Report

The dataset:

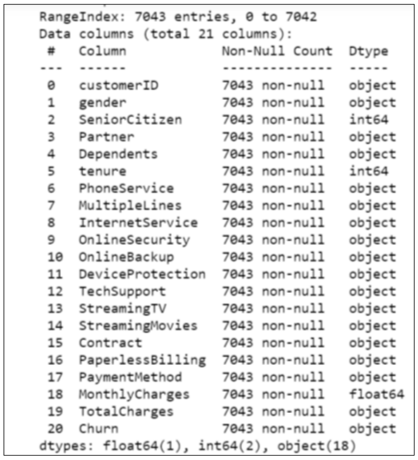
The raw dataset contains 7043 entries. All entries have several features and a column stating if the customer has churned or not.
To better understand the data we will first load it into pandas and explore it with the help of some very basic commands.

df.info(): gives us detailed information about every column. We can see that our data is divided into three types;

- **Object:** Object format means variables are categorical. Categorical variables in our dataset are: customerID, gender, partner, dependents, phone service, multiple lines, internet service, online security, online backup, device protection, tech support, streaming tv, streaming movies, contract, paperless billing, payment method, total charges, and churn.

- **int64**: It represents the integer variables. Senior citizen and tenure are of this format.

- **float64:** It represents the variables which have some decimal values involved. They are also numerical variables. There is only one variable with this format in our dataset which is monthly charges.

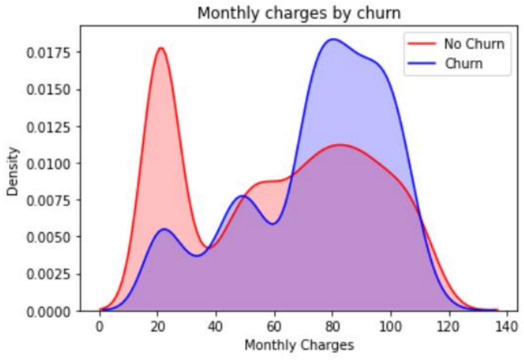


```
RangeIndex: 7043 entries, 0 to 7042
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        7043 non-null   object 
 1   gender            7043 non-null   object 
 2   SeniorCitizen     7043 non-null   int64  
 3   Partner           7043 non-null   object 
 4   Dependents        7043 non-null   object 
 5   tenure            7043 non-null   int64  
 6   PhoneService      7043 non-null   object 
 7   MultipleLines     7043 non-null   object 
 8   InternetService   7043 non-null   object 
 9   OnlineSecurity    7043 non-null   object 
 10  OnlineBackup      7043 non-null   object 
 11  DeviceProtection  7043 non-null   object 
 12  TechSupport       7043 non-null   object 
 13  StreamingTV       7043 non-null   object 
 14  StreamingMovies   7043 non-null   object 
 15  Contract          7043 non-null   object 
 16  PaperlessBilling  7043 non-null   object 
 17  PaymentMethod     7043 non-null   object 
 18  MonthlyCharges    7043 non-null   float64
 19  TotalCharges      7043 non-null   object 
 20  Churn             7043 non-null   object 
dtypes: float64(1), int64(2), object(18)
```

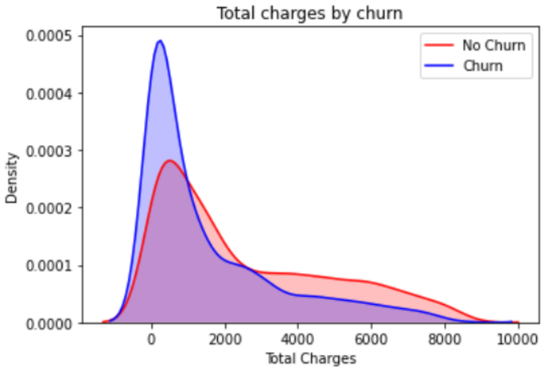
Findings:

1) (SeniorCitizen) is actually a categorical hence the 25%-50%-75% distribution is not proper 75% customers have tenure less than 55 months. Average Monthly charges are USD 64.76 whereas 25% customers pay more than USD 89.85 per month.

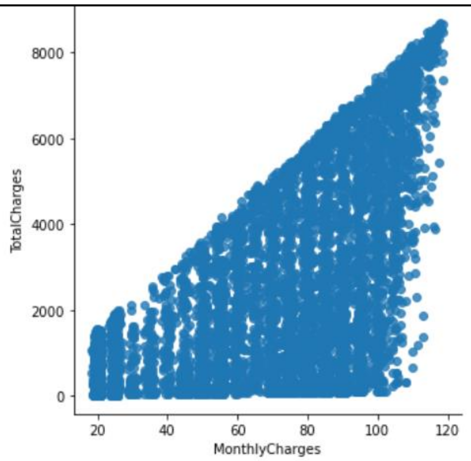
2) In data exploration we determined how each predictor variable is compared with the target variable (Churn).



Insight: Churn is high when Monthly Charges are high

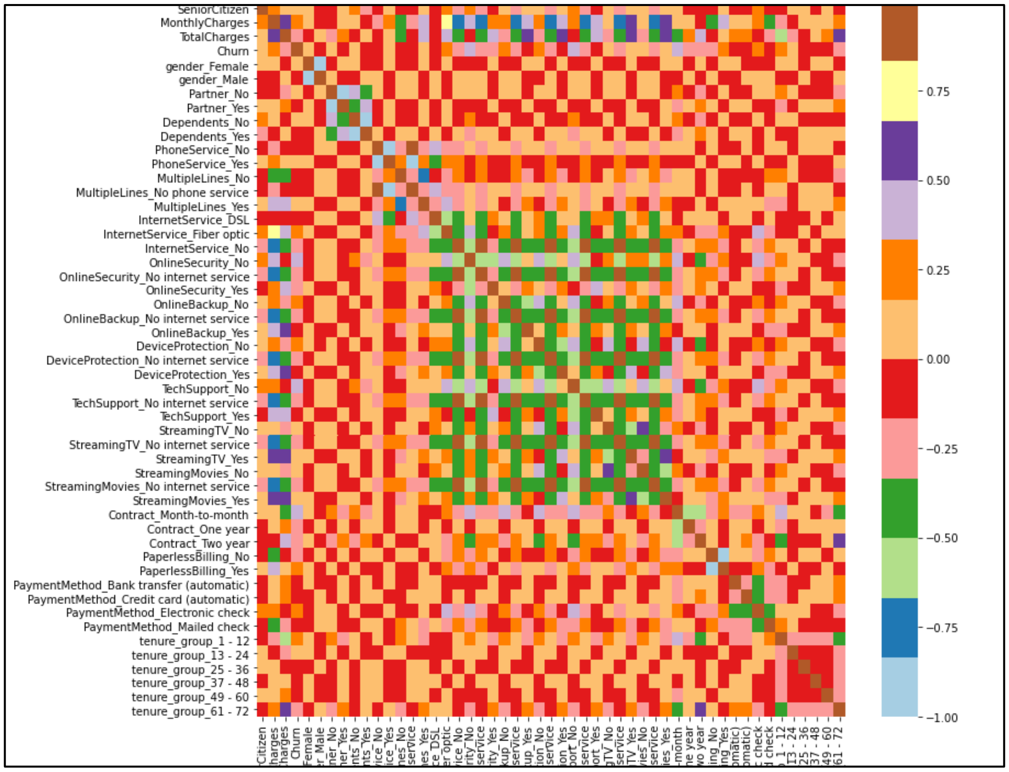


Surprising insight as higher Churn at lower Total Charges

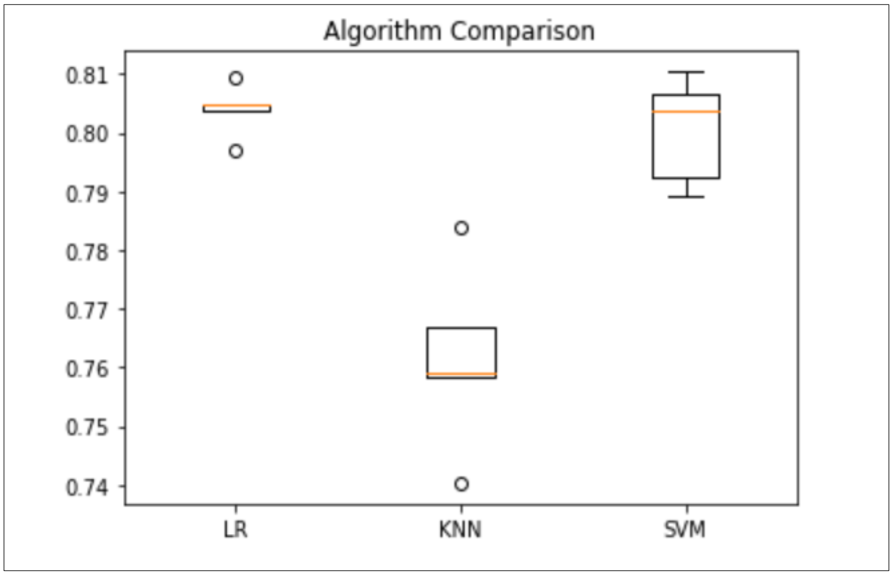


Total Charges increase as Monthly Charges increase - as expected.

3) HIGH Churn seen in case of Month to month contracts, no online security, No Tech support, first year of subscription and Fiber Optics Internet LOW Churn is seen in case of Long-term contracts, Subscriptions without internet service and the customers engaged for 5+ years. Factors like Gender, Availability of Phone Service and # of multiple lines have almost NO impact on Churn. This is also evident from the Heatmap and correlation table.



4) As a result the model that performed the best is Logistic Regression as shown in the boxplot and that’s because of its high evaluation performance.



5) For the analysis, it can be observed that some variables have a positive relation to our predicted variable and some have a negative relation. Customers with negative values show that they are unlikely to churn while those with positive values shows they are likely to churn.