## SCENARIO

A telecom industry is in a tight business with no option to move in new markets but only with the option of competing in the current market. The firm wants to increase the revenue and profit by analysing the customer data.

## PROBLEM STATEMENT

To enhance revenue and market share through suggesting possible crucial factors for customer satisfaction and by reducing churning of customers

## DATASET

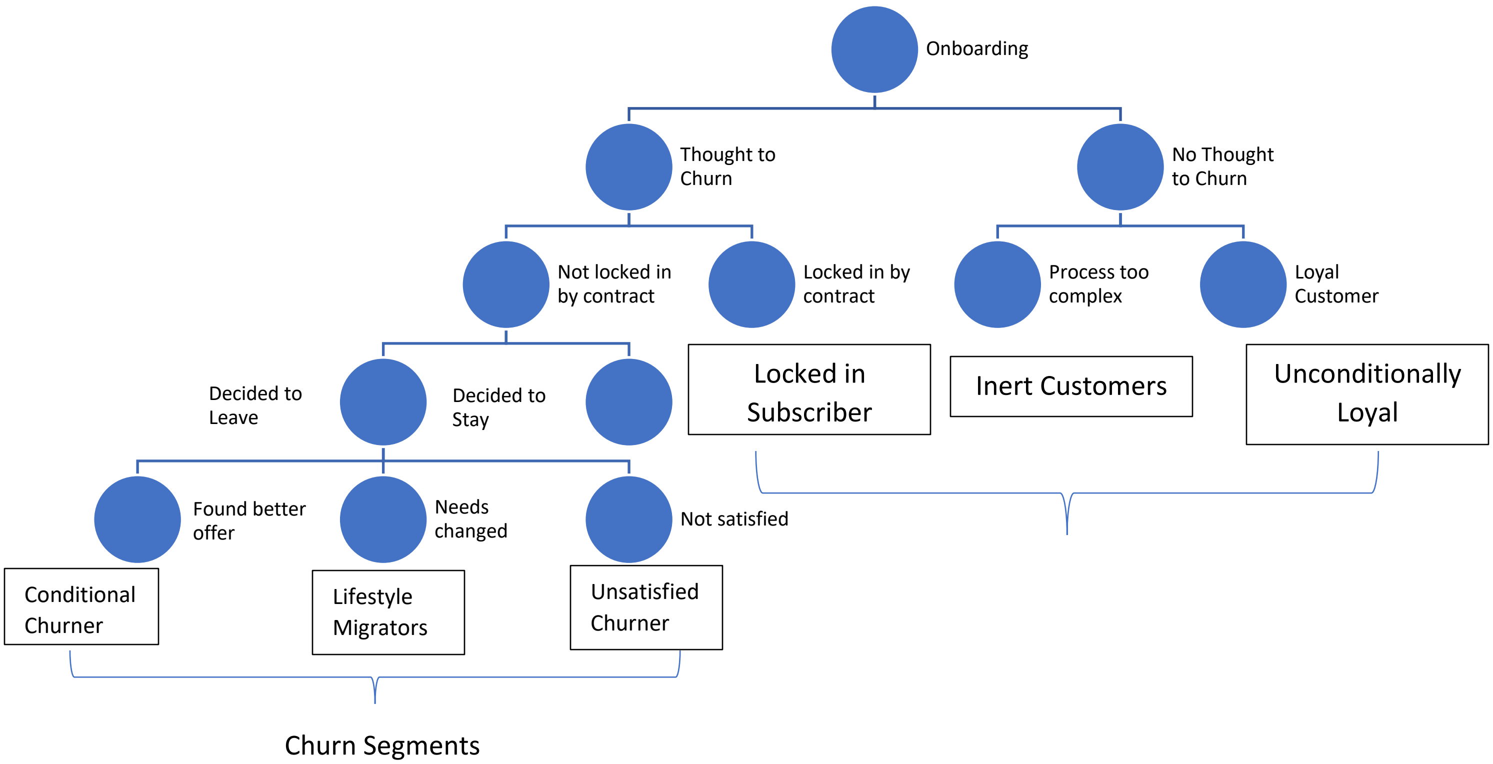
Each row represents a telecom customer, the columns contain customer's attributes and the factors responsible for customer satisfaction/dissatisfaction.

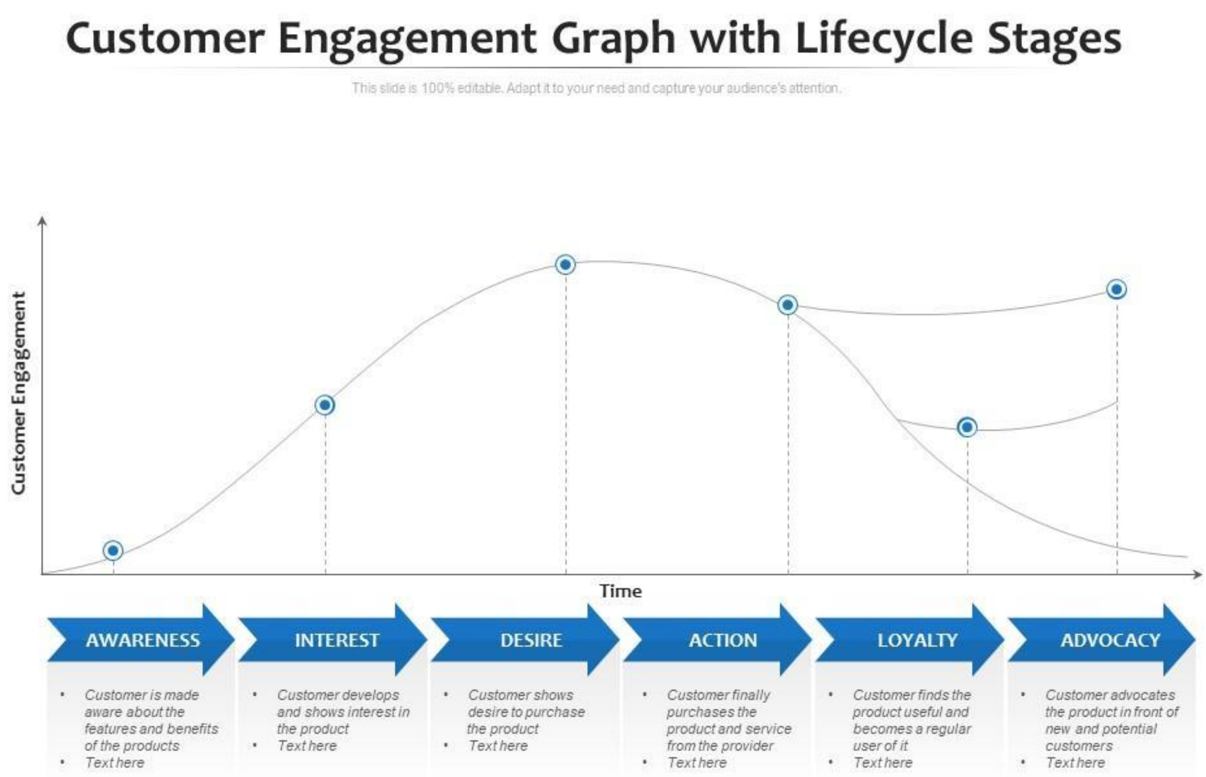
The data set includes information about:

- Customers who left within the last month - 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

(Source of Dataset - https://www.kaggle.com/datasets/blastchar/telco-customer-churn )

CUSTOMER JOURNEY IN TELECOM INDUSTRY






Source: https://www.slideteam.net/customer-engagement-graph-with-lifecycle-stages.html

**ABOUT CHURN IN TELCO INDUSTRY**

- Tariff Plan Churn
- Service Churn
- Product Churn
- Usage Churn

**TOOLS USED: Python – Jupyter Notebook - PANDAS, NumPy, Seaborn, Matplotlib**

## Data Cleaning and Wrangling

We will use Jupyter notebook for the analysis

1. Call in the different libraries used for data analysis

In [1]:
```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns # For creating plots
import matplotlib.ticker as mtick # For specifying the axes tick format
import matplotlib.pyplot as plt

sns.set(style = 'white')

# Input data files are available in the "../input/" directory.

import os
print(os.listdir("F:/MEM NU/Data Science/archive"))
```

2. Read the .csv data

In [2]:
```
cust_data = pd.read_csv('F:/MEM NU/Data Science/archive/WA_Fn-UseC_-Telco-Customer-Churn.csv')
```

In [3]:
```
cust_data.head()
```

Out[3]:

| | customerID | gender | SeniorCitizen | Partner | Dependents | tenure | PhoneService | MultipleLines | InternetService | OnlineSecurity | ... | DeviceProtection | TechSupp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7590-VHVEG | Female | 0 | Yes | No | 1 | No | No phone service | DSL | No | ... | No | |
| 1 | 5575-GNVDE | Male | 0 | No | No | 34 | Yes | No | DSL | Yes | ... | Yes | |
| 2 | 3668-QPYBK | Male | 0 | No | No | 2 | Yes | No | DSL | Yes | ... | No | |
| 3 | 7795-CFOCW | Male | 0 | No | No | 45 | No | No phone service | DSL | Yes | ... | Yes | |
| 4 | 9237-HQITU | Female | 0 | No | No | 2 | Yes | No | Fiber optic | No | ... | No | |

5 rows × 21 columns

3. Check for the data types

```
In [6]: cust_data.dtypes

Out[6]: customerID           object
        gender               object
        SeniorCitizen         int64
        Partner              object
        Dependents           object
        tenure                int64
        PhoneService         object
        MultipleLines        object
        InternetService      object
        OnlineSecurity       object
        OnlineBackup         object
        DeviceProtection     object
        TechSupport          object
        StreamingTV          object
        StreamingMovies      object
        Contract             object
        PaperlessBilling     object
        PaymentMethod        object
        MonthlyCharges      float64
        TotalCharges         object
        Churn                object
        dtype: object
```

Here we can observe that dataset has float, integer as well as object value

4. Convert data type to numeric and Check for any blank or null datatype

```
In [7]: cust_data.TotalCharges = pd.to_numeric(cust_data.TotalCharges, errors='coerce')
        cust_data.isnull().sum()

Out[7]: 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
```

We observe that Total Charges (TotalCharges) feature has 11 null datapoints

5. To Clean and categorize the data –
    - Remove the blanks and NA's using dropna command.
    - Remove the customer I'd column using iloc command as it is not needed for the analysis.
    - Bin the tenure data using cut command into separate groups of 12 months

```
In [8]: #Remove the blanks and NAs from the dataset
        cust_data.dropna(inplace = True)

In [9]: #Remove customer I'd because the data is not useful
        cust_data = cust_data.iloc[:,1:]

In [10]: print(cust_data['tenure'].max())

72

In [11]: labels = ["{0} - {1}".format(i, i + 11) for i in range(1, 72, 12)]

         cust_data['tenure_group'] = pd.cut(cust_data.tenure, range(1, 80, 12), right=False, labels=labels)

In [12]: cust_data['tenure_group'].value_counts()

Out[12]: 1 - 12     2175
         61 - 72    1407
         13 - 24    1024
         49 - 60     832
         25 - 36     832
         37 - 48     762
         Name: tenure_group, dtype: int64
```

6. Convert Target to binary form and all other categorical data into dummy variables

```
In [13]: #Convert the target into binary form to ease the analysis
         cust_data['Churn'].replace(to_replace='Yes', value=1, inplace=True)
         cust_data['Churn'].replace(to_replace='No',  value=0, inplace=True)

         #Convert all categorical variable into dummy variable
         df_dummies = pd.get_dummies(cust_data)
         df_dummies.head()

Out[13]:
```

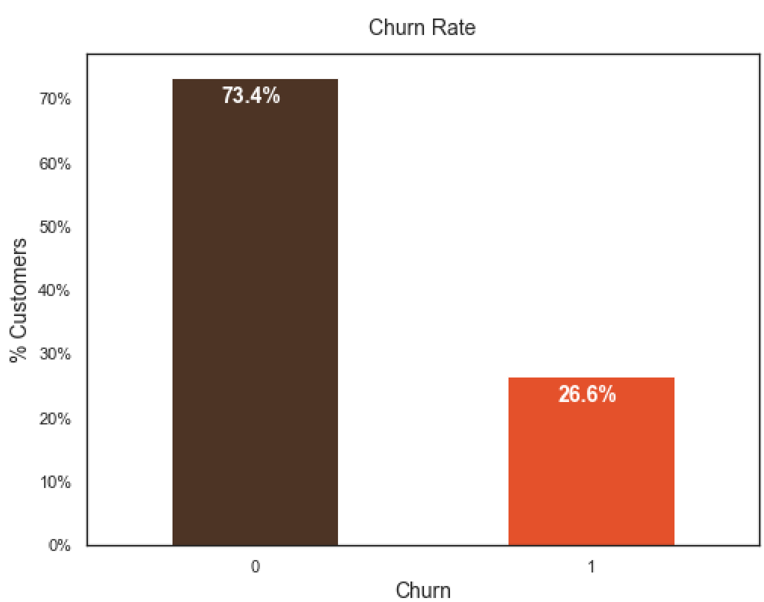
| | SeniorCitizen | tenure | MonthlyCharges | TotalCharges | Churn | gender_Female | gender_Male | Partner_No | Partner_Yes | Dependents_No | ... | PaymentMethod_Bai transfer (automati |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 29.85 | 29.85 | 0 | 1 | 0 | 0 | 1 | 1 | ... | |
| **1** | 0 | 34 | 56.95 | 1889.50 | 0 | 0 | 1 | 1 | 0 | 1 | ... | |
| **2** | 0 | 2 | 53.85 | 108.15 | 1 | 0 | 1 | 1 | 0 | 1 | ... | |
| **3** | 0 | 45 | 42.30 | 1840.75 | 0 | 0 | 1 | 1 | 0 | 1 | ... | |
| **4** | 0 | 2 | 70.70 | 151.65 | 1 | 1 | 0 | 1 | 0 | 1 | ... | |

5 rows × 52 columns

## Exploratory Data Analysis

1. Check the class imbalance of target

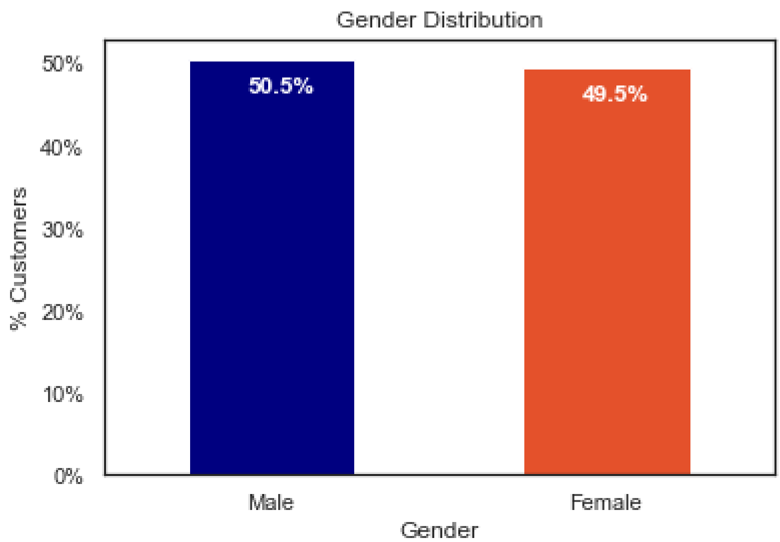
```
In [14]: colors = ['#4D3425','#E4512B']
ax = (cust_data['Churn'].value_counts()*100.0 /len(cust_data)).plot(kind='bar',
                                                                    stacked = True,rot = 0,
                                                                    color = colors,figsize = (8,6))
ax.yaxis.set_major_formatter(mtick.PercentFormatter())
ax.set_ylabel('% Customers',size = 14)
ax.set_xlabel('Churn',size = 14)
ax.set_title('Churn Rate', size = 14)

# create a list to collect the plt.patches data
totals = []

# find the values and append to list
for i in ax.patches:
    totals.append(i.get_width())

# set individual bar lables using above list
total = sum(totals)

for i in ax.patches:
    # get_width pulls left or right; get_y pushes up or down
    ax.text(i.get_x()+.15, i.get_height()-4.0, \
            str(round((i.get_height()/total), 1))+'%',
            fontsize=12,
            color='white',
           weight = 'bold',
           size = 14)
```



We find that there is high difference between Churned and non churned customers. Here it is not so important, but we need to ensure that either we add more datasets in the minority class (oversampling) or take out some datasets from majority class (under sampling) to reduce the high imbalance.

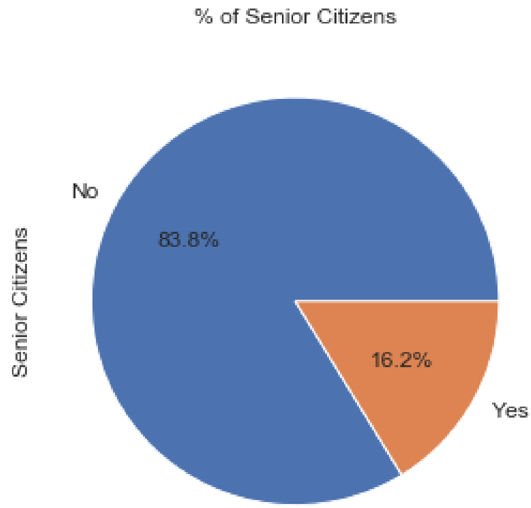
**Demographic Analysis of Population**

2. Check for the gender distribution in total population

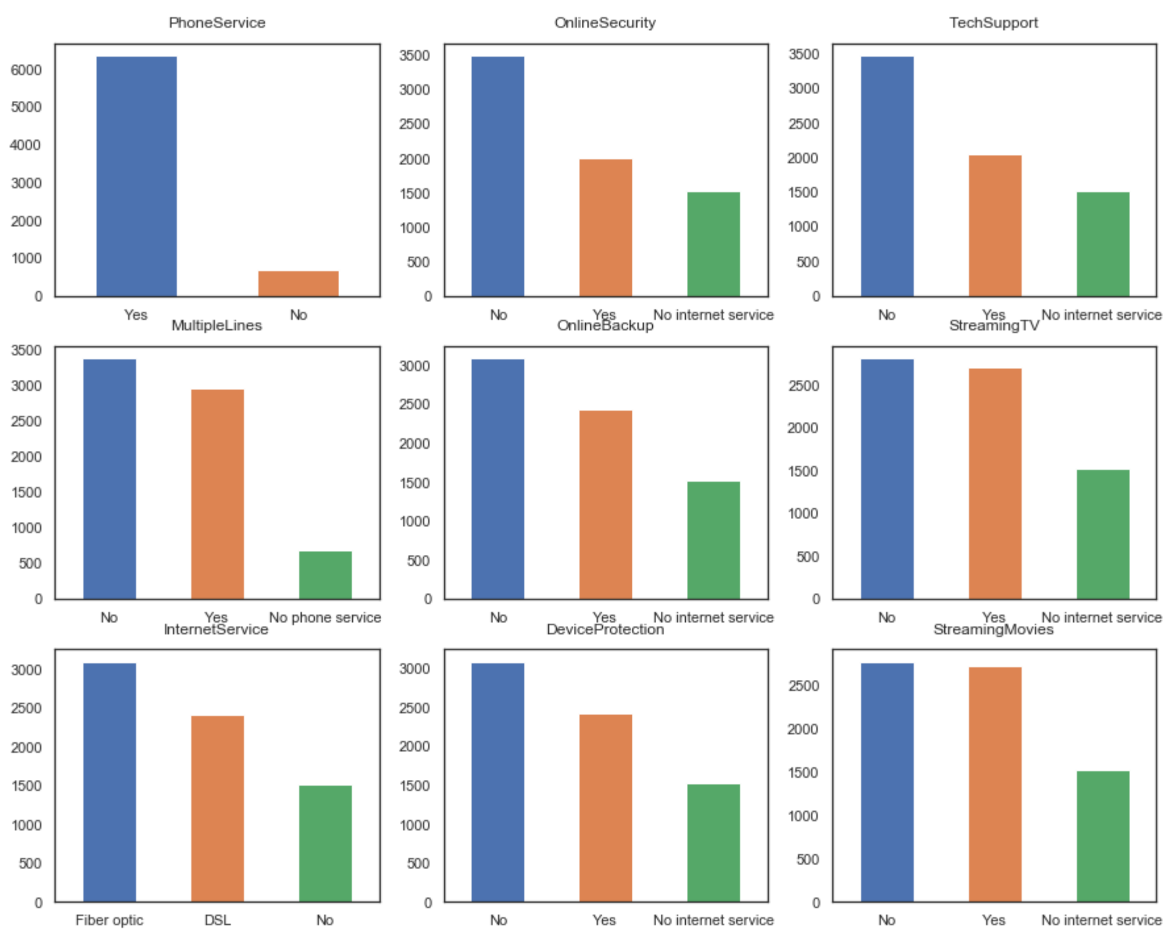



Observation: There is no major difference between the genders and both of them are relatively similar in numbers.

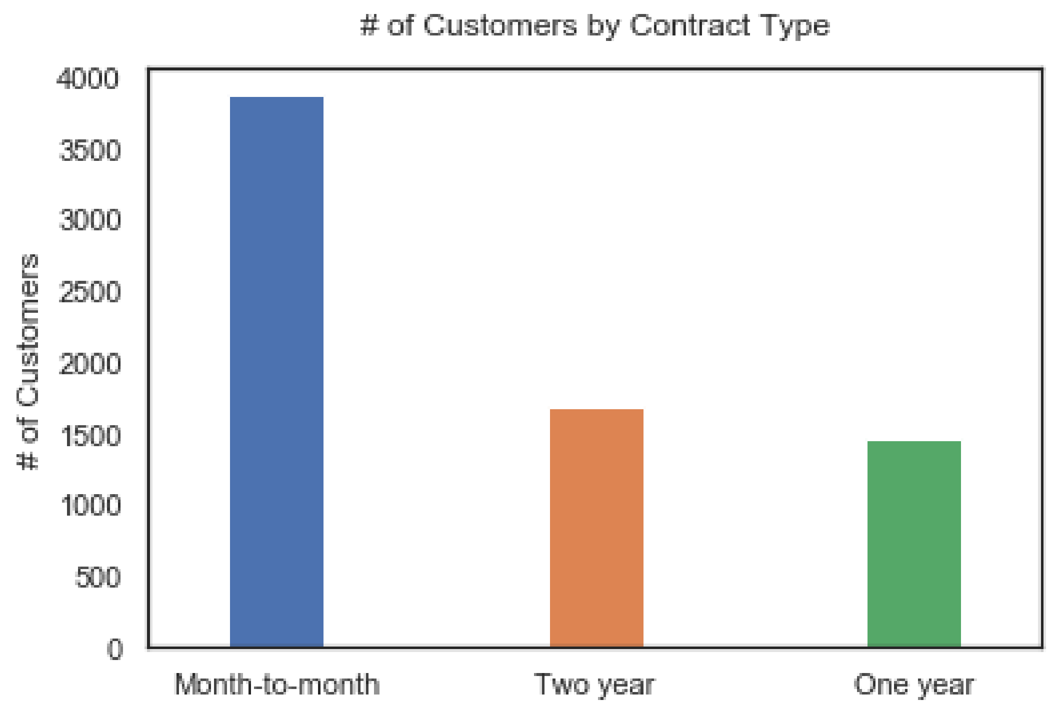
3. Senior Citizen




4. Analysis of Services availed by different customers

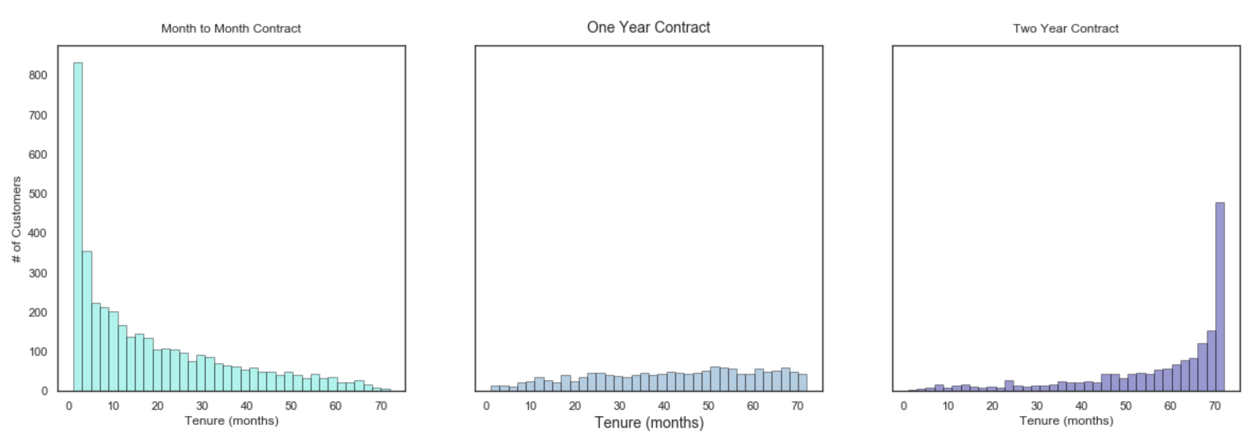


5. Analysis of customers based on the contract type



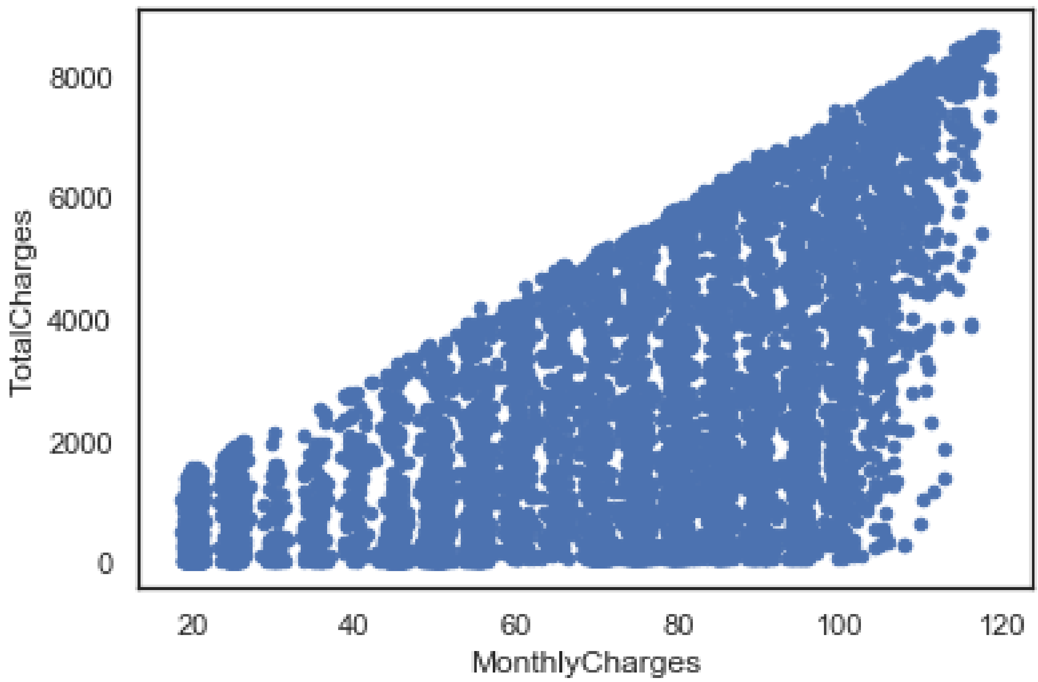

We can see that month-to-month contract type customers are maximum in number. Hence, we will have to compare the percentage data.

6. Distribution of customers within the contract period



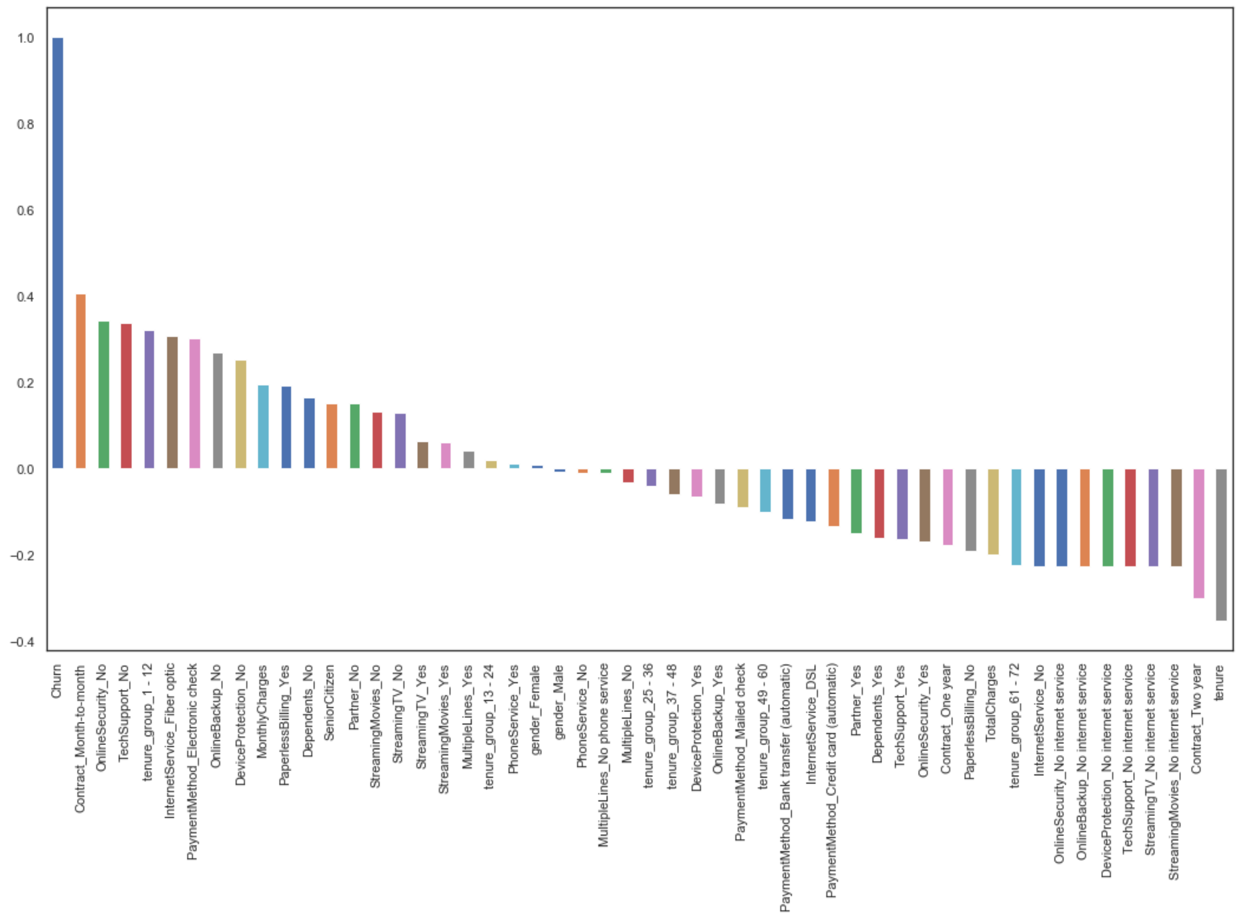
This shows that maximum number of people are either at the left-hand side or at the right most corner of contract. Either 1 month customer or 72 months customer

7. Distribution of Monthly vs Total Charges




Observation - As monthly charges increase, the total charges also tend to increase.

8. Check the correlation of different features with Churn

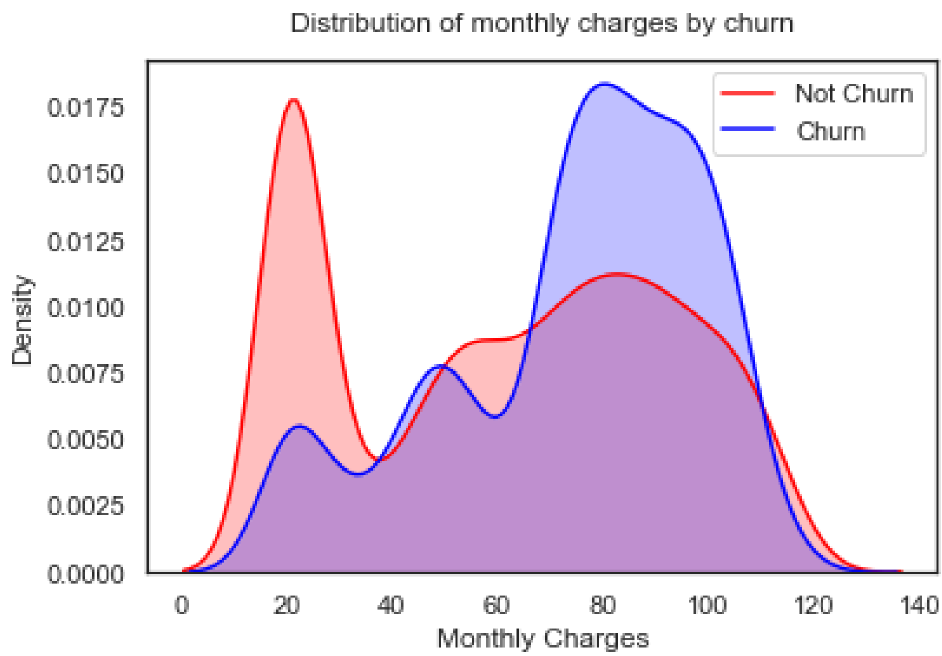


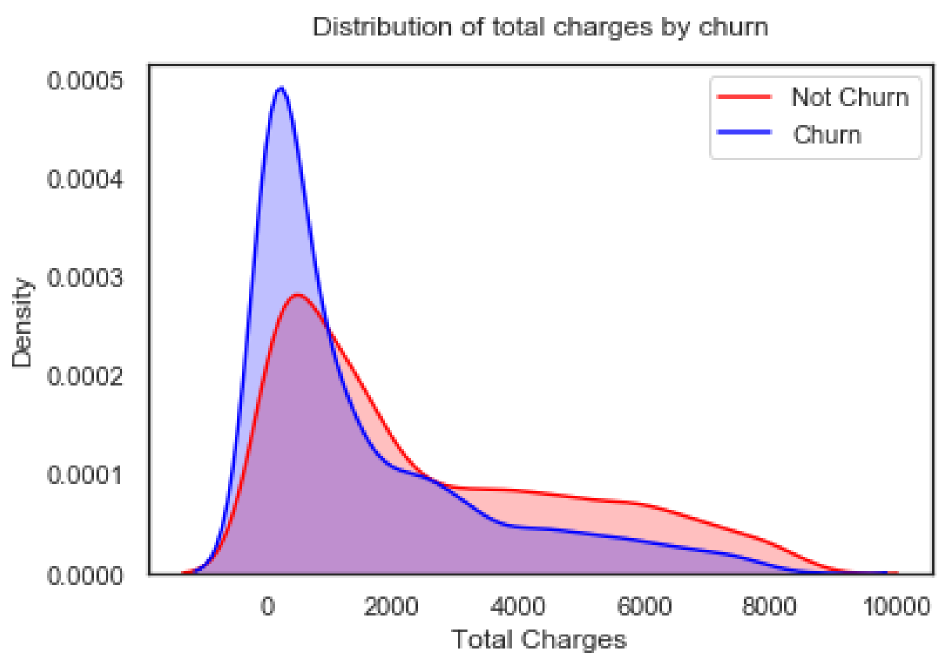

Observation: Churn is inversely proportional and heavily correlated to the tenure of customer. It is also directly proportional to the factor that customer has not availed any internet services. Factors like gender and phone service are not at all correlated with Churn.

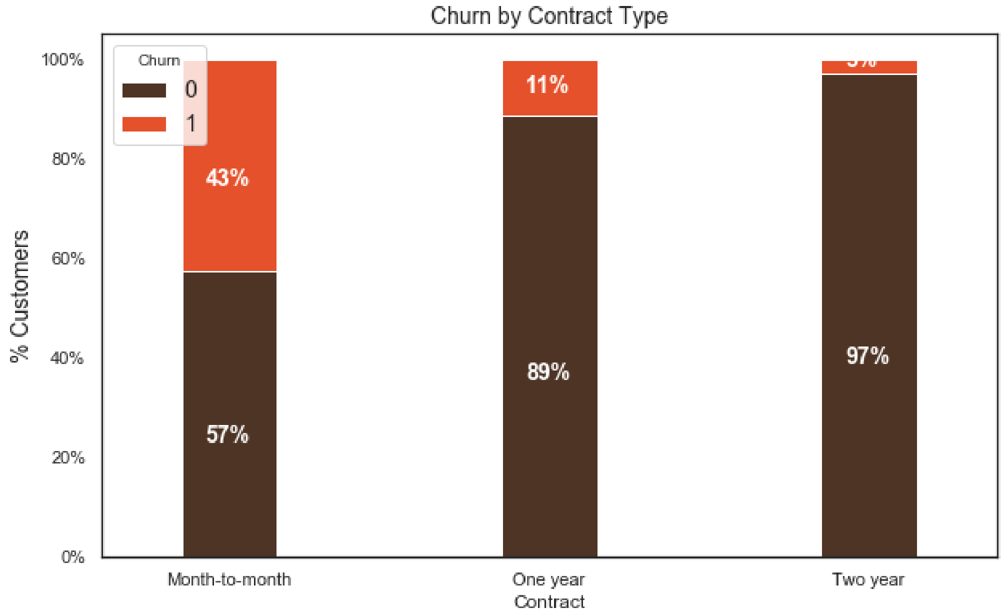

Heatmap can be another way to check the correlation. It can be little difficult to identify for datasets having high number of features but is very efficient for smaller datasets.

### 9. Univariate Analysis

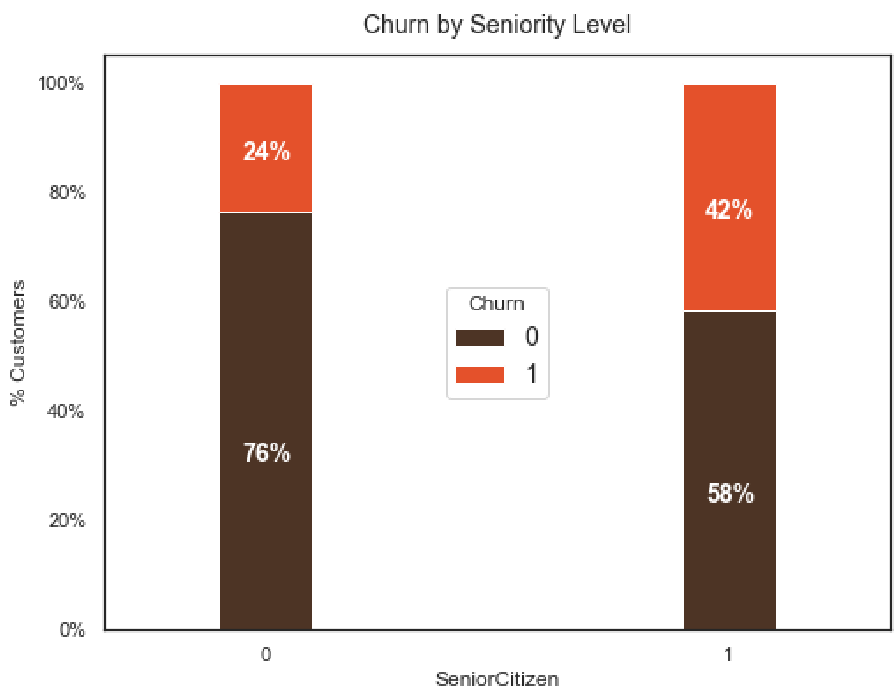
Conducting Univariate analysis to identify relation of different features with Churn



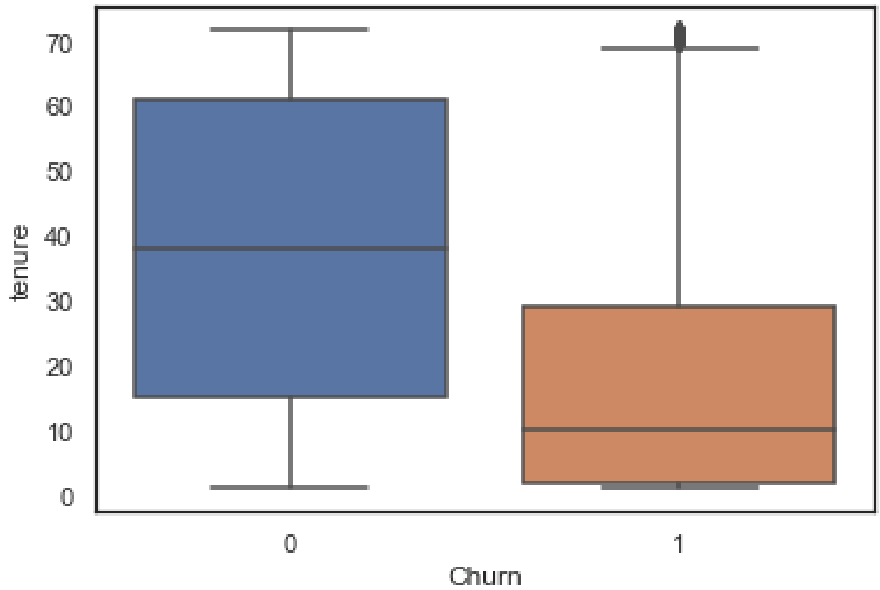



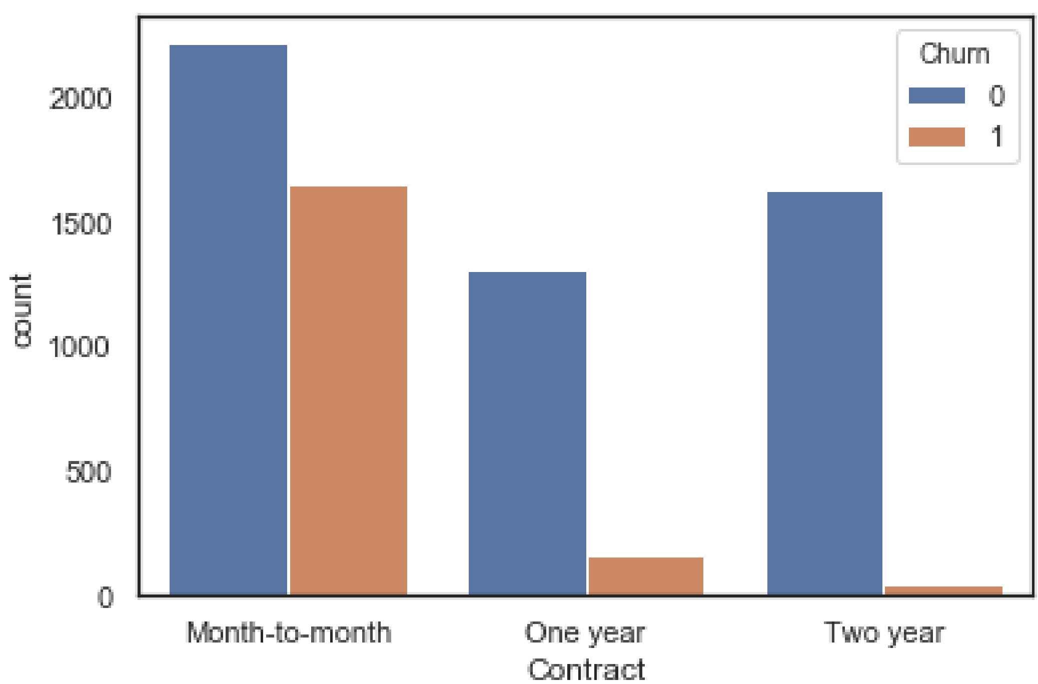


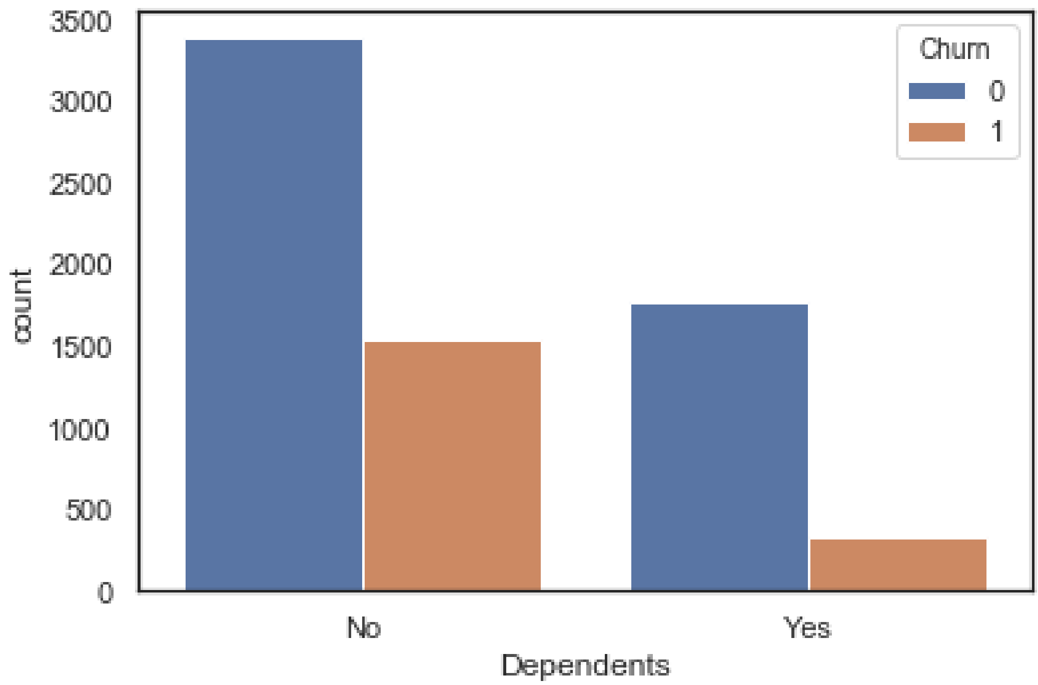
Churn by Contract Type
Churn
0
1
100%
80%
60%
40%
20%
0%
% Customers
43%
57%
11%
89%
97%
Month-to-month
One year
Two year
Contract

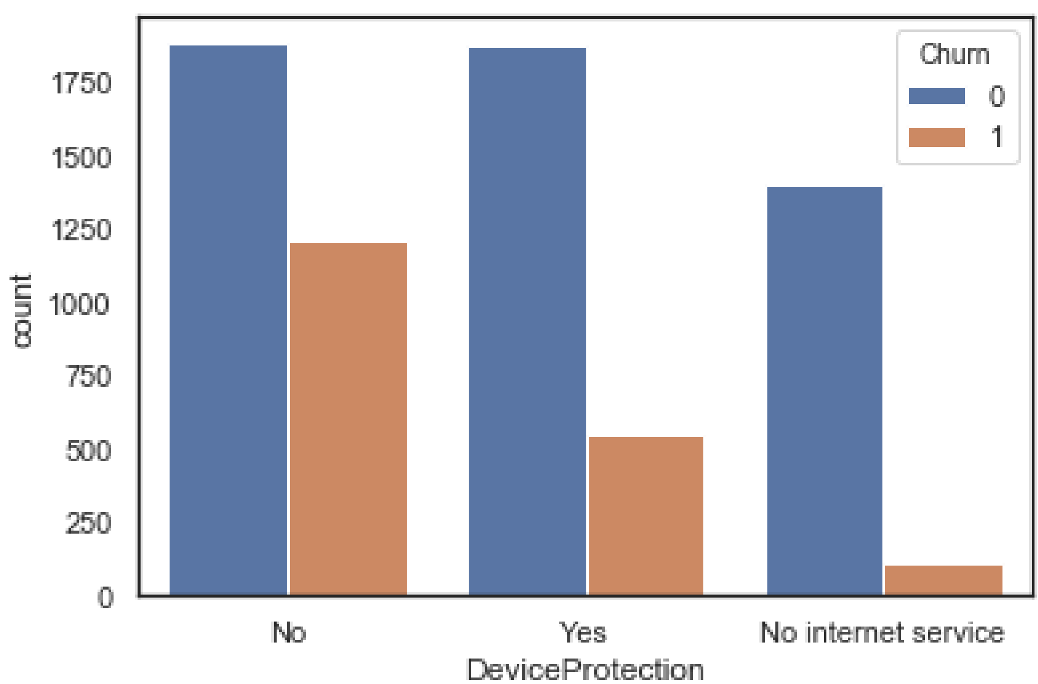



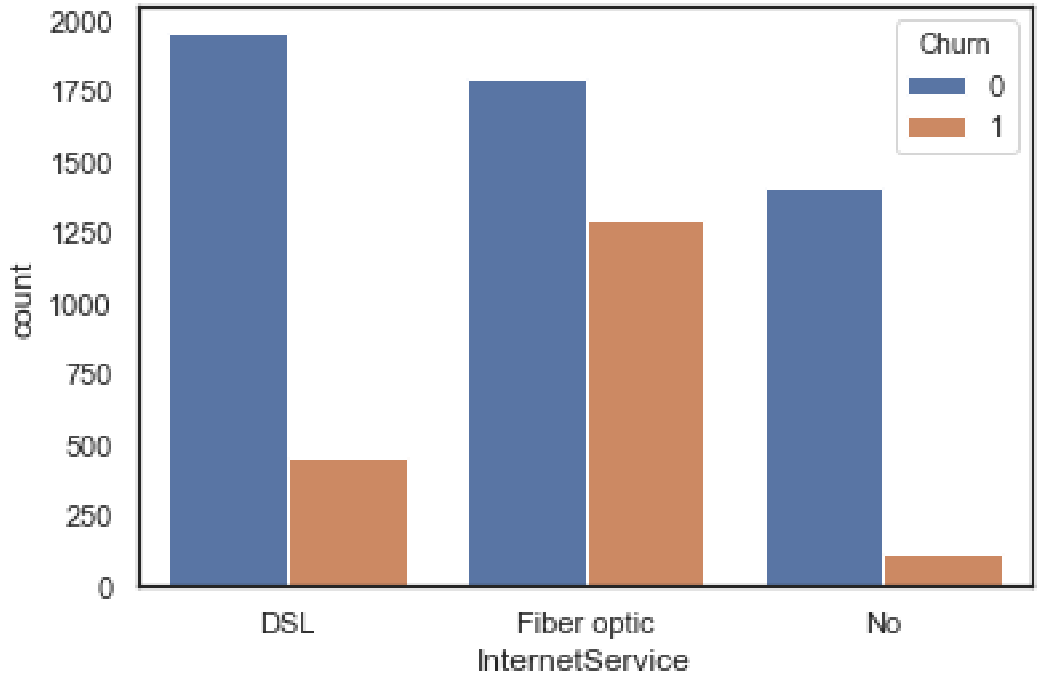
Churn by Seniority Level
100%
80%
60%
40%
20%
0%
% Customers
24%
76%
42%
58%
Churn
0
1
0
1
SeniorCitizen


70
60
50
40
30
20
10
0
tenure
0
1
Churn

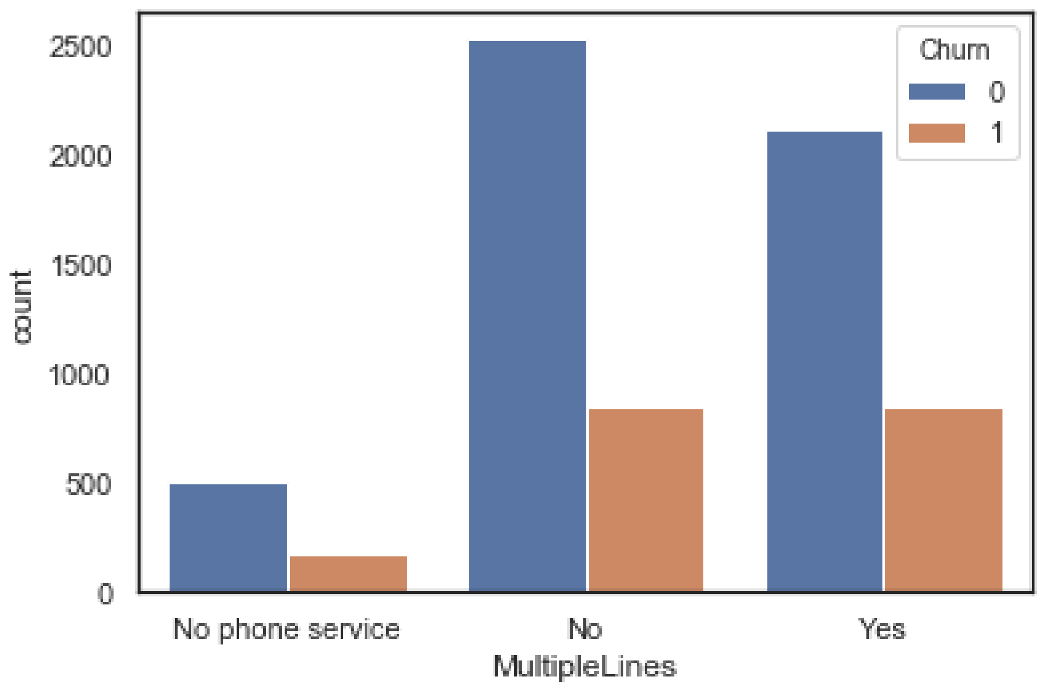

Churn
0
1
2000
1500
1000
500
0
count
Month-to-month
One year
Two year
Contract


Churn
0
1
3500
3000
2500
2000
1500
1000
500
0
count
No
Yes
Dependents

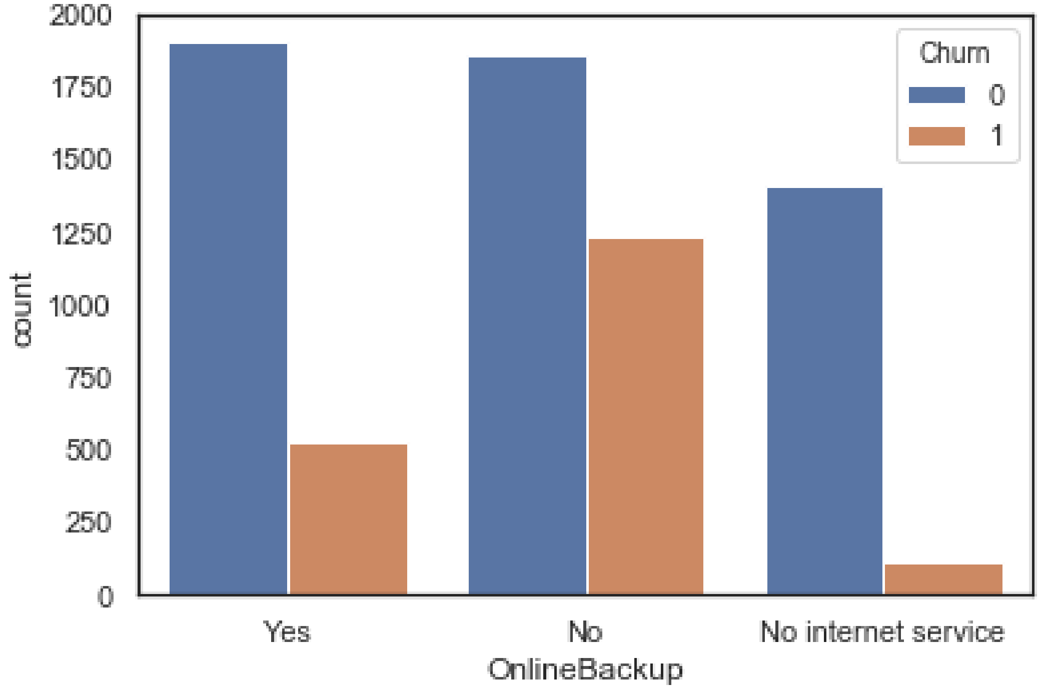

Churn
0
1
1750
1500
1250
1000
750
500
250
0
count
No
Yes
No internet service
DeviceProtection

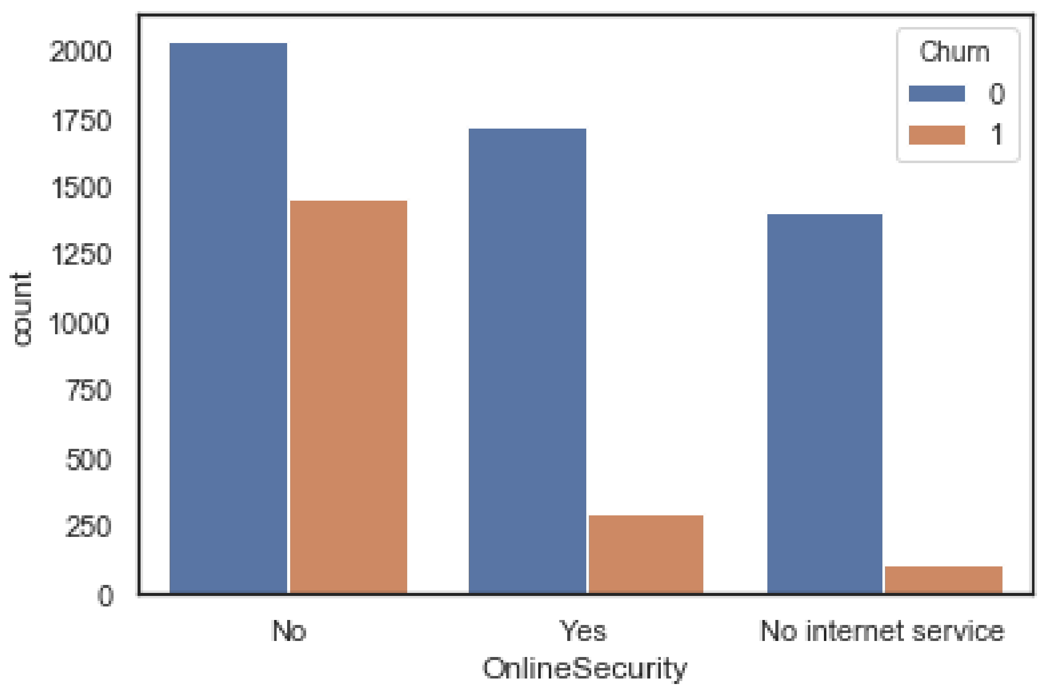


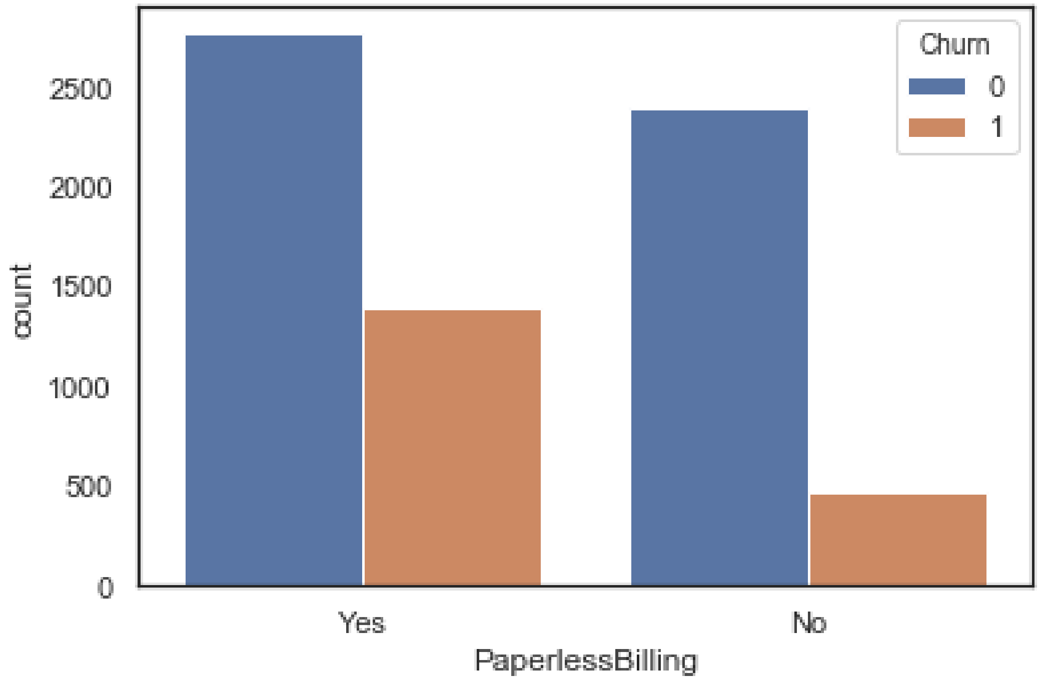
Churn
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
DSL
Fiber optic
No
InternetService

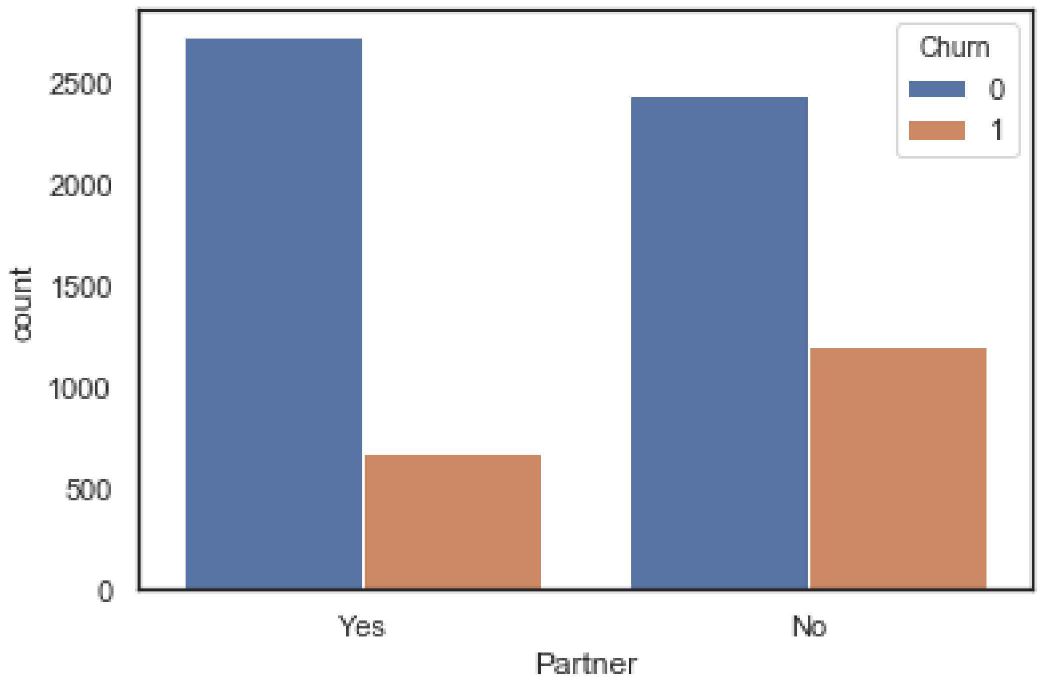



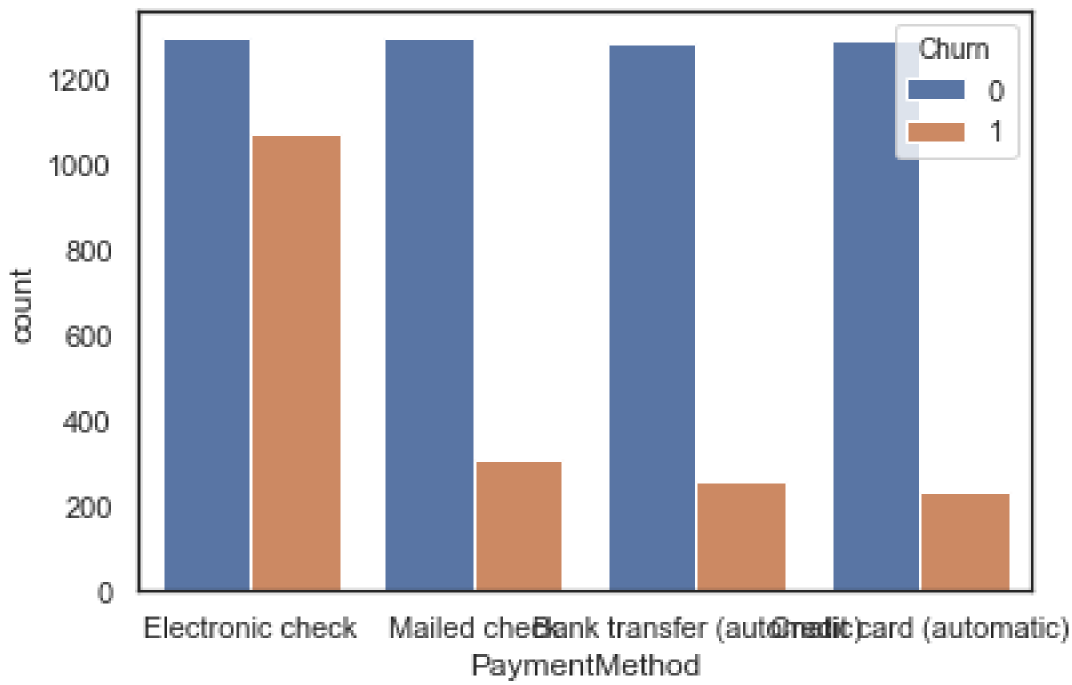
Churn
0
1
count
2500
2000
1500
1000
500
0
No phone service
No
Yes
MultipleLines


count
2000
1750
1500
1250
1000
750
500
250
0
Yes
No
No internet service
OnlineBackup
Churn
0
1

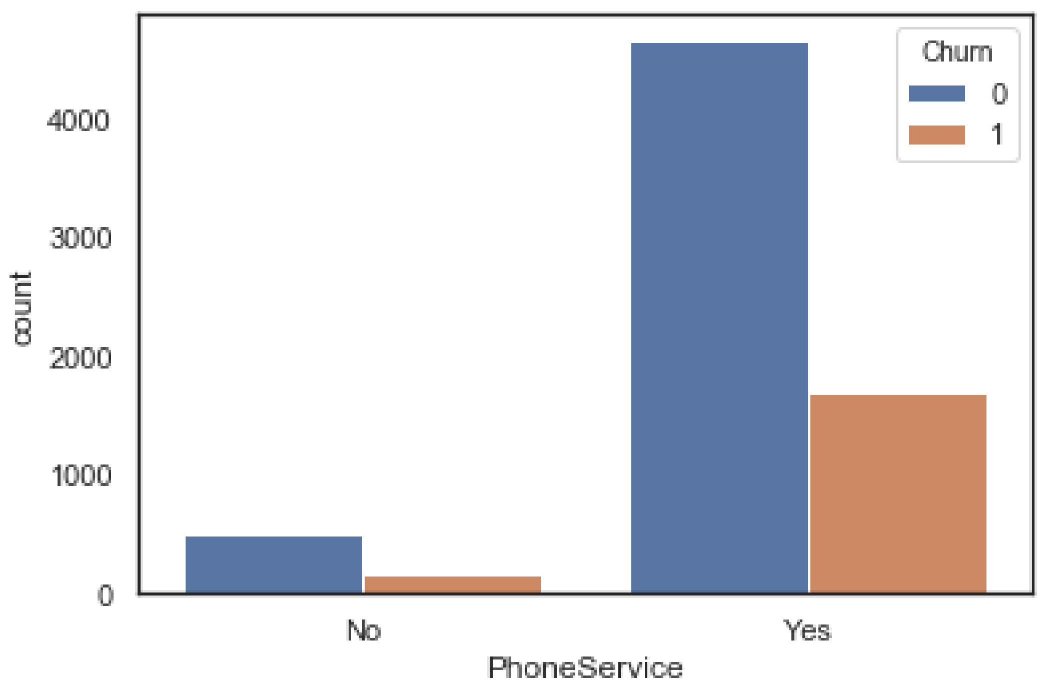

count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
OnlineSecurity
Churn
0
1


Churn
0
1
2500
2000
1500
1000
500
0
count
Yes
No
PaperlessBilling

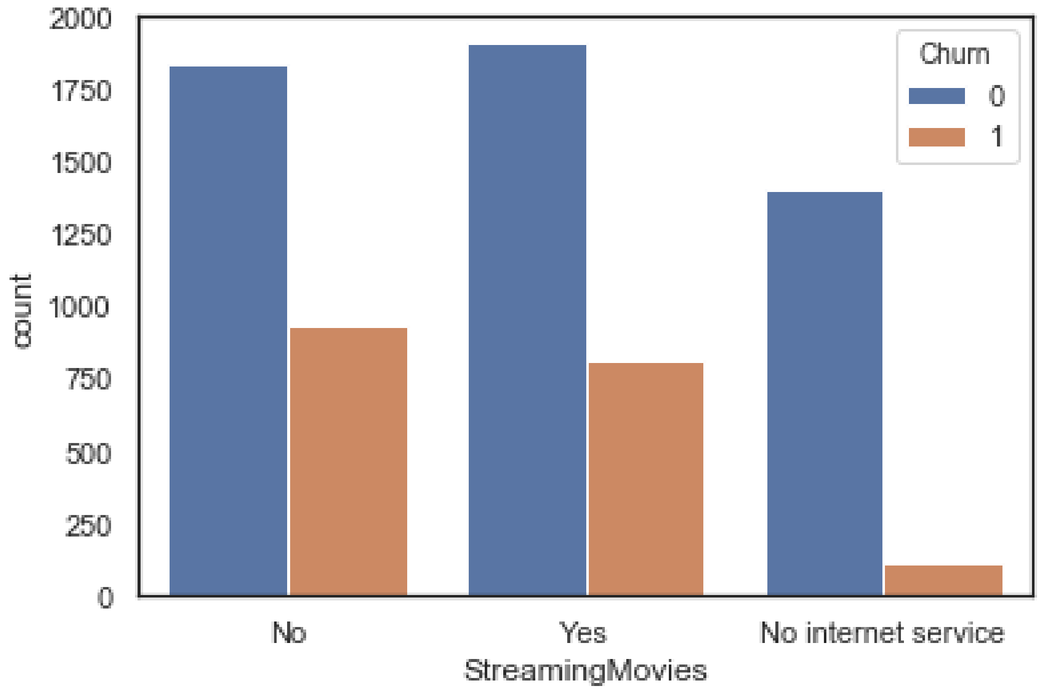

Churn
0
1
2500
2000
1500
1000
500
0
count
Yes
No
Partner

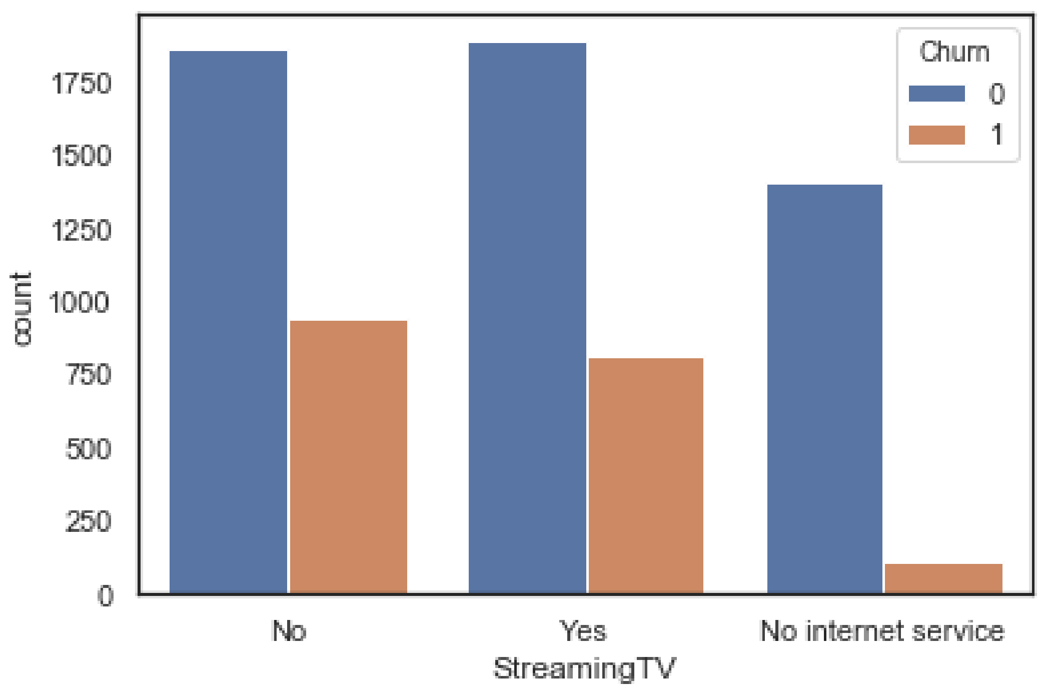


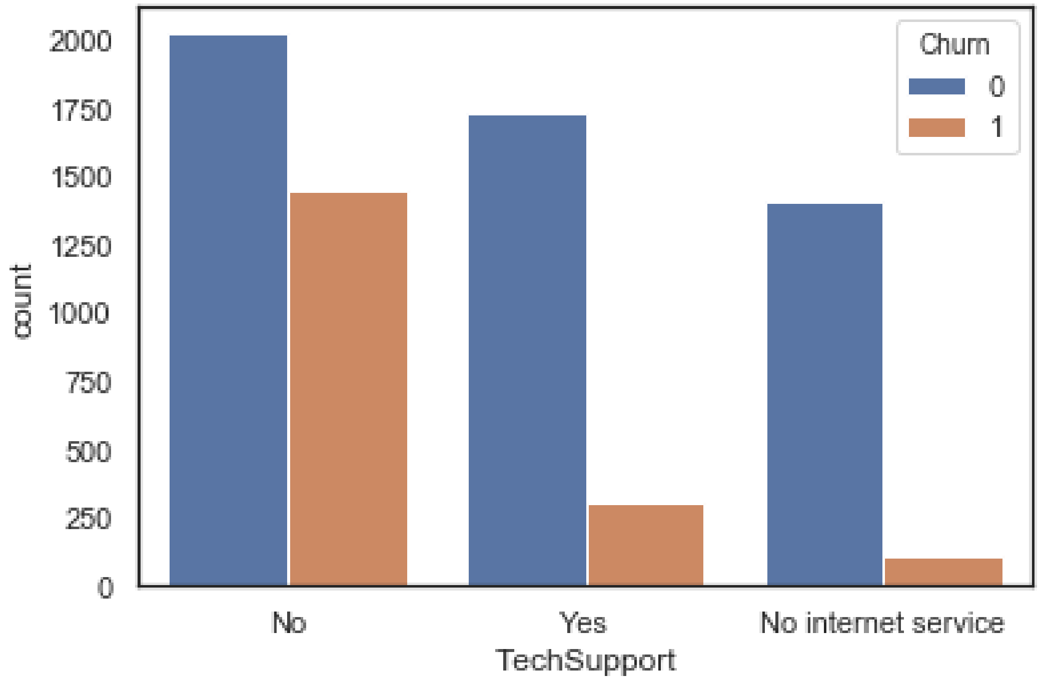
Churn
0
1
1200
1000
800
600
400
200
0
count
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod

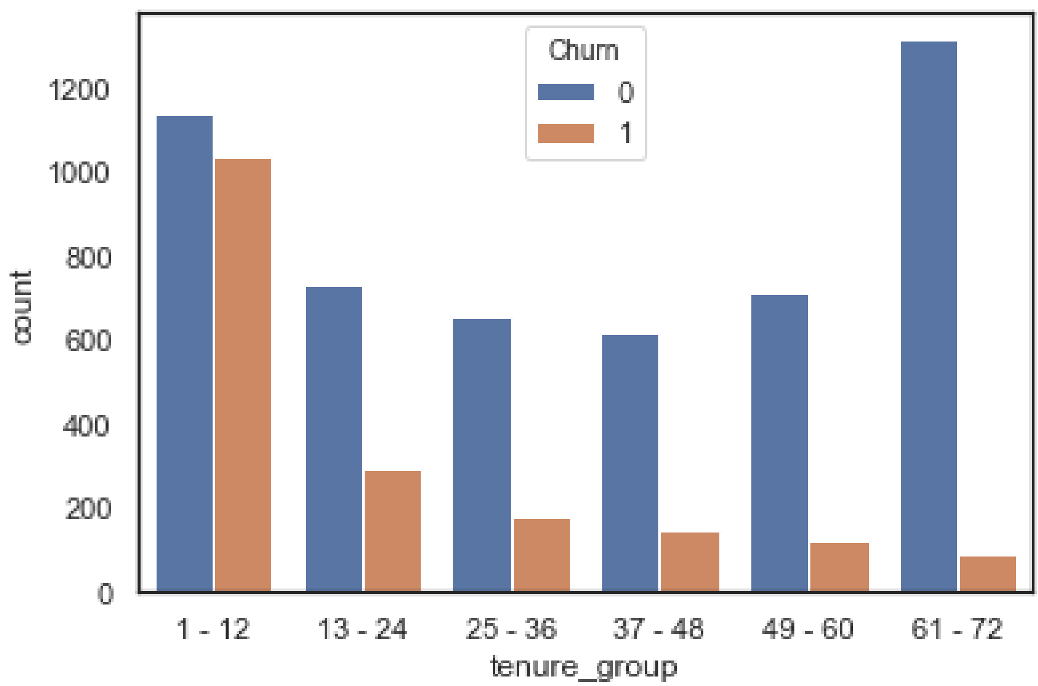
Churn
0
1
4000
3000
2000
1000
0
count
No
Yes
PhoneService


Churn
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
StreamingMovies

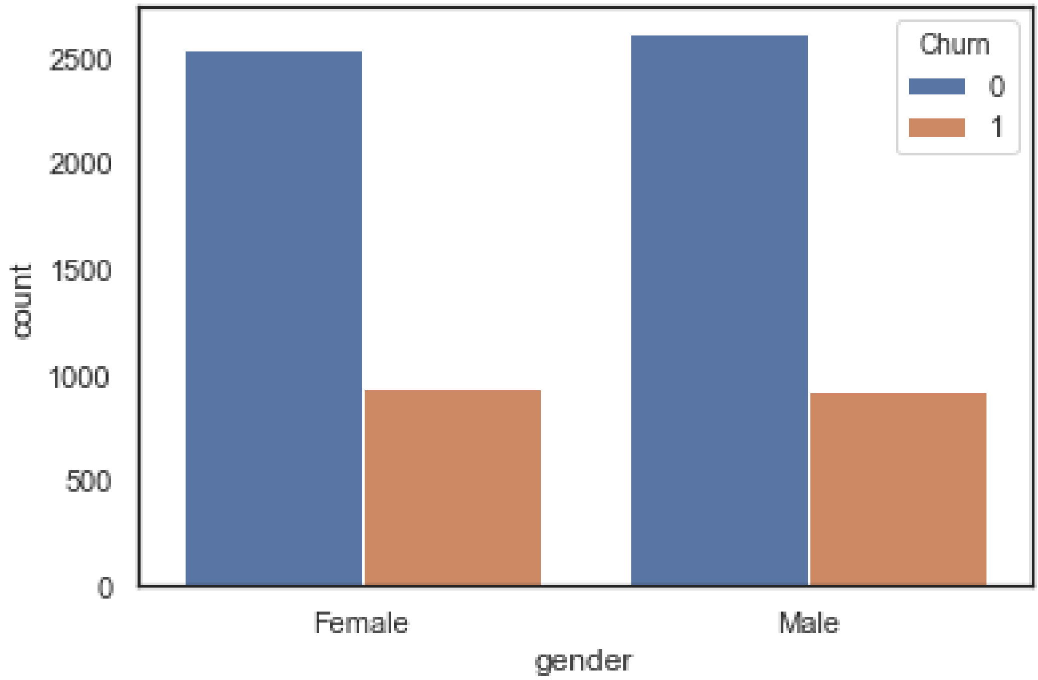

Churn
0
1
count
1750
1500
1250
1000
750
500
250
0
No
Yes
No internet service
StreamingTV

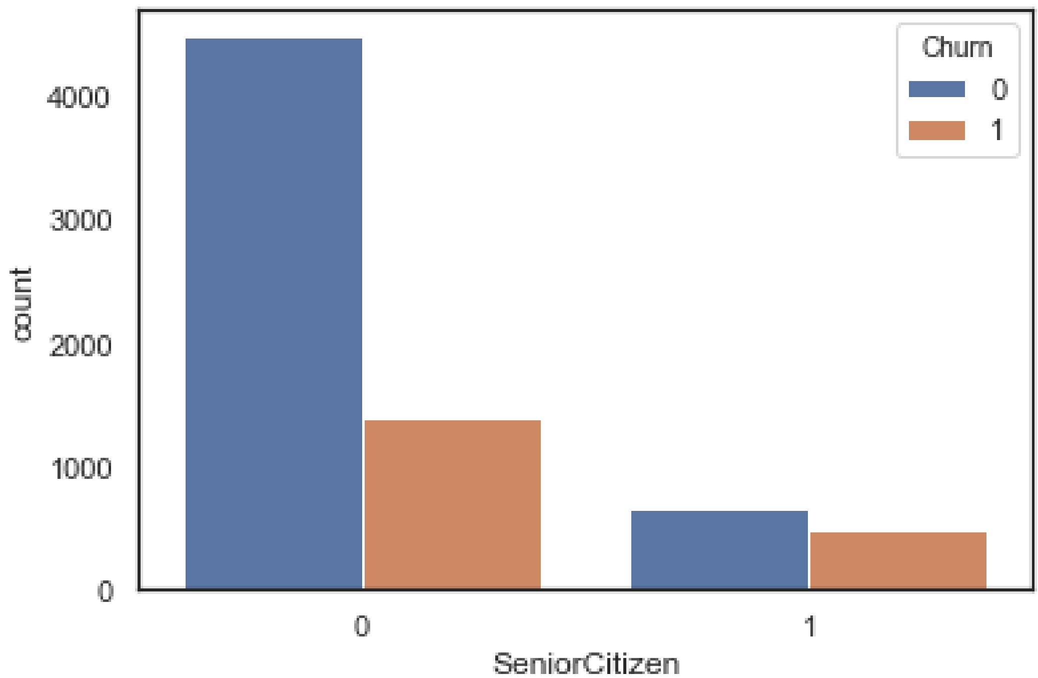


Churn
0
1
count
TechSupport
No
Yes
No internet service

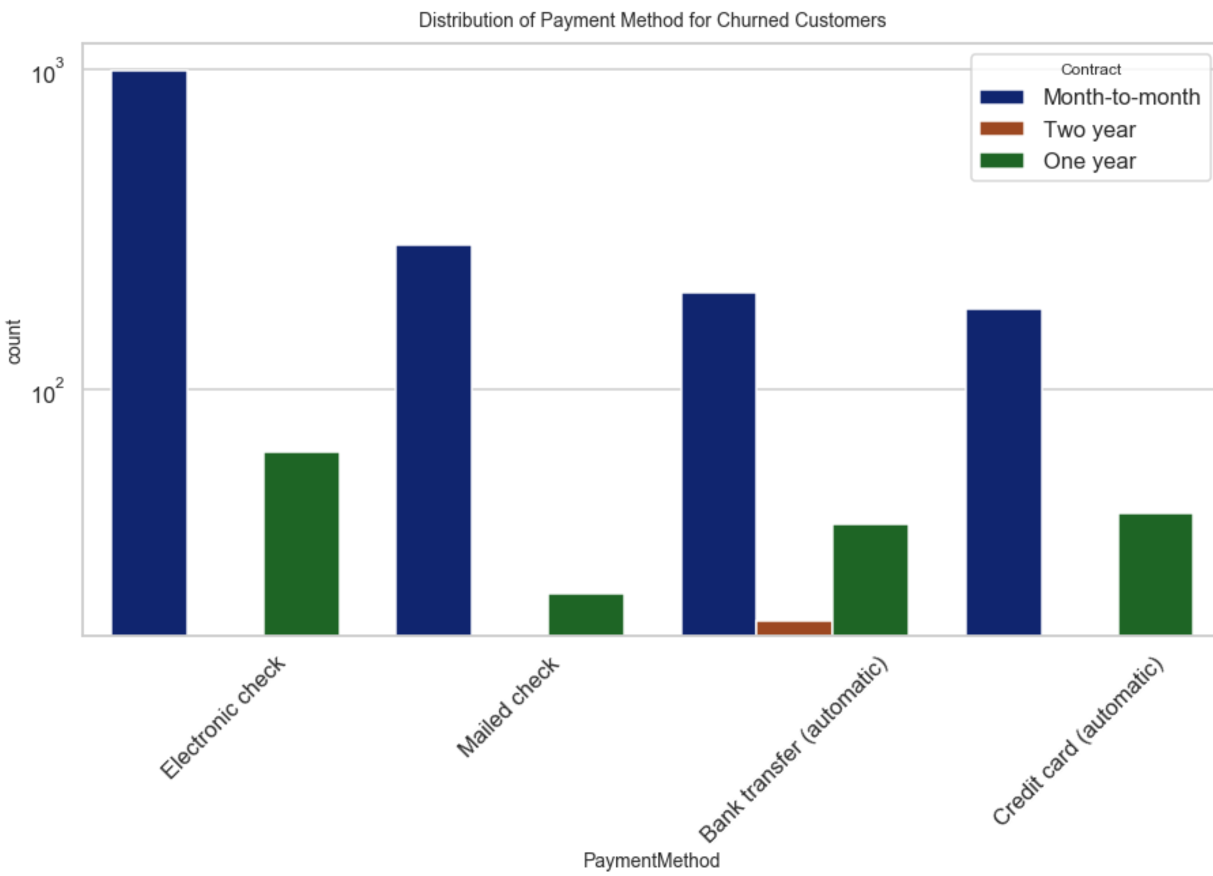



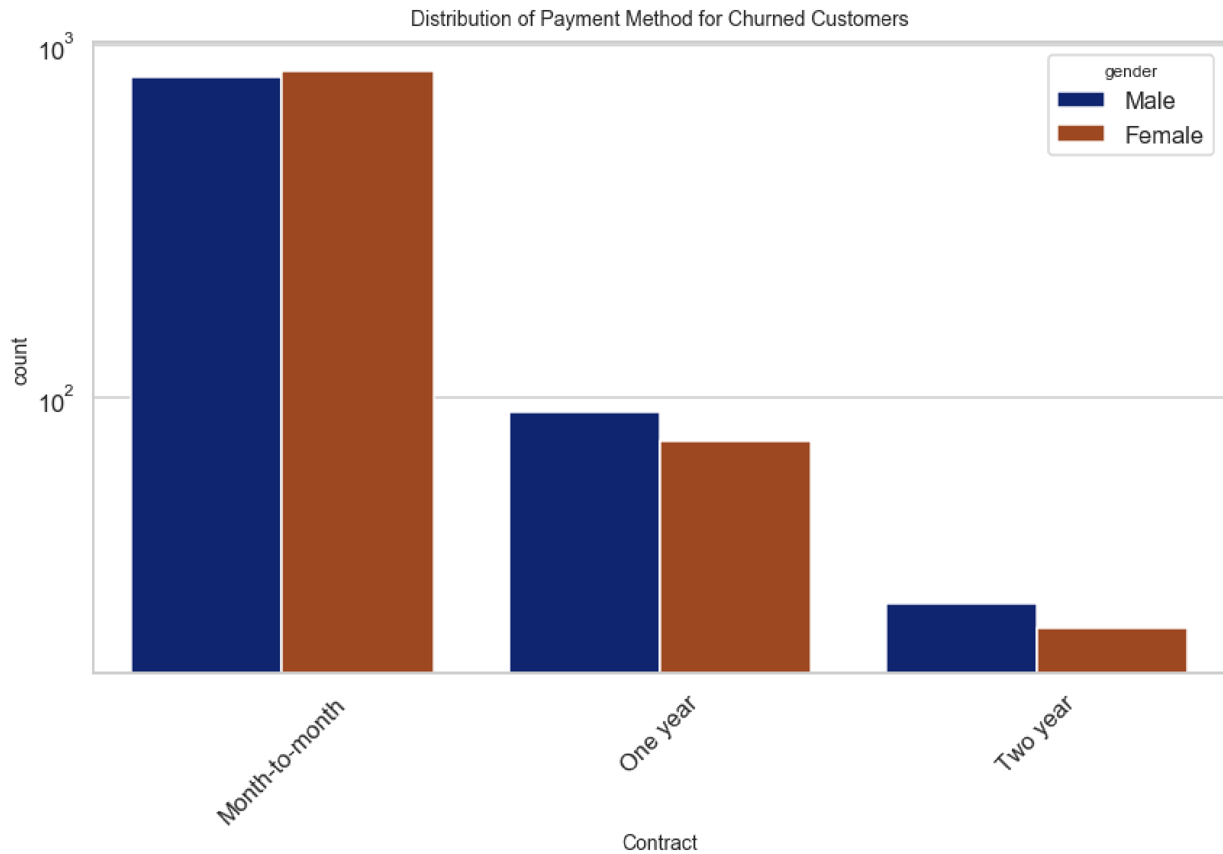
Churn
0
1
count
tenure_group
1 - 12
13 - 24
25 - 36
37 - 48
49 - 60
61 - 72


Churn
0
1
2500
2000
1500
1000
500
0
count
Female
Male
gender

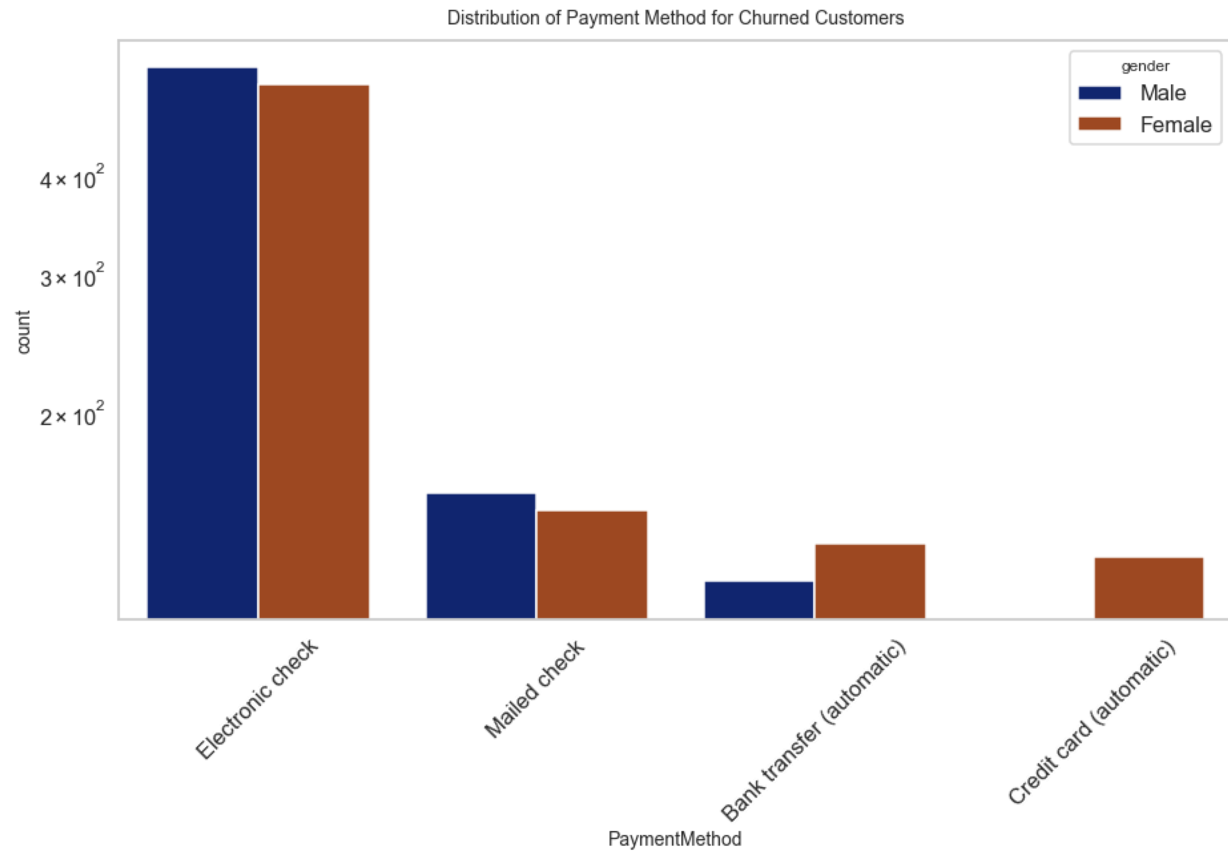

Churn
0
1
4000
3000
2000
1000
0
count
0
1
SeniorCitizen

**Bivariate Graphical Analysis**

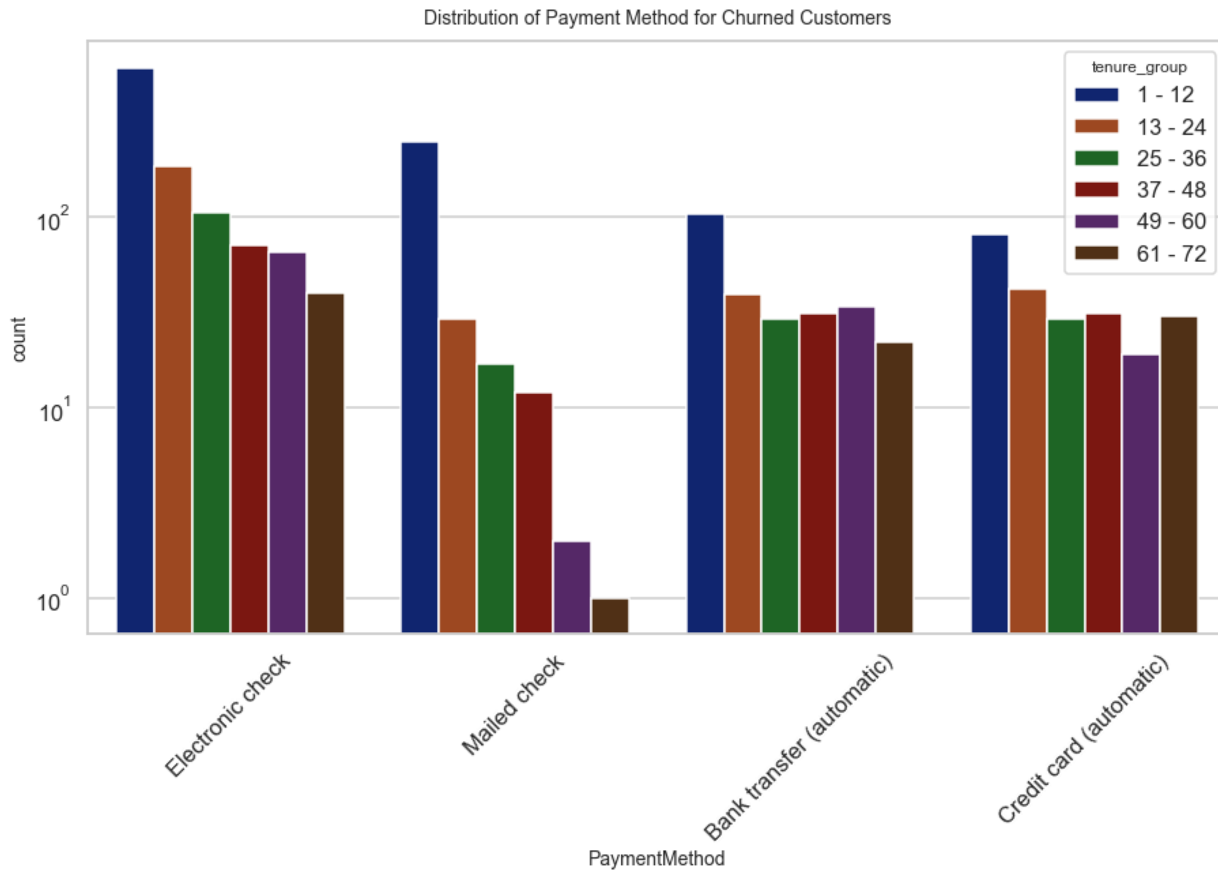




Distribution of Payment Method for Churned Customers
gender
Male
Female
count
$10^3$
$10^2$
Month-to-month
One year
Two year
Contract



Distribution of Payment Method for Churned Customers
gender
Male
Female
count
$4\times10^2$
$3\times10^2$
$2\times10^2$
Electronic check
Mailed check
Bank transfer (automatic)
Credit card (automatic)
PaymentMethod




**Conclusion**

1. Churn is directly linked with contract type. 43% of customers having month to month contract have churned whereas only 3% users having two-year contract churned
2. Customers having high monthly charges have churned whereas those having low monthly charges have stayed for longer duration
3. People paying with electronic check tend to churn more than any other customer type
4. Customers having No tech support tend to churn more often than customers availing tech support services
5. Through box plot we can identify that churn is inversely correlated to tenure of customer