Springboard Data Science Bootcamp – Capstone 2 – Predict Customer Churn

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 an 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.

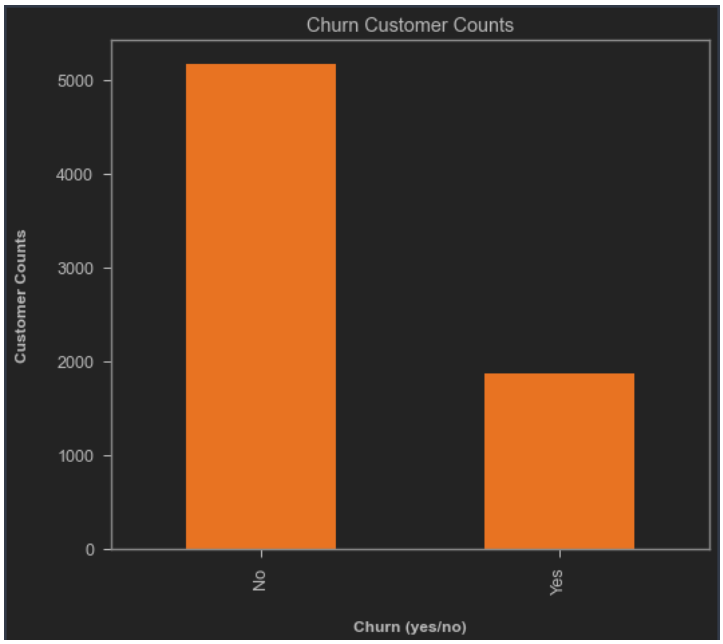
Descriptive Stats:

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

tenure
7043.000000
NaN
NaN
NaN
32.371149
24.559481
0.000000
9.000000
29.000000
55.000000
72.000000

MonthlyCharges
7043.000000
NaN
NaN
NaN
64.761692
30.090047
18.250000
35.500000
70.350000
89.850000
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:

Replacing the Churn string value (yes/no) to numbers (0 – no and 1 – yes)

```
# converting churn to numerical variable for analyzing
df.loc[df.Churn=='No','Churn'] = 0
df.loc[df.Churn=='Yes','Churn'] = 1
df['Churn'] = df['Churn'].astype(int)
```

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

```
#Total charges is obect 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)
```

Then investigate for missing values, it looks like TotalCharges has missing values.

```
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
```

Applying imputation to fix the issue with missing values with means of TotalCharges using SimpleImputer from sklearn.impute.

```python
# 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])
```

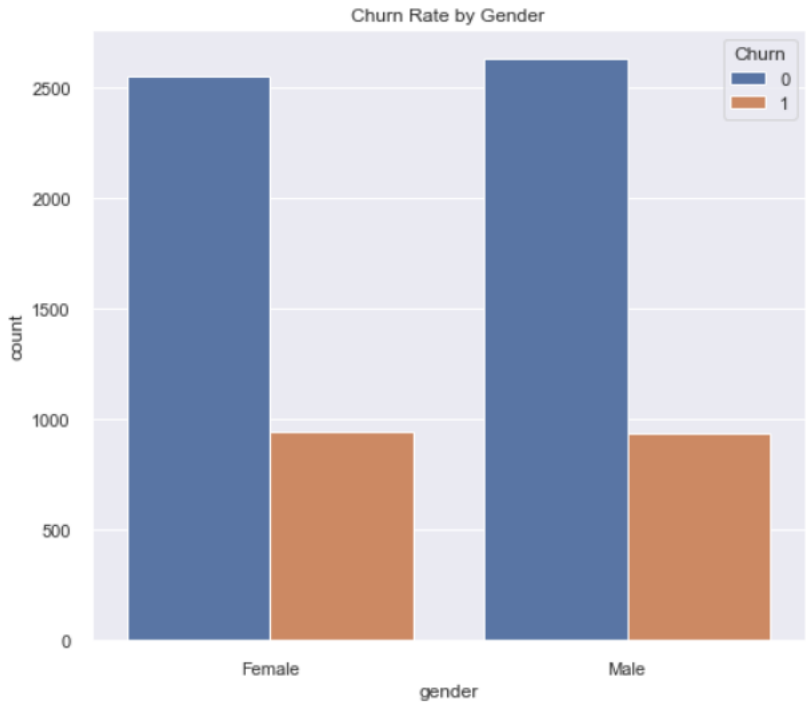
Identified the categorical, numerical and continuous features in the data set.

```python
columns_idx = np.s_[0:] # Slice of first row(header) with all columns.
first_record_idx = np.s_[0] # Index of first record

string_fields = [type(fld) is str for fld in df.iloc[first_record_idx, columns_idx]] # All string fields
all_features = [x for x in df.columns if x != 'Churn']
categorical_columns = list(np.array(df.columns)[columns_idx][string_fields])
categorical_features = [x for x in categorical_columns if x != 'Churn']
continuous_features = [x for x in all_features if x not in categorical_features]

print('All Features: ', all_features)
print('\nCategorical Features: ', categorical_features)
print('\nContinuous Features: ', continuous_features)
print('\nAll Categorical Columns: ', categorical_columns)
```

```
All Features:  ['customerID', 'gender', 'SeniorCitizen', 'Partner', 'Dependents', 'tenure', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtecti
on', 'TechSupport', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod', 'MonthlyCharges', 'TotalCharges']

Categorical Features:  ['customerID', 'gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport',
'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod']

Continuous Features:  ['SeniorCitizen', 'tenure', 'MonthlyCharges', 'TotalCharges']

All Categorical Columns:  ['customerID', 'gender', 'Partner', 'Dependents', 'PhoneService', 'MultipleLines', 'InternetService', 'OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSuppor
t', 'StreamingTV', 'StreamingMovies', 'Contract', 'PaperlessBilling', 'PaymentMethod']
```

EDA - Data visualization:

Looking at the churn rate by gender there seems to 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.

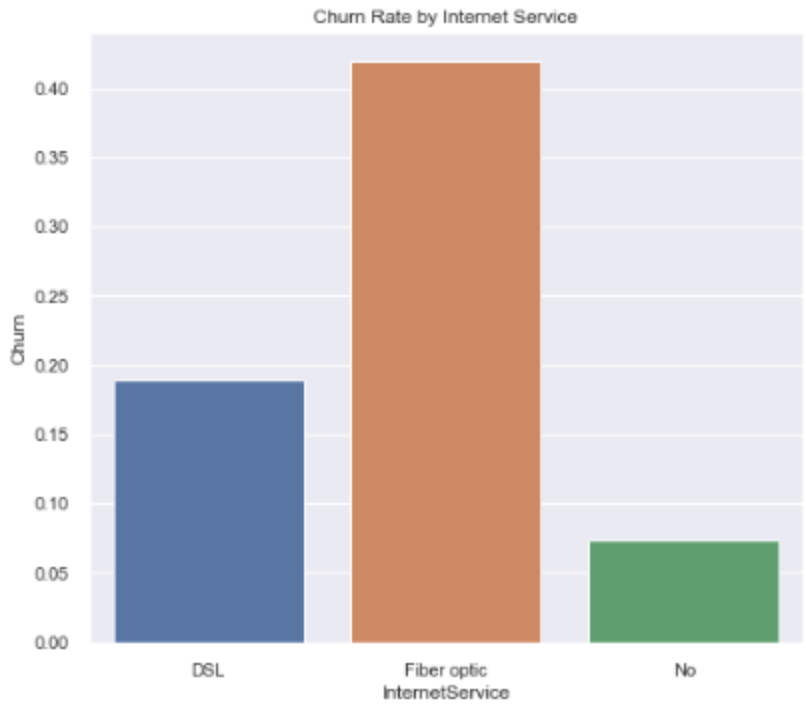



```
# plotting gender
mean_churn = df[['gender','Churn']].groupby('gender').mean()
print(mean_churn)
print("Female churn is little higher, but not that big difference")
```

```
           Churn
gender
Female  0.269209
Male    0.261603
Female churn is little higher, but not that big difference
```

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

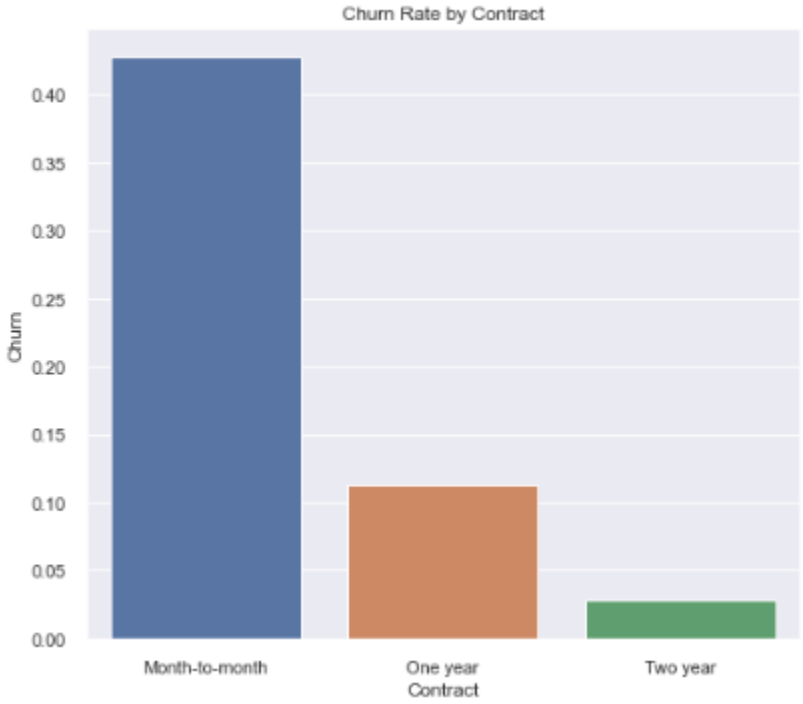
Fiber Optic customers are churning at a higher rate:




```
# plotting Internet Servive type
isp_churn = df[['InternetService','Churn']].groupby('InternetService').mean()
print(isp_churn)
print("Fiber Optic customers are churning at a higher rate")
```

```
                   Churn
InternetService
DSL             0.189591
Fiber optic     0.418928
No              0.074050
Fiber Optic customers are churning at a higher rate
```

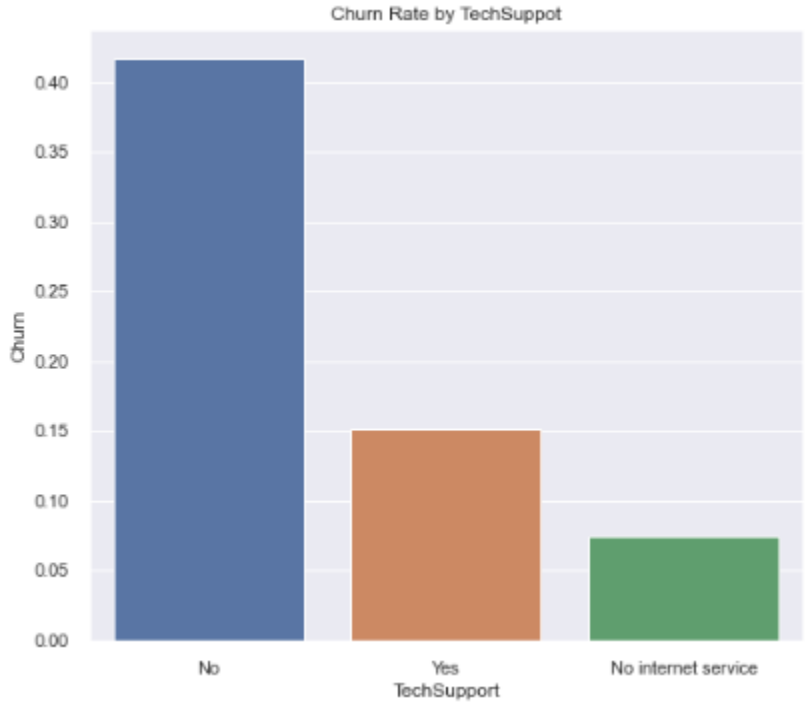
Month to Month subscribers are churning at a higher rate




```
# plotting Churn rate by Contract type
cont_churn = df[['Contract','Churn']].groupby('Contract').mean()
print(cont_churn)
print("Month to Month subscribers are churning at a higher rate")
```

```
                 Churn
Contract
Month-to-month  0.427097
One year        0.112695
Two year        0.028319
Month to Month subscribers are churning at a higher rate
```

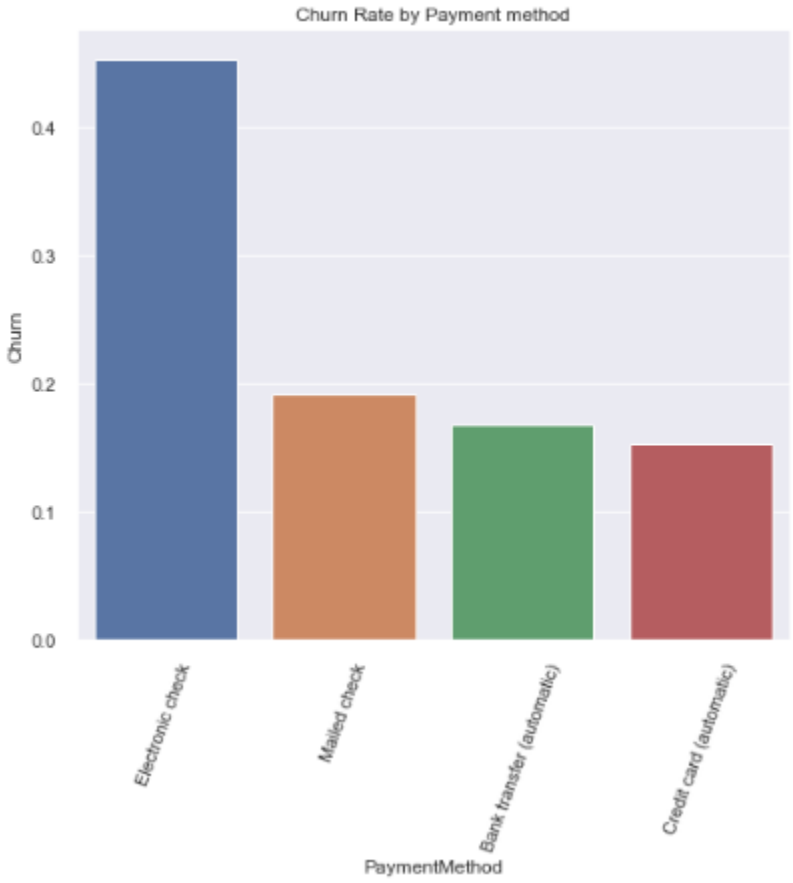
Customers who did not use Tech Support are churning at a higher rate




```
# plotting Churn rate by TechSupport
tech_churn = df[['TechSupport','Churn']].groupby('TechSupport').mean()
print(tech_churn)
print("Customers who did not use  Tech Support are churning at a higher rate")
```

```
                       Churn
TechSupport
No                  0.416355
No internet service  0.074050
Yes                 0.151663
Customers who did not use  Tech Support are churning at a higher rate
```

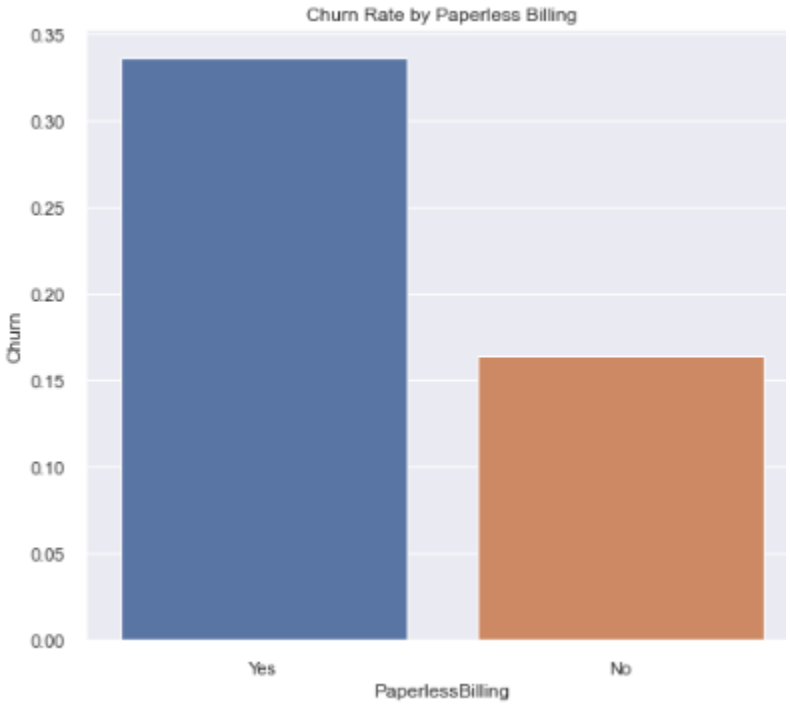
Customers who use checks as payment method are churning at higher rate




```
# plotting PaymentMethod
paym_churn = df[['PaymentMethod','Churn']].groupby('PaymentMethod').mean()
print(paym_churn)
print("Customers who use checks as payment method are churning at higher rate")
```

```
                           Churn
PaymentMethod
Bank transfer (automatic)  0.167098
Credit card (automatic)    0.152431
Electronic check           0.452854
Mailed check               0.191067
Customers who use checks as payment method are churning at higher rate
```

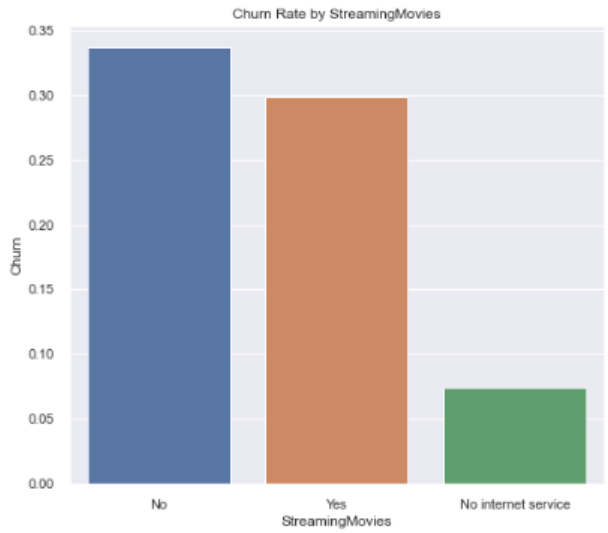
Customers who are on paperless billing have a higher churn rate. But this is could be very well because a lot of customers are enrolled into paperless billing.

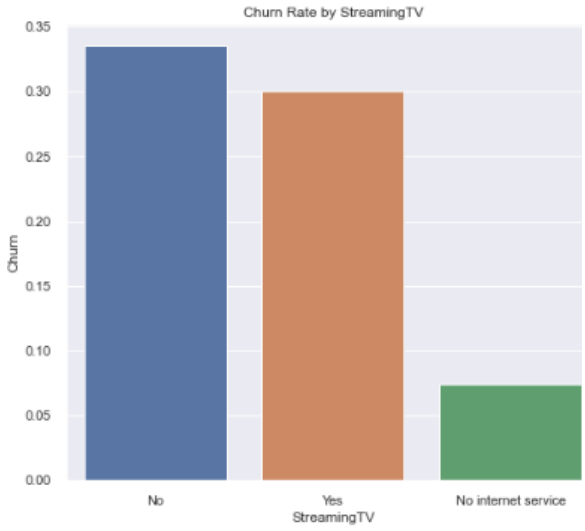



```
# plotting PaperlessBilling
ebill_churn = df[['PaperlessBilling','Churn']].groupby('PaperlessBilling').mean()
print(ebill_churn)
print("Customers who are on paperless billing are churning at higher rate")
```

```
                   Churn
PaperlessBilling
No              0.163301
Yes             0.335651
Customers who are on paperless billing are churning at higher rate
```

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

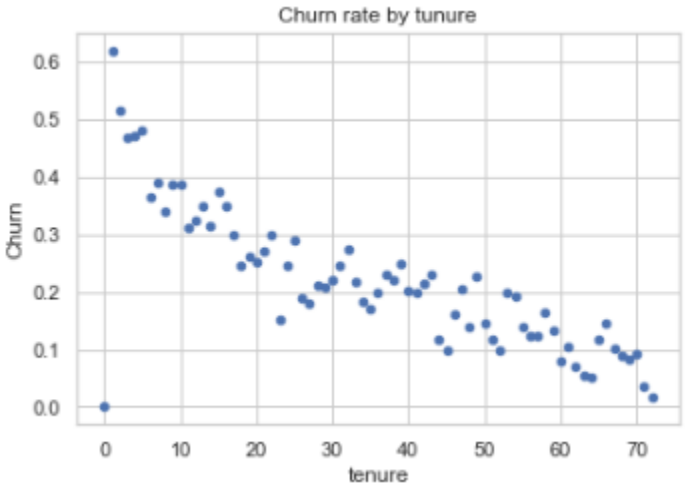




Let's look at some summary stats for tenure variable:

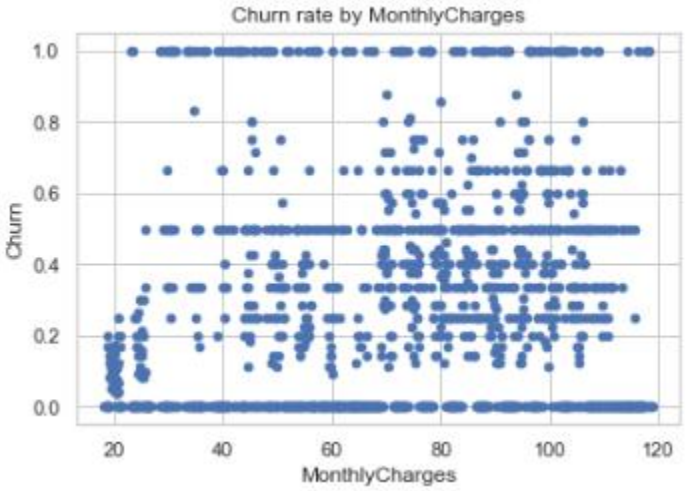
```
count    7043.000000
mean       32.371149
std        24.559481
min         0.000000
25%         9.000000
50%        29.000000
75%        55.000000
max        72.000000
Name: tenure, dtype: float64
```

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



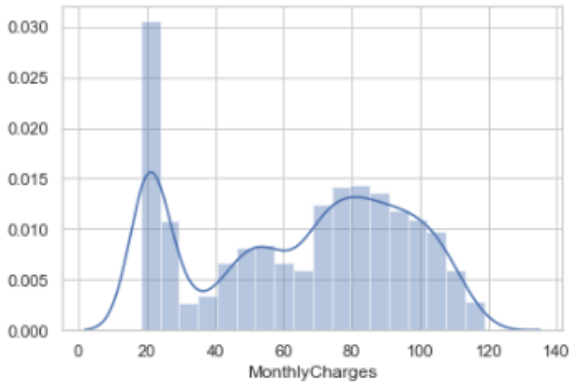
It shows that if the customer's tenure is long then churn rate is low.

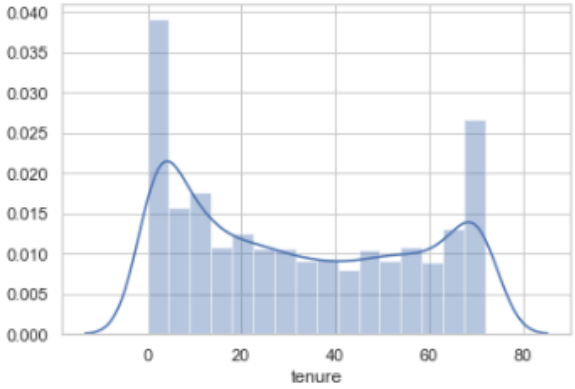
Let's also look at other numerical variables such as MonthlyCharges:



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.

EDA Summary:

- More customers are using Fiber Optic for Internet Service have left the company than compared to DSL.
- Customers who do not use online security have left the company.
- Customers not using technical support have left the company.
- Customers who pay month to month are the most who leave the company.
- Customer's gender has almost equal rates of churn between them.
- The Monthly Charges for customers who churned tends to pay higher monthly fees than those that stay.
- Customers that churn tend to be relatively new customers when looking at tenure distribution.