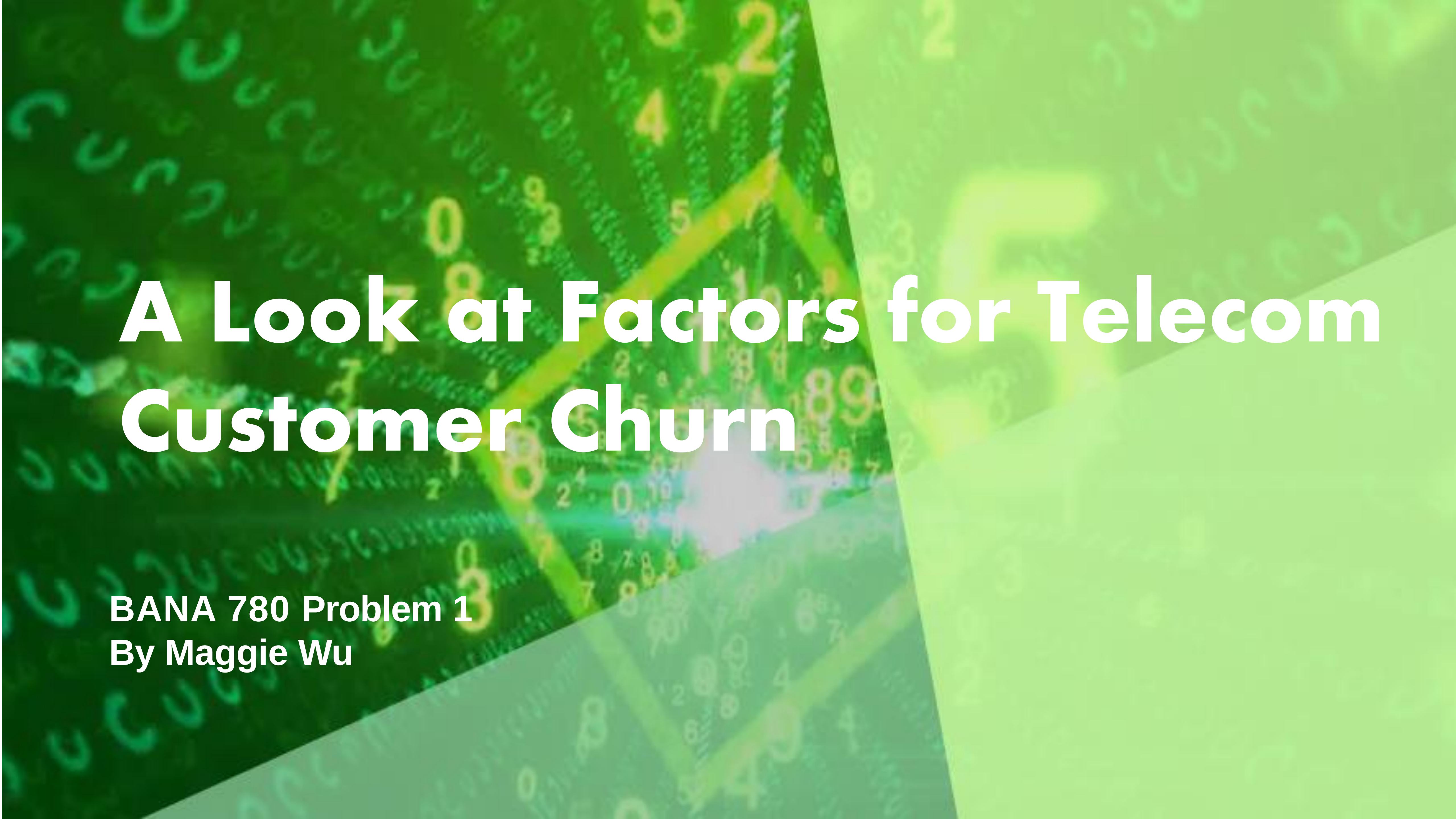


A Look at Factors for Telecom Customer Churn

BANA 780 Problem 1
By Maggie Wu

Agenda

1. Introduction to the Dataset
2. Explaining the Data
3. Data Clearning
4. Overview
5. Details
6. Summary of Findings
7. Analysis and Recommendations

1. Introduction to the Dataset

1. Introduction to the Dataset

This dataset comes from Kaggle, I choose it since the customer churn is a popular topic in many business. This dataset includes basic information on around 7000 users in California who use phone and internet service, and show their behaviours and payment situation, e.g. registered service, accounting information, churn information, etc. For better exploration, I divide them into three group:

- **Factor group 1：Customer information**

  Gender, Dependents, Partner, SeniorCitzen

- **Factor group 2：Servcie using behaviour**

  Onlinesecurity, onlinebackup, Phoneservice, etc

- **Factor group 3：Payment factors**

  MonthlyCharges, TotalCharges, tenure, contract, payment method, paperless billing

2. Explaining the Data

2. Explaining the Data

With these information, we can examine how these factors may affect the customers' churn, and think about how to reduce the churn rate and improve customer stickiness. Since in this dataset 21 variables are understandable according to their names, so I only give some examples:

- **Senior Citizen:** whether the customer is senior (yes/1, no/0);
- **Partner:** whether the customer has partner, (yes/1, no/0);
- **Dependents:** whether the customer has dependents, (yes/1, no/0);
- **Tenure:** how long has the customer registered in this telecom company, it's a discrete variable,

  we think it indicates the days calculated after people registered in this telecom company in Q3 which is mentioned

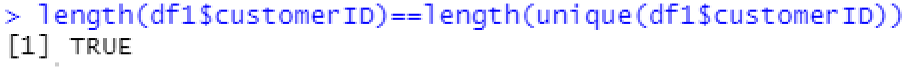
  in webpage, it covers from 0 to 72, so the minimum day is 0 day, the maximum day is 72 days.
- **Online Security:** whether the customer use online security service (Yes, No or No internet service);
- **Phone Service:** whether the customer use phone service (Yes / No)

3. Data Cleaning

3. Data Cleaning

1. Check if the ID column is duplicated

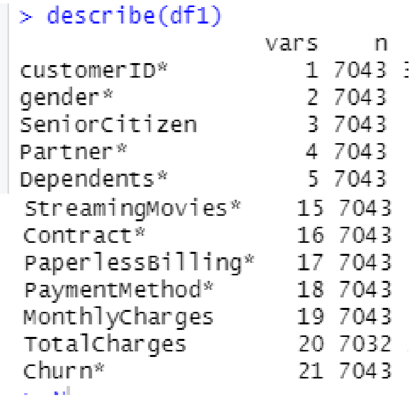
Result: unique



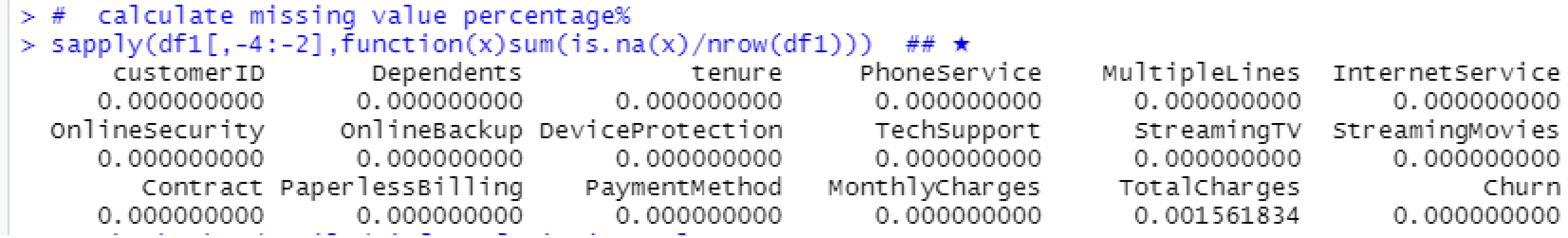
```
> length(df1$customerID)==length(unique(df1$customerID))
[1] TRUE
```

2. Check for missing values

Result: there are 11 missing values, and the missing percentage is very small. So I plan to use mean to fill in.



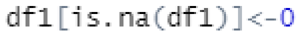
```
> describe(df1)
                 vars    n
customerID*         1 7043
gender*             2 7043
SeniorCitizen       3 7043
Partner*            4 7043
Dependents*         5 7043
StreamingMovies*   15 7043
Contract*          16 7043
PaperlessBilling*  17 7043
PaymentMethod*     18 7043
MonthlyCharges     19 7043
TotalCharges       20 7032
Churn*             21 7043
```



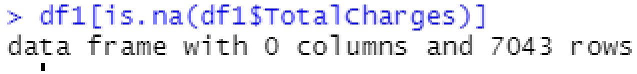
```
> #  calculate missing value percentage%
> sapply(df1[,-4:-2],function(x)sum(is.na(x)/nrow(df1)))  ## ★
      customerID       Dependents           tenure     PhoneService    MultipleLines  InternetService
     0.000000000      0.000000000      0.000000000      0.000000000      0.000000000      0.000000000
  OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV  StreamingMovies
     0.000000000      0.000000000      0.000000000      0.000000000      0.000000000      0.000000000
        Contract PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn
     0.000000000      0.000000000      0.000000000      0.000000000      0.001561834      0.000000000
```

3. Check data logic and fill in missing values

Result: the data with missing value show its tenue is 0, and no total charge, but has monthly charge, these customers are not lost, so we guess they just register in current month, they don't have historical total charge calculated from last month instead from current month. So I change my idea I plan to fill in 0 for TotalCharge.



```
df1[is.na(df1)]<-0
```

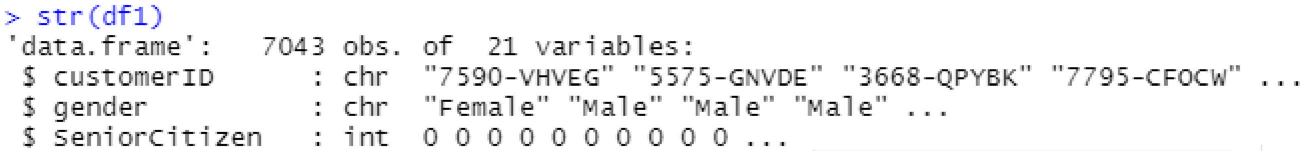


```
> df1[is.na(df1$TotalCharges)]
data frame with 0 columns and 7043 rows
```

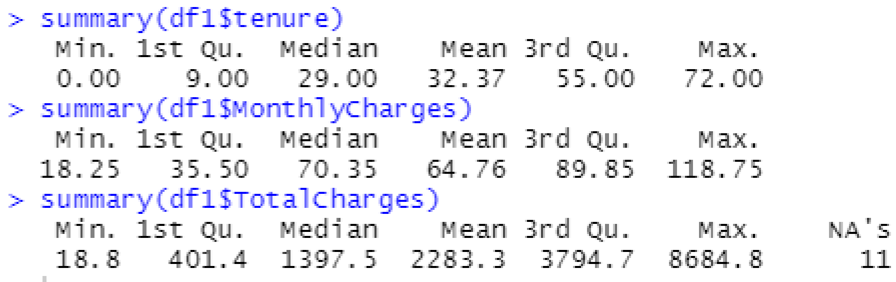
3. Data Cleaning

4. Check data type

Result: data type is correct, so we don’t need to make type transform.



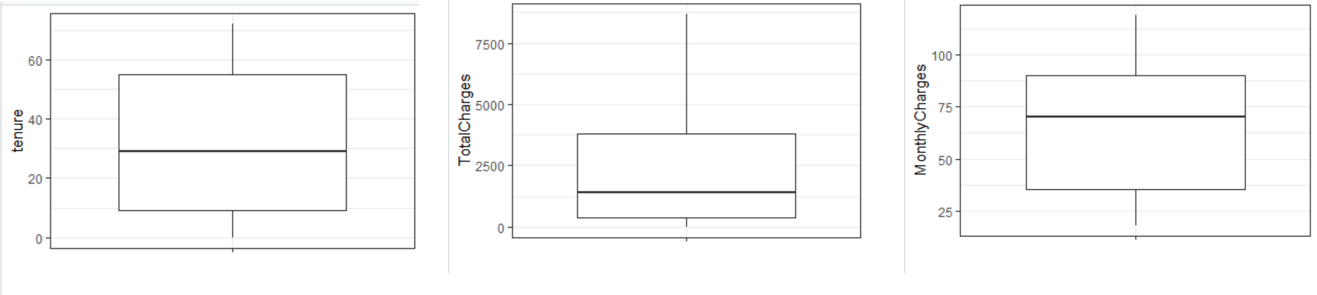
```
> str(df1)
'data.frame':	7043 obs. of  21 variables:
 $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
 $ gender          : chr  "Female" "Male" "Male" "Male" ...
 $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Partner         : chr  "Yes" "No" "No" "No"
```



```
> summary(df1$tenure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    9.00   29.00   32.37   55.00   72.00 
> summary(df1$MonthlyCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  18.25   35.50   70.35   64.76   89.85  118.75 
> summary(df1$TotalCharges)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   18.8   401.4  1397.5  2283.3  3794.7  8684.8      11 
```

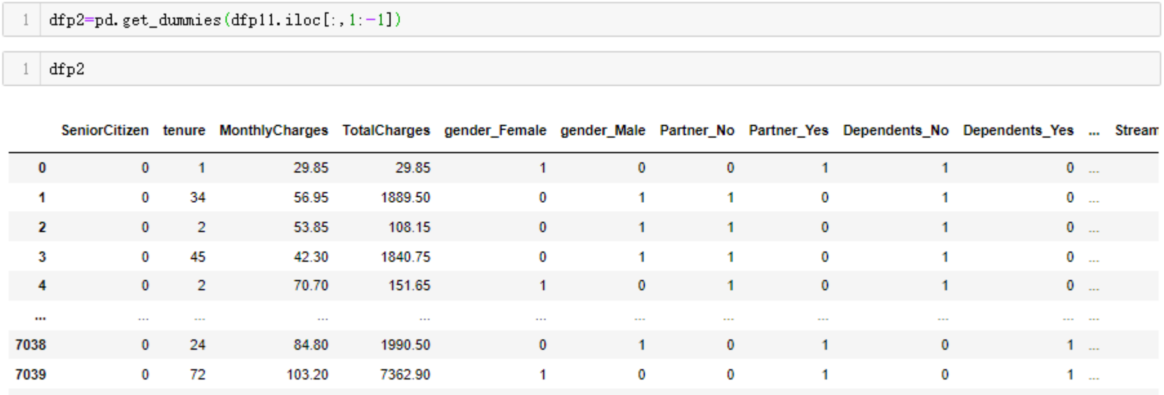
5. Check the numeric columns for outliers

Result: according to the boxplot, no outliers



6. Dummy transformation

Result: do dummy variable transformation



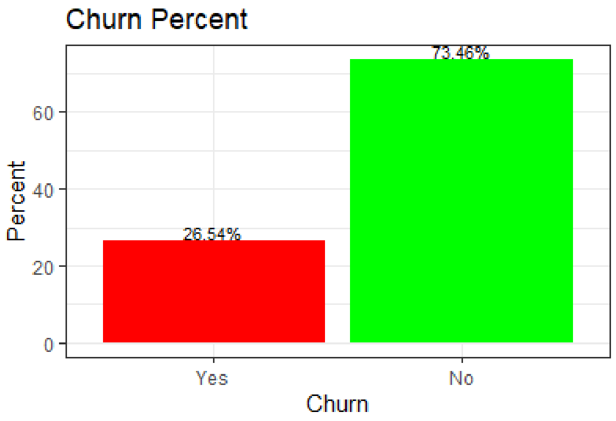
```
dfp2=pd.get_dummies(dfp11.iloc[:,1:-1])
```

```
dfp2
```

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges	gender_Female	gender_Male	Partner_No	Partner_Yes	Dependents_No	Dependents_Yes	...	Stream
0	0	1	29.85	29.85	1	0	0	1	1	0	...	
1	0	34	56.95	1889.50	0	1	1	0	1	0	...	
2	0	2	53.85	108.15	0	1	1	0	1	0	...	
3	0	45	42.30	1840.75	0	1	1	0	1	0	...	
4	0	2	70.70	151.65	1	0	1	0	1	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
7038	0	24	84.80	1990.50	0	1	0	1	0	1	...	
7039	0	72	103.20	7362.90	1	0	0	1	0	1	...	

4. Overview

4. Overview



Dependent variable:

- There are total around 7000 records, where the lost customers accounts for 26%, which is a relative high percentage. It is worthy to find out the impacting factors.

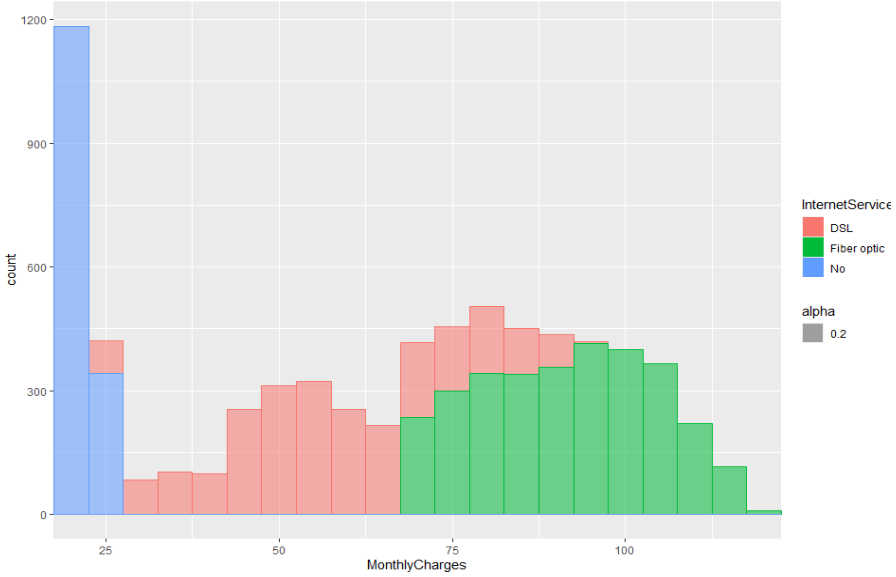
Independent variable (numeric ones):

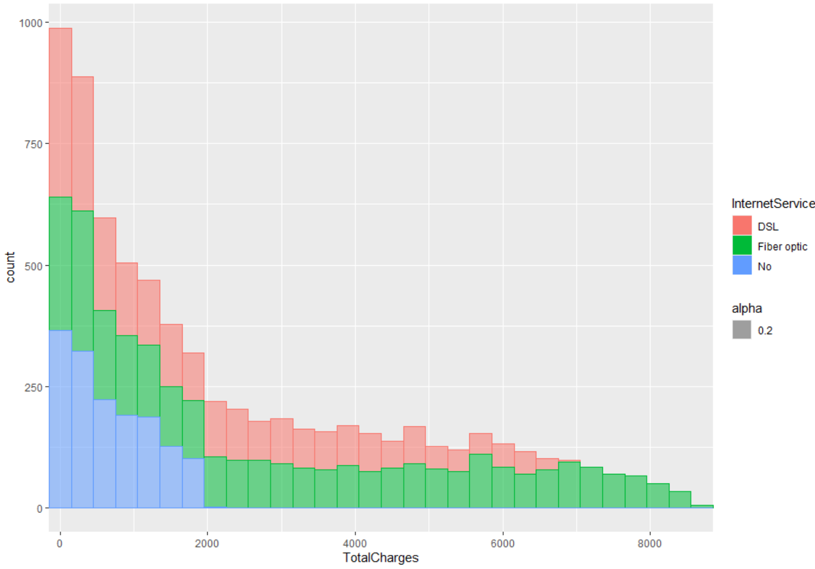
Q1: what’s the most common value?

In histogram, tall bars show the common values of a variable, and shorter bars show less-common values.

- **Monthly charge: 20**
- **Total charges: 300~600**

The common value indicate the low consumption.





5. Details

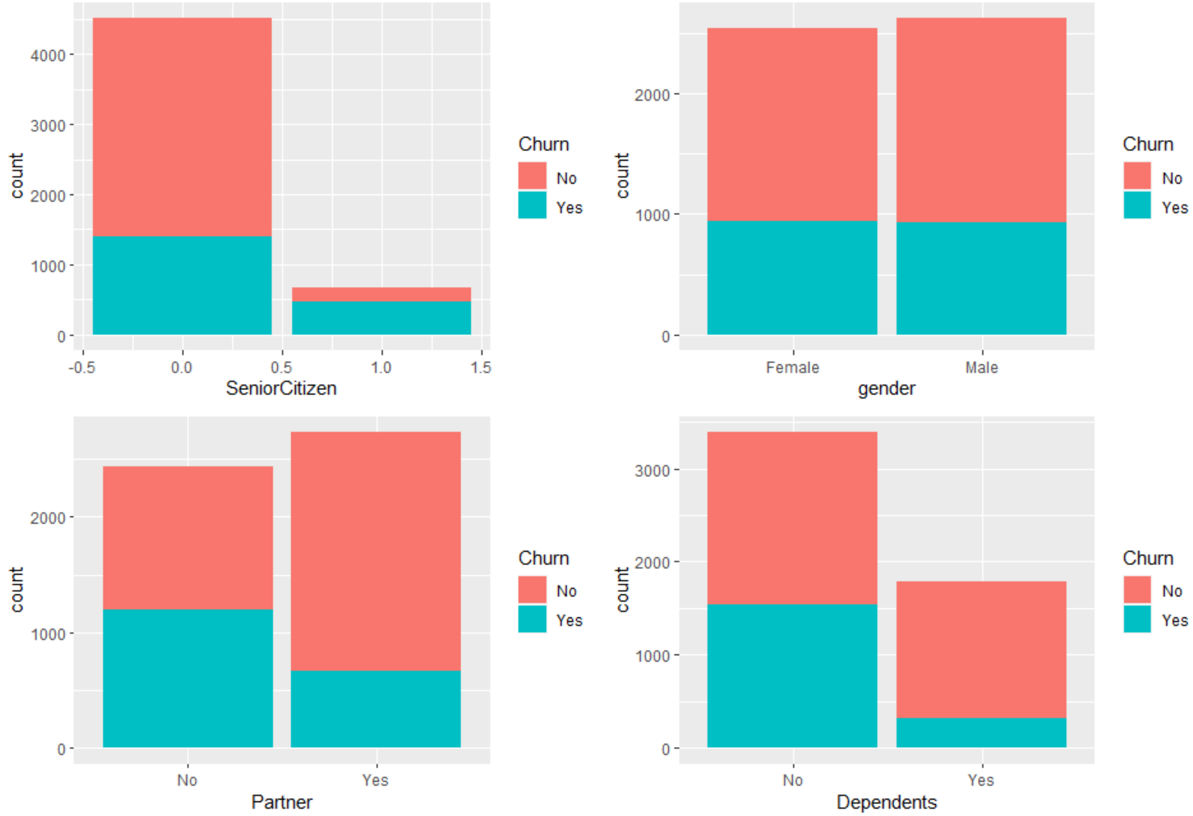
5. Details

We want to find out what are the impacting factors for churn rate, so through giving several questions which guide us explore the data in depth.

Q2. How do these three factor groups have impact to churn rate？

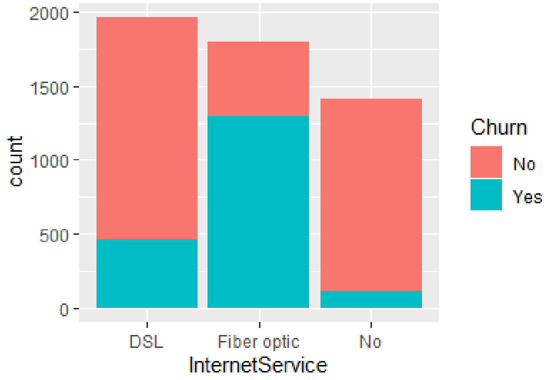
Factor group 1

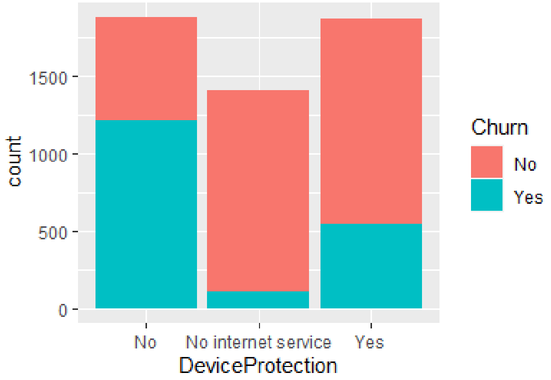
- According to the bar plot, we can find out that the senior customer, customer without dependents and customer without partner have high churn rate.
- While the gender doesn’t have big impact on it.

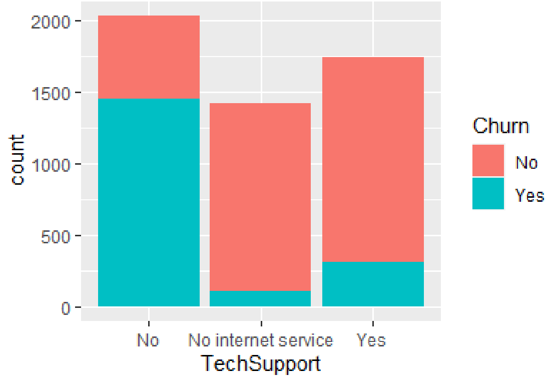


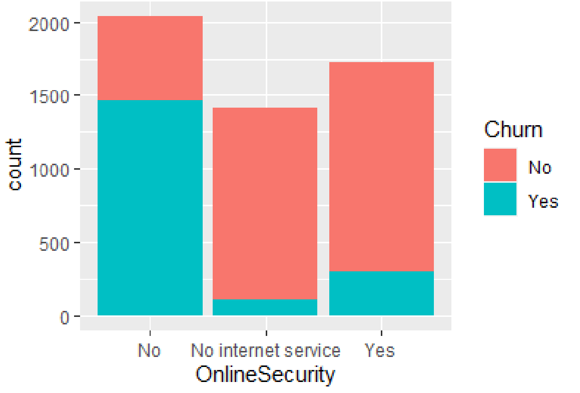
Factor group 2

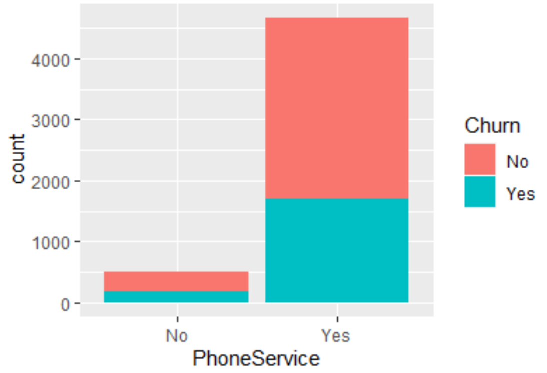
- According to the bar plot, we can find that the customer who haven’t registered for internet service or who have registered for internet service but meanwhile registered for many other added-value service (online security, device protection) is of low churn rate.
- While if the customer has registered for internet service but not for any other added-value service, they will have high churn rate.

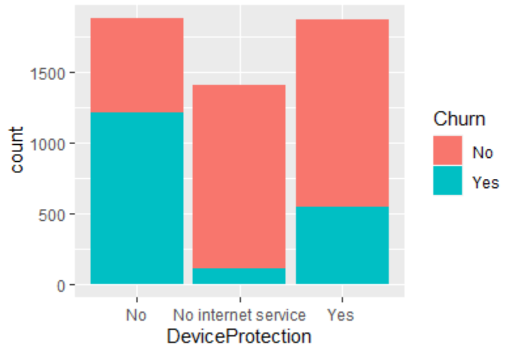


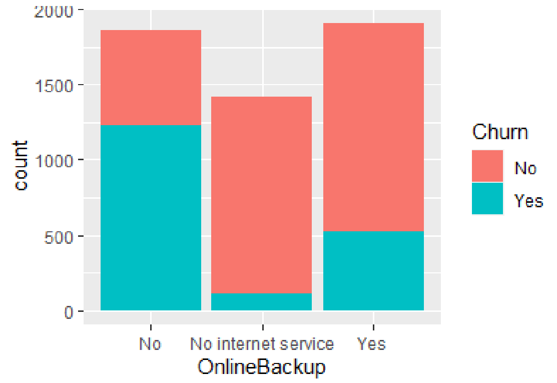


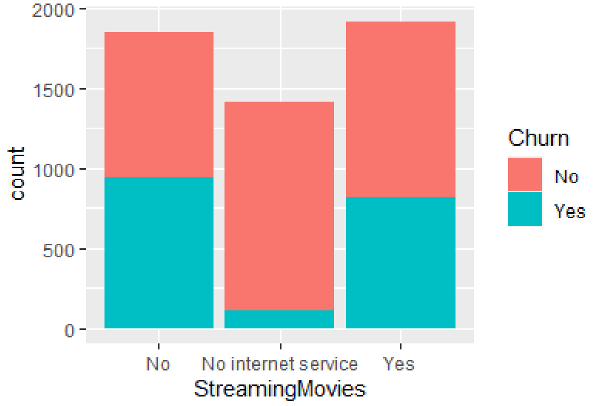






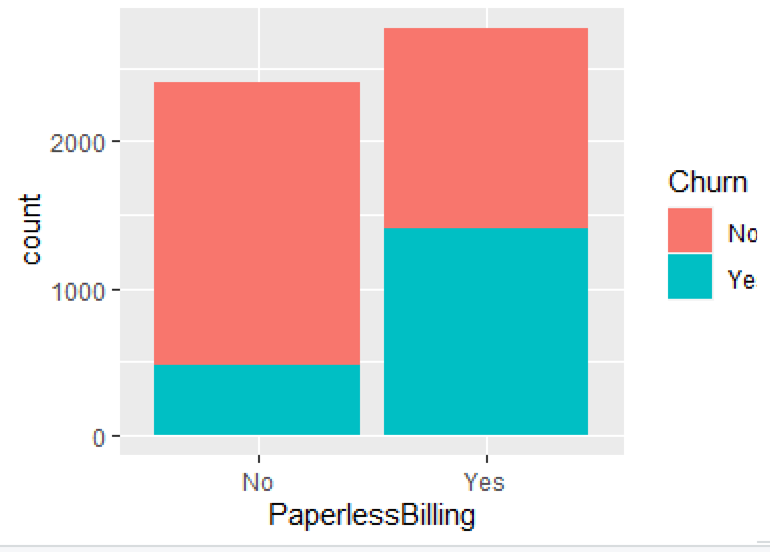


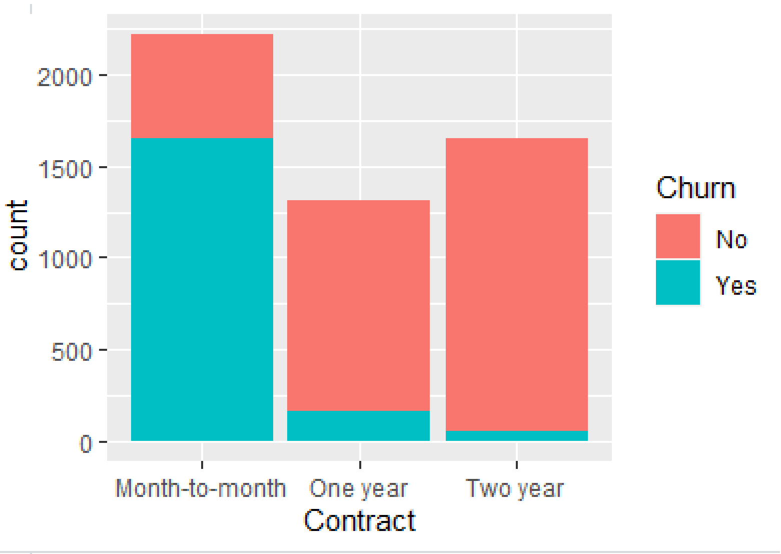


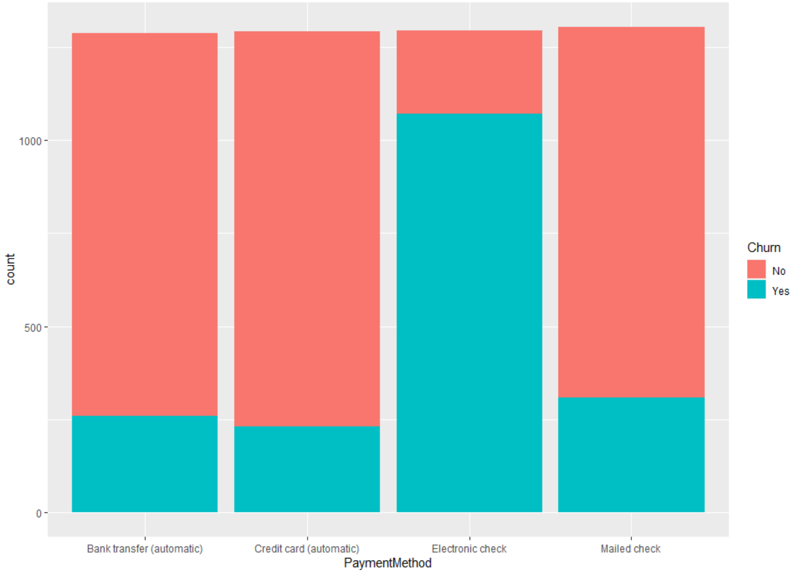


Factor group 3

According to the bar plot, we can find out that compared with other payment method, the customers with online payment (electronic check) ,or month to month contract or using paperless billing are easier to be lost.







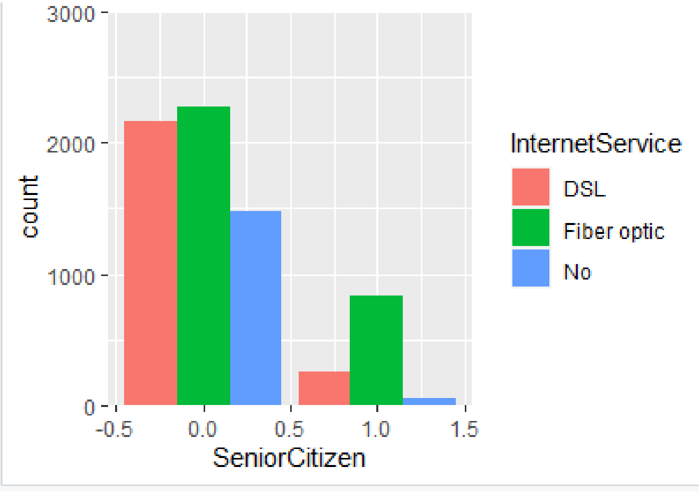
Q3. Concerning sub-factors found above, what’s the root reason for them, are they due to expensive price or poor service saticification?

Besides, since I think the internet service explain more about higher churn rate， so I focus on it to explore in depth as follows.

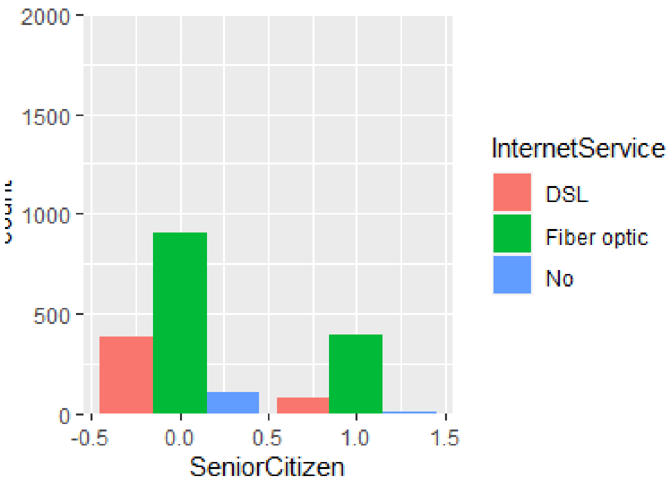
Factor group 1

- Concerning seniors, we can find out that out of our expectation, many seniors also like fiber optic. But finally whether they are senior or not, most of them with fiber optic type are lost.
- However, I also find out a great part of seniors not lost also choose fiber optic.

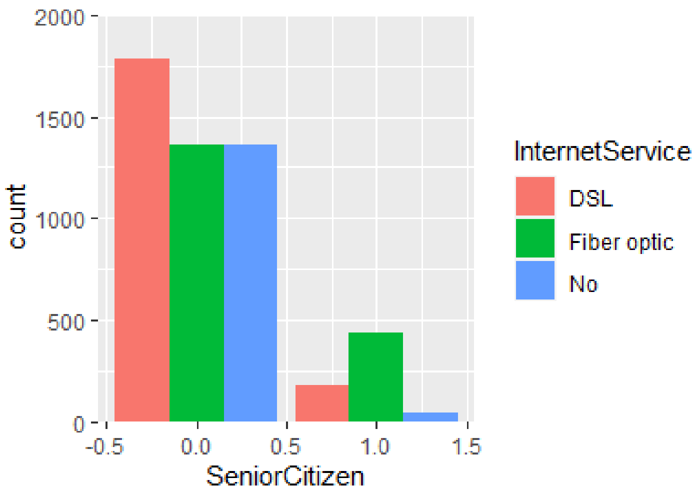
Overall



Churn= Yes



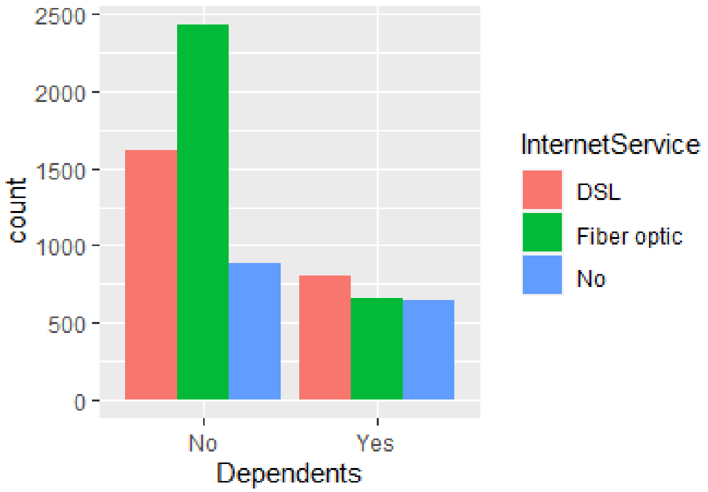
Churn= No



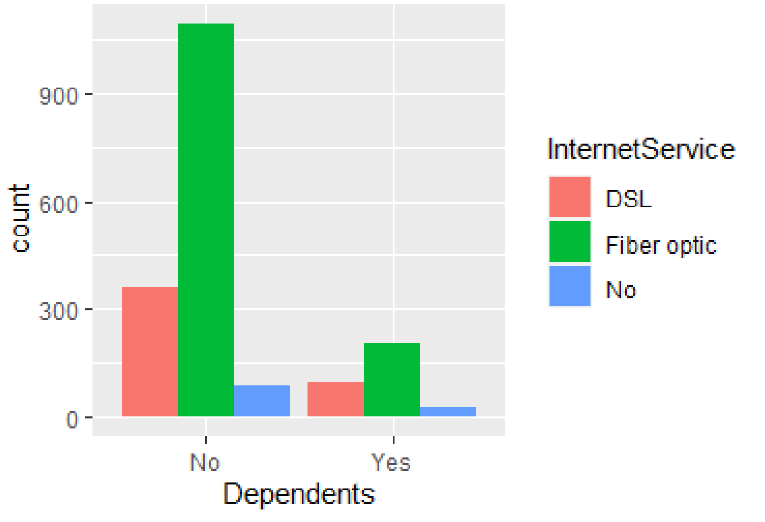
Factor group 1

- Concerning dependents, we can find out many people without dependents like fiber optic very much. But finally half of them are lost.

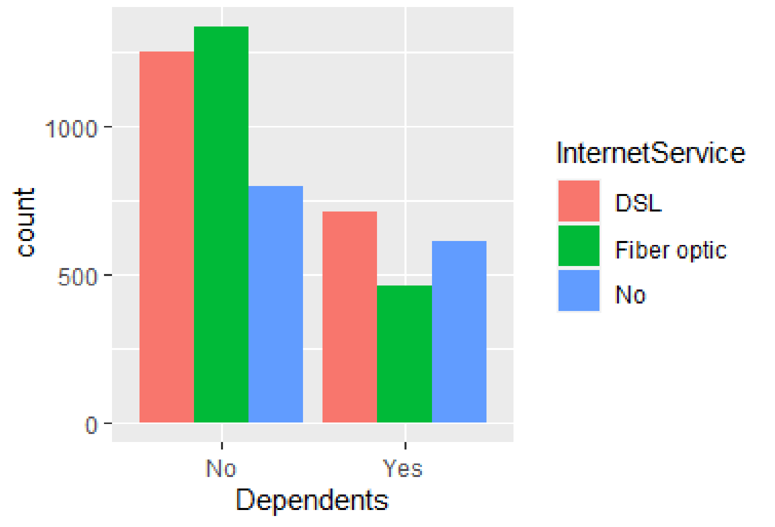
Overall



Churn= Yes



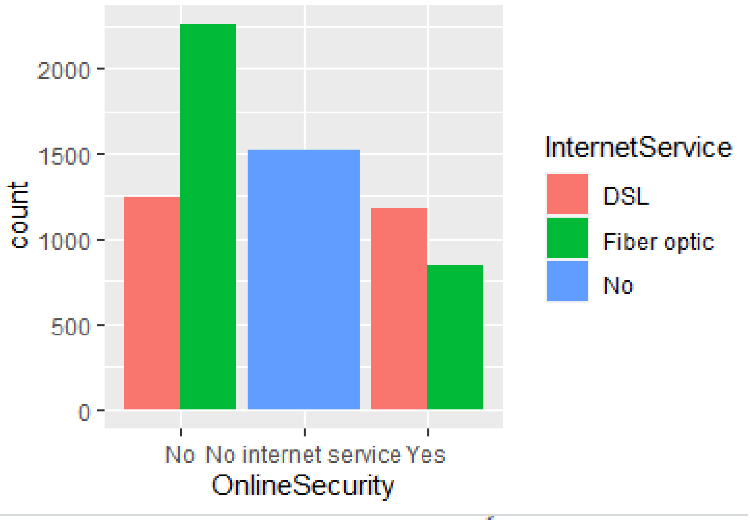
Churn= No



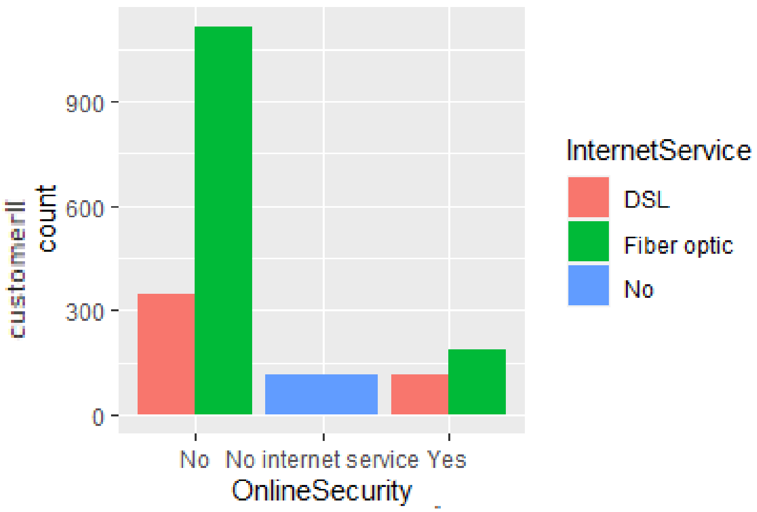
Factor group 2

Focusing on group 2 factors, since the service is similar, so I only take online security as example. I find out when churn is yes, customer without online security service use more fiber optic.

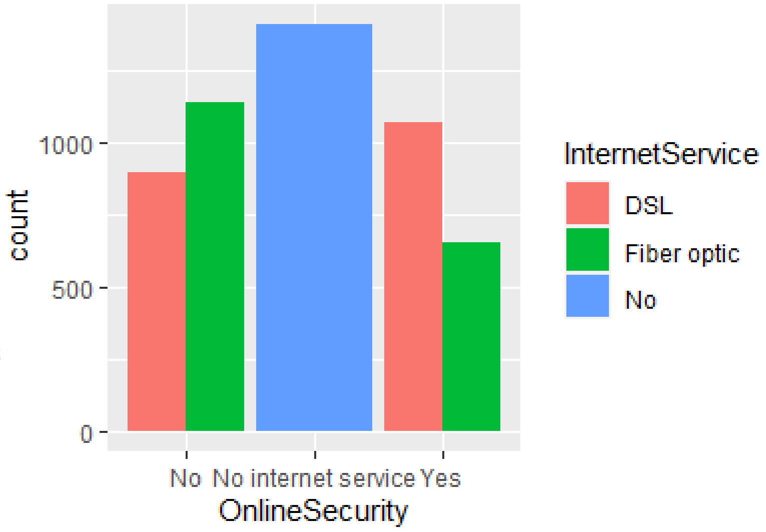
Overall



Churn= Yes

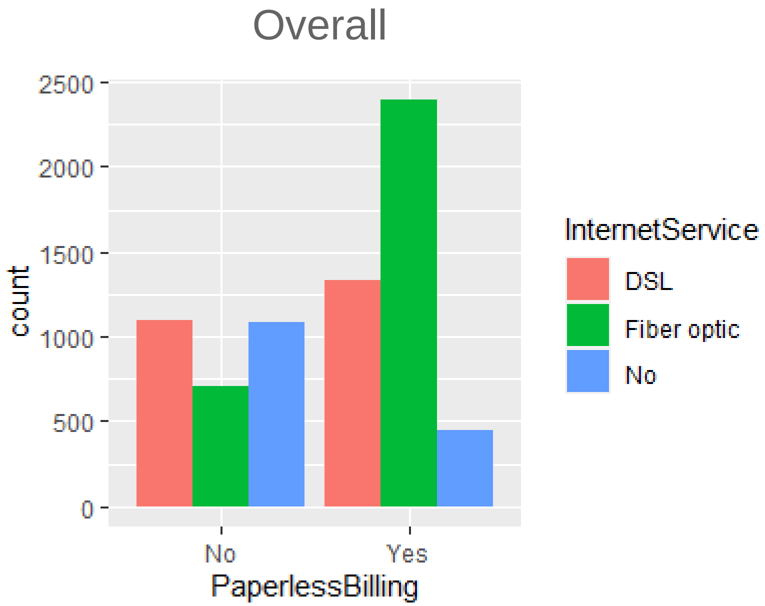


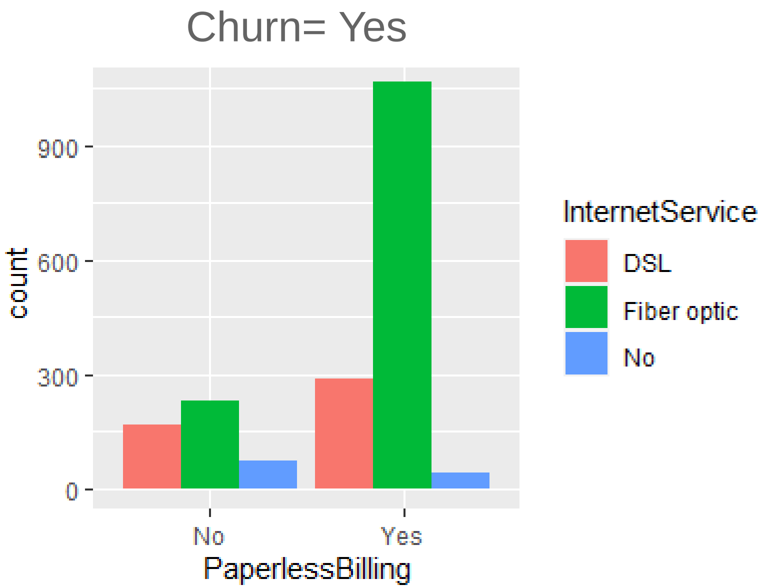
Churn= No

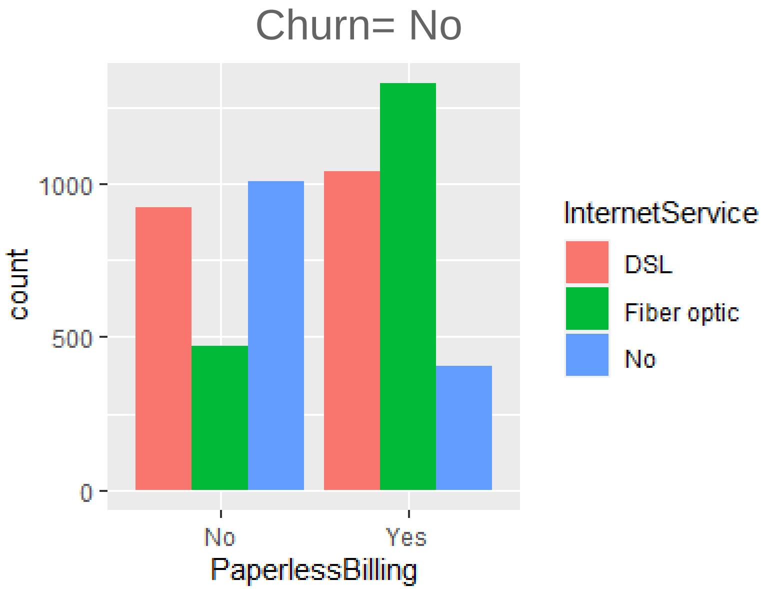


Factor group 3

Focusing on group 3 factors, since the result is similar, so I only take paperless billing as example. I find out when churn is yes, customer using paperless billing use more fiber optic.

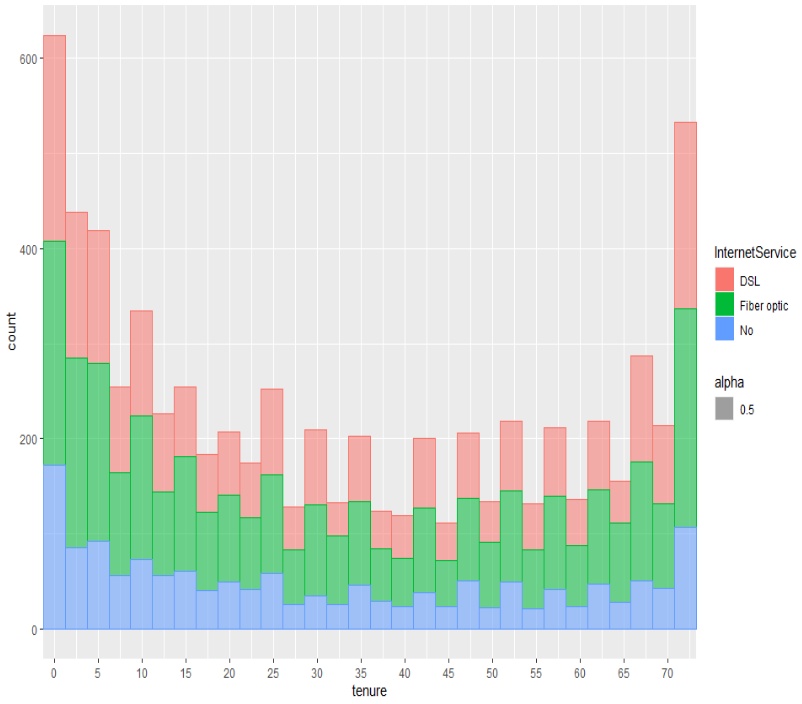

Overall


Churn= Yes

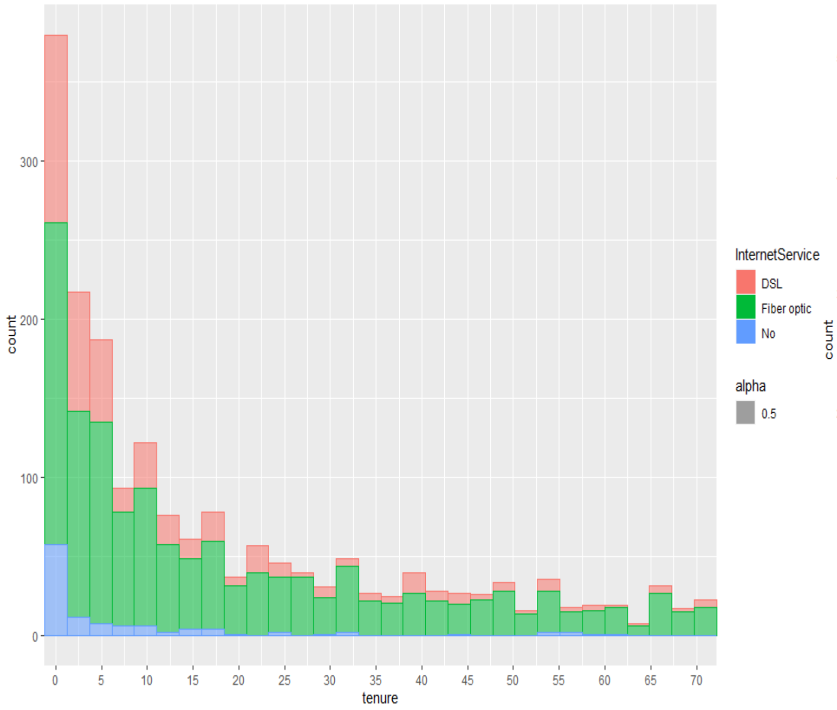

Churn= No

Factor group 3

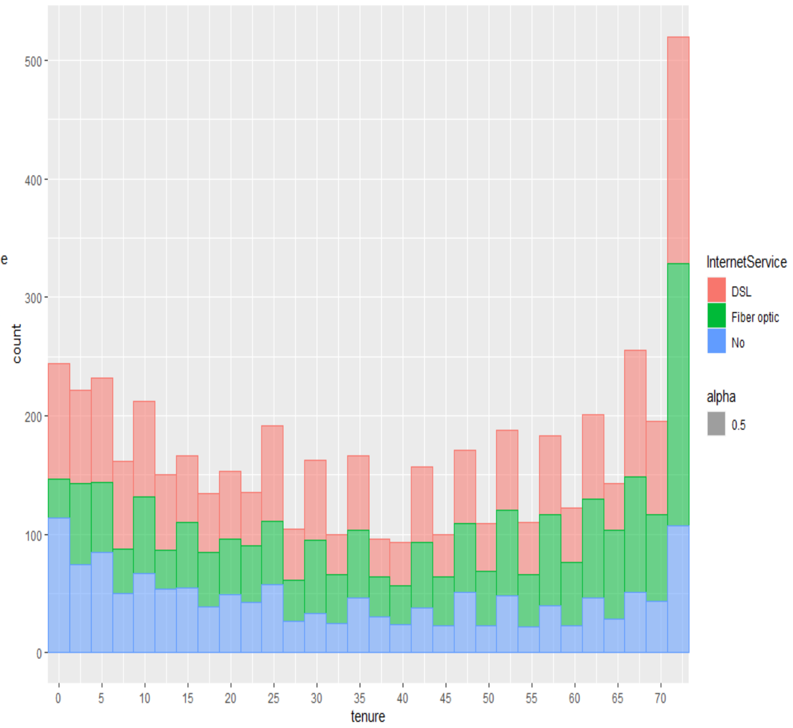
- According to the histogram, we can find out when churn is yes, over the whole range it is covered by fiber optic. When churn is no, the distribution is equal.
- Besides, in the first 5 enrolling days, the customers with fiber optic accounts for high percentage and also during these days they high churn rate is incurred.



Overall



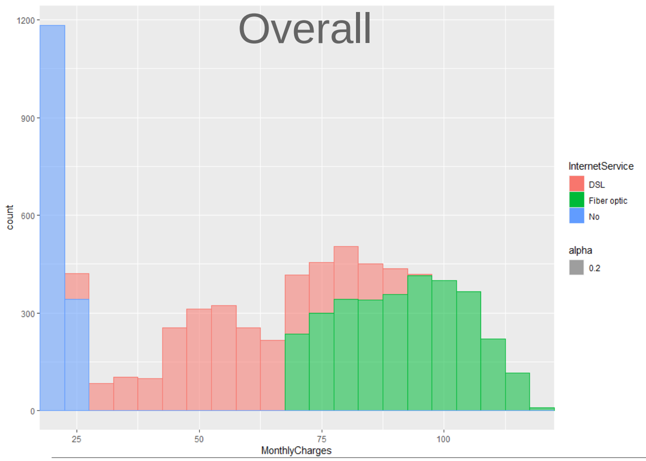
Churn= Yes

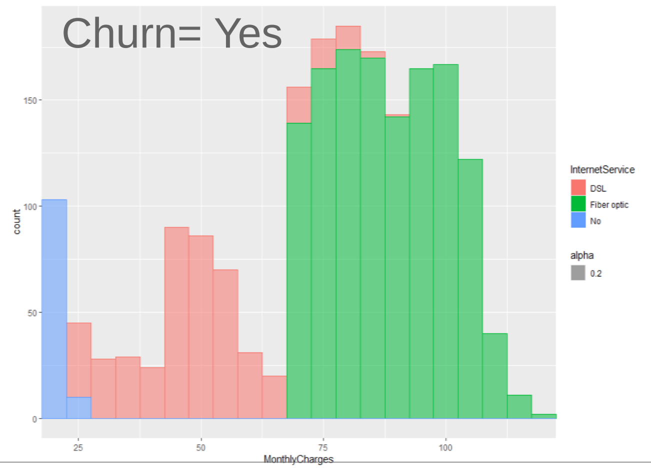


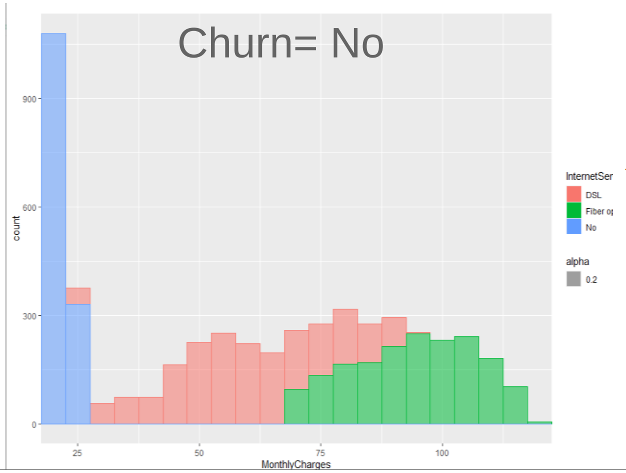
Churn= No

Factor group 3

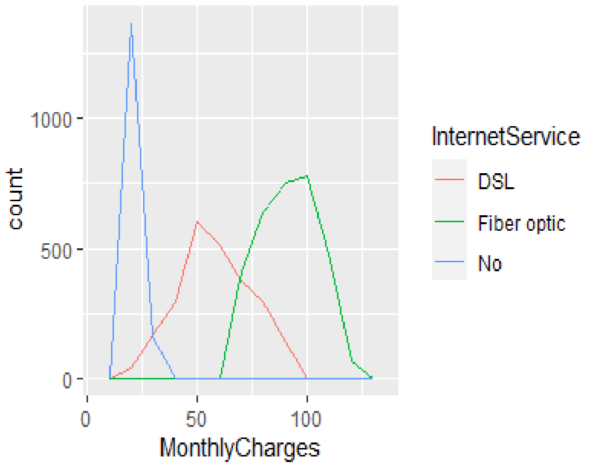
- According to the histogram, the high monthly charges and total charges are fiber optic which lead to churn, while the low monthly charges is no internet or DSL.
- When people using fiber optic, and the monthly charge is above 75, they are easy to be lost.

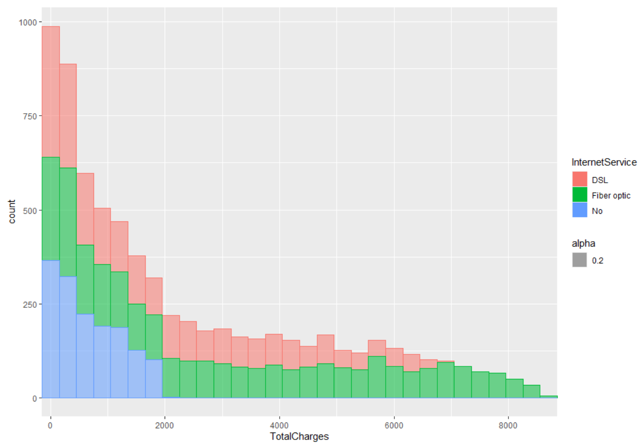

Overall

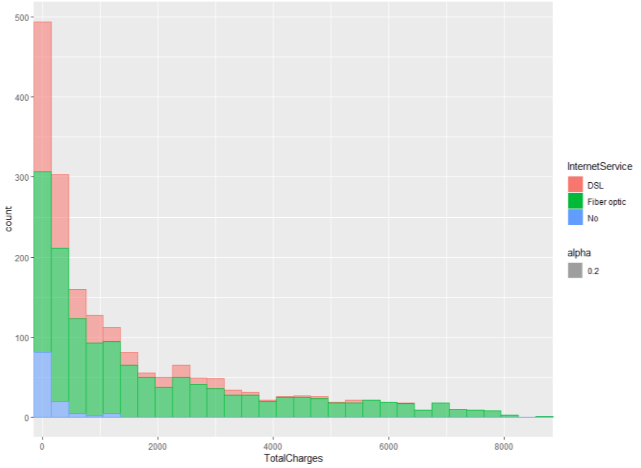

Churn= Yes

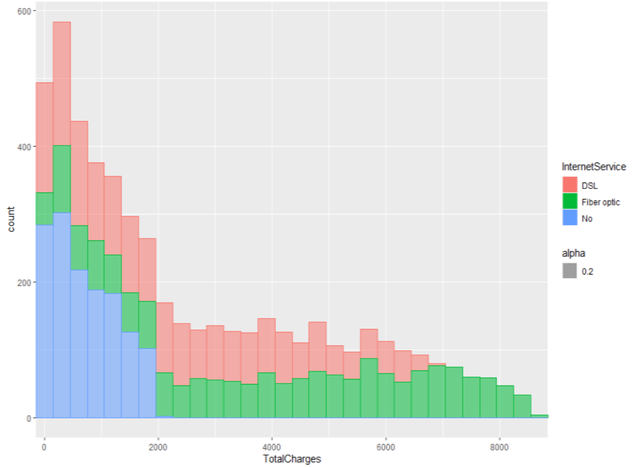

Churn= No

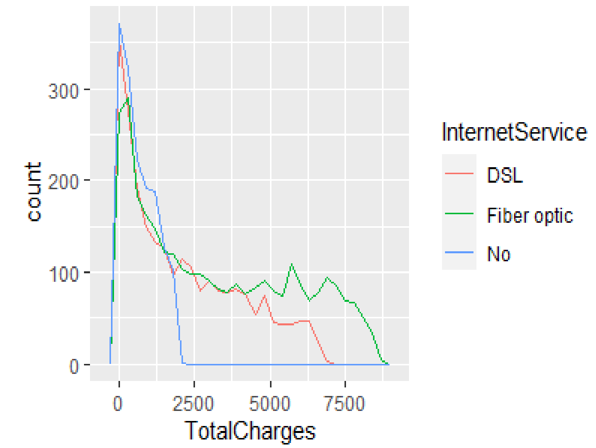
Overall





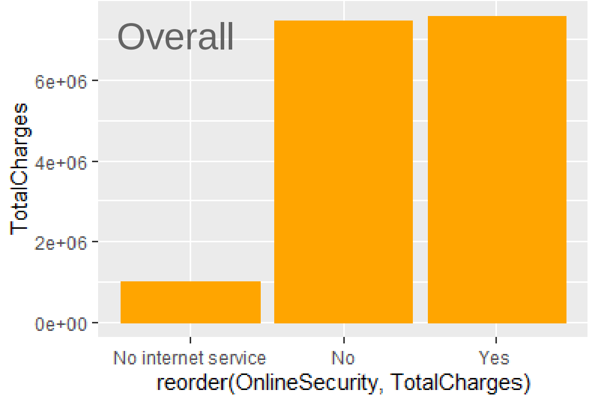


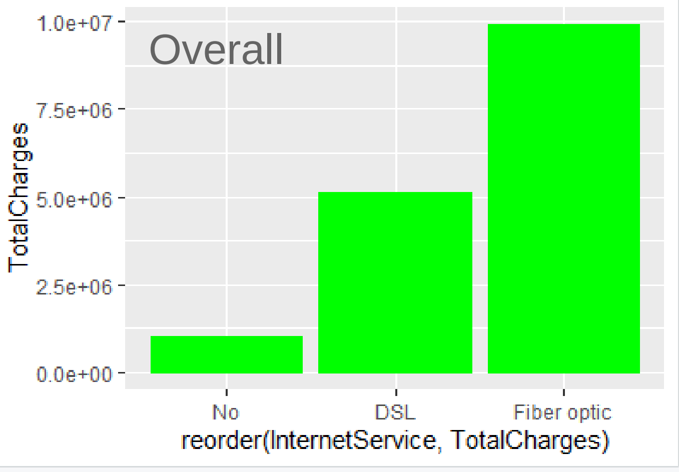




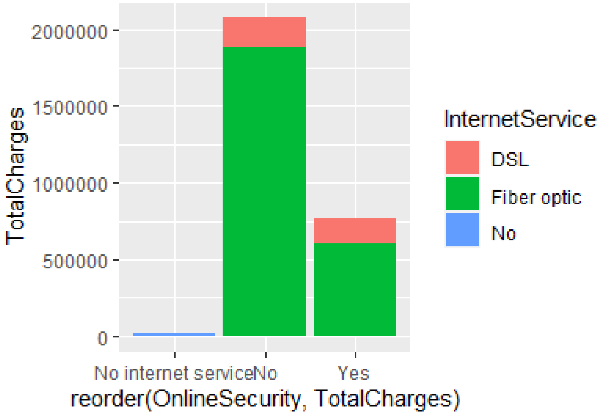
Q4. What’s the character of fiber optic and added value service?

- According to the bar plot, we can see the total charges with fiber optic is the most expensive, while without internet is the cheapest.
- While the total charge whether using added value service or not is similar. Besides, when churn is yes, without online security with fiber optic are of higher total charges, which let us have a feeling that such service is with lower cost performance.
- So I think root reason is expensive price of fiber optic. So for fiber optic, how to balance the price and speed is a question, which is a key factor for improving churn rate.

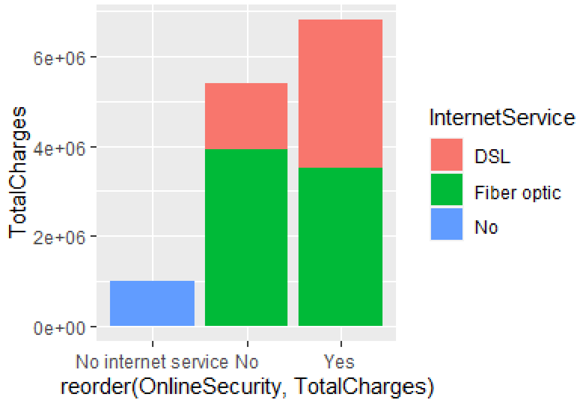




Churn= Yes



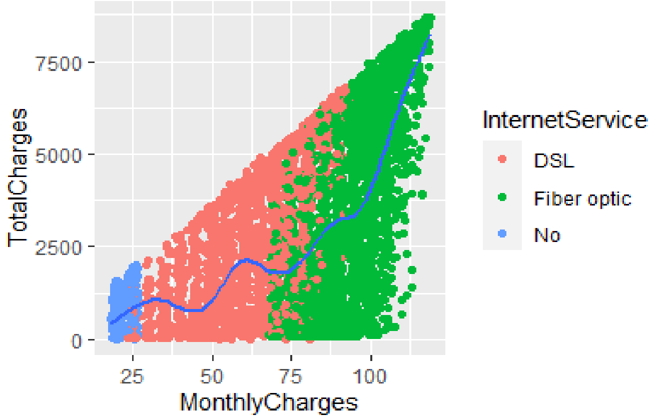
Churn= No

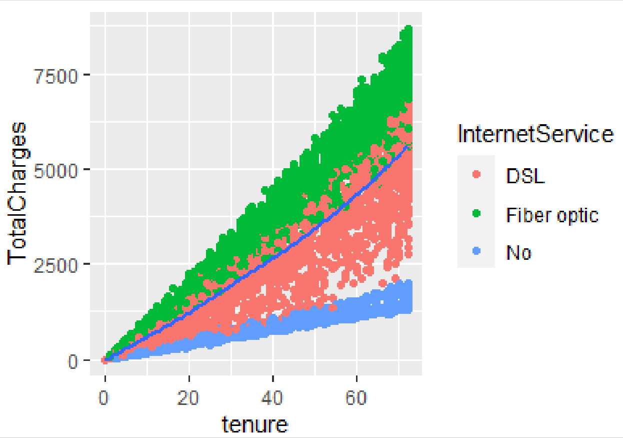


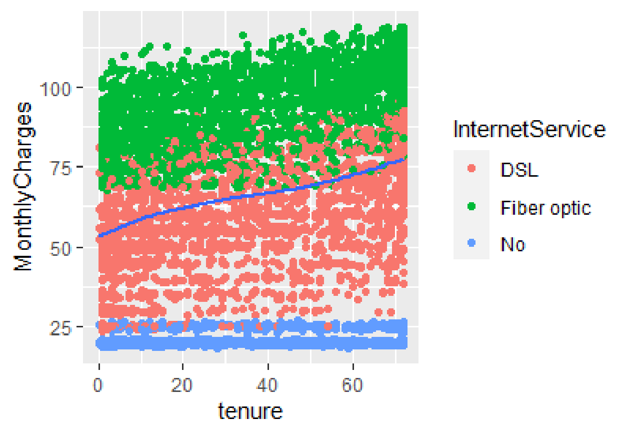
Q5. What’s the relationship between these 2 numeric variables?

- We can see the price trend of 3 types of internet service more clearly. Fiber optic is the most expensive way.
- Besides, we can find out when churn is yes, only the fiber optic type is distributed more densely.

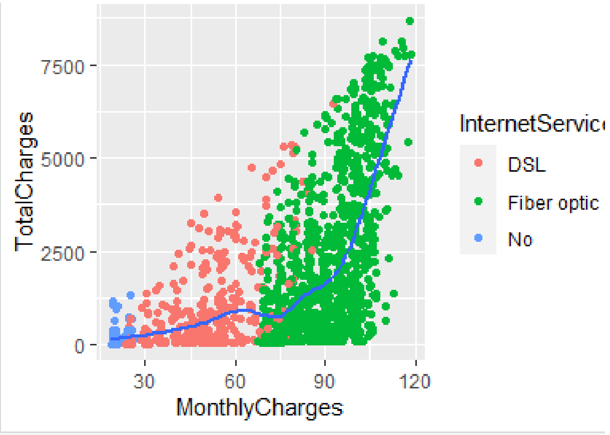
Overall

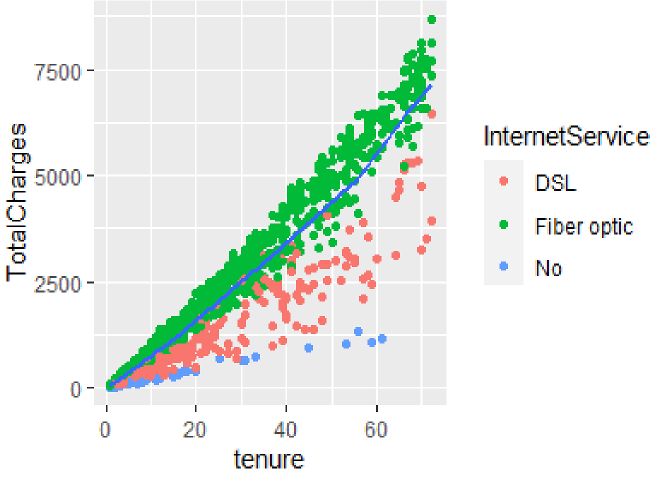


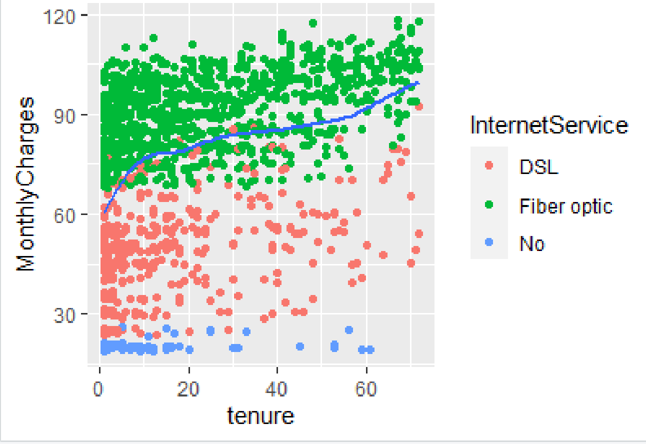




Churn= Yes



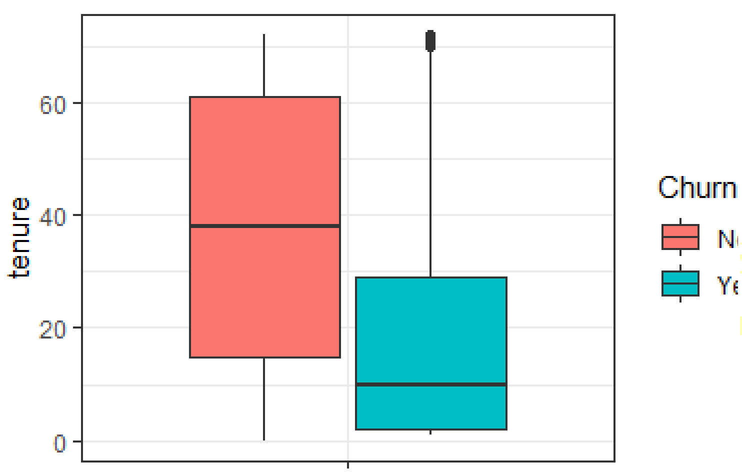


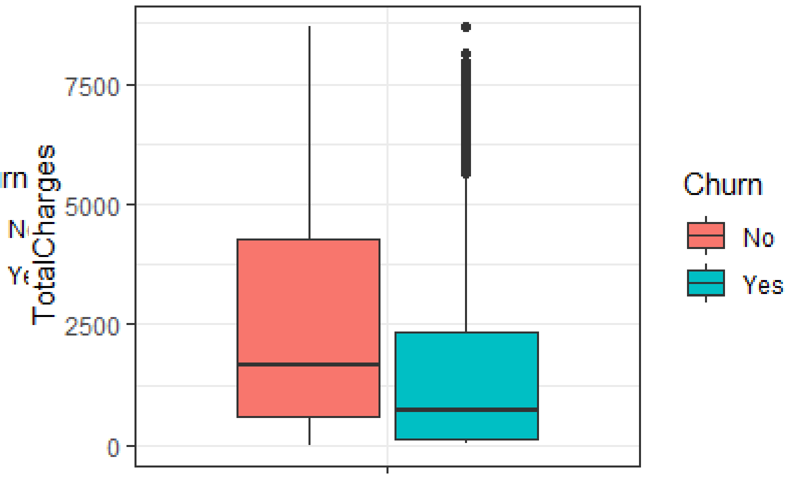


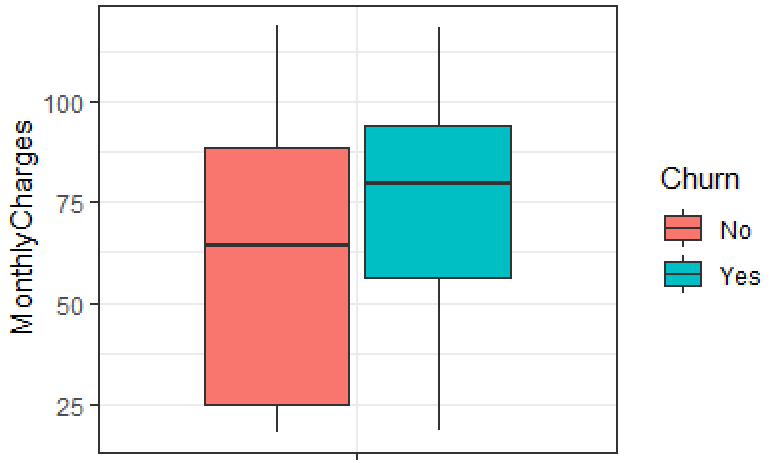
Q6. How are these numeric variables distributed in box plot for 2 churn types?

- According to the boxplot, the lost customers are with lower total charge and tenure but higher monthly charges.
- I think it’s not strange, the lost customers may try the expensive monthly package and then not satisfied with it, and finally lost.

Factor group 3

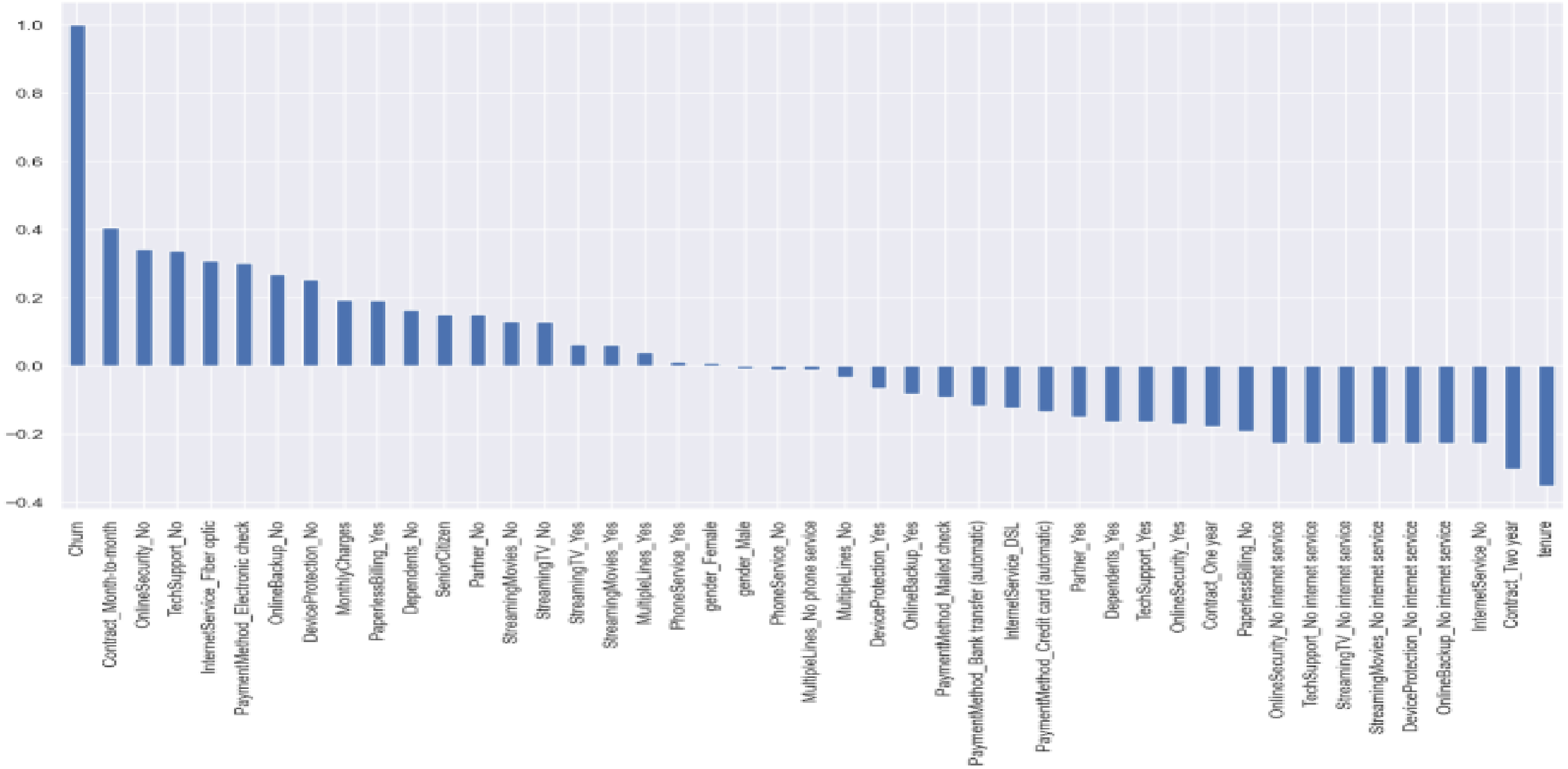






Q7. What’s correlation relationship between them?

According to the bar, I show the correlation coefficient of independent variable with Churn in descent order.
We can find out the high positive relationship, e.g. contact month to month, online security no, internet service fiber optic.
We can find out the high negative relationship, e.g. tenue, contract two year, etc.



6. Summary of Findings

6. Summary of Findings

Positive Factors for low Churn Rate:

- People without using internet service
- Using internet service meanwhile with multiple added value service (e.g. online security)
- People having registered for long time
- People signing contract for long time
- People using paper contract or non electronic payment

Negative Factors for low Churn Rate:

- Internet service using fiber optic
- Senior citizen, people who don’t have dependents, people who don’t have partner
- People using paperless billing, or with electronic check
- People using month to month contract
- People using internet service but without added value service

7. Analysis and Recommendations

7. Analysis and Recommendations

Factor group 1

Concerning negative factor, such as senior citizen, people who don’t have dependents. Their churn rate is high. But it seems they like to try fiber optic (**which is faster and more stable for surfing, but more expensive**), maybe they try one month, and find it has lower cost performance, and then terminate the contract. Meanwhile, they try less added value service, such as online security function.

I suggest Telecom Company plan special fiber optic internet service package for these kind customer to retain these customers. Secondly, maybe don’t promote fiber optic service for the first time, they can promote DSL service and plus the added value service, which have higher cost performance.

Factor group 2

We can find out many lost customers, they use fiber optic, but they don’t use added value service, such as online security, so they use fiber optic only for surfing, which will lead to the lower cost performance.

I suggest Telecom Company give promotion activities, when people use fiber optic, the added value service will be entitled automatically free of charge for certain period, so the customers will feel the cost performance is improved, and think this service is necessary, they will continue to pay.

Factor group 3

We can also find out that when fiber optic pop up, the online payment become more popular way as well, so when people is not satisfied with the service, the termination of contract also become more convenient, which may be also a reason for higher churn rate.

Compared with other payment method, the customers with online payment (electronic check) or using paperless billing, or with short er period contact are easier to be lost.

I think the paperless billing is the trend, we can’t change it. So I suggest the supplier can consider to adopt promotion activities, such as discount or giving a reward when consumption above certain value, so as to guide the customer to conclude the contract with longer period

.

To sum up

The fiber-optic is the most important factor for high churn rate, or we can say its service price. It’s a double edged sword. This new technology bring about many advantages, such as faster speed and more stable signal, however, it is very expensive.

The supplier have made great publicity for this technology, everyone shows great interest. So whether the young or seniors, whether male or female, everyone would like to try this new surfing way. Nevertheless, after 1 month experience, people find its cost performance is low, and gradually give up the trying. They even didn’t use the added value service during this period, or they maybe don’t know this function.

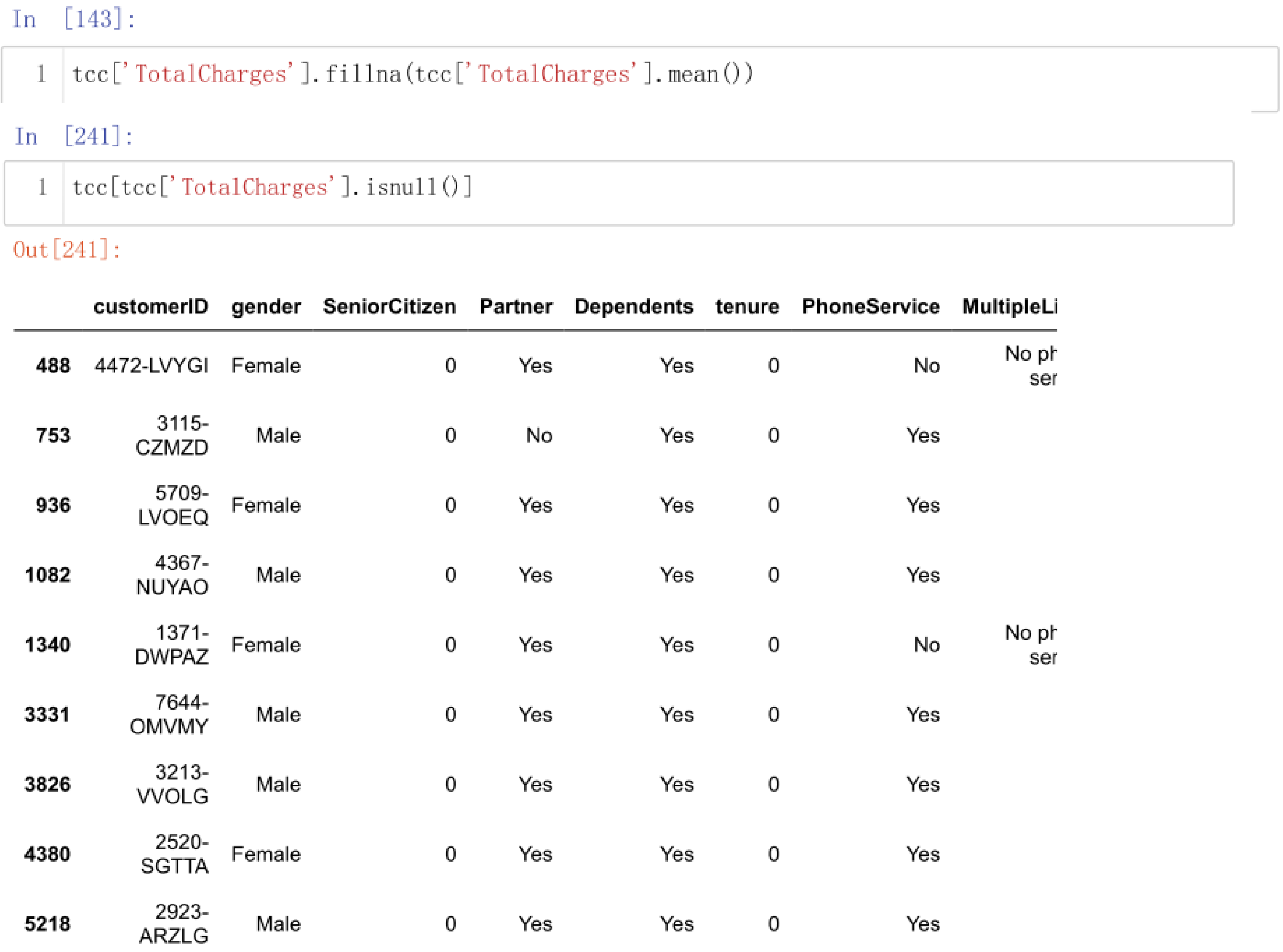
From my perspective, the most effective way to improve churn rate is to use cost-volume-profit (PVC) analysis to find a balance price for fiber optic internet service. The supplier should try to reduce the fiber optic service price to attract more customers. As a result, the competitive price will enable these customers more sticker.

In other words, the Telecom company should improve the cost performance of fiber optic, maybe when customers choose this way, they can entitled to certain added value service free of charge, so as to improve their satisfaction.

Furthermore, the electronic monthly payment may lead to their easy termination, I think if we can’t reduce the price directly, it would be better to use reward policy, such as when you charge for at least one year, will be rewarded for another year, or maybe can entitled to some added value service, such a mobile phone free of charge.

Appendix：Partial Coding

Data exploring：



In [143]:

```
tcc['TotalCharges'].fillna(tcc['TotalCharges'].mean())
```

In [241]:

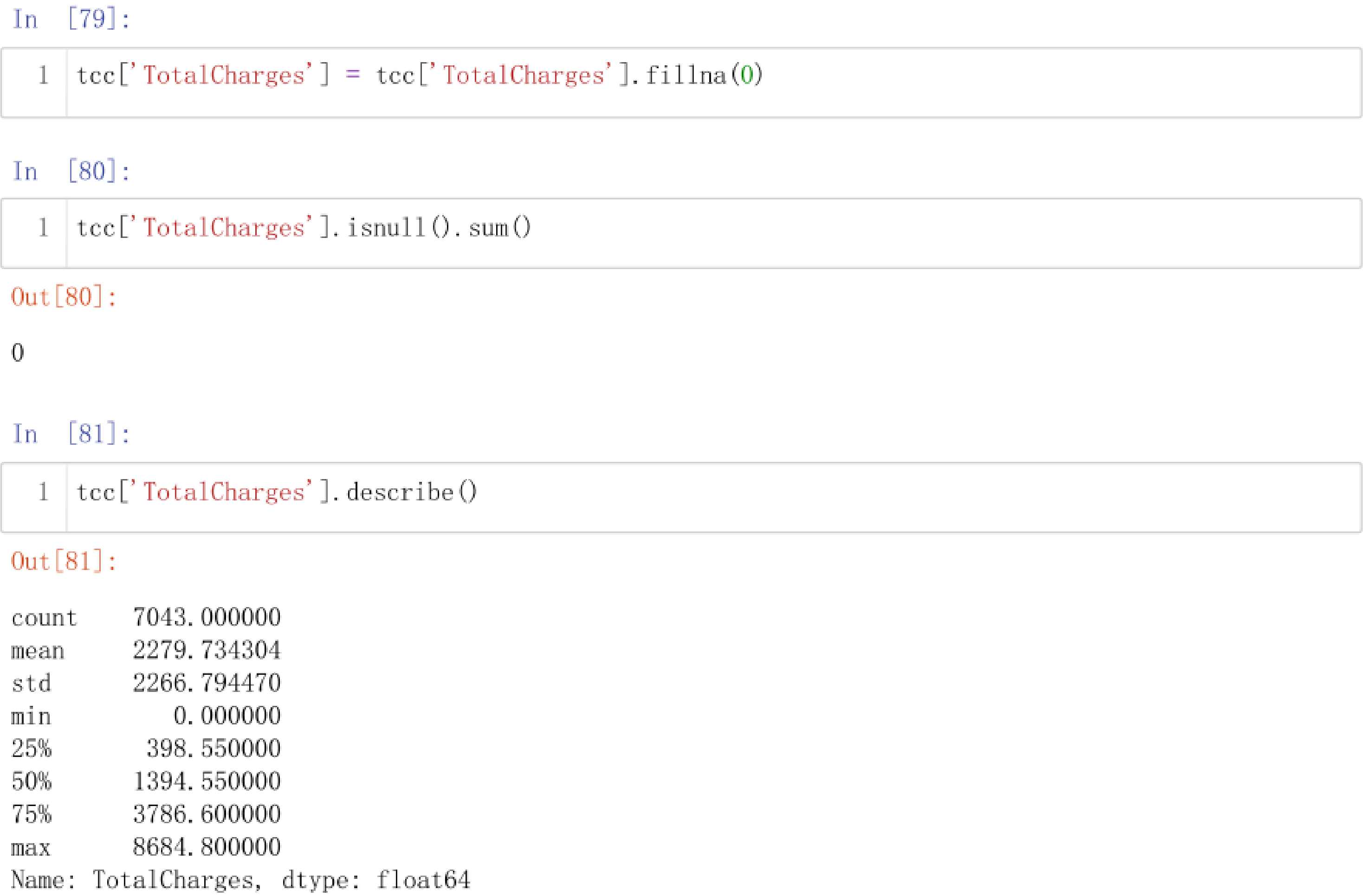
```
tcc[tcc['TotalCharges'].isnull()]
```

Out[241]:

	customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLi
488	4472-LVYGI	Female	0	Yes	Yes	0	No	No ph ser
753	3115-CZMZD	Male	0	No	Yes	0	Yes	
936	5709-LVOEQ	Female	0	Yes	Yes	0	Yes	
1082	4367-NUYAO	Male	0	Yes	Yes	0	Yes	
1340	1371-DWPAZ	Female	0	Yes	Yes	0	No	No ph ser
3331	7644-OMVMY	Male	0	Yes	Yes	0	Yes	
3826	3213-VVOLG	Male	0	Yes	Yes	0	Yes	
4380	2520-SGTTA	Female	0	Yes	Yes	0	Yes	
5218	2923-ARZLG	Male	0	Yes	Yes	0	Yes	

Appendix：Partial Coding

Data exploring：



```
In  [79]:

tcc['TotalCharges'] = tcc['TotalCharges'].fillna(0)
```

```
In  [80]:

tcc['TotalCharges'].isnull().sum()
```

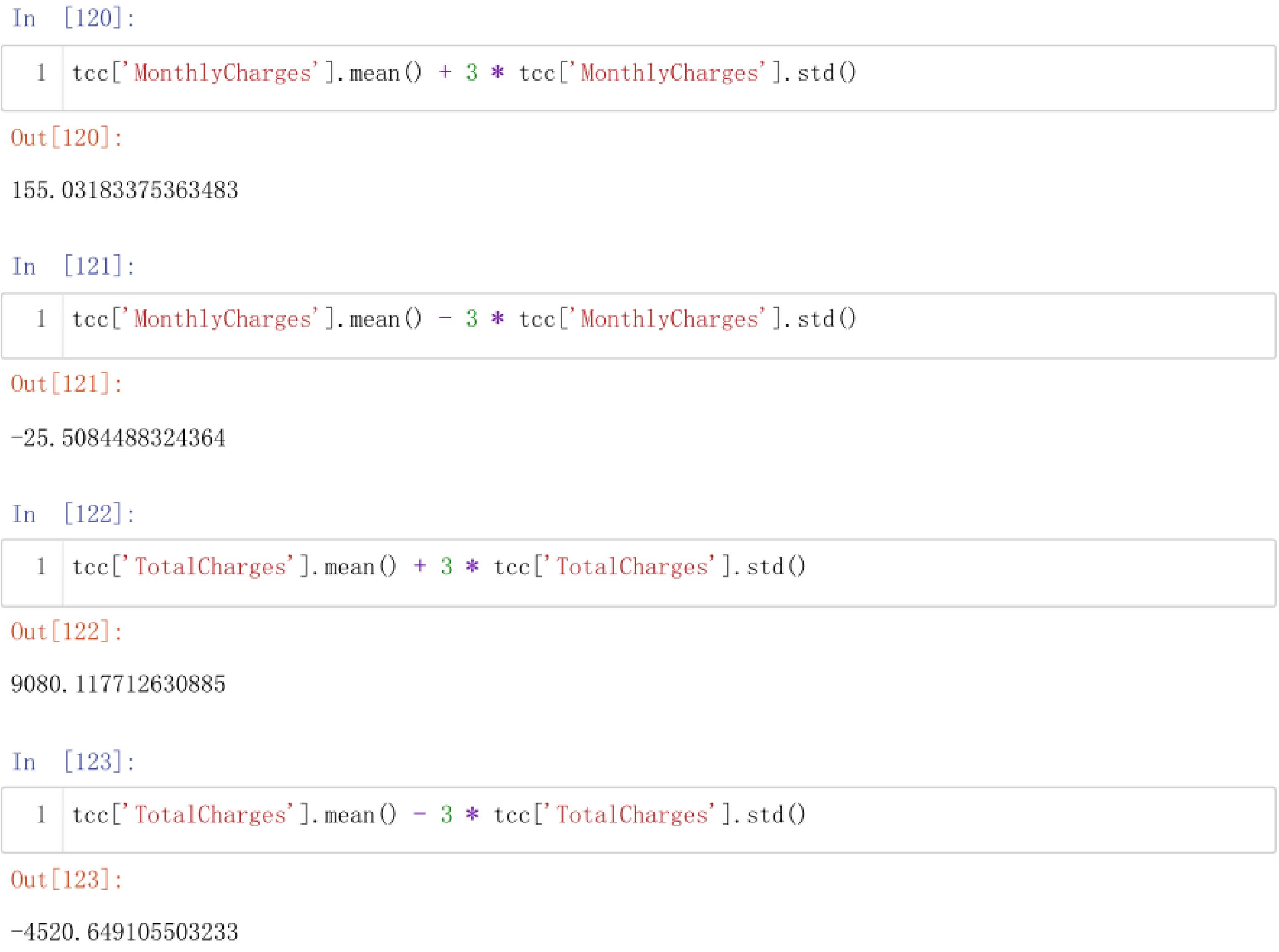
```
Out[80]:

0
```

```
In  [81]:

tcc['TotalCharges'].describe()
```

```
Out[81]:

count    7043.000000
mean     2279.734304
std      2266.794470
min         0.000000
25%       398.550000
50%      1394.550000
75%      3786.600000
max      8684.800000
Name: TotalCharges, dtype: float64
```

Appendix：Partial Coding

Abnormal check：



In [120]:

```
tcc['MonthlyCharges'].mean() + 3 * tcc['MonthlyCharges'].std()
```

Out[120]:

```
155.03183375363483
```

In [121]:

```
tcc['MonthlyCharges'].mean() - 3 * tcc['MonthlyCharges'].std()
```

Out[121]:

```
-25.5084488324364
```

In [122]:

```
tcc['TotalCharges'].mean() + 3 * tcc['TotalCharges'].std()
```

Out[122]:

```
9080.117712630885
```

In [123]:

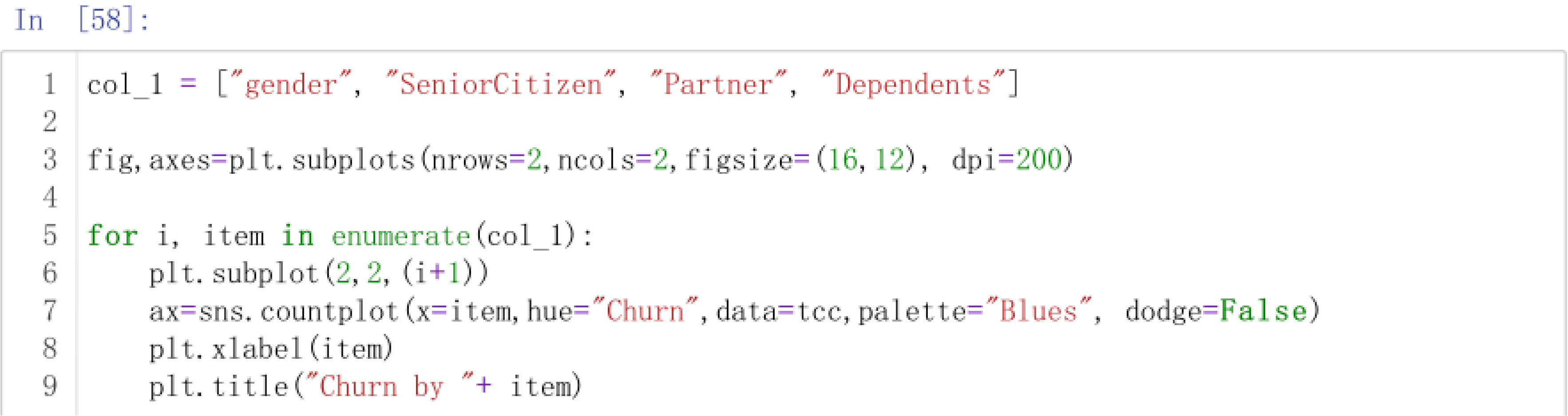
```
tcc['TotalCharges'].mean() - 3 * tcc['TotalCharges'].std()
```

Out[123]:

```
-4520.649105503233
```

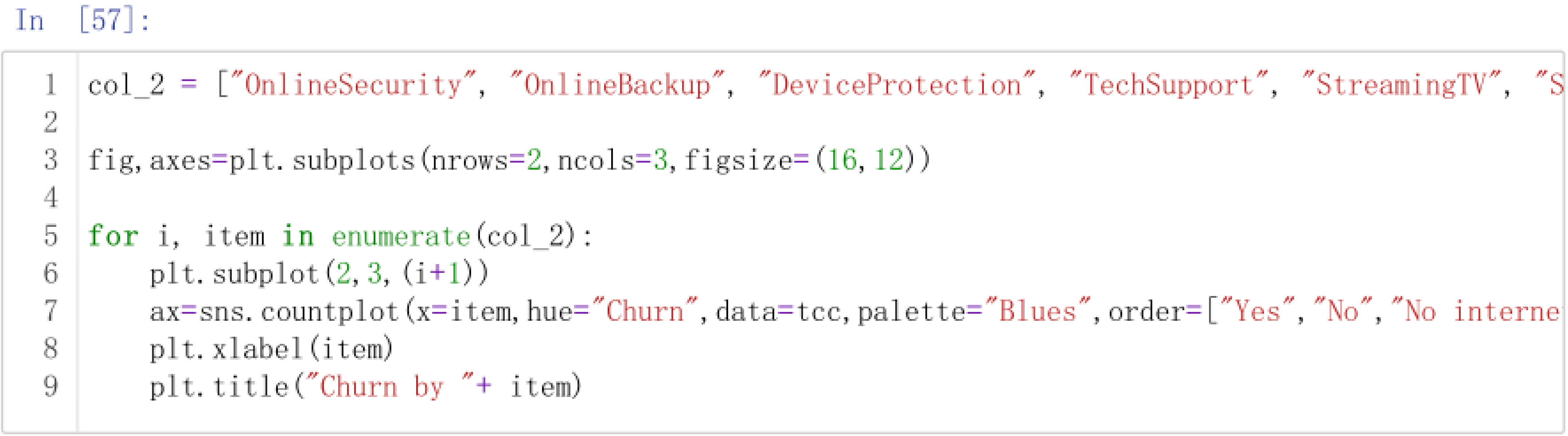
Appendix：Partial Coding

Further data exploring：



In [58]:

```
col_1 = ["gender", "SeniorCitizen", "Partner", "Dependents"]

fig,axes=plt.subplots(nrows=2,ncols=2,figsize=(16,12), dpi=200)

for i, item in enumerate(col_1):
    plt.subplot(2,2,(i+1))
    ax=sns.countplot(x=item,hue="Churn",data=tcc,palette="Blues", dodge=False)
    plt.xlabel(item)
    plt.title("Churn by "+ item)
```

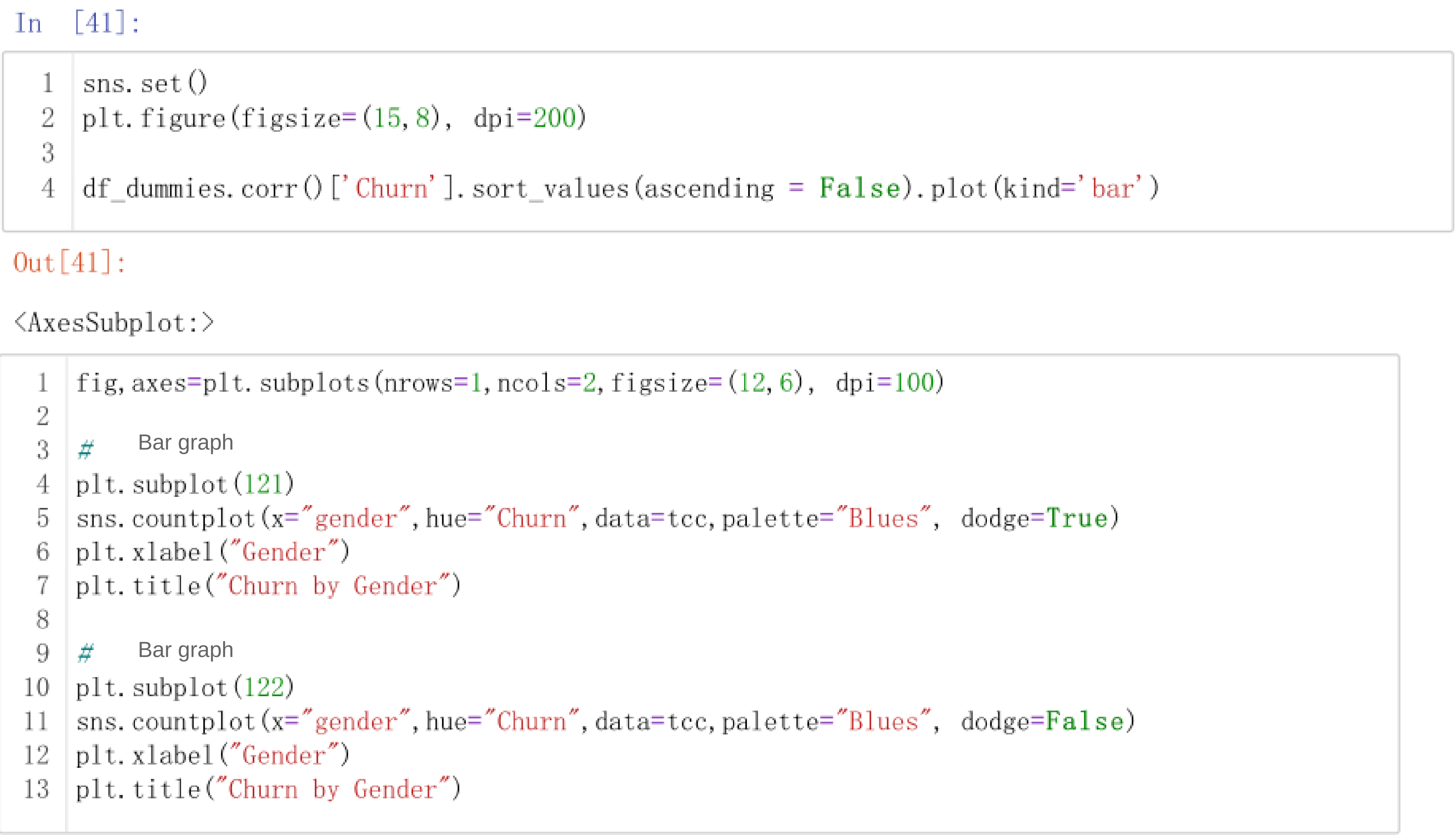


In [57]:

```
col_2 = ["OnlineSecurity", "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV", "S

fig,axes=plt.subplots(nrows=2,ncols=3,figsize=(16,12))

for i, item in enumerate(col_2):
    plt.subplot(2,3,(i+1))
    ax=sns.countplot(x=item,hue="Churn",data=tcc,palette="Blues",order=["Yes","No","No interne
    plt.xlabel(item)
    plt.title("Churn by "+ item)
```

Appendix：Partial Coding

Feature correlation：



In [41]:

```
sns.set()
plt.figure(figsize=(15,8), dpi=200)

df_dummies.corr()['Churn'].sort_values(ascending = False).plot(kind='bar')
```

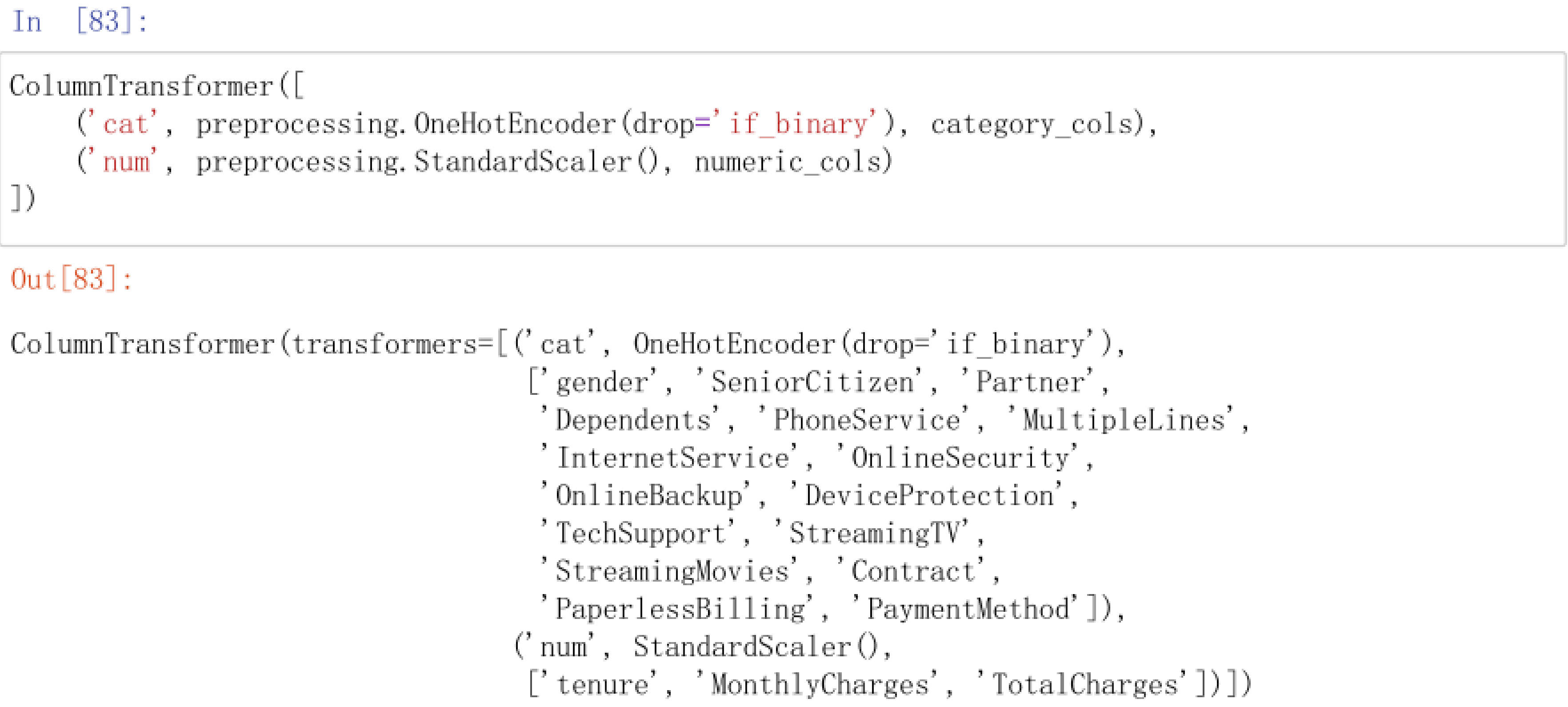
Out[41]:

```
<AxesSubplot:>
```

```
fig,axes=plt.subplots(nrows=1,ncols=2,figsize=(12,6), dpi=100)

#     Bar graph
plt.subplot(121)
sns.countplot(x="gender",hue="Churn",data=tcc,palette="Blues", dodge=True)
plt.xlabel("Gender")
plt.title("Churn by Gender")

#     Bar graph
plt.subplot(122)
sns.countplot(x="gender",hue="Churn",data=tcc,palette="Blues", dodge=False)
plt.xlabel("Gender")
plt.title("Churn by Gender")
```

Appendix：Partial Coding

Column transforming:



In [83]:

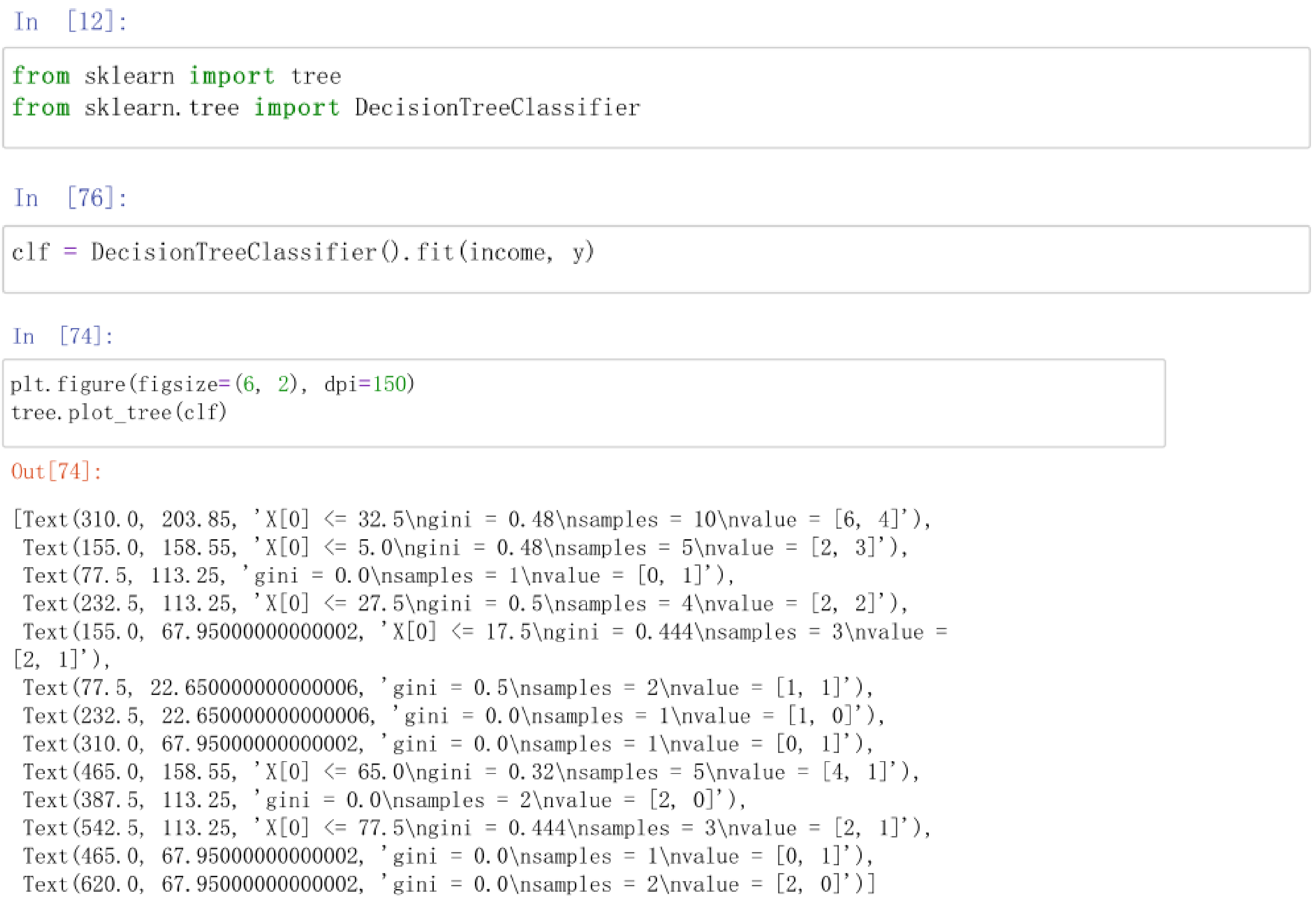
```
ColumnTransformer([
    ('cat', preprocessing.OneHotEncoder(drop='if_binary'), category_cols),
    ('num', preprocessing.StandardScaler(), numeric_cols)
])
```

Out[83]:

```
ColumnTransformer(transformers=[('cat', OneHotEncoder(drop='if_binary'),
                                 ['gender', 'SeniorCitizen', 'Partner',
                                  'Dependents', 'PhoneService', 'MultipleLines',
                                  'InternetService', 'OnlineSecurity',
                                  'OnlineBackup', 'DeviceProtection',
                                  'TechSupport', 'StreamingTV',
                                  'StreamingMovies', 'Contract',
                                  'PaperlessBilling', 'PaymentMethod']),
                                ('num', StandardScaler(),
                                 ['tenure', 'MonthlyCharges', 'TotalCharges'])])
```

Appendix： Partial Coding

Take income as feature, y as label to train the decision tree：



In [12]:

```
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
```

In [76]:

```
clf = DecisionTreeClassifier().fit(income, y)
```

In [74]:

```
plt.figure(figsize=(6, 2), dpi=150)
tree.plot_tree(clf)
```

Out[74]:

```
[Text(310.0, 203.85, 'X[0] <= 32.5\ngini = 0.48\nsamples = 10\nvalue = [6, 4]'),
 Text(155.0, 158.55, 'X[0] <= 5.0\ngini = 0.48\nsamples = 5\nvalue = [2, 3]'),
 Text(77.5, 113.25, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(232.5, 113.25, 'X[0] <= 27.5\ngini = 0.5\nsamples = 4\nvalue = [2, 2]'),
 Text(155.0, 67.95000000000002, 'X[0] <= 17.5\ngini = 0.444\nsamples = 3\nvalue =
[2, 1]'),
 Text(77.5, 22.650000000000006, 'gini = 0.5\nsamples = 2\nvalue = [1, 1]'),
 Text(232.5, 22.650000000000006, 'gini = 0.0\nsamples = 1\nvalue = [1, 0]'),
 Text(310.0, 67.95000000000002, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(465.0, 158.55, 'X[0] <= 65.0\ngini = 0.32\nsamples = 5\nvalue = [4, 1]'),
 Text(387.5, 113.25, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]'),
 Text(542.5, 113.25, 'X[0] <= 77.5\ngini = 0.444\nsamples = 3\nvalue = [2, 1]'),
 Text(465.0, 67.95000000000002, 'gini = 0.0\nsamples = 1\nvalue = [0, 1]'),
 Text(620.0, 67.95000000000002, 'gini = 0.0\nsamples = 2\nvalue = [2, 0]')]
```

Thank you