**Customer Churn Analysis**

**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.

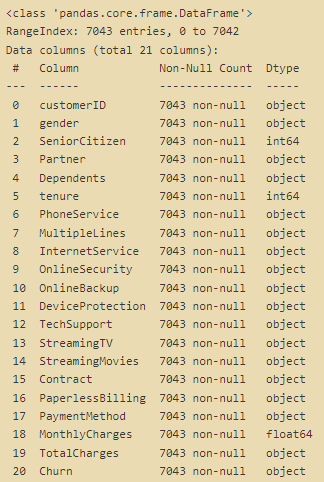
A high churn rate will adversely affect a company’s profits and impede growth. Our churn prediction would be able to provide clarity to the company on how well it is retaining its existing customers and understand what are the underlying reasons that are causing existing customers to terminate their contract .

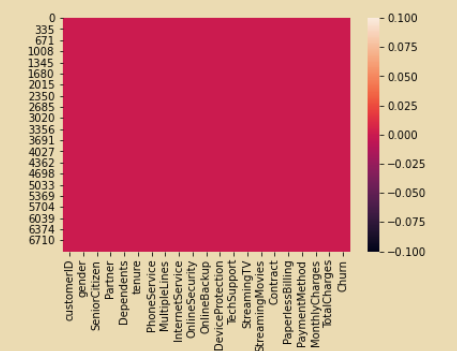
Customer retention can be achieved with good customer service and products. But the most effective way for a company to prevent attrition of customers is to truly know them. The vast volumes of data collected about customers can be used to build churn prediction models. Knowing who is most likely to 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.

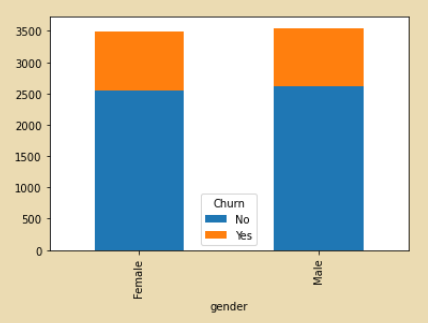
**Data Analysis:**

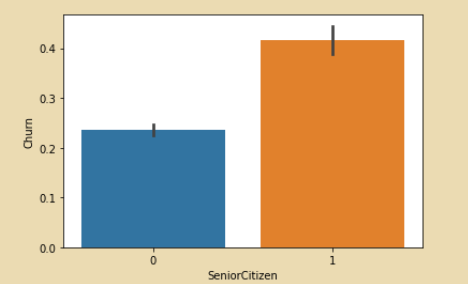
As I read the data using Pandas in Python, I found that there was no missing data from the raw data set and most of the features were categorical. Where I found that the TotalCharges column was of object type.

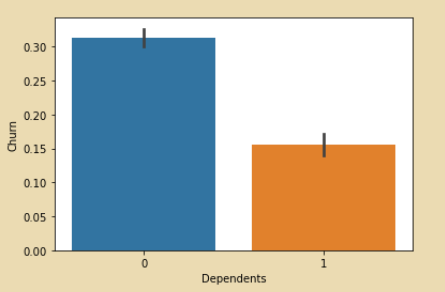
****



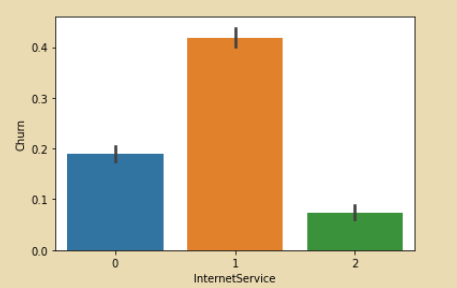
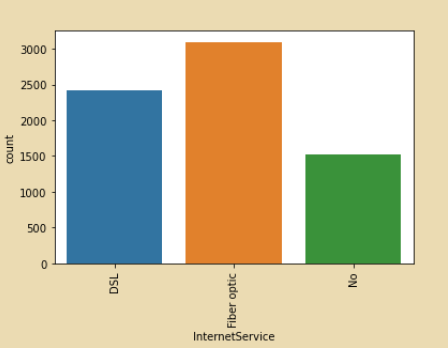
No null values.



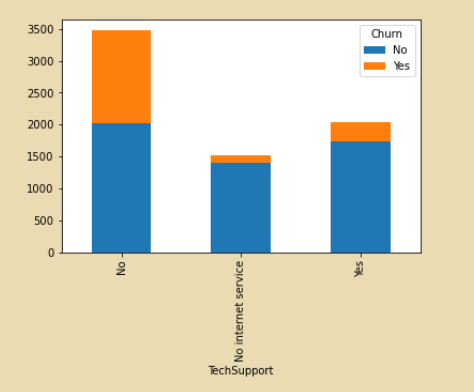
* Female and male tend to churn on an equal amount. Gender Equality!! 
* Churn is higher in case of Senior Citizen. We can say that, the customer who is a senior citizen will be more likely to churn for the service.



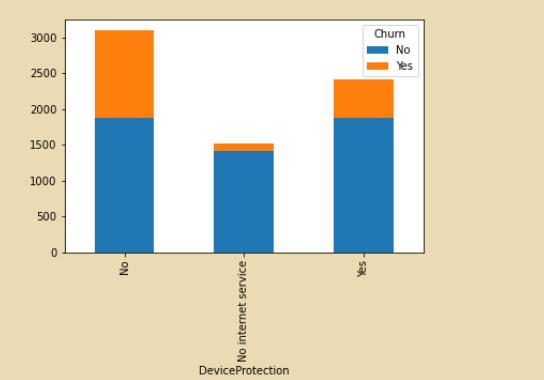
* The customer having no Dependents is more likely to churn for the service.



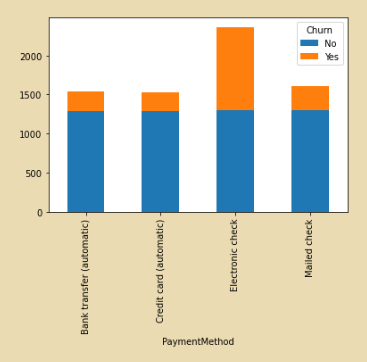
* Maximum number of people is using fibre optic as internet service and the people using fibre optic are more likely to churn.



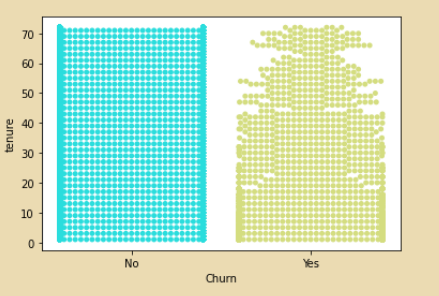
* People having no tech support are more likely to churn.



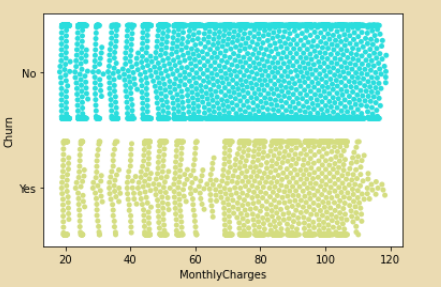
* People having no device protection are more likely to churn.



* In this graph it is clearly visible that the Electronic check mediums are the highest churners

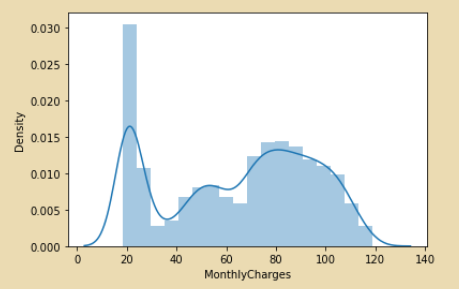
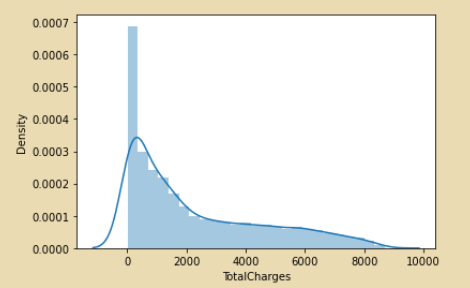


* Customers having shorter tenure are more likely to churn as compared to customer having longer tenure.



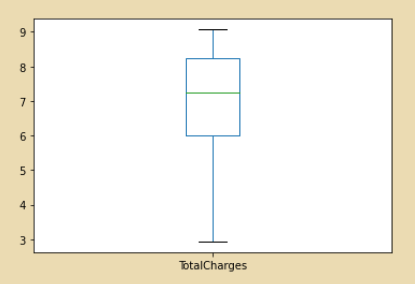
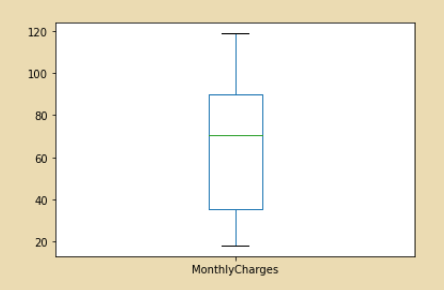
* Customers paying monthly charges in the mid-range are more likely to churn as compared with customer paying higher monthly charges.

After doing visualisation I checked skew-ness and outliers in the numerical columns.

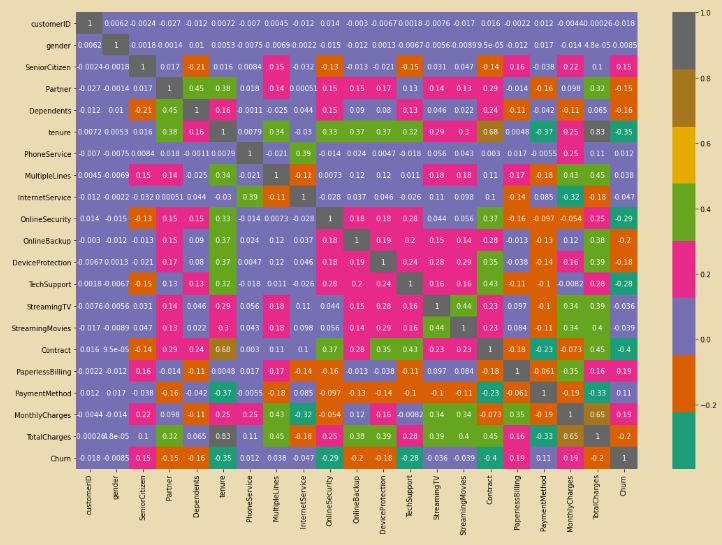
Monthly Charges -0.222103 Total Charges 0.961642

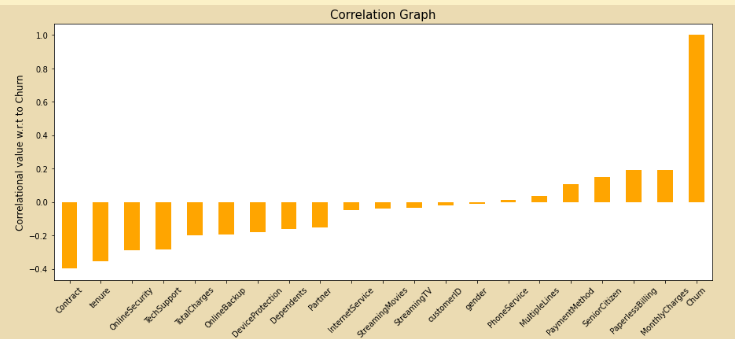
* We can see that there is skew-ness present the columns.

* There are no outliers in the columns.

After checking Outliers and Skew-ness I have encoded the categorical columns to numerical for building model. Correlation for the same is:

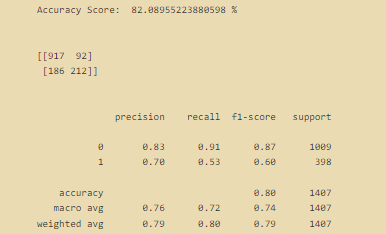




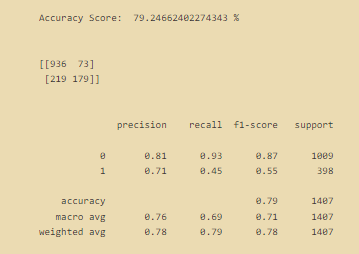
**Building Machine Learning Model:**

As our target variable is categorical in nature ,the problem is of classification type. I have chosen 5 classification models to build a predictive model:

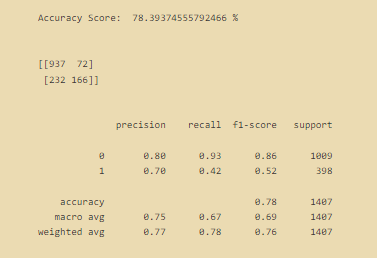
1. Logistic Regression



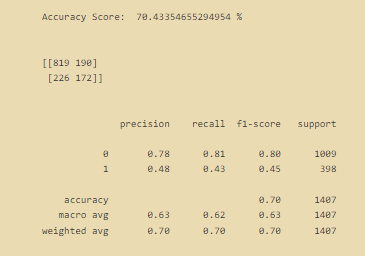
1. Random Forest Classifier



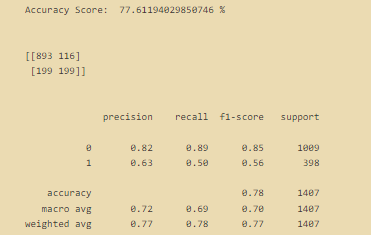
1. Support Vector Classifier



1. Decision Tree Classifier



1. K Neighbors Classifier

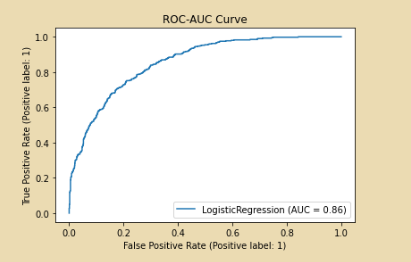


I did cross validation score for two models:

* Logistic regression-> 
* Random Forest classifier-> 

After performing hyper parameter tuning of the best performing model so far(Logistic Regression). I did plotting of ROC AUC Curve and after plotting I saved the model.





**Conclusion:**

During our analysis, we managed to identify key factors which led potential customers to churn using descriptive analysis.  
We proposed a Logistic regression model with an accuracy of 80.6%.