

Objective:

Customer Churn measures the loss of customers and service provider companies use this metric to understand the customer retention. The objective is to predict behavior to retain customers by analyzing all relevant customer data and develop focused customer retention programs.

Problem:

Is the customer going churn?

Outcome:

When a customer stops service or company losing customer is referred to as Customer Churn. This is an important measure for any service-based company. The model predictions can provide the propensity of churning and gives the companies with the feature's importance that leads the customer to churn. With the list of potential customers who are likely to churn, the marketing/retention teams can then take measure to reduce their churn probability. This project helps companies in identifying customer who are at risk of churning and we have used this IBM sample data set provided for a telecom company. We will be using statistical analysis to understand variables that are associated with customer churn.

Dataset:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Data set contains 7043 rows and 21 columns, see below for more information:

- *customerID*: Customer ID
- *genderCustomer*: gender (female, male)
- *SeniorCitizen*: Whether the customer is a senior citizen or not (1, 0)

- *PartnerWhether*: the customer has a partner or not (Yes, No)
- *Dependents*: Whether the customer has dependents or not (Yes, No)
- *tenure*: Number of months the customer has stayed with the company
- *PhoneService*: Whether the customer has a phone service or not (Yes, No)
- *MultipleLines*: Whether the customer has multiple lines or not (Yes, No, No phone service)
- *InternetService*: Customer's internet service provider (DSL, Fiber optic, No)
- *OnlineSecurity*: Whether the customer has online security or not (Yes, No, No internet service)
- *OnlineBackup*: Whether the customer has online backup or not (Yes, No, No internet service)
- *DeviceProtection*: Whether the customer has device protection or not (Yes, No, No internet service)
- *TechSupport*: Whether the customer has tech support or not (Yes, No, No internet service)
- *StreamingTV*: Whether the customer has streaming TV or not (Yes, No, No internet service)
- *StreamingMovies*: Whether the customer has streaming movies or not (Yes, No, No internet service)
- *Contract*: The contract term of the customer (Month-to-month, One year, Two year)
- *PaperlessBilling*: Whether the customer has paperless billing or not (Yes, No)
- *PaymentMethod*: The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- *MonthlyCharges*: The amount charged to the customer monthly
- *TotalCharges*: The total amount charged to the customer
- *Churn*: Whether the customer churned or not (Yes or No)

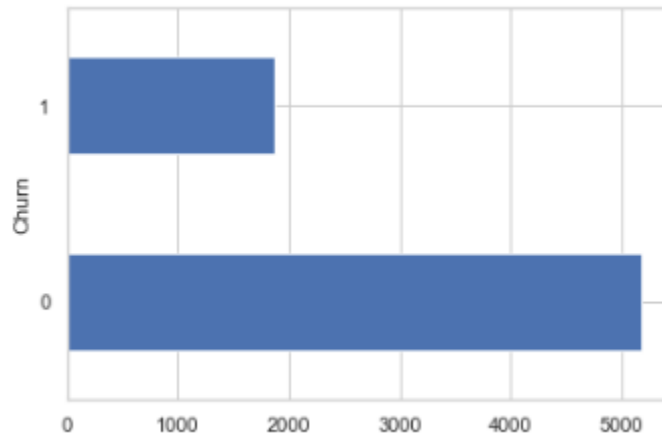
There are many categorical variables in this data set. The numerical features are Tenure, MonthlyCharges and TotalCharges.

Descriptive Stats:

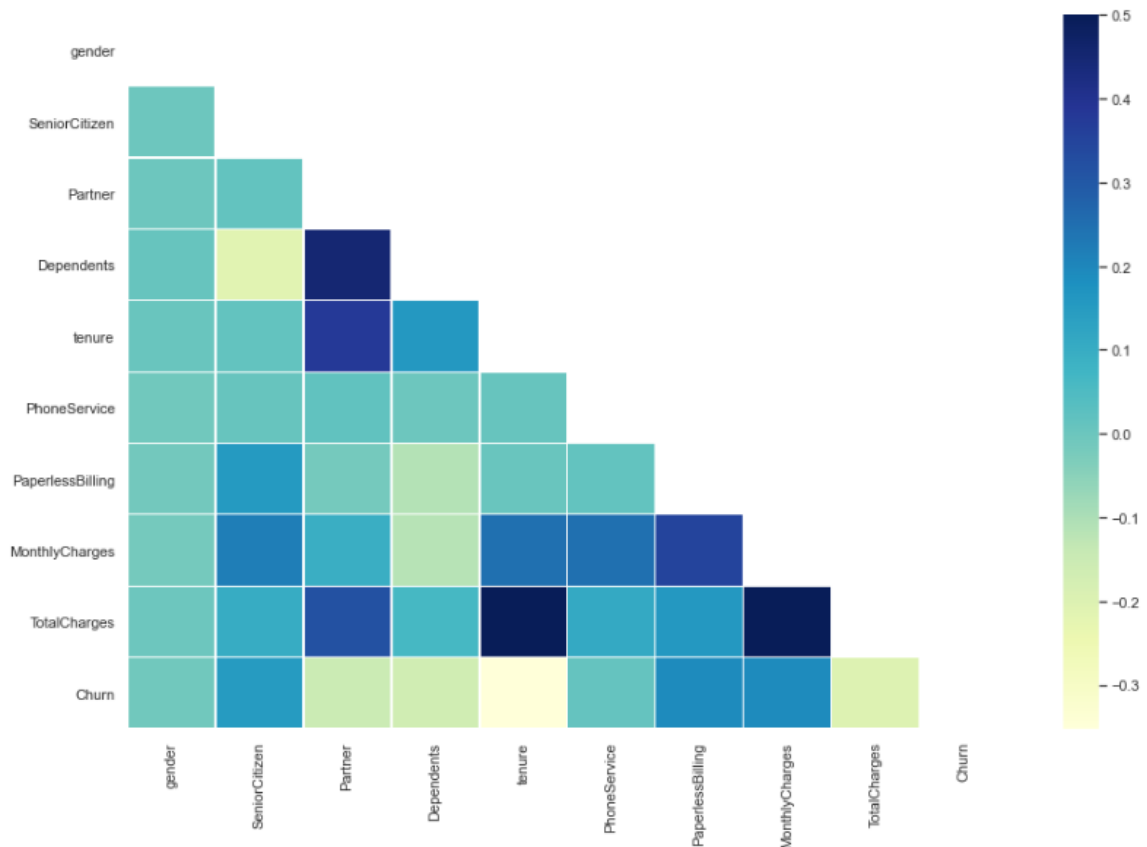
We see that Tenure ranges from 0 (new customer) to 6 years, Monthly charges range from \$18 to \$118, etc

tenure	MonthlyCharges
7043.000000	7043.000000
NaN	NaN
NaN	NaN
NaN	NaN
32.371149	64.761692
24.559481	30.090047
0.000000	18.250000
9.000000	35.500000
29.000000	70.350000
55.000000	89.850000
72.000000	118.750000

Roughly a quarter of the customers have churned in this data set.



Looking at the coloration matrix, there seems to be some positive correlation between Monthly Charges and Churn and some negative correlation between tenure and Churn.



Data Wrangling:

Total Charges column was read as object, to fix this, we are converting it's data type.

```
#Total charges is object in the data types converting to number
totalCharges = df.columns.get_loc("TotalCharges")
new_col = pd.to_numeric(df.iloc[:, totalCharges], errors='coerce')
df.iloc[:, totalCharges] = pd.Series(new_col)
```

Handling missing values:

```

1 print(df.isnull().values.any())
  df.isnull().sum()
  |
  # Looks Like Total charges has missing values

True

: customerID      0
  gender          0
  SeniorCitizen   0
  Partner         0
  Dependents      0
  tenure          0
  PhoneService    0
  MultipleLines   0
  InternetService 0
  OnlineSecurity  0
  OnlineBackup    0
  DeviceProtection 0
  TechSupport     0
  StreamingTV     0
  StreamingMovies 0
  Contract        0
  PaperlessBilling 0
  PaymentMethod   0
  MonthlyCharges  0
  TotalCharges    11
  Churn           0
dtype: int64

```

TotalCharges column has missing values.

Applying imputation to fix the issue with missing values:

```

# Handle missing values for nan_column (TotalCharges)
from sklearn.impute import SimpleImputer

# Find the column number for TotalCharges (starting at 0).
total_charges_idx = df.columns.get_loc("TotalCharges")
imputer = SimpleImputer(missing_values=np.nan, strategy='mean')

df.iloc[:, total_charges_idx] = imputer.fit_transform(df.iloc[:, total_charges_idx].values.reshape(-1, 1))
df.iloc[:, total_charges_idx] = pd.Series(df.iloc[:, total_charges_idx])

```

Dropping non-essential features:

```

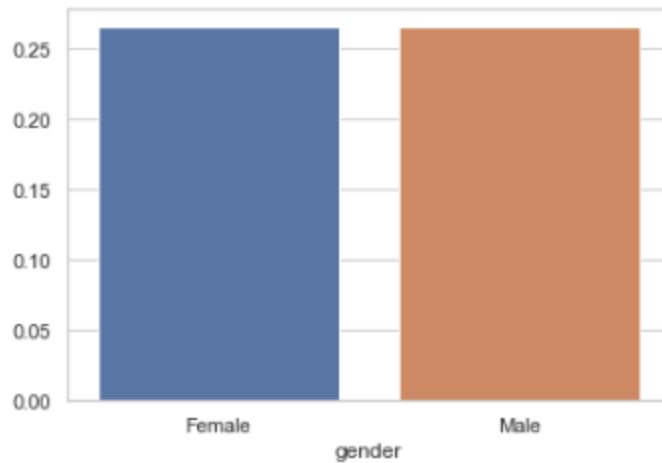
# don't need customerID for modelling

df = df.drop(['customerID'], axis=1)

```

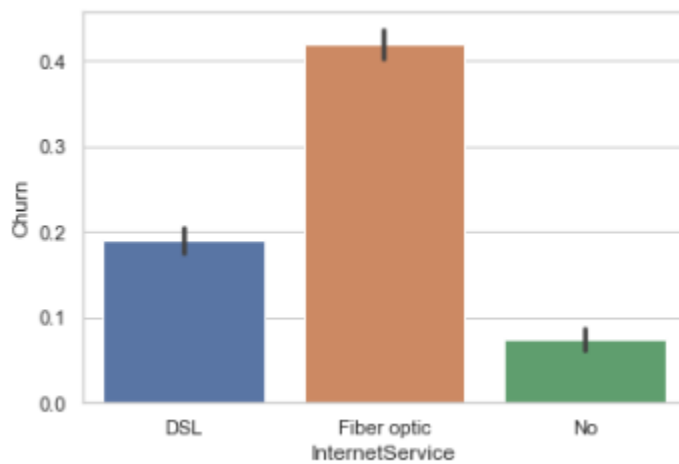
Data visualization:

Looking at the churn rate by gender there seems to be a very little higher % of female customer churning than compared to male customers, however, I don't think this is a big enough difference to be considered.

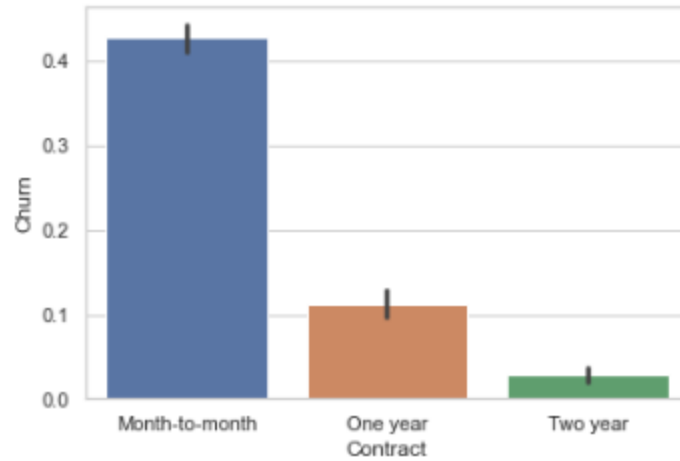


Looking at the churn rate by Internet service type there seems to be a high % of customers churning that are Fiber Optic service. This also could be due to Fiber optic being the most used service type.

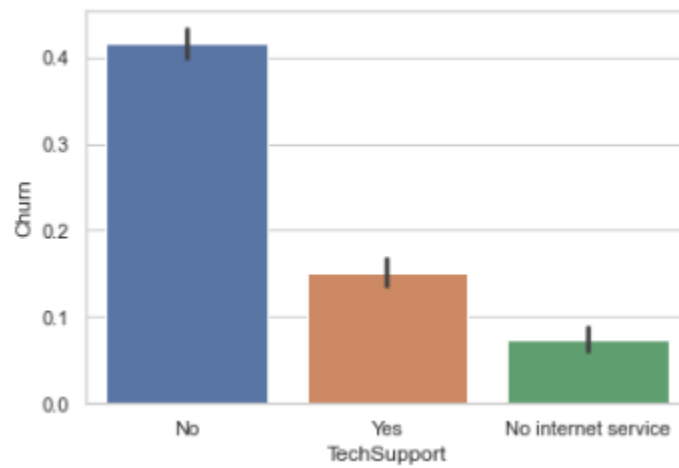
Fiber Optic customers are churning at a higher rate:



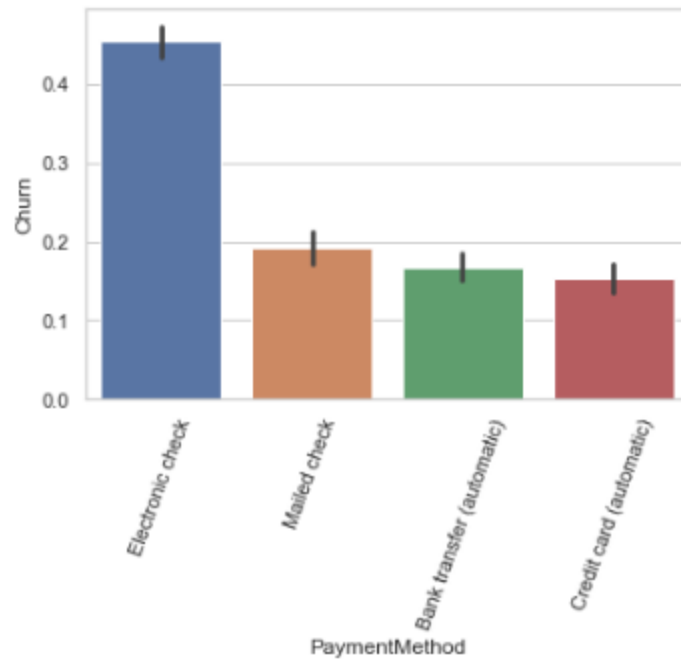
Month to Month subscribers are churning at a higher rate



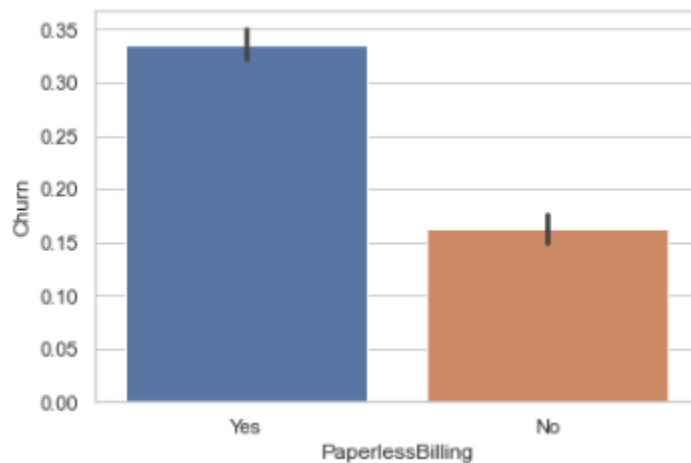
Customers who did not use Tech Support are churning at a higher rate



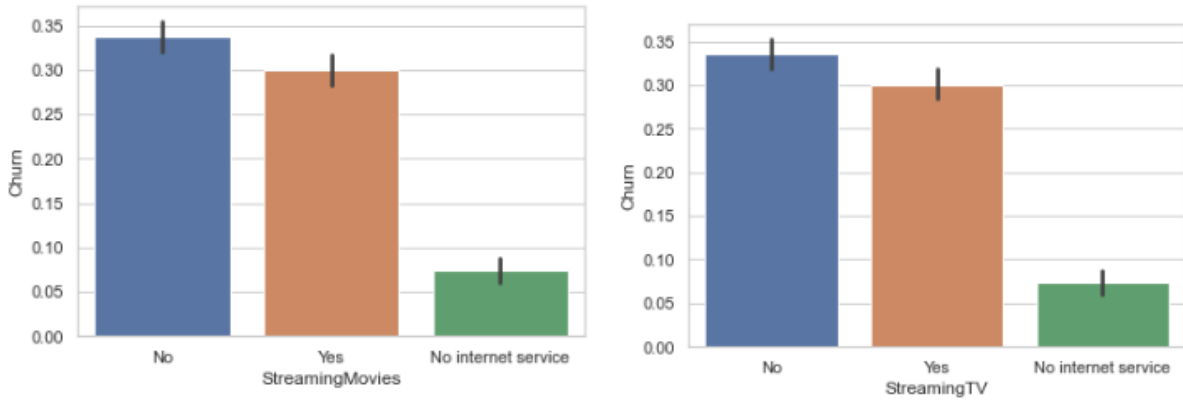
Customers who use checks as payment method are churning at higher rate



Customers who are on paperless billing have a higher churn rate. But this is could be very well because a lot of customers are enrolled into paperless billing.



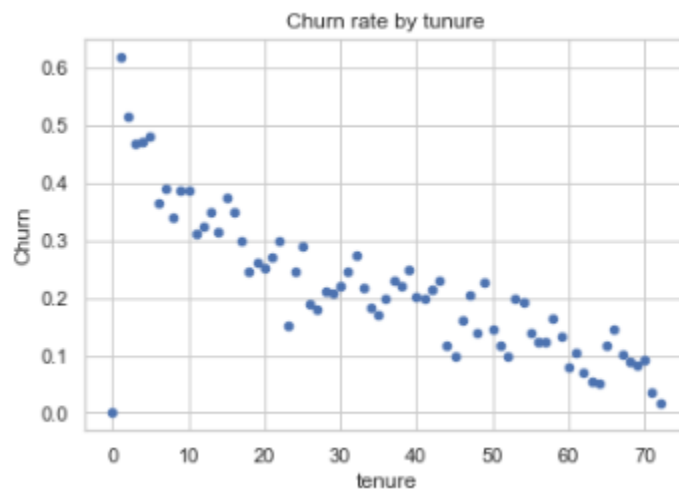
Customer who are Steaming movies vs Steaming tv seems to have very similar Churn rates:



Let's look at some summary stats for tenure variable:

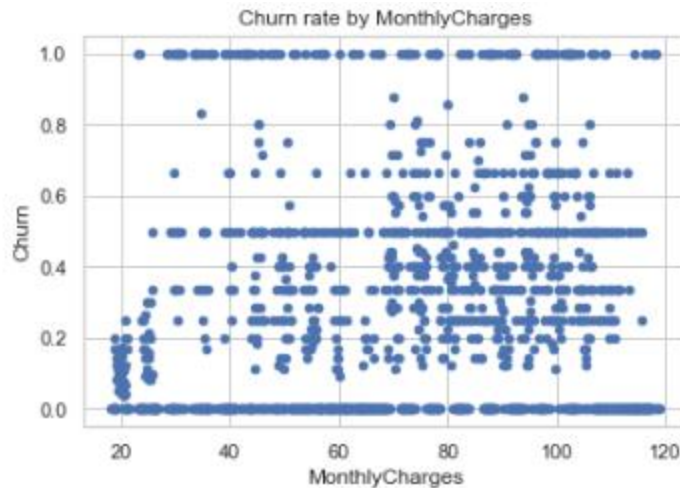
```
count    7043.000000
mean     32.371149
std      24.559481
min       0.000000
25%       9.000000
50%      29.000000
75%      55.000000
max      72.000000
Name: tenure, dtype: float64
```

Now, we will take a look at how mean 10/5/2020 churn rates are doing when compared with tenure.



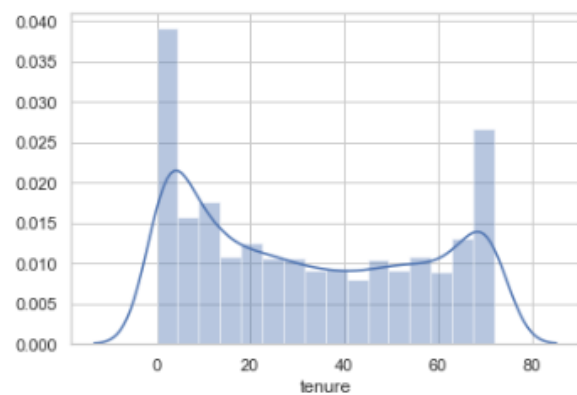
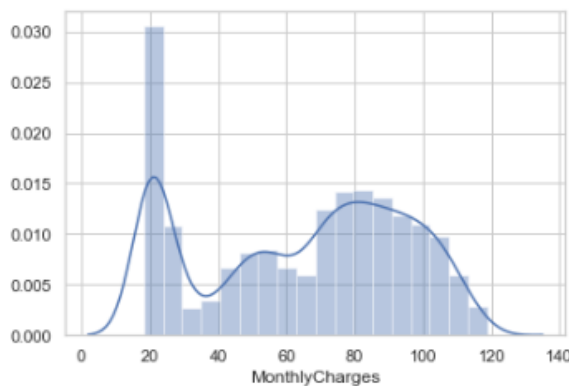
It shows that if the customer's tenure is long then churn rate is low.

Let's also look at other numerical variables such as MonthlyCharges:



There seems to be no relation for Churn and MonthlyCharges

Distribution plots for MonthlyCharges and tenure:



Monthly Charges seems to be roughly normal distribution and Tenure Distribution seems to be high at the ends, so a portion of the customers have either had lowest and highest tenure periods.