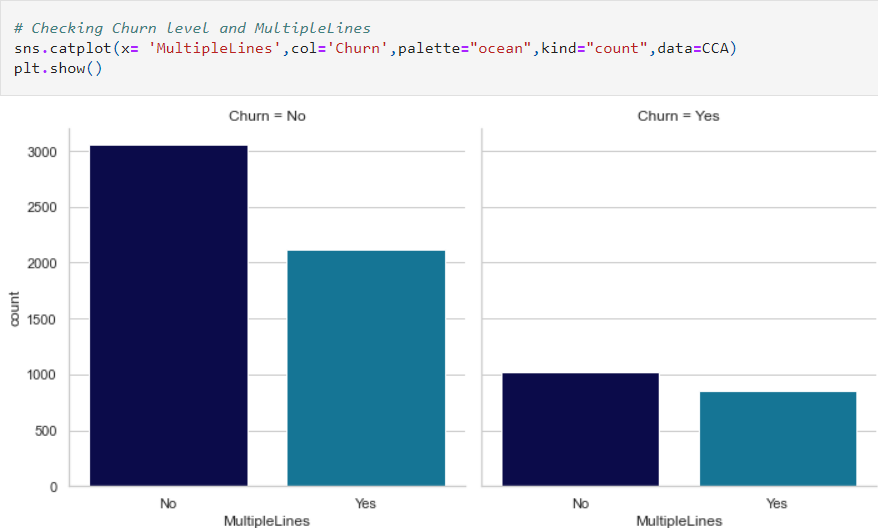


* Observations:
* The customers who have dependents with high tenure, then the churned level is high80-110.
* The customers who have total charges in the range of 0-2000 with dependents then the chance of getting churned is high.
* The customers having Monthly charges between 80-110 with dependents have high churn rate and when the customers have no dependents and having monthly charges around 20 then teh ratio of churn is very high.
* If the customer is a senior citizen and have no dependents then there is a tendancy of grtting churned.



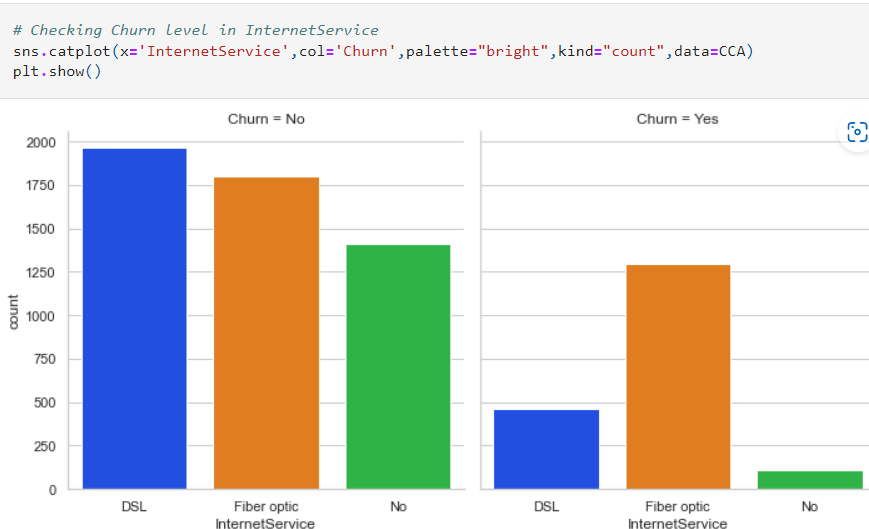
**Observations:**

* 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.
* Level and multiple lines:



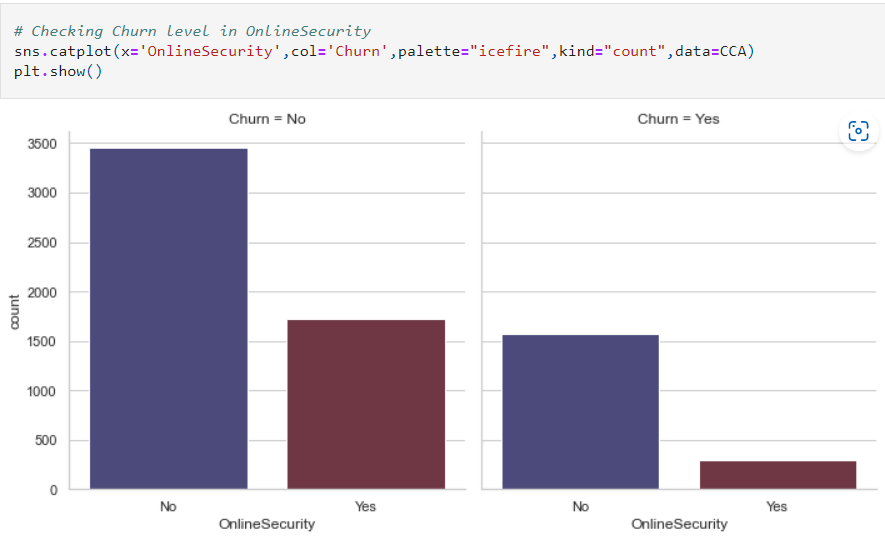
The customers who have phone services from single line have high churn rate compared to the customers having phone services from multiple lines, also there are very less number of customers who do not have phone services.

* **Level in internet service:**



The ratio of churn is high when the customers prefer Fiber optic internet services compared to other services, may be this type of service is bad and and need to be focused on. And the customers who own DSL service they have very less churn rate

* **Level of Online security:**



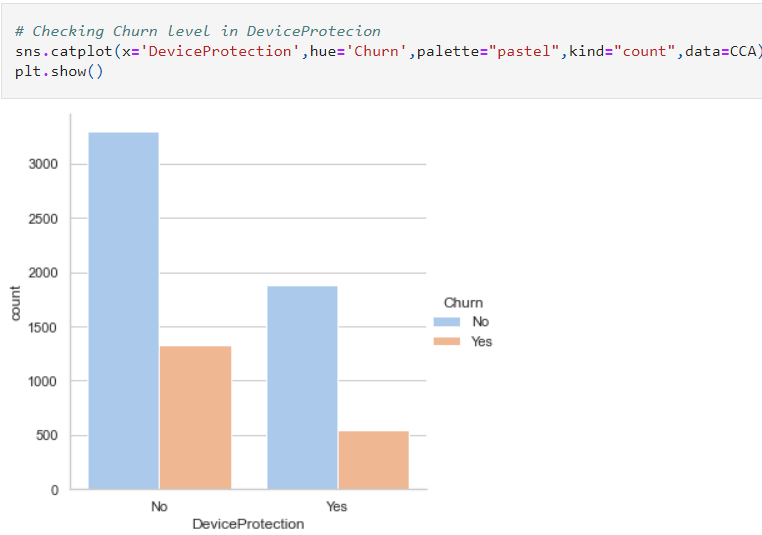
The customers who have no internet service have very less churn rate and the customers who do not have online security services have high tendency to getting churned.

* **Level in Online backup:**



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.

* **Level in device protection:**



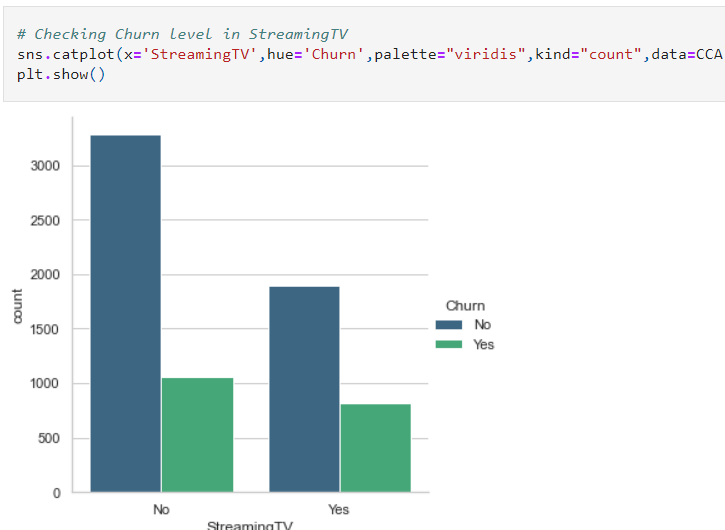
The customers who do not own any Device protection have very high churn rate compared to others.

* **Level in Tech Support:**



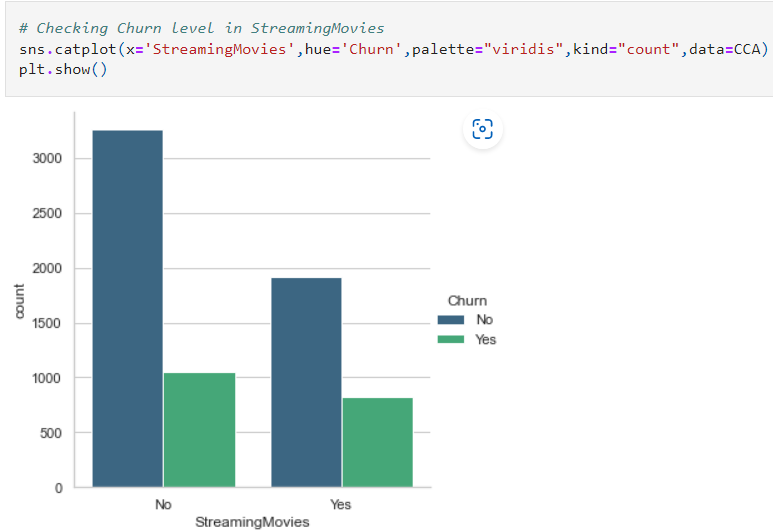
Here we can clearly see that the customers who do not have any techsupport then they have high churn ratio

* **Streaming TV**



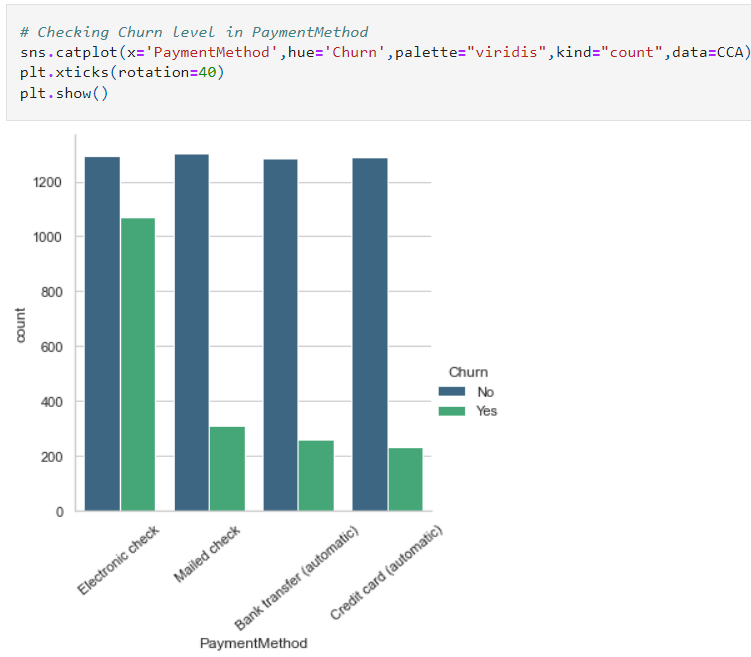
The churn rate is nearly same if the customer own StreamingTV or not.

* **Streaming Movies:**

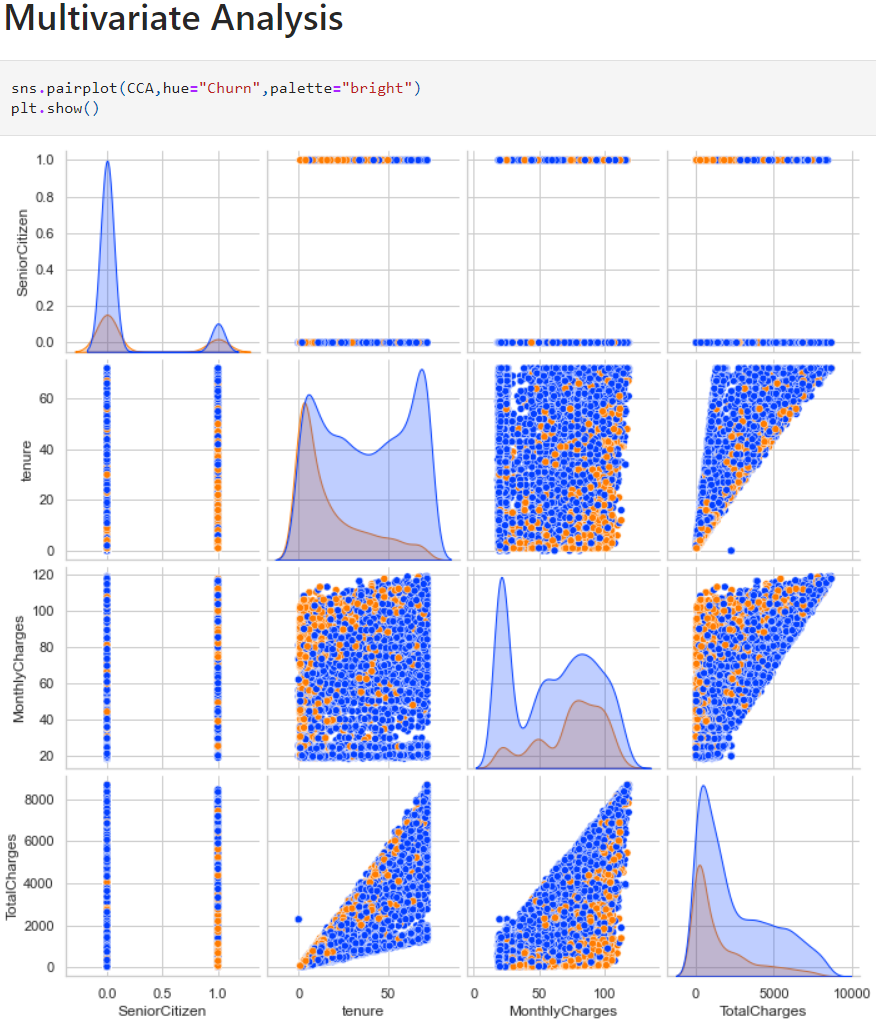


The customers who are existing in the company they do not own StreamingMovies in their devices. And the churn rate is low when the customer do not have internet services.

* **Payment method:**

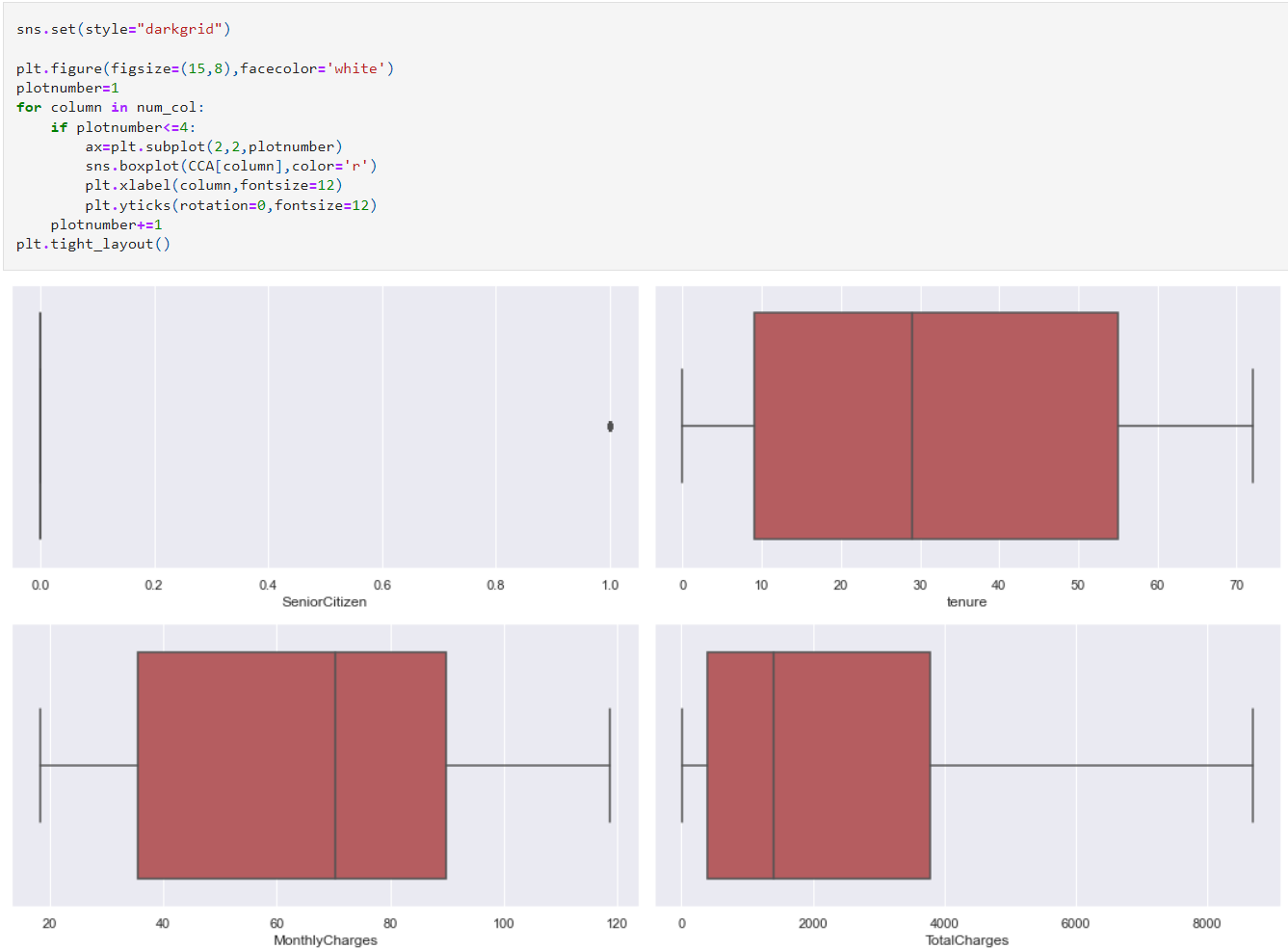


The customers who prefer Electronic check have high churn rate also the customers who existing in the company uses equal payment method.

****

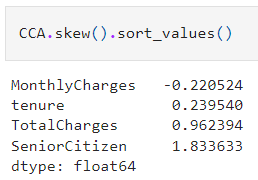
* **Observation:**
* The pairplot gives the pairwise relation between the features on the basis of the target "Churn". On the diagonal we can notice the distribution plots.
* The features tenure and TotalCharges, Monthlycharges and TotalCharges have strong linear relation with each other.
* There are no outliers in any of the columns but let's plot box plot to identify the outliers.
* **Let us check for Outliers:**

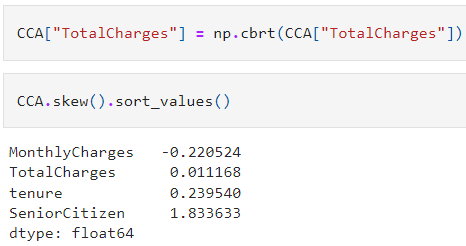
There are two graphical techniques for identifying outliers, scatter plots and box plots. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.

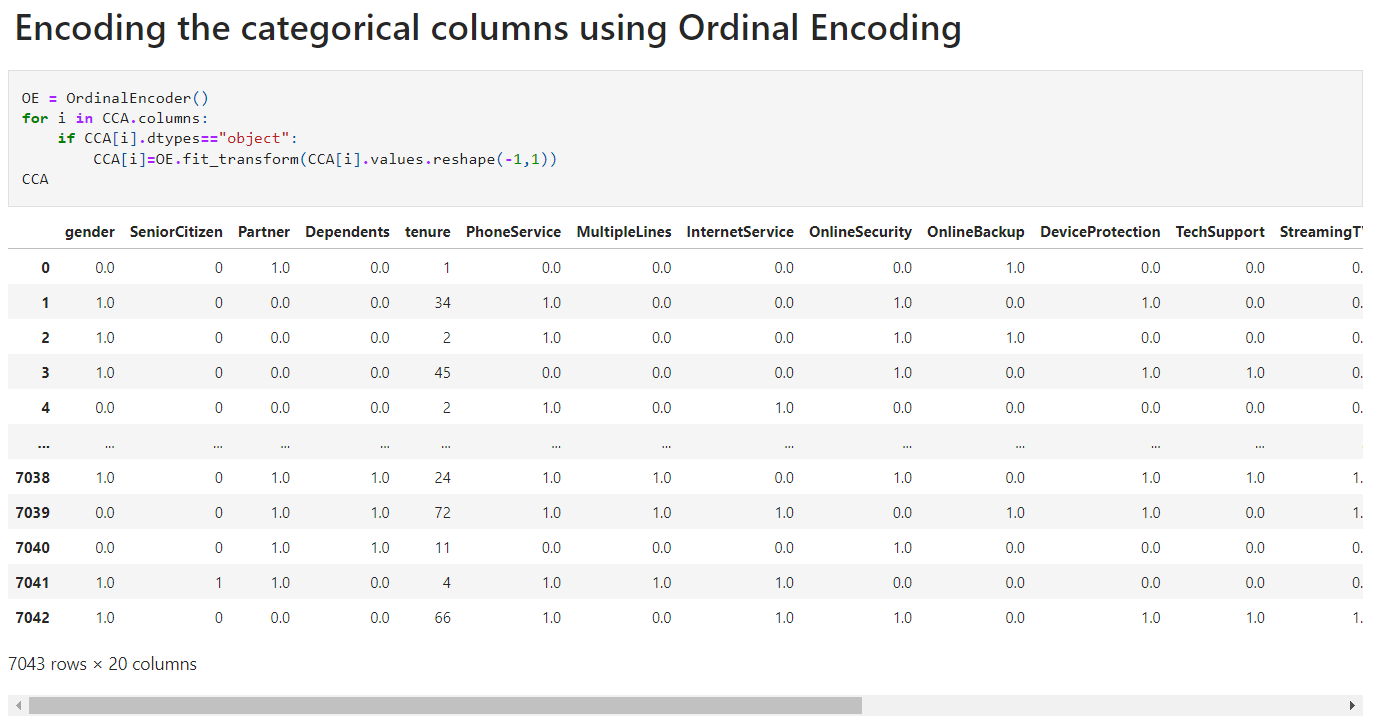
****

The column Senior Citizen has outliers but it contains categorical data so no need to remove outliers. Apart from this none of the columns have outliers.

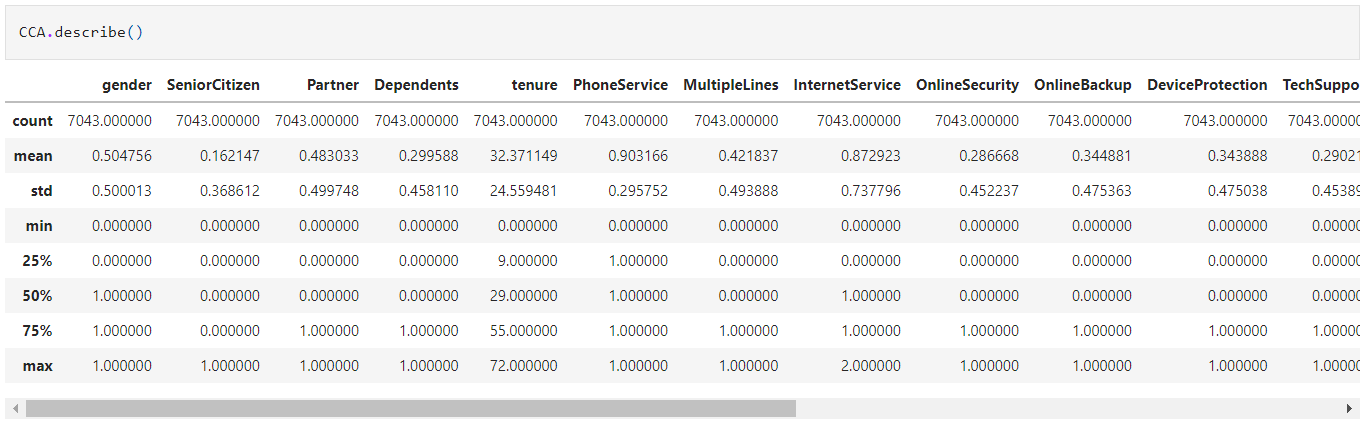
* Checking and Removing Skewness

 The columns SeniorCitizen and TotalCharges have skewness in the data. Since SeniorCitizen is categorical no need to remove skewness but in TotalCharges. Since TotalCharges is continuous in nature, lets use cube root method to remove skewness.



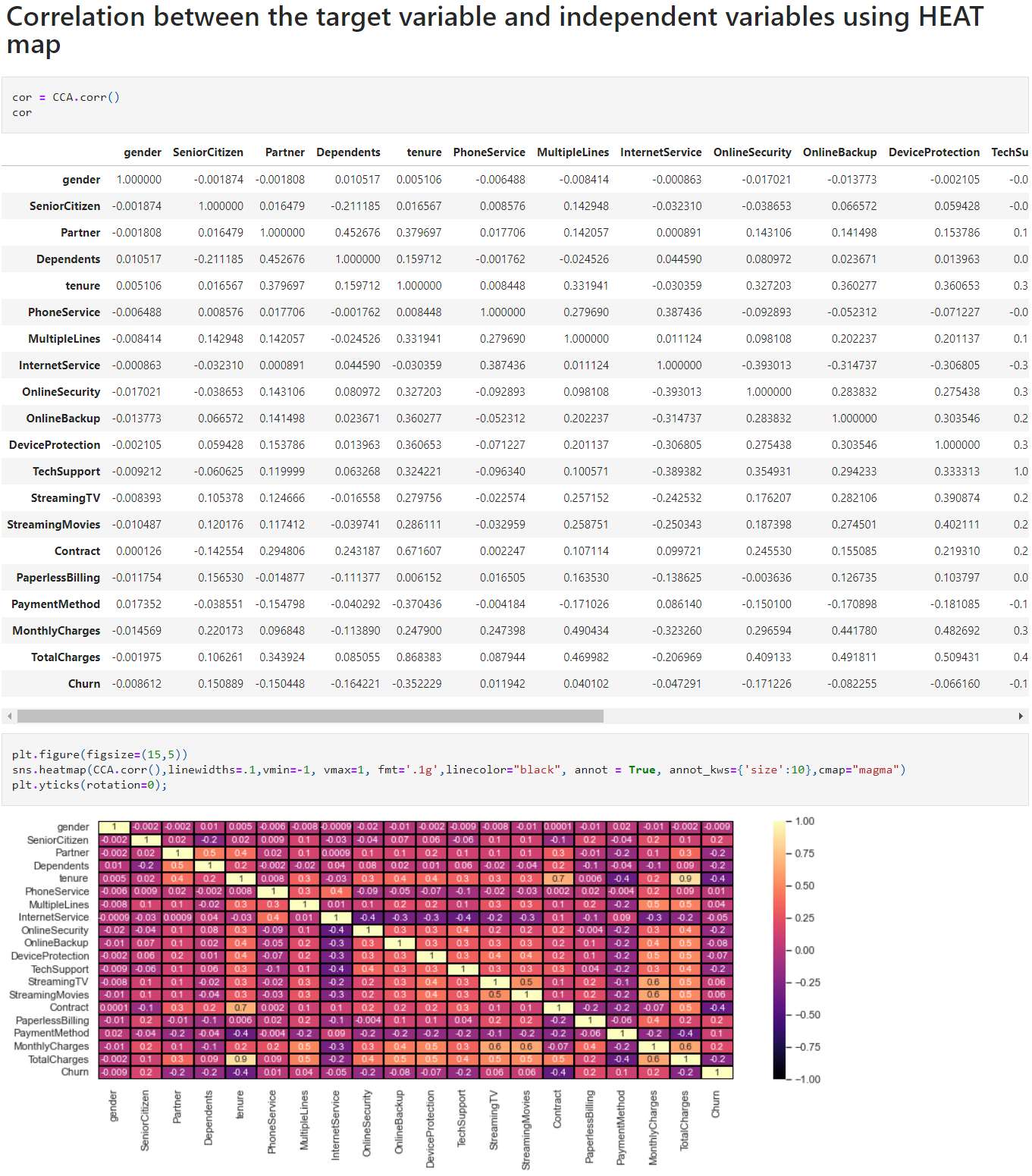


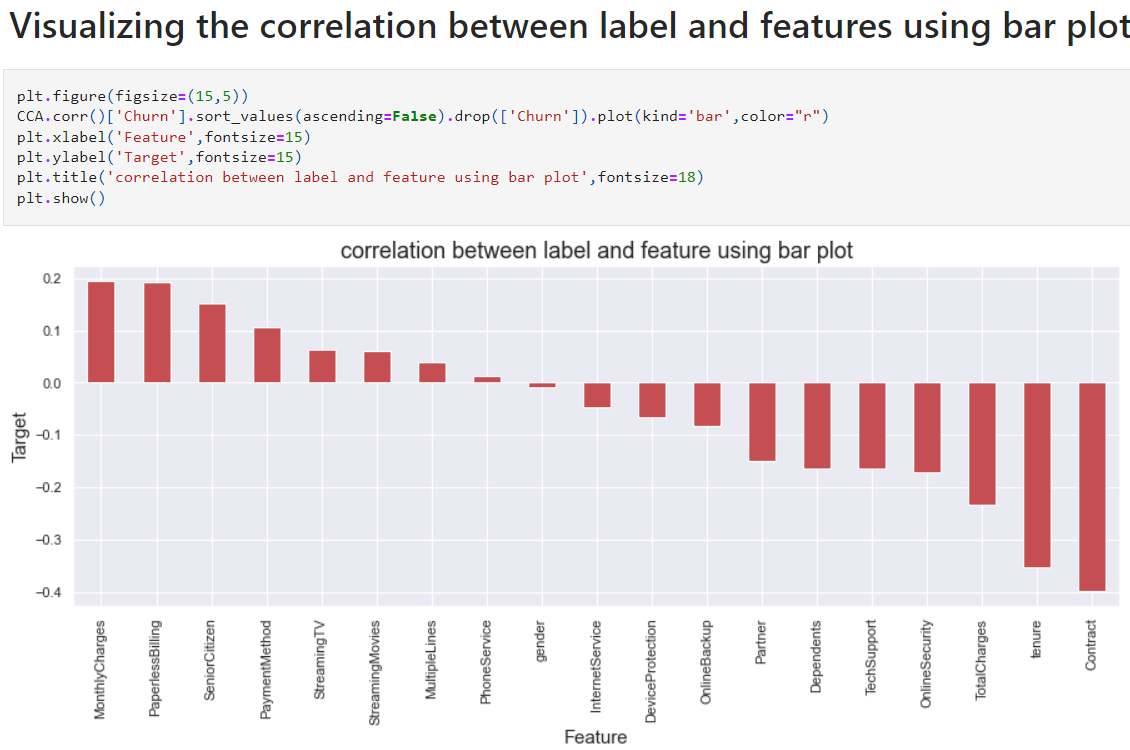
Now we have converted the categorical columns into numerical columns using Ordinal Encoding method.



After encoding the categorical column we are able to see all the columns details here. The counts of all the columns are same that means no null values in the dataset. This describe method descrbes the count, mean, standard deviation, min, IQR and max values of all the columns.

* **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 mapcontains 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 collrelation 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.
* The columns TotalCharges and tenure, Contract and tenure, TotalCharges and MonthlyCharges and many other columns have high correlation with each other. This leads to multicolllinearity issue, to overcome with this problem we will check VIF values and then we will drop the columns having VIF above 10.

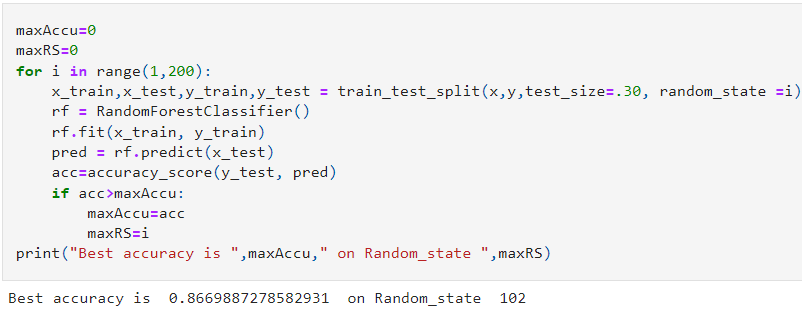




From the above bar plot we can notice the positive and negative correlation between the features and the target. Here the features gender and PhoneService have very less correlation with the column.

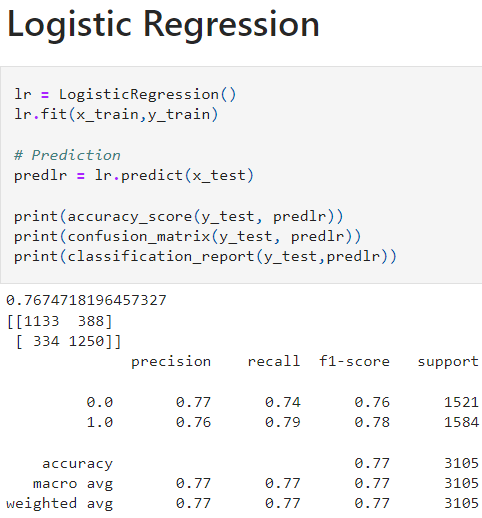
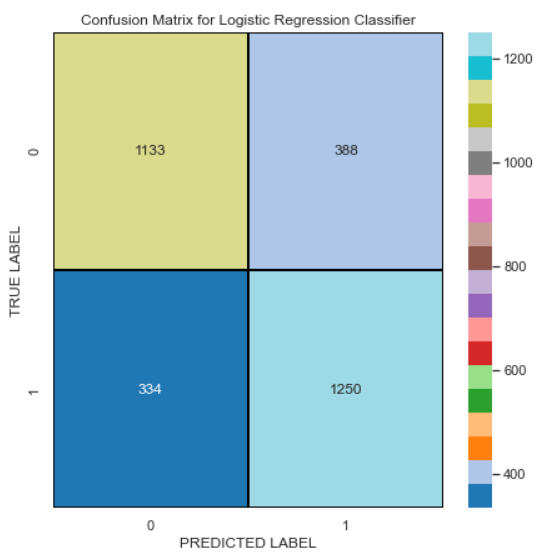
**Building Machine Learning Models:**

We find the best random state to split the data into training set and test set, so that models give the best score.

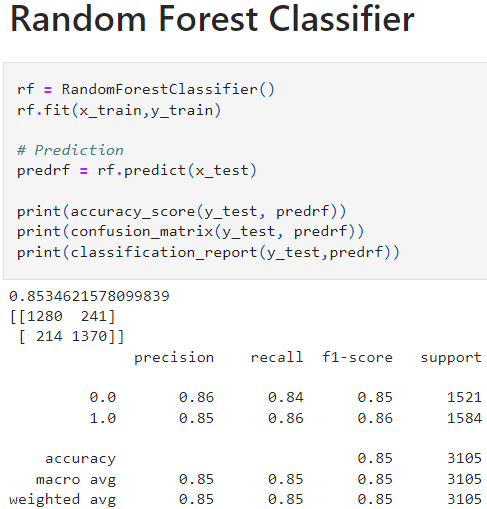
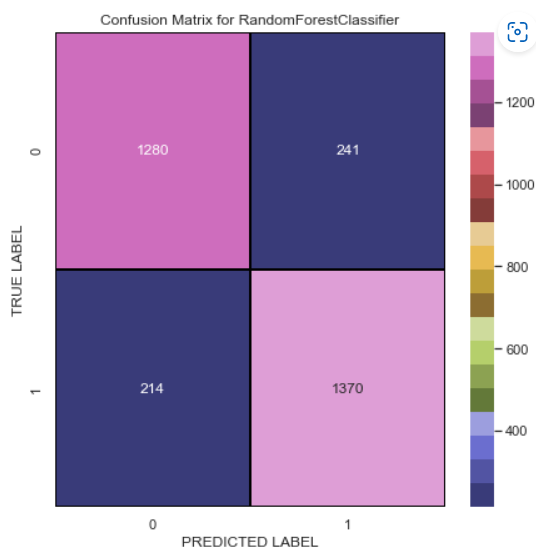


We have got the best random state and maximum accuracy.

We are using Classification machine learning algorithms like , Logistic Regression, Random Forest Classifier, Support Vector Machine Classifier, AdaBoost Classifier, Gradient Boosting Classifier, Naive Bias Classifier, Extra Trees Classifier.

**We have got an accuracy score of 78.67% with the base line model**

**We have got the accuracy score of 86.40% with Random Forest Classifier**

# 

# We have got an accuracy score of 81.73% with the SVC model

# 

# We have got an accuracy score of 82.80% with the AdaBoost Classifier

# 

# We have got an accuracy score of 85.24% with the Gradient Boosting Classifier

# 

# We have got an accuracy score of 78.61% with GaussianNB

# 

# We have got an accuracy score of 86.66% with the extra trees classifier model

# 

# 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.

# 

# 

# This is the ROC curve of the best model which is the extra trees classifier (AUC = 0.93)

# 

# The model is now saved.

* **Conclusion:**

Telecommunication industry has suffered from high churn rates and enormous churning loss. Good methods need to be developed and existing methods must be enhanced to prevent the telecommunication industry to face challenges. In this article we discussed the various prediction models and compared the quality measures of prediction models. We found that the accuracy achieved with SVM Classifier is far much higher than the logistic regression technique which clearly states that decision tree is an efficient technique.