Exploratory Data Analysis (EDA) and Descriptive Statistics for Telecom Users Data

<u>STATISTICS-I PROJECT</u>

Prepared by

Sandip Kumar Saha

(C21012)

Praxis Business School (PGPDS-21-JAN-KOL)

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to the Praxis Business School for providing me the opportunity to do an elementary level project of Statistics using Python and SPSS.First, I wish to express my special thanks to Prof Sayantani Roy Choudhury who has been very helpful throughout the project and for her enthusiasm,patience,helpful comments,practical suggestions and unceasing ideas that have helped me tremendously throughout the project.Her immense knowledge, profound experience and professional expertise in Data Quality Control has enabled me to complete this research successfully. Without her support and guidance, this project would not have been possible. I could not have imagined having a better supervisor in my study.

Finally, last but by no means least; also to everyone in the PGPDS-21-JAN batch. It was a great sharing experience with all of you during these three months.

Thanks for all the Encouragement!
Sandip K. Saha

MOTIVATION

With the rapid development of the telecommunication industry, the service providers are inclined more towards expansion of the subscriber base. To meet the need of surviving in the competitive environment, the retention of existing customers has become a huge challenge. In the survey done in the Telecom industry, it is stated that the cost of acquiring a new customer is far more than retaining the existing one. Therefore, collecting knowledge from the telecom industries can help in predicting the association of the customers as to whether or not they will leave the company. The required action needs to be undertaken by the telecom industries in order to initiate the acquisition of their associated customers to make their market value stagnant.

Due to the rapid growth in the data communication network and advancement in Information Technology, a massive amount of data is available. With the increase in the competition in the market, companies have devoted their time more in making their previous clients associated with them rather than convincing the new clients. Since the major source of profit is customers, customer churn plays a significant role in the survival and development of the telecommunication industry.

INTRODUCTION

Any business wants to maximize the number of customers. To achieve this goal, it is important not only to try to attract new ones, but also to retain existing ones. Retaining a client will cost the company less than attracting a new one. In addition, a new client may be weakly interested in business services and it will be difficult to work with them, while old clients already have the necessary data on interaction with the service.

Accordingly, predicting the churn, we can react in time and try to keep the client who wants to leave. Based on the data about the services that the client uses, we can make him a special offer, trying to change his decision to leave the operator. This will make the task of retention easier to implement than the task of attracting new users, about which we do not know anything yet.

I have downloaded the data file in csv format from Kaggle as it provides reliable data of large volumes and is perfect for the Analysis.The data contains information about almost six thousand users, their demographic characteristics, the services they use, the duration of using the operator's services, the method of payment, and the amount of payment.

Customer churn rate =No of customers lost(over a time period)/total customers(at the beginning of the time period)*100.

Our goal is to extract as much information as we can and derive insights from it using Exploratory Data Analysis i.e. descriptive Statistics.I have also explored the correlation between different features to find the dependency using SPSS.

DATASET DESCRIPTION

The data has 5987 entries and 22 features.It has no null values so we don’t require any data cleaning though it is an important part in data analysis.

Following are the salient features of the data set:

customerID - customer id

gender - client gender (male / female)

SeniorCitizen - is the client retired (1, 0)

Partner - is the client married (Yes, No)

tenure - how many months a person has been a client of the company

PhoneService - is the telephone service connected (Yes, No)

MultipleLines - are multiple phone lines connected (Yes, No, No phone service)

InternetService - client's Internet service provider (DSL, Fiber optic, No)

OnlineSecurity - is the online security service connected (Yes, No, No internet service)

OnlineBackup - is the online backup service activated (Yes, No, No internet service)

DeviceProtection - does the client have equipment insurance (Yes, No, No internet service)

TechSupport - is the technical support service connected (Yes, No, No internet service)

StreamingTV - is the streaming TV service connected (Yes, No, No internet service)

StreamingMovies - is the streaming cinema service activated (Yes, No, No internet service)

Contract - type of customer contract (Month-to-month, One year, Two year)

PaperlessBilling - whether the client uses paperless billing (Yes, No)

PaymentMethod - payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

MonthlyCharges - current monthly payment

TotalCharges - the total amount that the client paid for the services for the entire time

Churn - whether there was a churn (Yes or No) : Predictor Class

Exploratory Data Analysis

Getting started:

Initially the raw data had 5986 columns and 22 columns.There is one column which shows customer number and i removed that column as it has no relevance for data analysis.Basically it is data cleaning process.Now it has 21 rows.There are some categorical features like-Partner,Dependents,PhoneService,MultipleLines,Internet Service(which has one categorical variable No and other two types are Fibre Optic and DSL),OnlineSecurity,OnlineBackup,DeviceProtection,Tech Support,streaming TV,streaming Movies,Paperless billing and finally the target variable i.e. churn(whether the subscriber renew subscription or not)

There are no missing values and I have applied dummification for the categorical values (Yes=1 and No=0).The feature Senior Citizen includes 1 for yes and 0 for not senior citizen is inbuilt in the data itself.

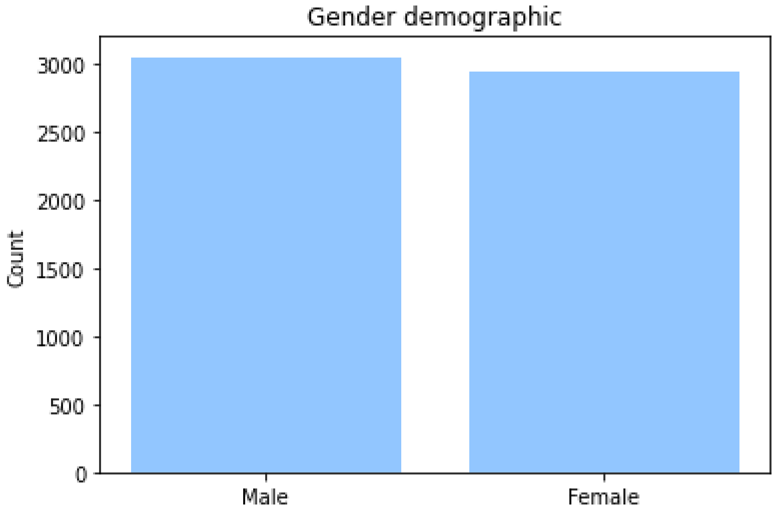
Here is the snapshot of the data information after cleaning and dummification.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5986 entries, 0 to 5985
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype    Note: each row has 5986
---  ------            --------------  -----         entries and 21
 0   customerID        5986 non-null   object        columns.All other
 1   gender            5986 non-null   object    information is given.
 2   SeniorCitizen     5986 non-null   int64
 3   Partner           5986 non-null   int64
 4   Dependents        5986 non-null   int64
 5   tenure            5986 non-null   int64
 6   PhoneService      5986 non-null   int64
 7   MultipleLines     5986 non-null   object
 8   InternetService   5986 non-null   object
 9   OnlineSecurity    5986 non-null   int64
 10  OnlineBackup      5986 non-null   int64
 11  DeviceProtection  5986 non-null   int64
 12  TechSupport       5986 non-null   int64
 13  StreamingTV       5986 non-null   int64
 14  StreamingMovies   5986 non-null   int64
 15  Contract          5986 non-null   int64
 16  PaperlessBilling  5986 non-null   int64
 17  PaymentMethod     5986 non-null   object
 18  MonthlyCharges    5986 non-null   float64
 19  TotalCharges      5986 non-null   object
 20  Churn             5986 non-null   int64  dtypes: float64(1), int64(14), object(6)
```

Analysis:

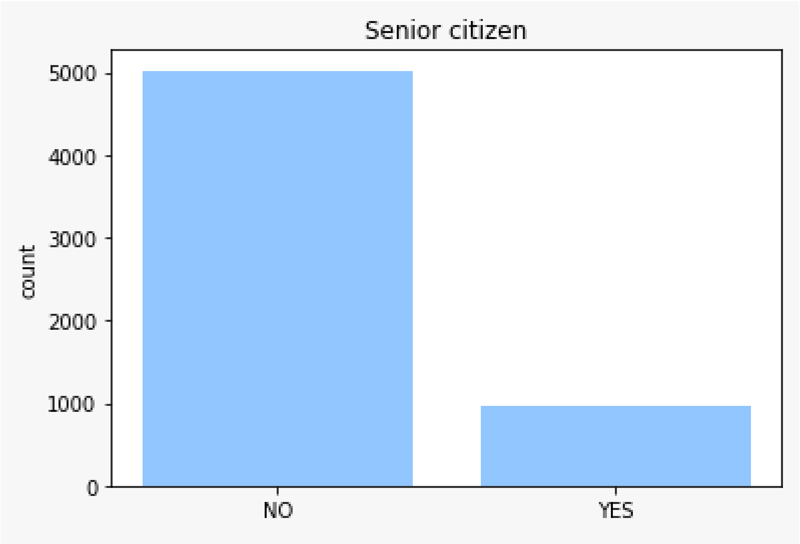
At first we will explore the relationship between features except with the target variable (i.e. Customer churn)to get a basic understanding of the Telecom Users data like-demographics,whether client uses phone service or not etc.

Gender Demographics:



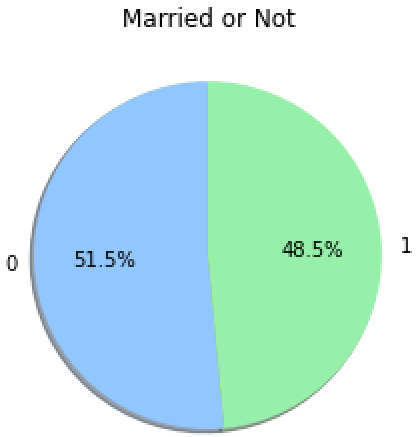
We can see that male customers are more than female.

Whether the client is senior Citizen Or not:



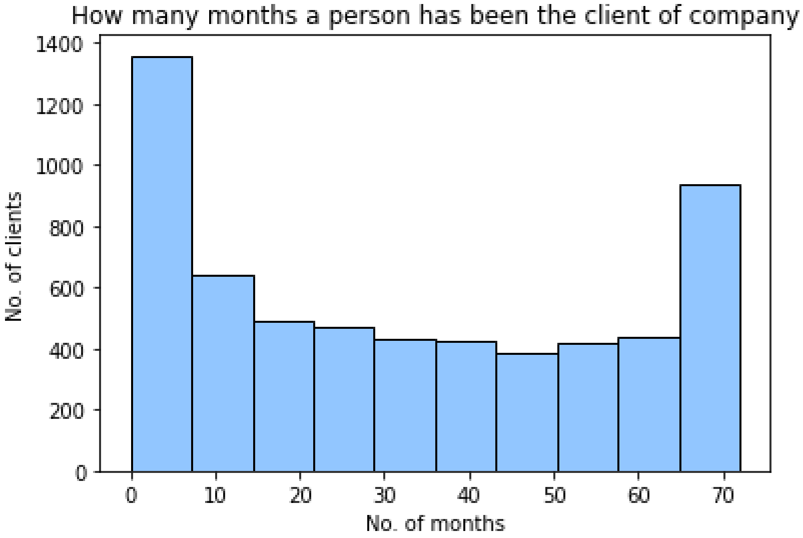
16.14% of clients are senior citizens.

How many clients are married?



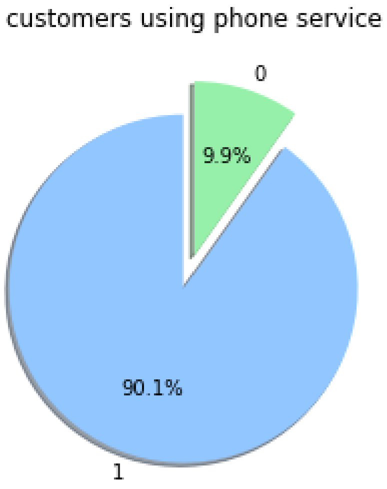
48.5% of customers are married.

Customer Tenure(Histogram):



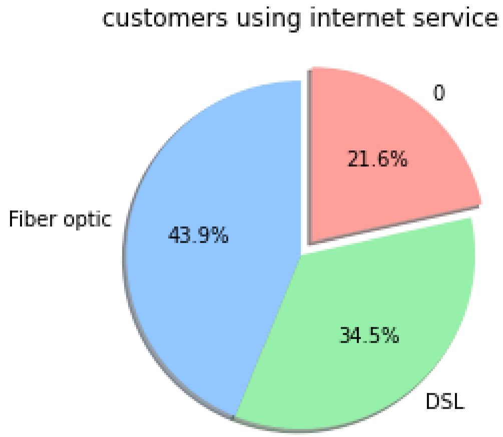
Max customer tenure for the company is 72 months and average tenure is 32 months.We can see after 7 months there has been a sharp decline of customers.

Number of customers using Phone Service:



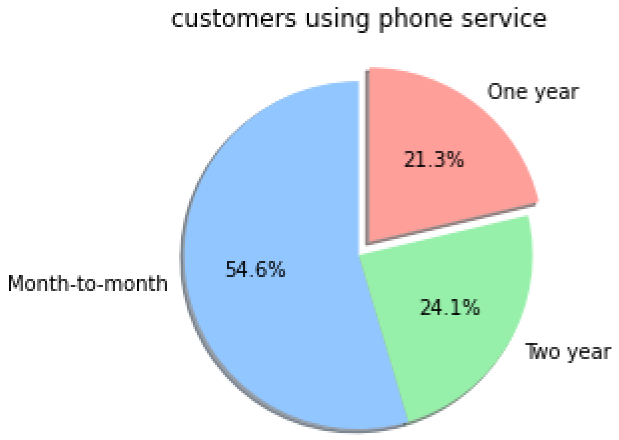
So the majority of customers are using phone service of the respective company.

Number of customers using Internet Service:



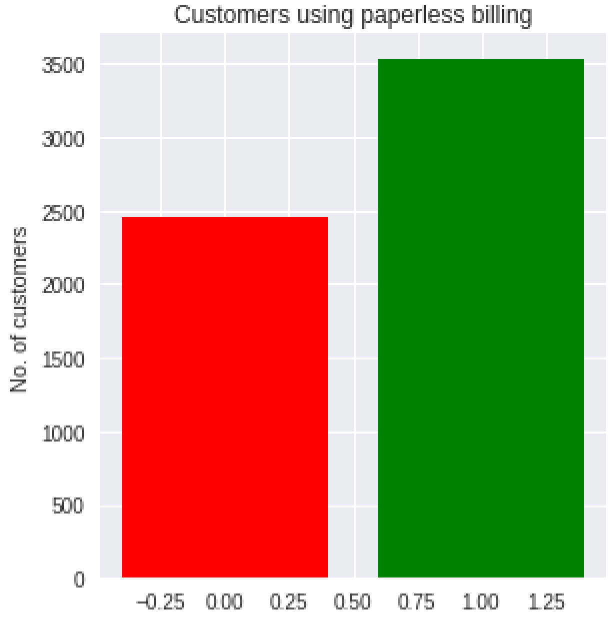
It is clearly shown that 43.9% of the customers have Fiber optic connection,34.5% customers have DSL connection and the rest have no internet service.

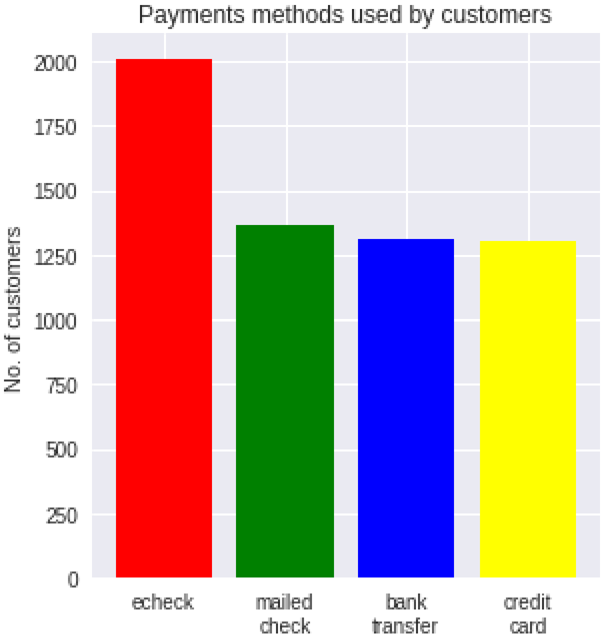
Which type of Contract is most preferred by the client?



We can infer from the Pie-Diagram that 54.6% of the customers prefer the monthly contract for the phone service and there is approximately a 3% increase in the two year plan compared to one year plan.

Next we can check how many customers are using paperless billing and what kind of payment methods are they using if the payment mode is online:

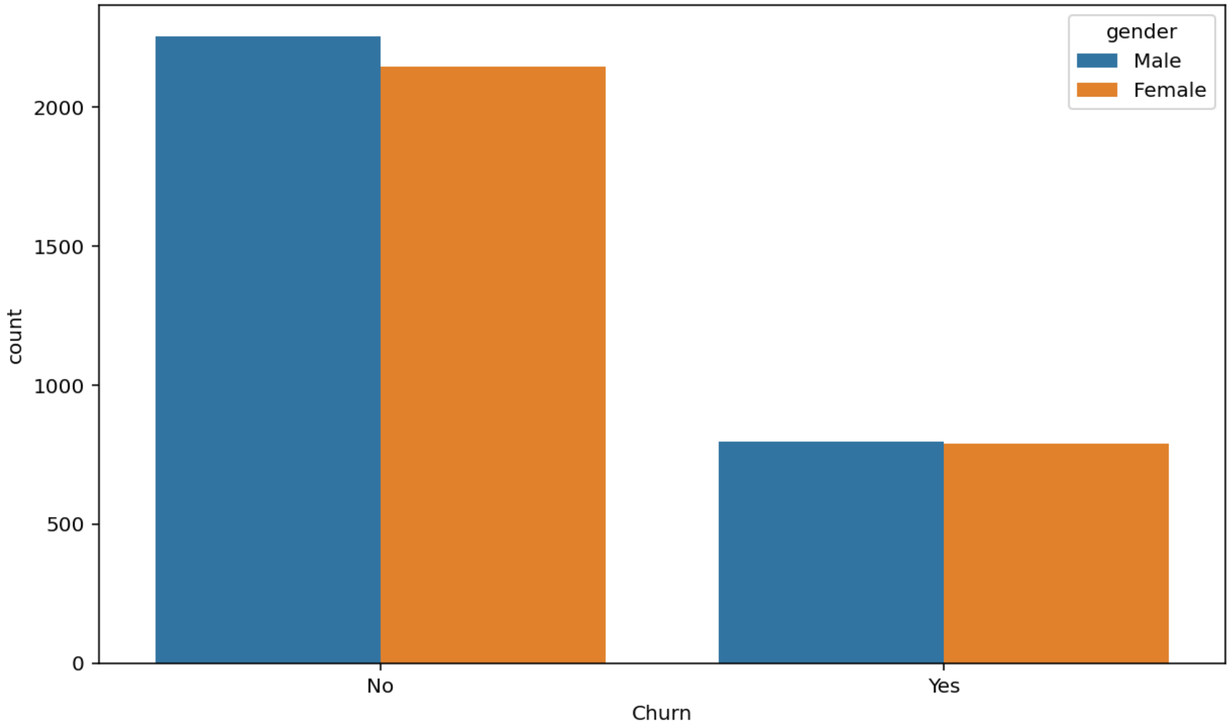




- Approximately 40% customers are using paperless billing and the remaining 60% i.e. the majority of them are still using paper billing.
- Among the payment methods there are typically four kinds of mode where the customers are divided into.
- Like-credit card Automatic,Bank Transfer(Automatic),Electronic Check and Mailed Check.
- It is interesting to note that 33.41% of customers are using the Electronic Check method to pay their bills.

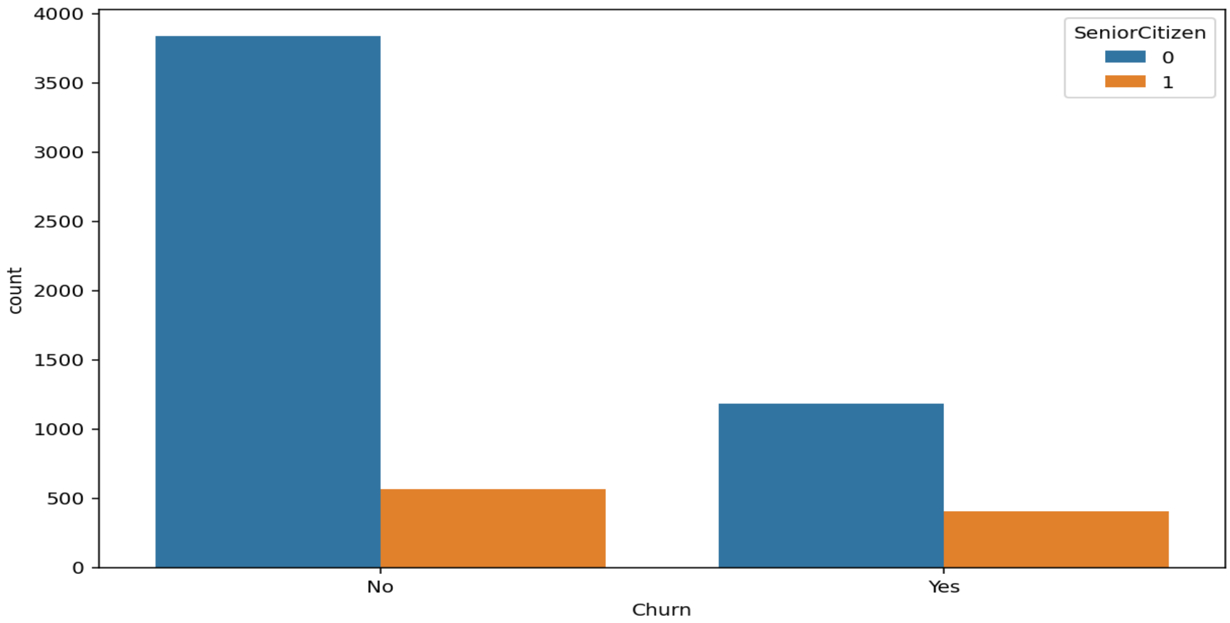
As we are going to predict the Churn patterns of telecom users, let's explore the relationship between various features with Predictor Class Churn.

1. Relationship of churn rate with gender



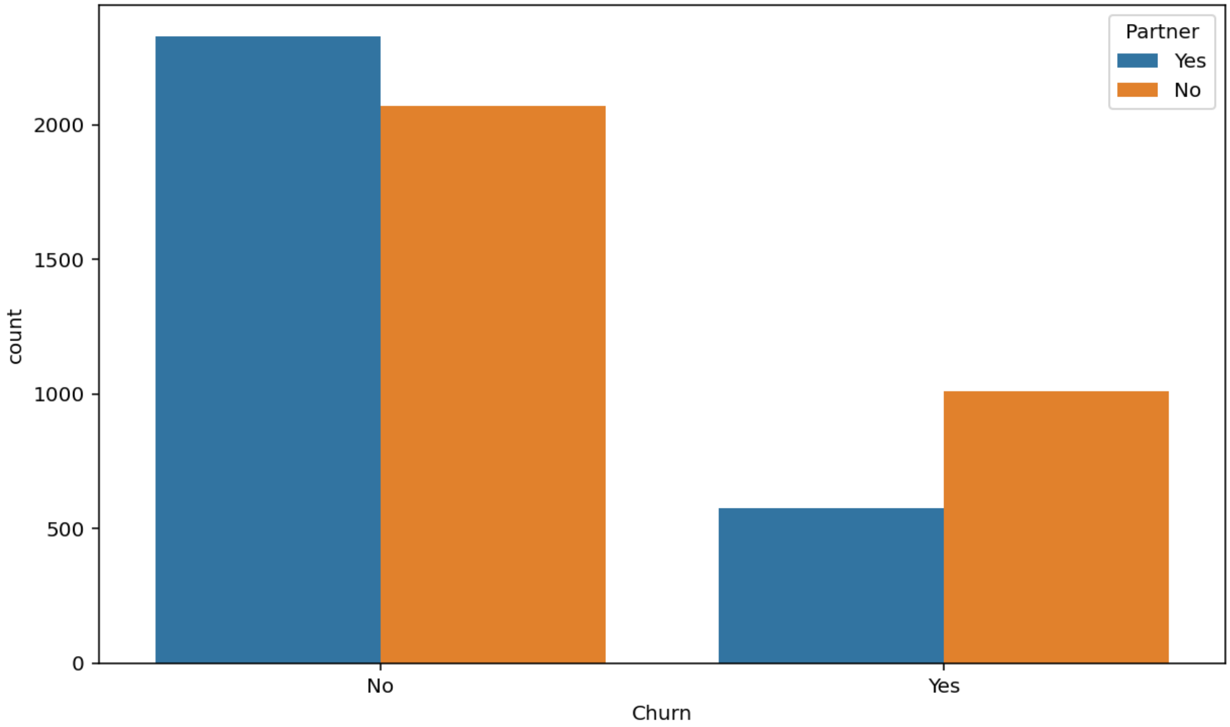
As we can see that in terms of customer churn(No or Yes) there is no major difference between male and female i.e. distribution is symmetrical except male Non-churn rate is slightly higher than female.

2.Relationship of churn rate with senior Citizen:



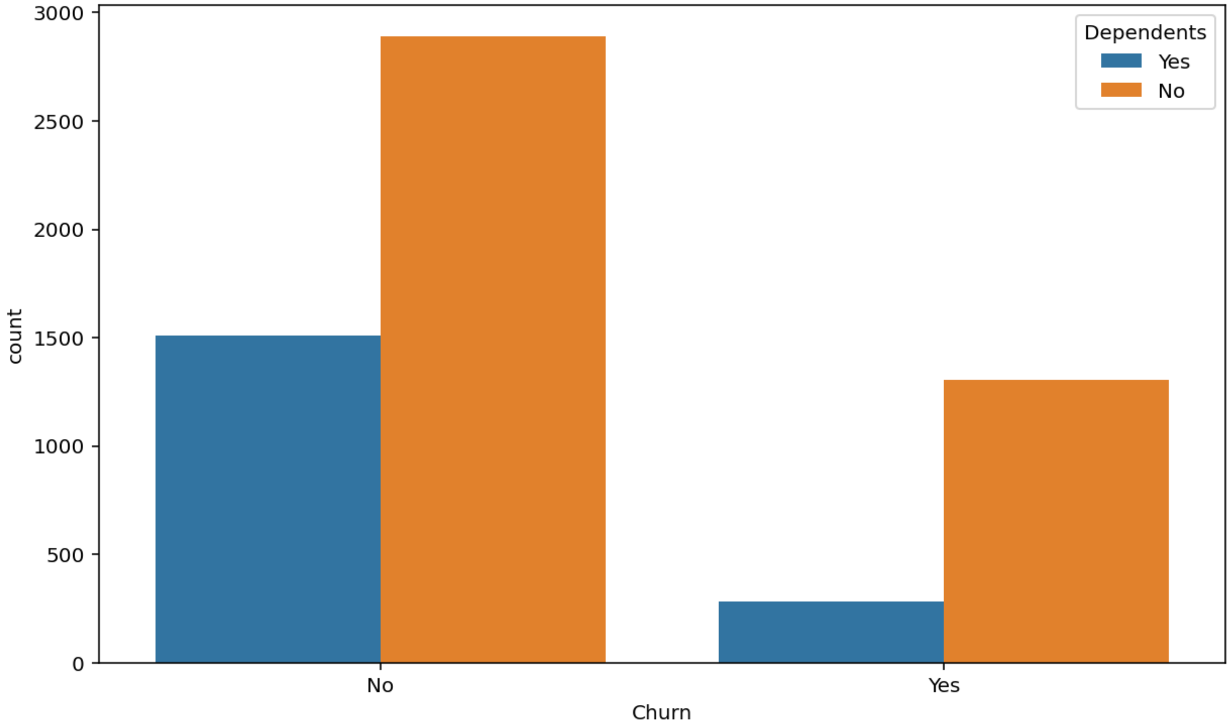
Here we can see the non-attrition rate among senior citizens is higher compared to the adults and the same trend goes with the attrition rate also.

3.**Relationship between churn rate and partner:**



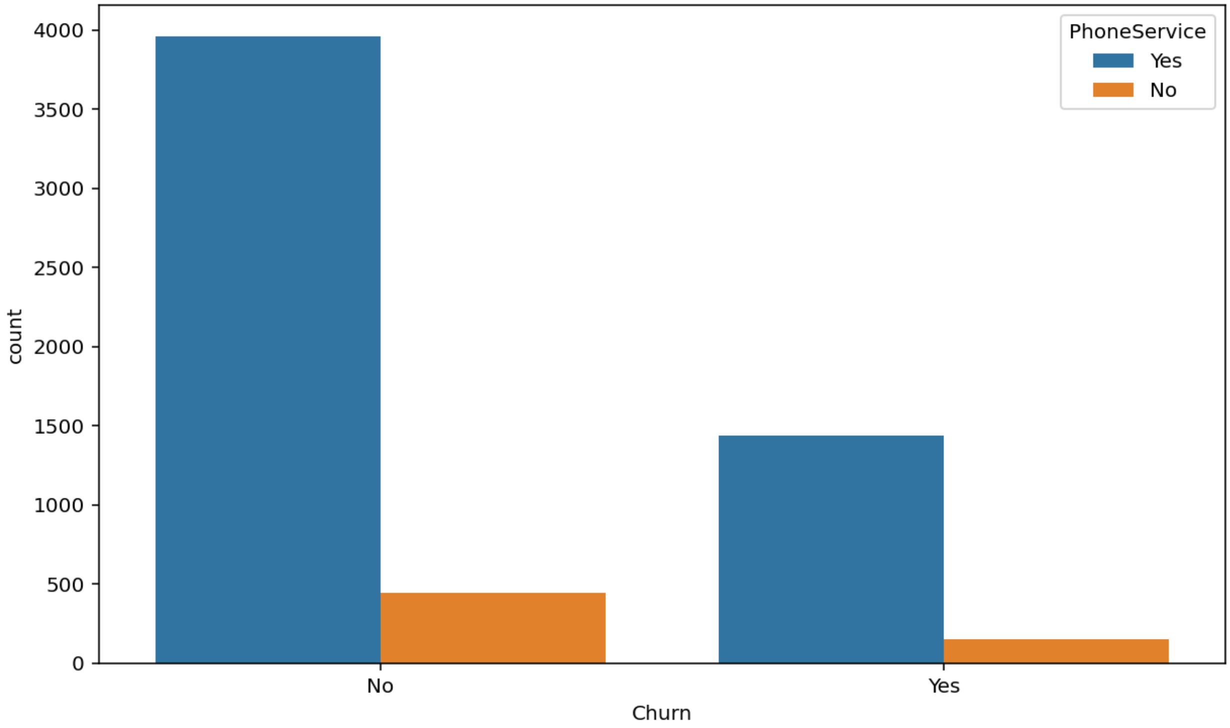
This case the non-attrition rate of the customers who are married or have a partner is more than the single customers and it is interesting to note that the attrition rate of customers who are single more than the later.

4.**Relationship between churn rate and Dependents:**



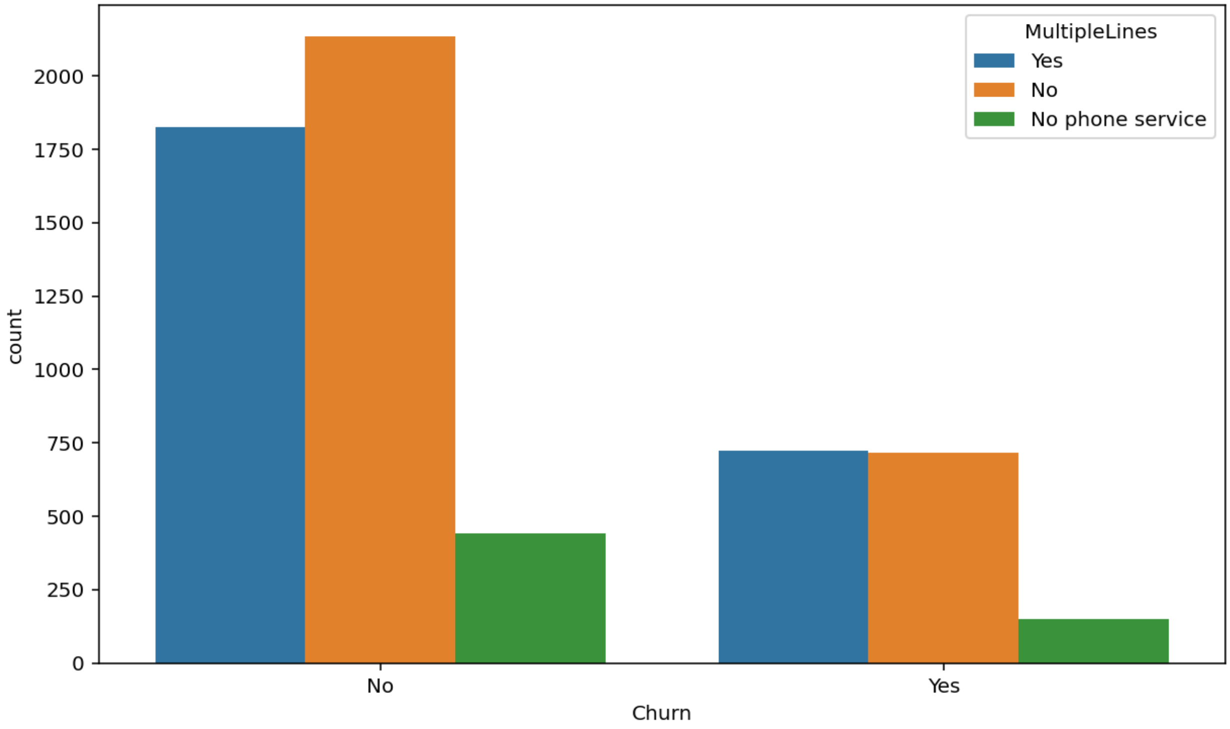
Nothing to say particular in this case as both churn rates are inversely related if they have no dependents.

5.Relationship between churn rate and Phone Service:



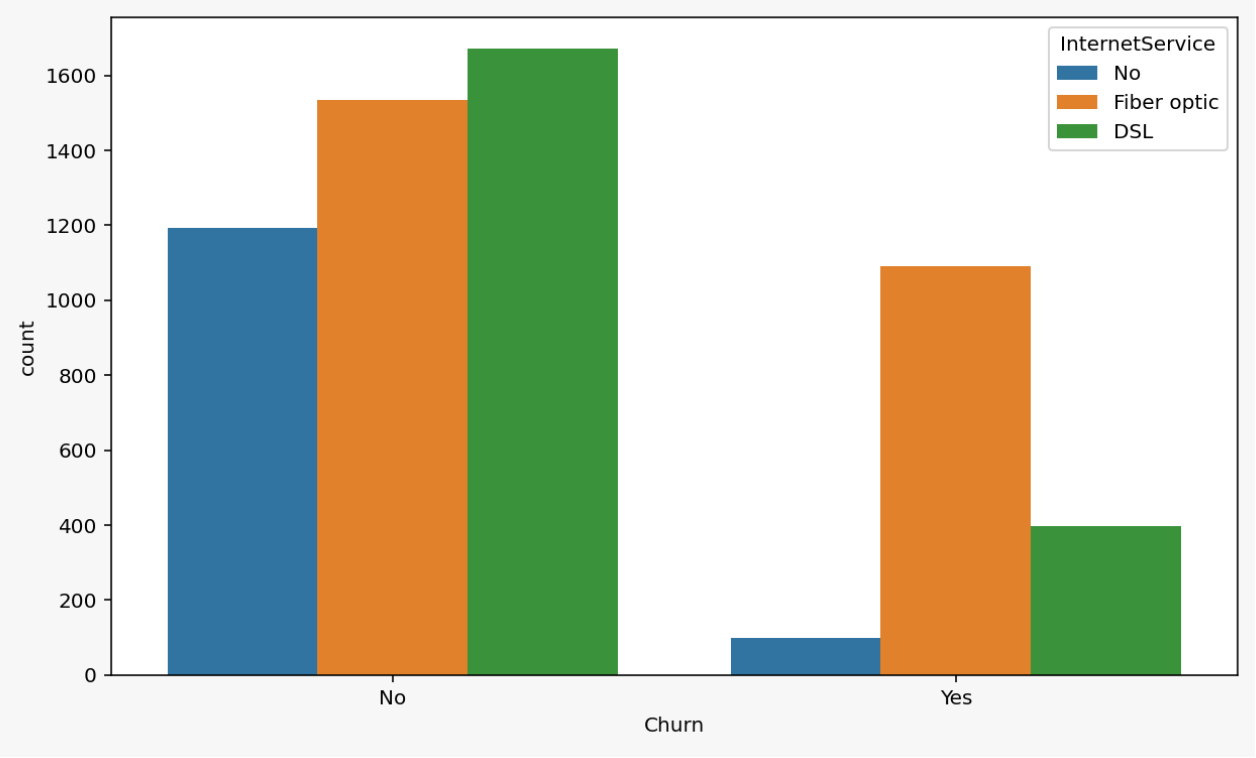
Both churn and non-churn rate are linearly proportional with Phone service i.e. those customers who have included phone service with this particular telecom operator are less likely to cancel their subscription.Approximately 65% customers who included phone service with internet don’t likely to churn in near term.

6.Relationship between churn rate and Multiple lines connection:



It can also be thought logically that the customers who did not have multiple lines are more likely to stay with the service.

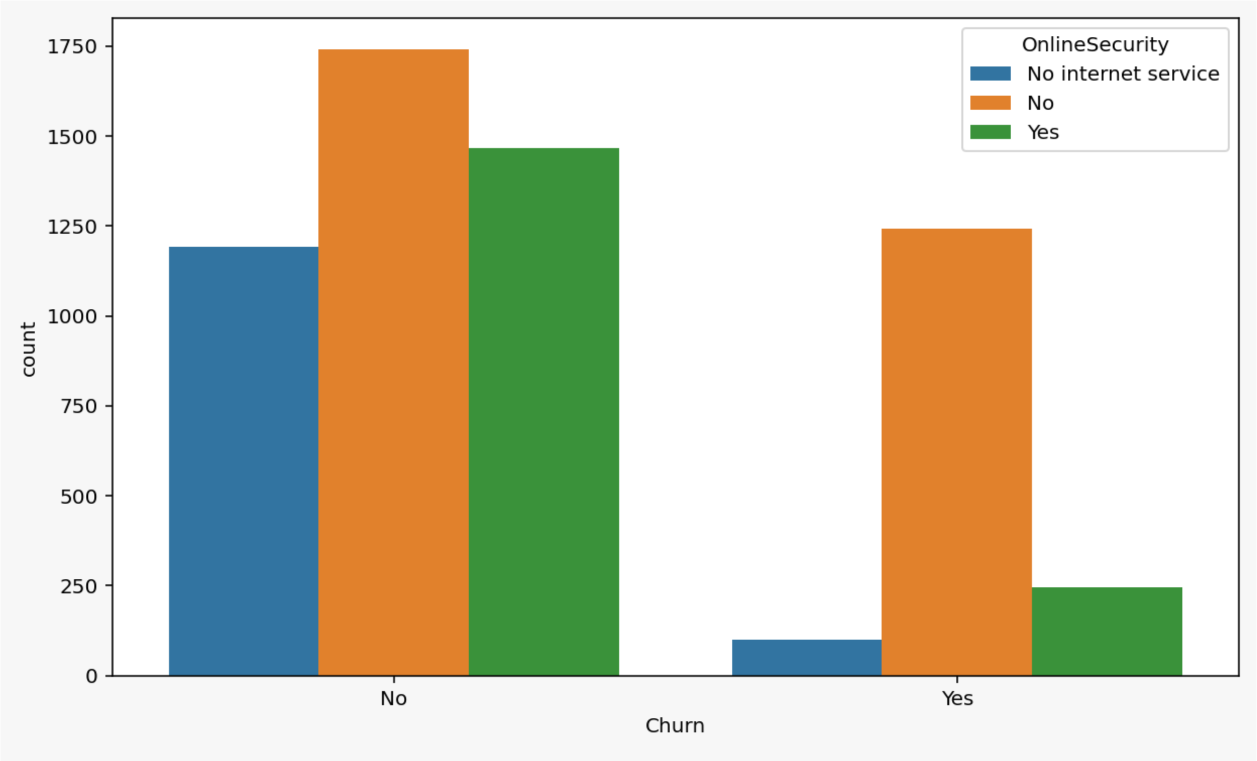
7.Relationship between attrition rate and internet service:



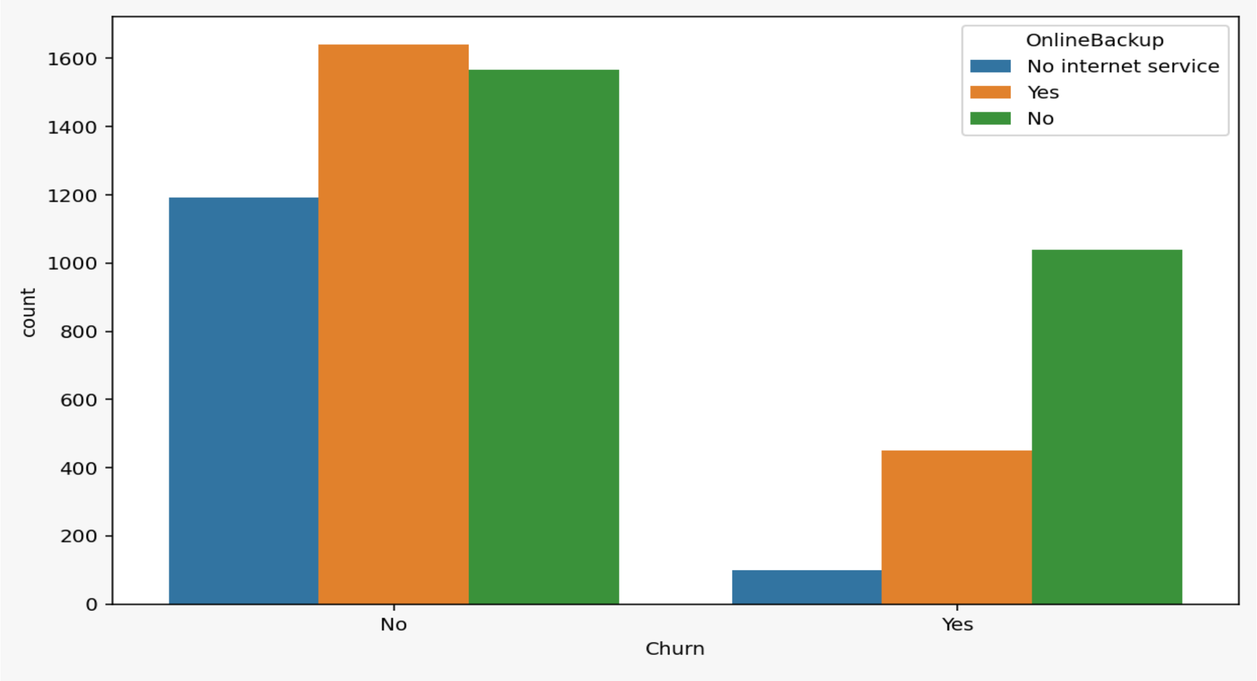
We can get very useful information that the customers who have fiber optic connection are more likely to cancel their internet service but it is opposite in case of DSL.

8.Relationship between Churn rate and online security:

- We can infer from the below diagram that customers who have not included online security service with their telecom operator are likely to withdraw from the service but astonishingly the customers retained by the company have not included Online security as a service.
- So the company needs to improve its Online Security service in the near future to attract more customers as the target of the company to be an integrated service provider.
- But 23% of customers have opted for the online security service compared to 28% who did not.



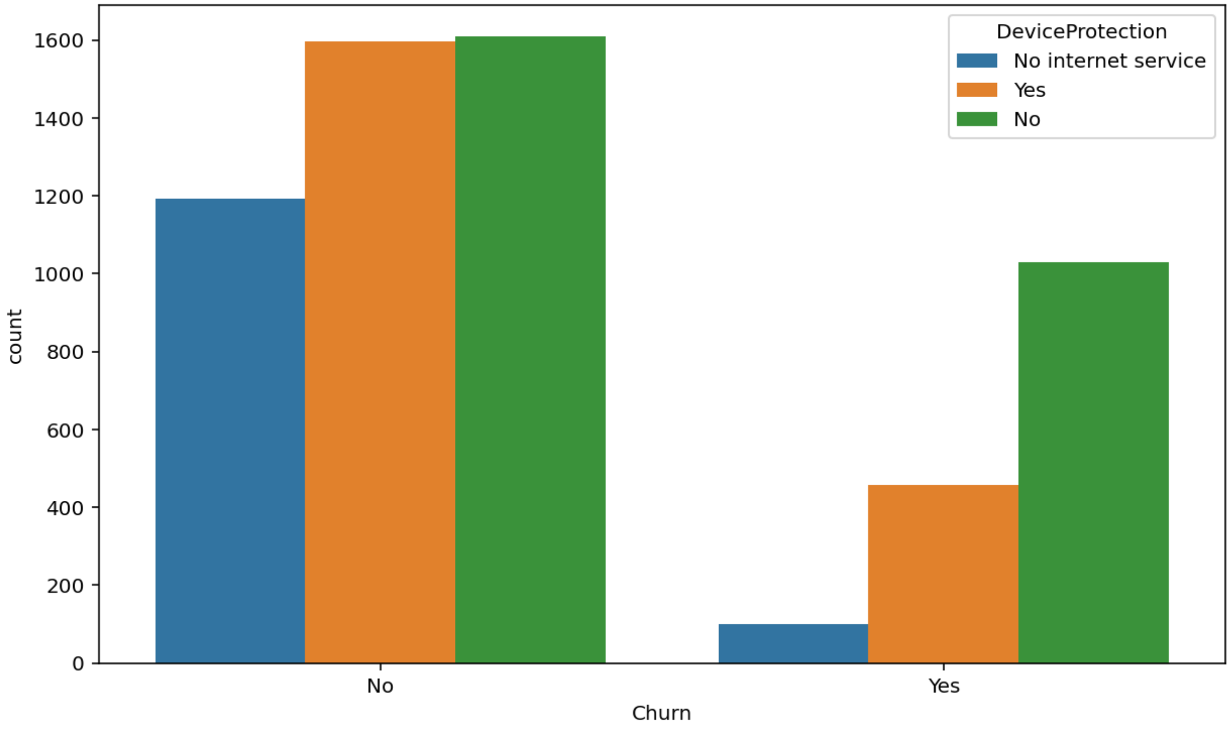
9.Relationship between churn rate and online backup:



- Approximately 27% of the customers are using online backup service and 25% of the customers have not opted for this service.As it is calculated for those who are likely to stay with the company, we can conclude that some customers found this service very important compared to others.

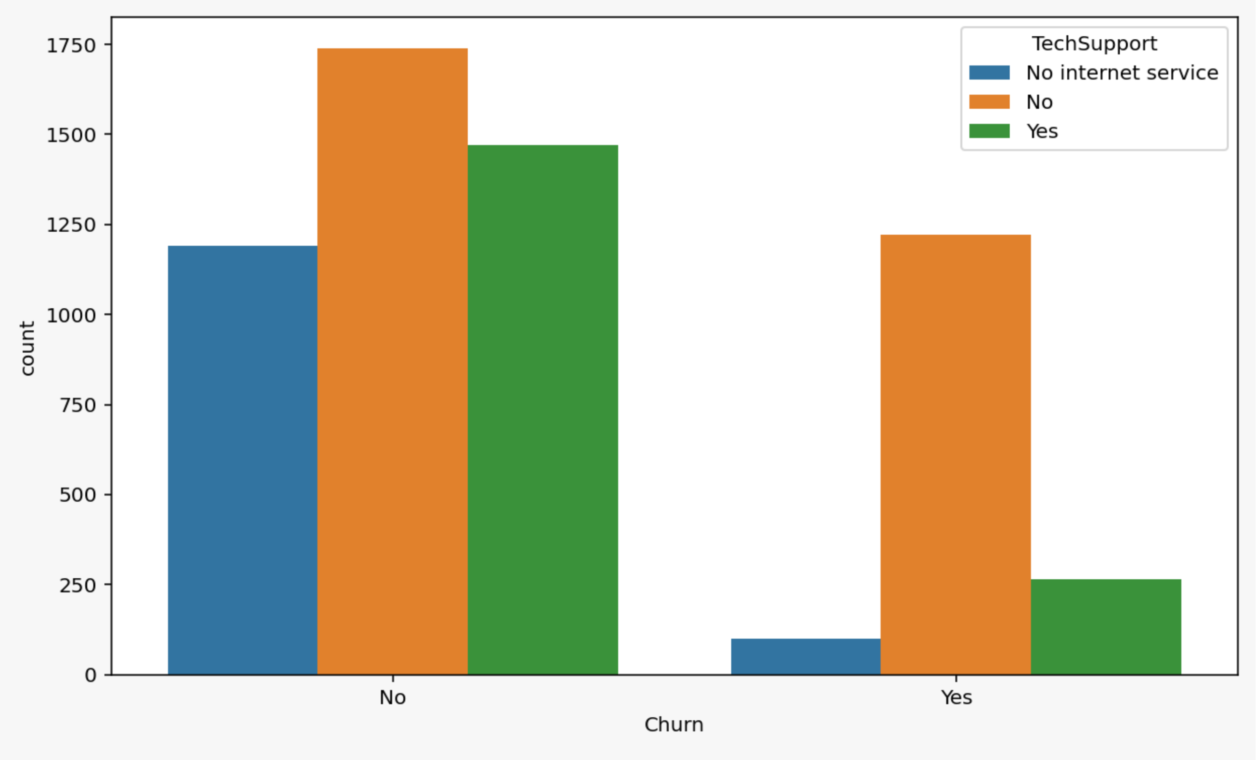
- Majority of the churned customers have not included online backup service (18%).

10.Relationship between churn rate and device protection:



- We find a mix of insights in case of customers who have included device protection.
- Close to 25% customers who didn't churn uses device protection and 26% of them don't use device protection

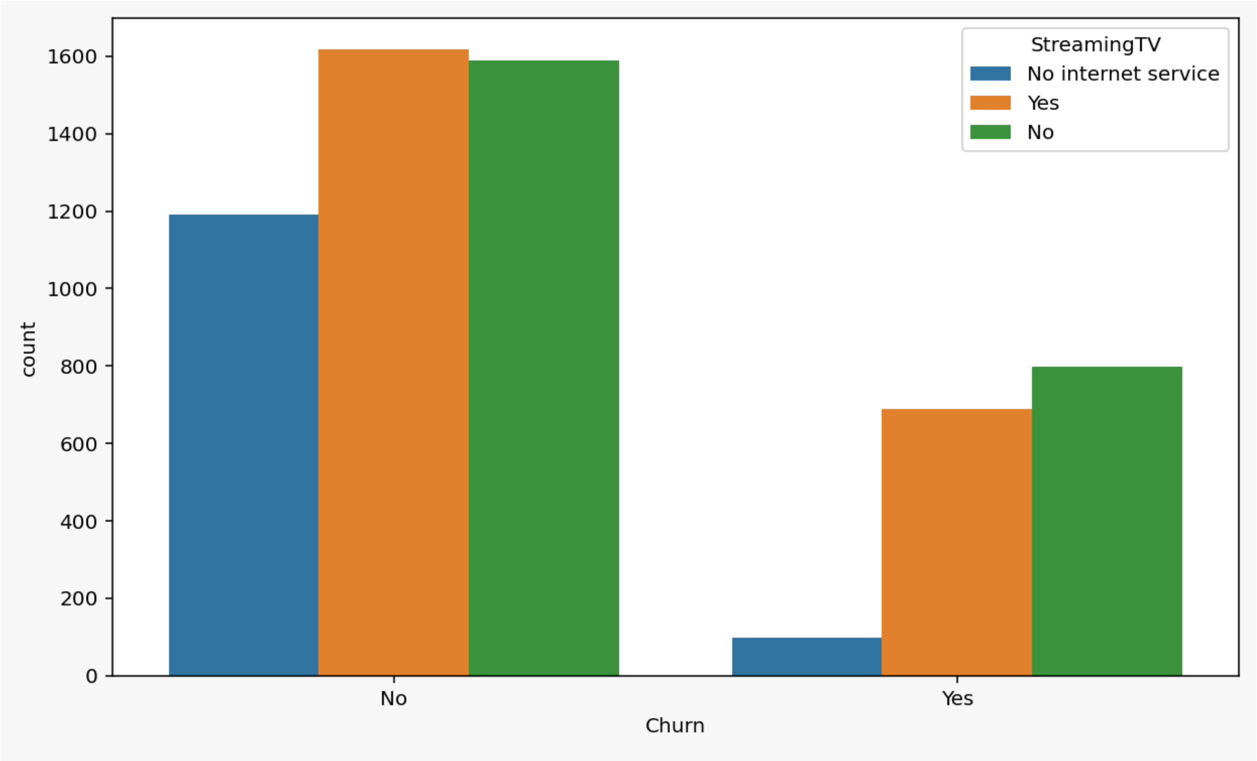
11.Relationship between churn rate and tech support:



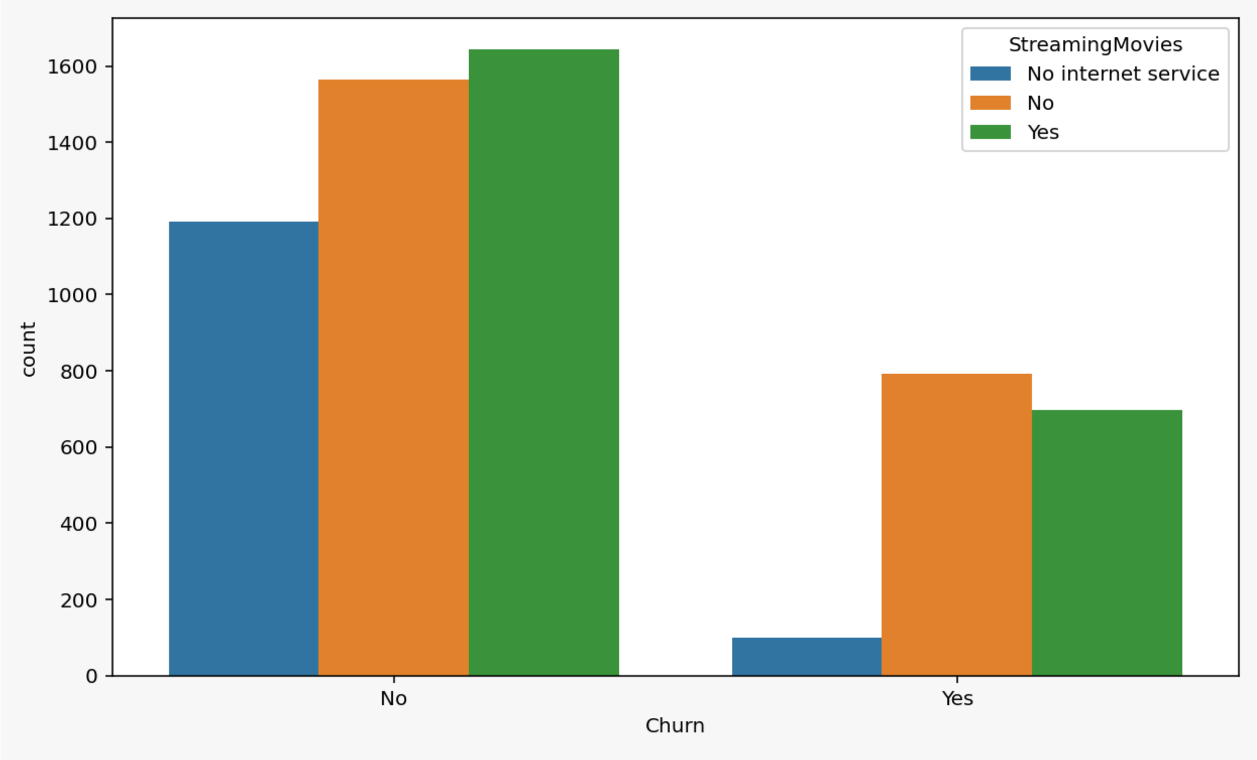
- Close to 28% customers have not availed tech support compared to 23% who subscribed tech support.(in case of retained customers)
- 4% customers who churned had also included device protection as a service.

12.Relationship between churn rate and streaming TV service:

- Churn : Yes
- 11% of customers did use streaming TV services with the telecom operator and 12 % of the customers didn't use streaming TV service.
- Churn : No
- There is a very thin margin who use and don’t use this service and the difference may be less than 1% that means this service is more of a choice for the customer.

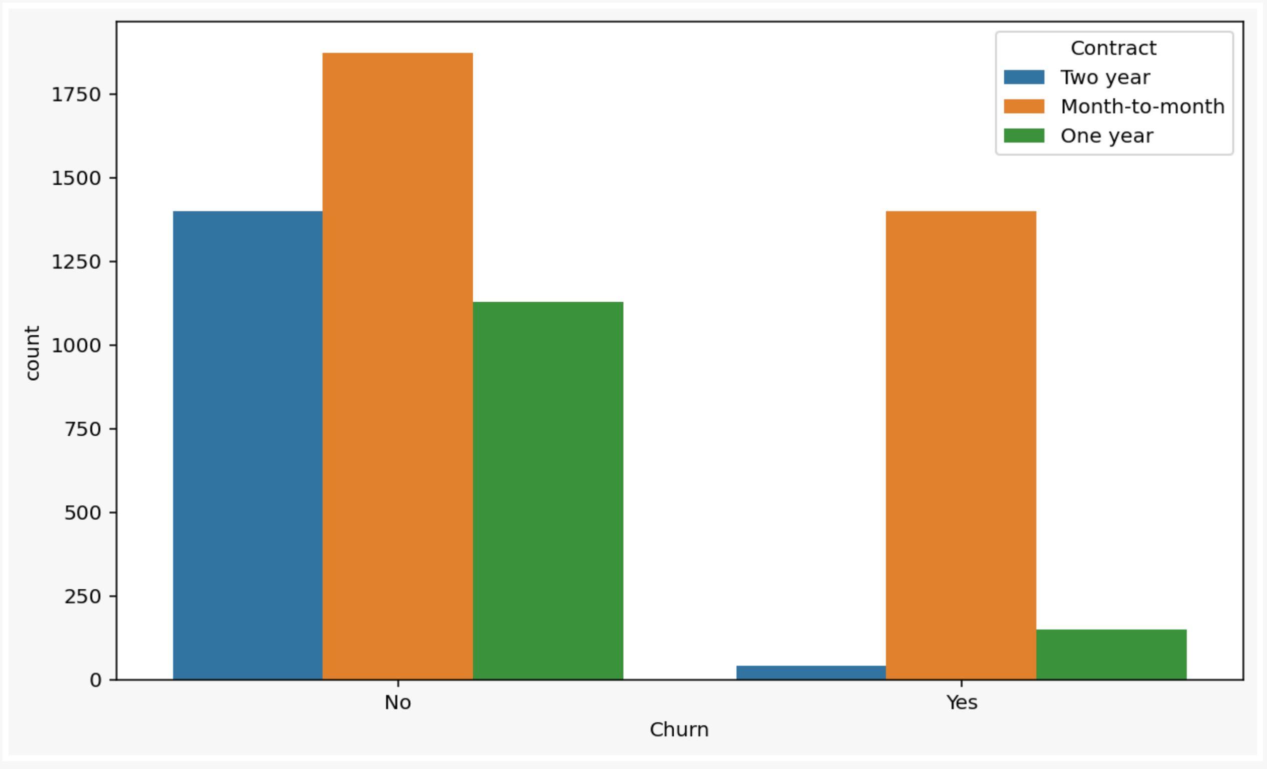


13.Relationship between churn rate and Streaming Movie Service:



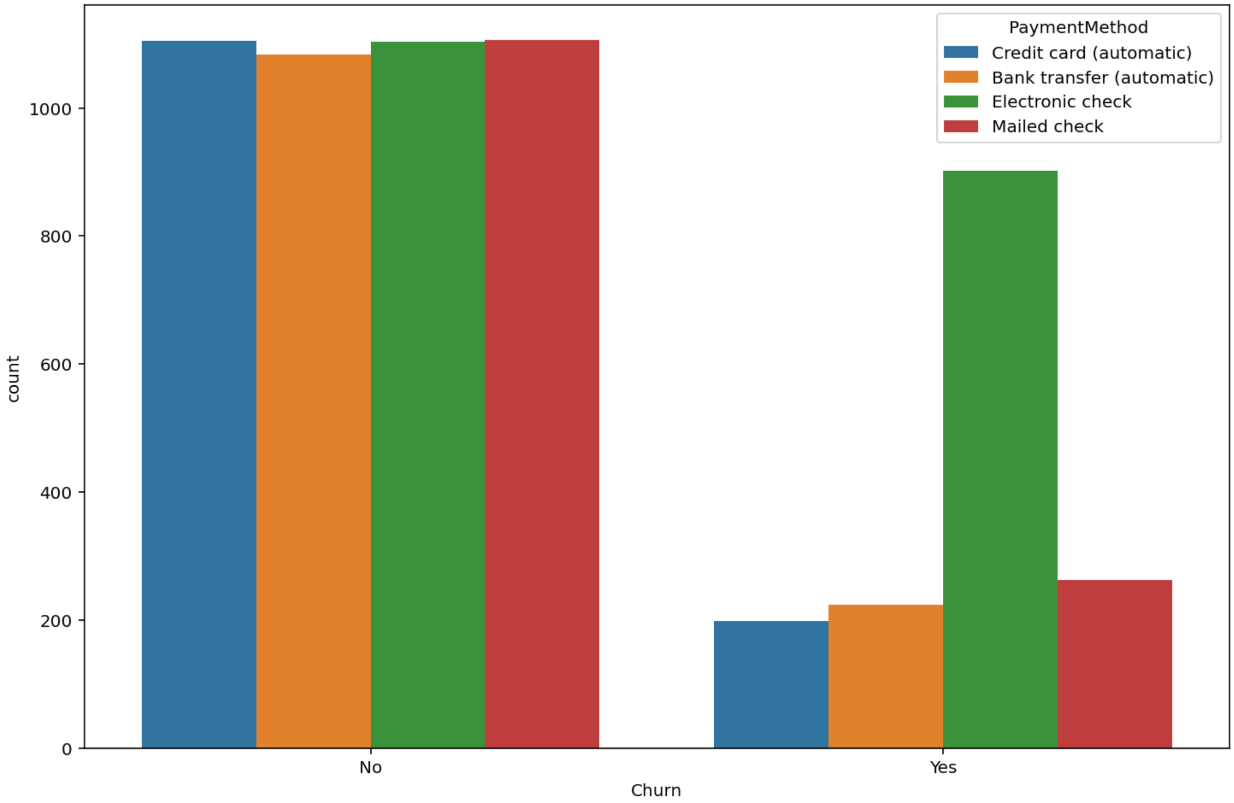
- Customers who continued the subscription more likely to avail the streaming movie services compared to the customers who discontinued the service.

14.Relationship between churn rate and contract agreement:



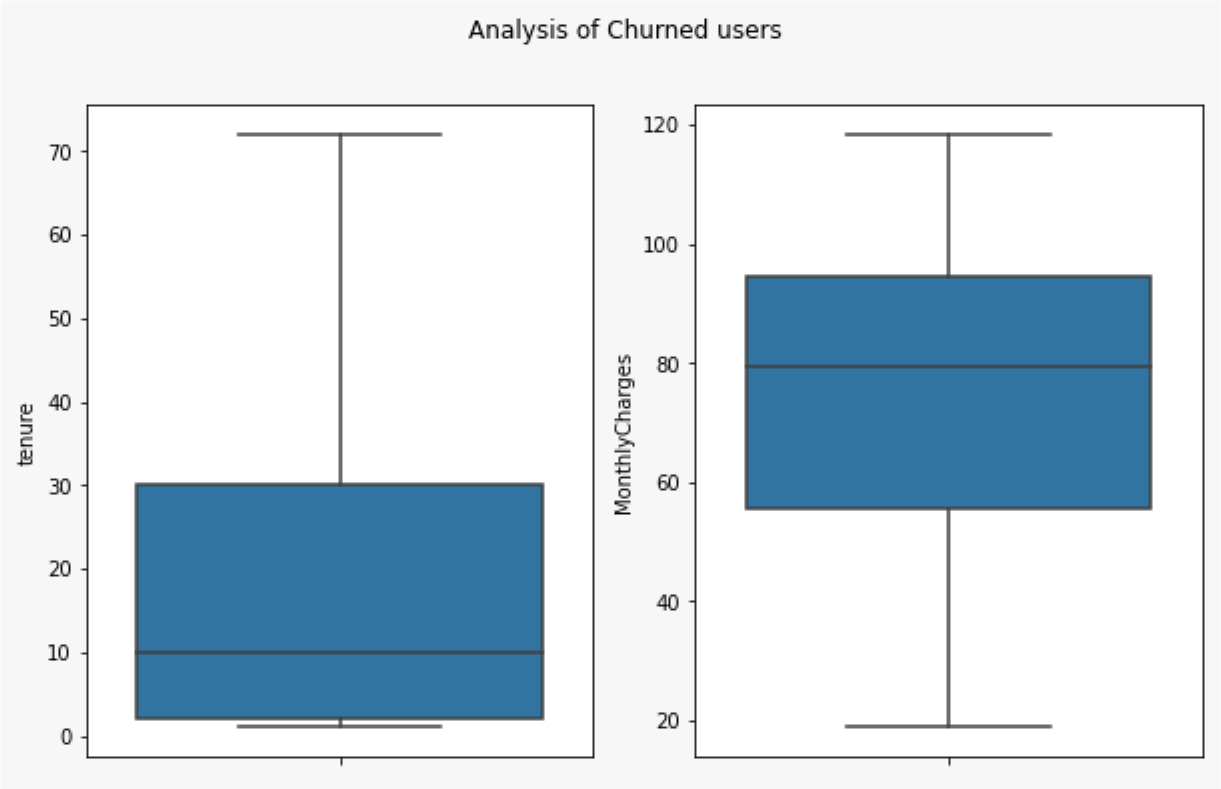
- In case of churned customers majority of them have opted for month-to-month contract agreement and this contract is also dominant in case of the customers who continued their service.

15.Relationship between churn rate and Payment Method:



Analyzing Tenure and Monthly Charges incurred based on Churn:

Boxplot of churned Users:



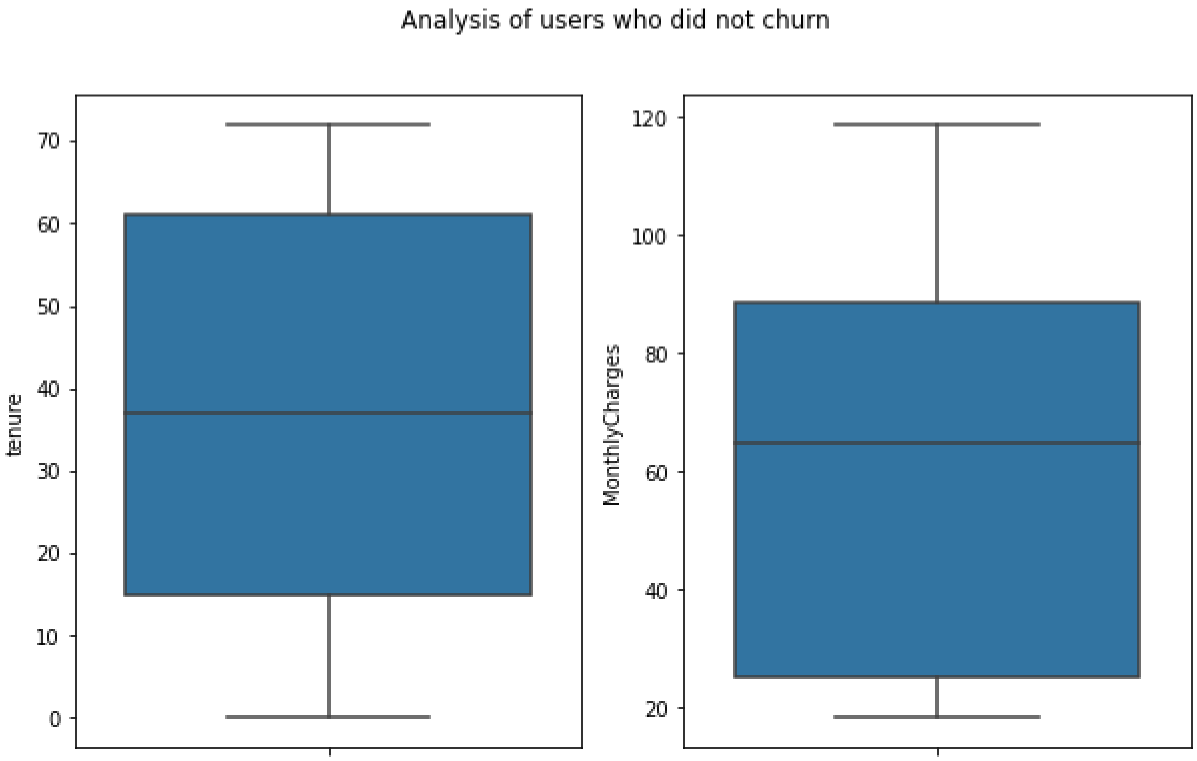
Now we can get the whole information using .describe() function in Python i.e. complete analysis of this boxplot.

tenure	MonthlyCharges	
count	1587.000000	1587.00000 0
mean	18.246377	74.164871
std	19.667262	24.965002
min	1.000000	18.850000
25%	2.000000	55.675000
50%	10.000000	79.500000
75%	30.000000	94.400000

max	72.000000	118.350000

The average tenure of users who have churned is 18 months.The median is 10 months.A user who churns incurs a maximum monthly charge of 118 currency units and a minimum of 18 currency units.Average users of the service incurs 74 currency units.

Boxplot of users who didn’t churn:



The corresponding information regarding this plot is given below:

	tenure	**MonthlyCharges**
count	4399.000000	4399.000000
mean	37.599682	61.424506

std	24.065131	31.086101
min	0.000000	18.250000
25%	15.000000	25.125000
50%	37.000000	64.750000
75%	61.000000	88.700000
max	72.000000	118.750000

The average tenure of users who didn't churn is 37 months and a maximum tenure of 72 months.Average users incurs a monthly cost of 62 to 64 currency units.

Correlation of the most related two features of the dataset:(using SPSS)

Correlation between Tenure and Total Charges:

```
Correlations
                    tenure    TotalCharges
tenure    Pearson Correlation        1          .827**
          Sig. (2-tailed)            .000
          N         5986      5976
TotalCharges        Pearson Correlation         .827**    1
          Sig. (2-tailed)   .000
          N         5976      5976
** Correlation is significant at the 0.01 level (2-tailed).
```

- So we can conclude that there is a reasonably strong relationship between the number of months the customer has been a client to the company and the total charges the customer incurs.

Correlation between Tenure and Monthly Charges:

```
Correlations
                    tenure    MonthlyCharges
tenure    Pearson Correlation        1          .257**
```

Sig. (2-tailed) .000

N 5986 5986

MonthlyCharges Pearson Correlation .257** 1

Sig. (2-tailed) .000

N 5986 5986

** Correlation is significant at the 0.01 level (2-tailed).

- We conclude that there is no significant relationship between the monthly charge a customer incurs and the no of months the customer has been a client to the company.

Correlation between tenure and churn rate:

Correlations

tenure Churn

tenure Pearson Correlation 1 -.348**

Sig. (2-tailed) .000

N 5986 5986

Churn Pearson Correlation -.348** 1

Sig. (2-tailed) .000

N 5986 5986

** Correlation is significant at the 0.01 level (2-tailed).

- A negative correlation is a relationship between two variables in which an increase in one variable is associated with a decrease in the other.As the tenure of a customer increases the churn rate decreases over a period of time.

Conclusion:

It has been clear from the above analysis that different customers have different choices and preferences based on different features of the dataset.In case B2C customers churn rate is due to poor signal issues or associated services like streaming movie service,device protection etc and one customer may be among one of hundreds and thousands so revenue per customer is comparatively low but in case of B2B customer experience plays a very crucial role in preventing leaving as the number of customers is significantly lower and acquiring new customer is higher but revenue from each customer is also large.

In today’s growing marketplace, customer attrition is commonplace. Consumers have a wide variety of options to choose from, each one offering something different — a better customer experience, lower pricing, or better products and services. So, it’s vital for organizations to perform customer churn analysis to retain their customers, be it a B2C or a B2B scenario.

References:

- ➢ Kaggle.com
- ➢ http://www.ijstr.org/final-print/jan2021/Customer-Churn-Prediction-In-Telecom-Sector-A-Survey-And-Way-A-Head.pdf
- ➢ https://blog.gramener.com/churn-analysis-customer-retention/