Analysis Project - Customer Churn

Jonilyto Jean Georges Junior JEAN LOUIS

June, 2020

1. Introduction

1.1 Background

Customers churn occurs when customers or subscribers stop business with a company. Businesses are very keen on measuring churn because keeping an existing customers is far less expensive than acquiring a new customer. New business involves working leads through a sales funnel, using marketing and sales budgets to gain additional customers. Existing customers will often have a higher volume of service consumption and can generate additional customer referrals

Preventing customer churn is critically important to the Telecommunications sector, as the barriers to entry for switching services are so low.

1.2 Problem

The loss of customers is the company's main problem.

Why is this problem important to the organization?

Losing customers continuously reduces the company's income, moreover, the loss of customers may cause the company to lose its leading position in the market.

1.3 Interest

The marketing team and the customer service team.

Each team has its own reason for wanting the analysis. The marketing team wants to find out who the most likely people to churn are and create content that suits their interests. The customer service team would like to proactively reach out to customers who are about to churn, and try to encourage them to stay.

1.4 Context

Analysis of Telecom company customer database, with information about the attributes of its customers. The intention customers with greater potential to leave company.

2. Data Acquisition

2.1 Data Acquisition

The data acquired for this project, is at the following address
https://www.kaggle.com/blastchar/telco-customer-churn/data#

2.2 Content

Each row represents a customer, each column contains customers attributes described on the column Metadata.

2.3 The dataset includes information about:

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

3. <u>Methodology</u>

3.1 Exploratory Data Analysis

3.1.1 Statistical summary of Senior Citizen, tenure, Total Charges, Monthly Charges

The describe function in python is used to get statistics of the Customer-Churn, this returns the mean, standard deviation, minimum, maximum, 1st quartile (25%), 2nd quartile (50%), and the 3rd quartile (75%).(See fig 3.1.1)

Entrée [277]:
```
df.describe()
```

Out[277]:

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7032.000000	7032.000000	7032.000000	7032.000000
mean	0.162400	32.421786	64.798208	2283.300441
std	0.368844	24.545260	30.085974	2266.771362
min	0.000000	1.000000	18.250000	18.800000
25%	0.000000	9.000000	35.587500	401.450000
50%	0.000000	29.000000	70.350000	1397.475000
75%	0.000000	55.000000	89.862500	3794.737500
max	1.000000	72.000000	118.750000	8684.800000

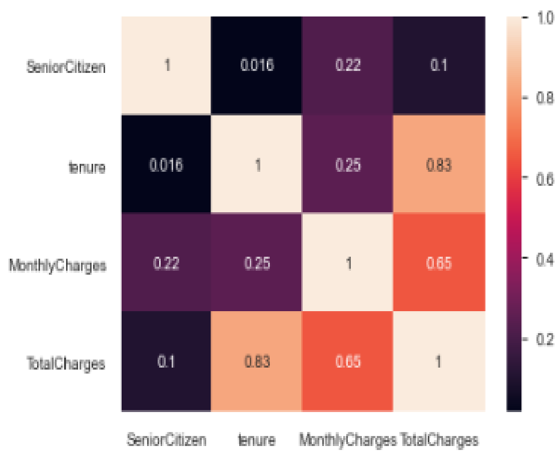
Fig 3.1.1 Statistical description of the Customer-Churn

Correlation Matrix

Correlation Matrix

Entrée [37]:
```
# Let's check for the Correlation Matrix in seaborn
sns.heatmap(df.corr(),xticklabels=df.corr().columns.values,yticklabels=df.corr().columns.values, annot=True)
```

Out[37]: <matplotlib.axes._subplots.AxesSubplot at 0x181cec10>



Entrée [38]:
```
df.corr()
```

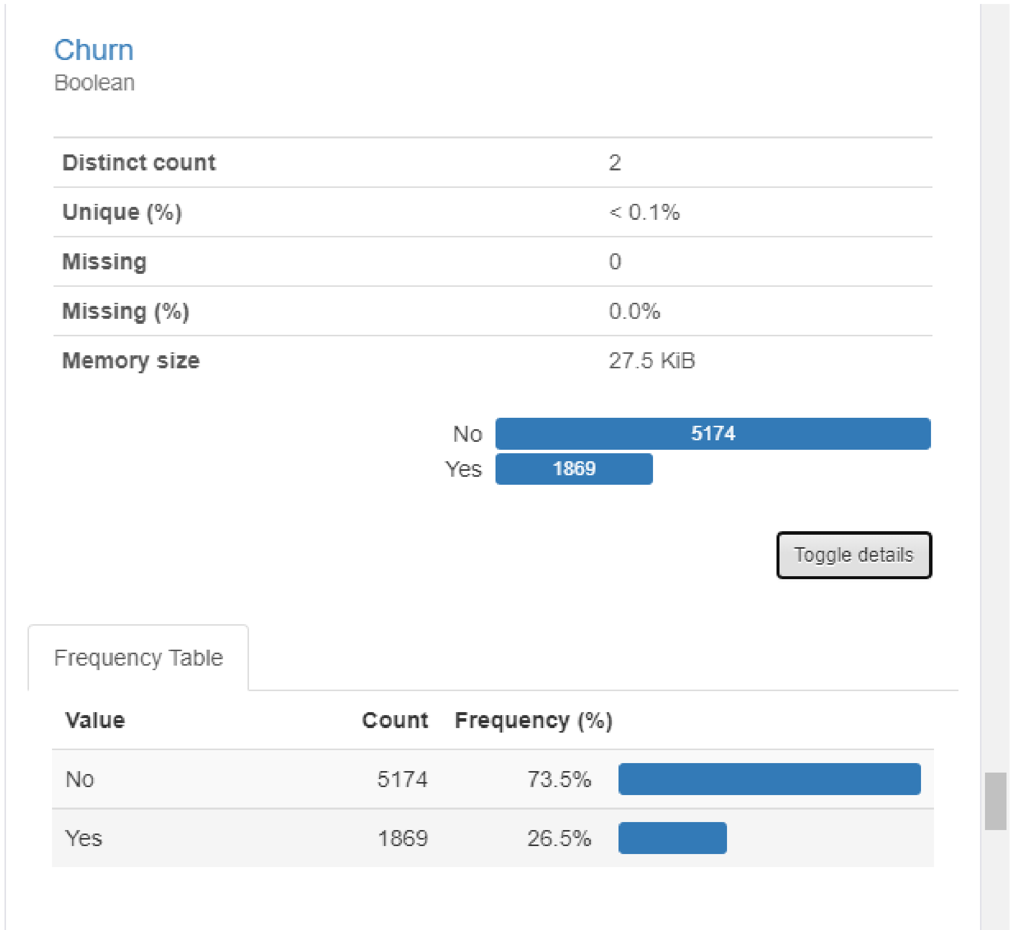
Out[38]:

	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
SeniorCitizen	1.000000	0.015683	0.219874	0.102411
tenure	0.015683	1.000000	0.246862	0.825880
MonthlyCharges	0.219874	0.246862	1.000000	0.651065
TotalCharges	0.102411	0.825880	0.651065	1.000000

We can see tenure and Total Charge are correlate and also Monthly Charges and Total Charges are also correlate each other. So this is proving our first Hypothesis right of considering Total Charges = Monthly Charges * tenure + Additional Tax that we had taken above.

We will use the variable "Churn" to do our analysis and also to make our prediction.

Now let's look at how many customers have churn and the percentage of churn. (See Fig 3.1.2)


Churn
Boolean

Distinct count	2
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	27.5 KiB

Value	Count	Frequency (%)
No	5174	73.5%
Yes	1869	26.5%

Fig 3.1.2 customers have churn and the percentage of churn

As we can see, more than 26% of company's population have churned.

We will visualize others variables as we will perform our analysis.

Now let's start comparing

Gender vs Churn

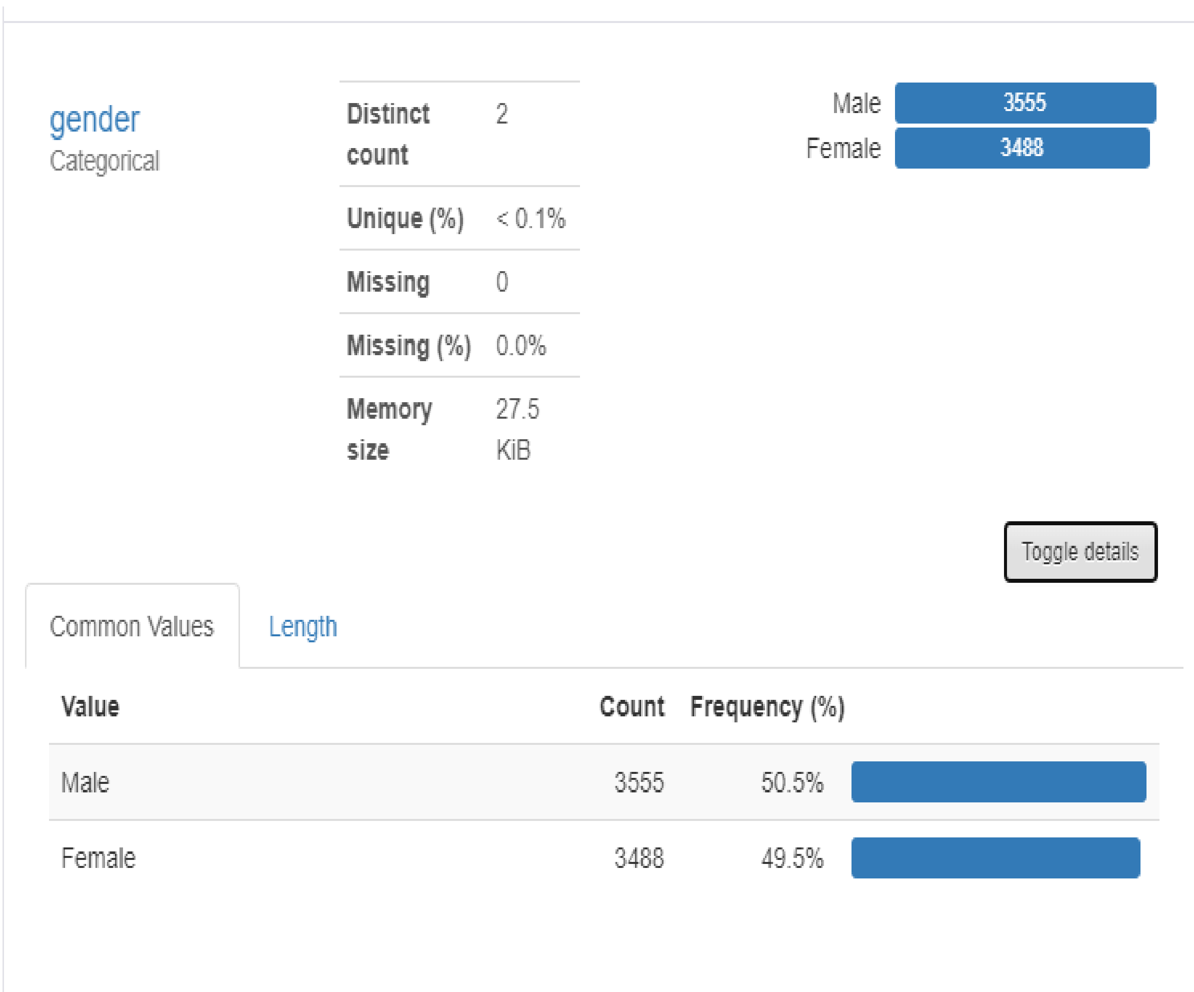
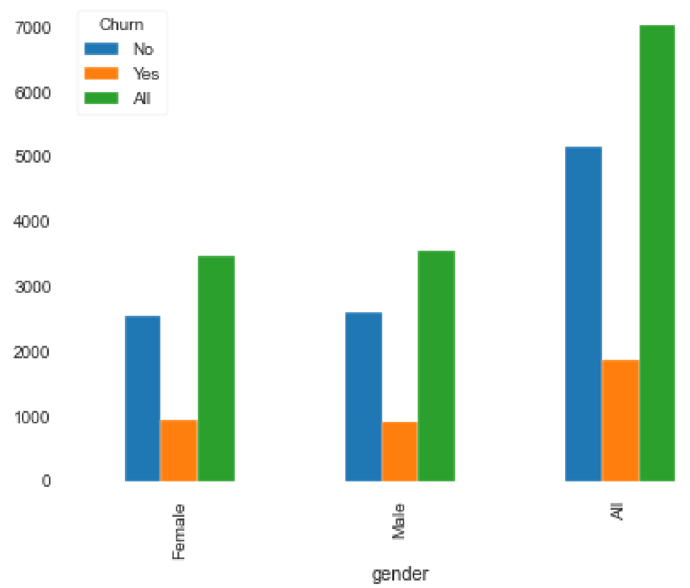


Fig 3.1.2 total customers and the percentage by gender

Entrée [262]:
```
# Now let's start comparing
# Gender vs Churn
print (pd.crosstab(df.gender, df.Churn, margins=True))
pd.crosstab(df.gender, df.Churn, margins=True).plot(kind='bar', figsize=(7,5))
```

```
Churn     No   Yes   All
gender
Female  2544   939  3483
Male    2619   930  3549
All     5163  1869  7032
```

Out[262]: <matplotlib.axes._subplots.AxesSubplot at 0x209ed988>



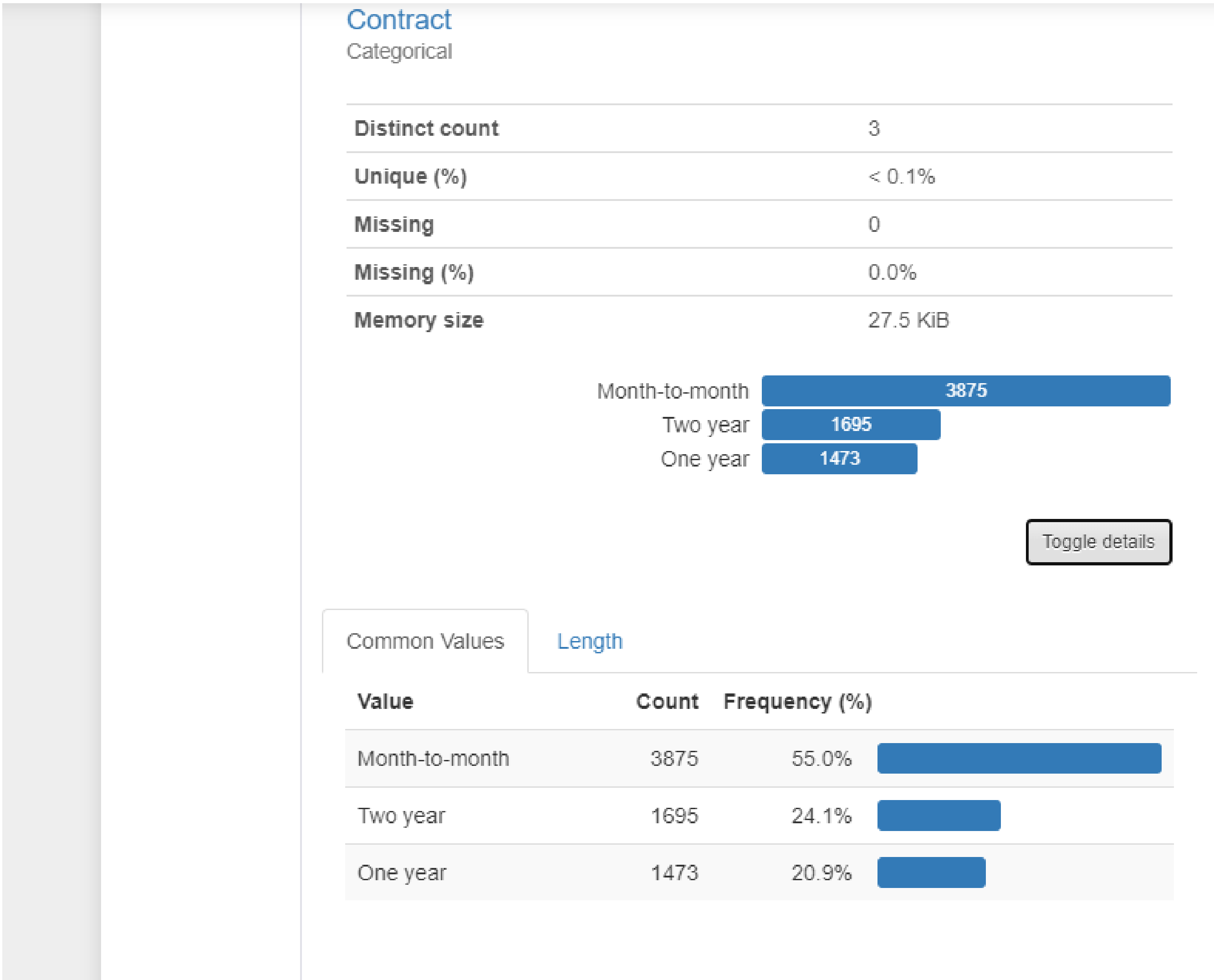
Entrée [263]:
```
print(' Percent of females that left the company  {0}'.format((939/1869)*100))
print(' Percent of males that left the company  {0}'.format((930/1869)*100))
```

```
Percent of females that left the company  50.24077046548957
Percent of males that left the company  49.75922953451043
```

Fig 3.1.3 customers have churn and the percentage of churn by gender

We can see that gender doesn't play an important role in predicting our target variable.

Contact vs Churn



Contract
Categorical

Distinct count	3
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	27.5 KiB

Month-to-month	3875
Two year	1695
One year	1473

Toggle details

Common Values | Length

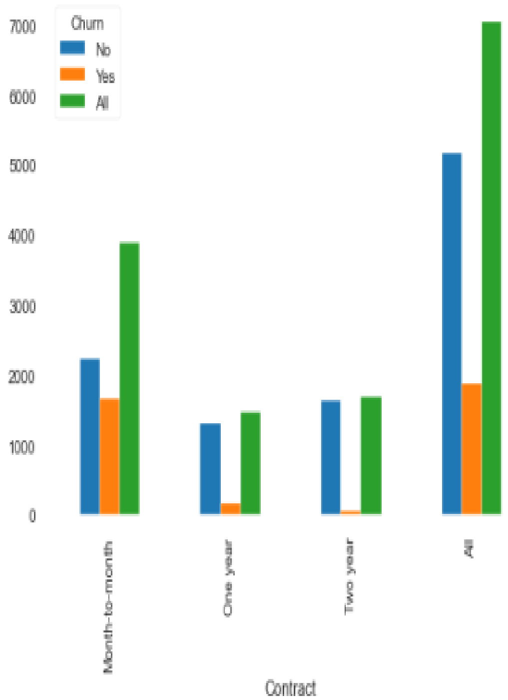
Value	Count	Frequency (%)
Month-to-month	3875	55.0%
Two year	1695	24.1%
One year	1473	20.9%

Fig 3.1.4 total customers and the percentage by contract

```
Entrée [264]: # Contact vs Churn
print (pd.crosstab(df.Contract, df.Churn, margins=True))
pd.crosstab(df.Contract, df.Churn, margins=True).plot(kind='bar', figsize=(7,5))
```

```
Churn             No   Yes   All
Contract
Month-to-month  2220  1655  3875
One year        1306   166  1472
Two year        1637    48  1685
All             5163  1869  7032
```

```
Out[264]: <matplotlib.axes._subplots.AxesSubplot at 0x20983b98>
```




```
Entrée [265]: print(' Percent of Month-to-Month Contract that left the company  {0}'.format((1655 /1869)*100))
print(' Percent of one-year that left the company  {0}'.format((166 /1869)*100))
print(' Percent of two-years that left the company  {0}'.format((48/1869)*100))
```

```
Percent of Month-to-Month Contract that left the company  88.55002675227395
Percent of one-year that left the company  8.881754949170679
Percent of two-years that left the company  2.568218298555377
```

Fig 3.1.5 customers have churn and the percentage of churn by Contract

Most of the people that left were the Ones who had Month-to-Month Contract

Internet Service vs Churn

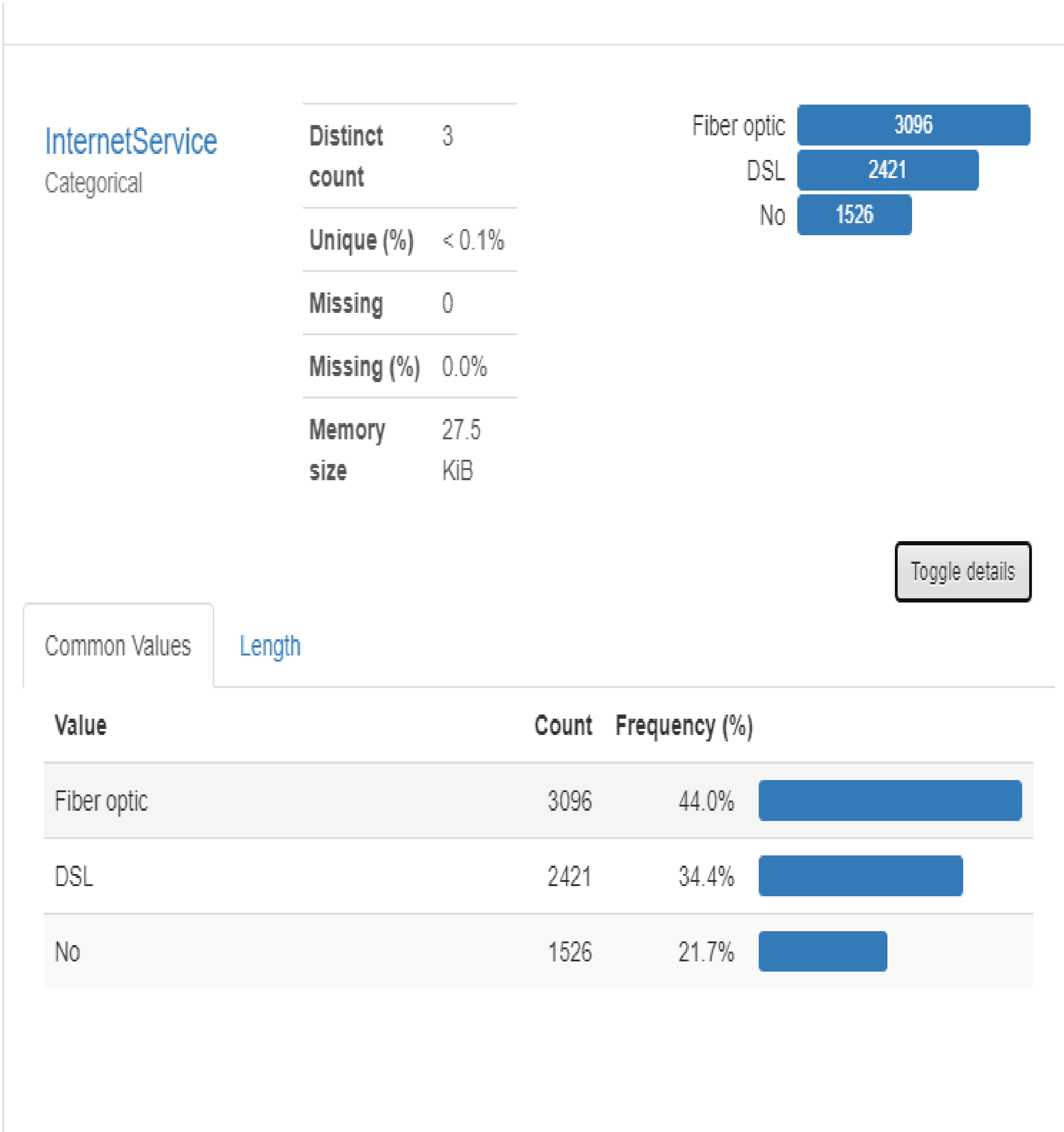
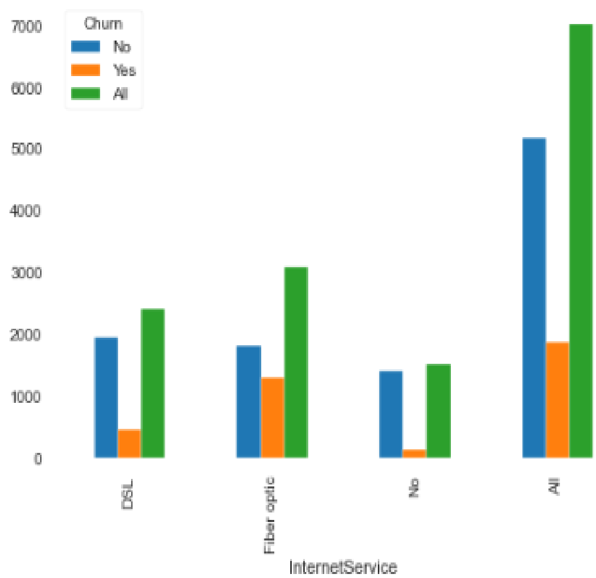


Fig 3.1.6 total customers and the percentage by Internet Service

Entrée [25]:
```
# Internet Service vs Churn
print (pd.crosstab(df.InternetService, df.Churn, margins=True))
pd.crosstab(df.InternetService, df.Churn, margins=True).plot(kind='bar', figsize=(7,5))
```

```
Churn              No   Yes   All
InternetService
DSL              1957   459  2416
Fiber optic      1799  1297  3096
No               1407   113  1520
All              5163  1869  7032
```

Out[25]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1691a430>
```



Entrée [26]:
```
print(' Percent of DSL InternetService that left the company  {0}'.format((459 /1869)*100))
print(' Percent of Fiber Optic InternetService that left the company {0}'.format((1297 /1869)*100))
print(' Percent of No InternetService that left the company  {0}'.format((113/1869)*100))
```

```
 Percent of DSL InternetService that left the company  24.558587479935795
 Percent of Fiber Optic InternetService that left the company 69.39539860888175
 Percent of No InternetService that left the company  6.046013911182451
```

Fig 3.1.7 customers have churn and the percentage of churn by Internet Service.

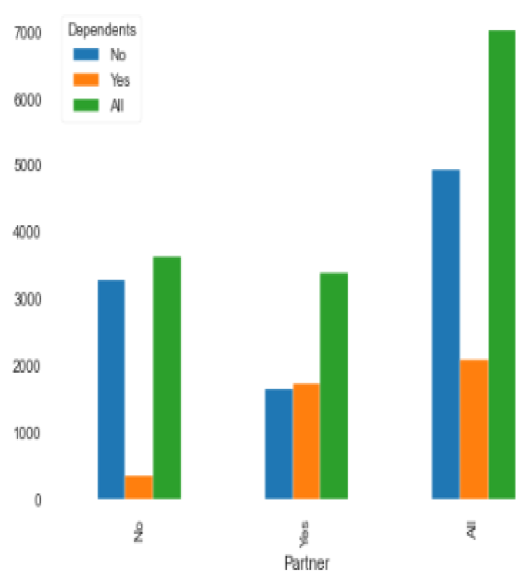
Most of the people that left had Fiber Optic Internet-Service.

Partner vs Dependents

Entrée [28]:
```
# Partner vs Dependents
print (pd.crosstab(df.Partner, df.Dependents, margins=True))
pd.crosstab(df.Partner, df.Dependents, margins=True).plot(kind='bar', figsize=(7,5))
```

```
Dependents    No   Yes   All
Partner
No          3280   359  3639
Yes         1653  1740  3393
All         4933  2099  7032
```

Out[28]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x17cc3430>
```



Entrée [29]:
```
print(' Percent of Partner that had Dependents  {0}'.format((1749 /2110)*100))
print(' Percent of Non-Partner that had Dependents  {0}'.format((361 /2110)*100))
```

```
Percent of Partner that had Dependents  82.8909952606635
Percent of Non-Partner that had Dependents  17.10900473933649
```

Fig 3.1.7

We can see Partners had much larger percent of Dependents than Non-Partner this tells us that most Partners might be married.

Senior Citizen vs Churn

Entrée [33]:
```
# SeniorCitizen vs Churn
print (pd.crosstab(df.SeniorCitizen, df.Churn, margins=True))
pd.crosstab(df.SeniorCitizen, df.Churn, margins=True).plot(kind='bar', figsize=(7,5))
```

```
Churn            No   Yes   All
SeniorCitizen
0              4497  1393  5890
1               666   476  1142
All            5163  1869  7032
```

Out[33]: <matplotlib.axes._subplots.AxesSubplot at 0x1696c148>

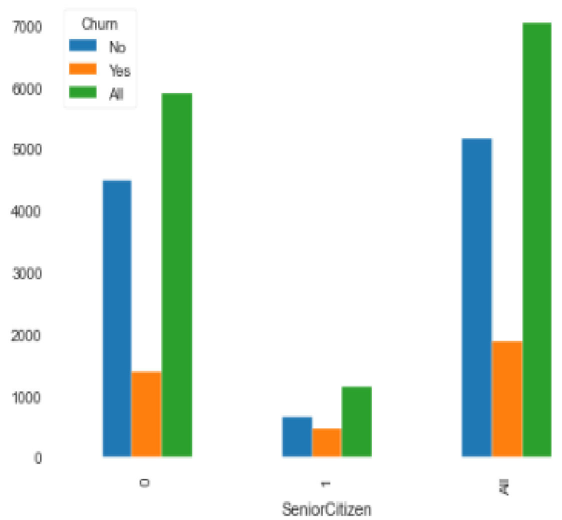


Fig 3.1.7

When Churn ="Yes"

```
Entrée [285]: df.loc[(df.Churn== 'Yes'), 'MonthlyCharges'].median()

   Out[285]: 79.65

Entrée [286]: df.loc[(df.Churn== 'Yes'), 'TotalCharges'].median()

   Out[286]: 703.55

Entrée [287]: df.loc[(df.Churn== 'Yes'), 'tenure'].median()

   Out[287]: 10.0

Entrée [288]: df.loc[(df.Churn== 'Yes'), 'PaymentMethod'].value_counts(normalize = True)

   Out[288]: Electronic check             0.573034
             Mailed check                 0.164794
             Bank transfer (automatic)    0.138042
             Credit card (automatic)      0.124131
             Name: PaymentMethod, dtype: float64
```

Fig 3.1.8

Most of the people that left are the ones who had Payment Method as Electronic check, so let's make a separate variable for it so that the model can easily predict our target variable.

```
Entrée [289]: df['Is_Electronic_check']= np.where(df['PaymentMethod'])=='Electronic check'
```

```
Entrée [290]: df.loc[(df.Churn== 'Yes'), 'PaperlessBilling'].value_counts(normalize = True)

   Out[290]: Yes    0.749064
             No     0.250936
             Name: PaperlessBilling, dtype: float64
```

```
Entrée [291]: df.loc[(df.Churn== 'Yes'), 'DeviceProtection'].value_counts(normalize = True)

   Out[291]: No                     0.64794
             Yes                    0.29160
             No internet service    0.06046
             Name: DeviceProtection, dtype: float64
```

```
Entrée [292]: df.loc[(df.Churn== 'Yes'), 'OnlineBackup'].value_counts(normalize = True)

   Out[292]: No                     0.659711
             Yes                    0.279829
             No internet service    0.060460
             Name: OnlineBackup, dtype: float64
```

```
Entrée [293]: df.loc[(df.Churn== 'Yes'), 'TechSupport'].value_counts(normalize = True)

   Out[293]: No                     0.773676
             Yes                    0.165864
             No internet service    0.060460
             Name: TechSupport, dtype: float64
```

```
Entrée [294]: df.loc[(df.Churn== 'Yes'), 'OnlineSecurity'].value_counts(normalize = True)

   Out[294]: No                     0.781701
             Yes                    0.157838
             No internet service    0.060460
             Name: OnlineSecurity, dtype: float64
```

Fig 3.1.9

We can see that people that left the company didn't use Services like Online Security, Device Protection, Tech Support and Online Backup quite often. Hence for our Prediction these variables will not be much importance. We will drop them in the end.

Modelling Part

Entrée [129]:

```
# Compare several models according to their Accuracies
Model_Comparison =pd.DataFrame({
    'Model': ['Logistic Regression', 'Support Vector Machine', 'K-Nearest Neighbors',
              'Decision Tree', 'Random Forest'],
    'Score': [logmodel_accuracy, svc_accuracy, knn_accuracy, dt_accuracy, rf_accuracy ]})
Model_Comparison_df = Model_Comparison.sort_values(by= 'Score', ascending = False)
Model_Comparison_df = Model_Comparison.set_index('Score')
Model_Comparison_df.reset_index()
```

Out[129]:

	Score	Model
0	79.91	Logistic Regression
1	79.91	Support Vector Machine
2	79.91	K-Nearest Neighbors
3	79.91	Decision Tree
4	79.91	Random Forest

Entrée [ ]:

Entrée [130]:

```
# Generate  Confusion Matrix for logistics regression model as it has maximum accuracy
from sklearn.metrics import confusion_matrix
conf_mat_logmodel = confusion_matrix(y_test, pred)
conf_mat_logmodel
```

Out[130]:

```
array([[1395,  166],
       [ 258,  291]], dtype=int64)
```

Predict the probability of Churn of each customer

Entrée [131]:
```
# Predict the probability of Churn of each customer
df['Probability_of_Churn'] = logmodel.predict_proba(df[X_test.columns])[:,1]
```

Entrée [132]:
```
# Create a Dataframe showcasing probability of Churn of each customer
df[['customerID','Probability_of_Churn']].head(15)
```

Out[132]:

	customerID	Probability_of_Churn
0	7590-VHVEG	0.402667
1	5575-GNVDE	0.062546
2	3668-QPYBK	0.351959
3	7795-CFOCW	0.081756
4	9237-HQITU	0.626622
5	9305-CDSKC	0.784992
6	1452-KIOVK	0.579886
7	6713-OKOMC	0.300425
8	7892-POOKP	0.596614
9	6388-TABGU	0.030536
10	9763-GRSKD	0.262717
11	7469-LKBCI	0.023724
12	8091-TTVAX	0.117109
13	0280-XJGEX	0.435657
14	5129-JLPIS	0.583301

Discussion

1. How much is churn affecting the business? How big is churn compared to the existing customer base?

R) As we have seen on a total of 7043 clients, 1869 churn or a value of 26.57% in the last month. This is really a lot, especially when we know that churn is only acceptable if it is less than 10%.

On we look at in terms of revenue how much we will see for the 26.5% of customers who churn the company has lost 30.53% of its revenue in one month and for its Total Charges 17.83%. So we can see that the company is really affected.

```
Entrée [51]: df[['Churn','MonthlyCharges']].groupby(['Churn']).MonthlyCharges.sum().to_frame()
```

Out[51]:

Churn	MonthlyCharges
No	316530.15
Yes	139130.85

```
Entrée [52]: df[['Churn','MonthlyCharges']].groupby(['Churn']).MonthlyCharges.sum().to_frame()/df['MonthlyCharges'].sum()
```

Out[52]:

Churn	MonthlyCharges
No	0.694661
Yes	0.305339

```
Entrée [53]: df[['Churn','TotalCharges']].groupby(['Churn']).TotalCharges.sum().to_frame()
```

Out[53]:

Churn	TotalCharges
No	13193241.8
Yes	2862926.9

```
Entrée [55]: df[['Churn','TotalCharges']].groupby(['Churn']).TotalCharges.sum().to_frame()/df['TotalCharges'].sum()
```

Out[55]:

Churn	TotalCharges
No	0.821693
Yes	0.178307

2. Explain churn by the below categories. Are there any factors that combine to be especially impactful?

a. Customer demographics like age and gender

Sex is not a very important factor to talk about churn, because you can see that as many women as men have churn, even if the women are a little bit superior, which is not a problem. On the other hand, if we analyse it according to age, we will see that the older the clients are, the more likely they are to churn, moreover, we can see that the majority of the clients are minors.

And also, we can see Partners had much larger percent of Dependents than Non-Partner this tells us that most Partners might be married.

Most of people that were Partner will stay longer with the company. So being a Partner is a plus-point for the company as they will stay with them.

b. Services used

Most of the people that left had Fiber Optic Internet-Service.

They prefer to use DSL Internet Service on the other hand, customers who do not use the internet service are much less likely to leave. Is it because of the quality of service? Or is the price high? Or maybe both, say, a poor quality of service for an inflated price?

We can see that people that left the company didn't use Services like Online Security, Device Protection, Tech Support and Online Backup quite often.

c. Billing information

Most of the people that left are the ones who had Payment Method as Electronic check (57.30%).

3. What services are typically purchased by customers who churned? Are any services especially helpful in retaining customers?

R) Customers who churn bought the Fiber Optic Internet service, out of a total of all those who took the service, 69.40% of those who took the fiber optic service opted out.

Second, PaperlessBilling is not advantageous for the company, 74.90% of customers who use PaperlessBilling churn with a score of 42.91 for the Electronic check.

In spite of everything, the company's strength lies in DSL internet services.

4. Bonus! How long will it take for the company to lose all its customers? Which demographics will they lose first?

R) Mathematically speaking, the company will lose all its customers in about 29 months, that is, 2 years and 5 months.

The company will primarily lose its adult customers, especially those who use the fiber optic internet service, those who serve PaperlessBilling especially those who pay with Electronic check and those who also have a Month-to-Month contract.

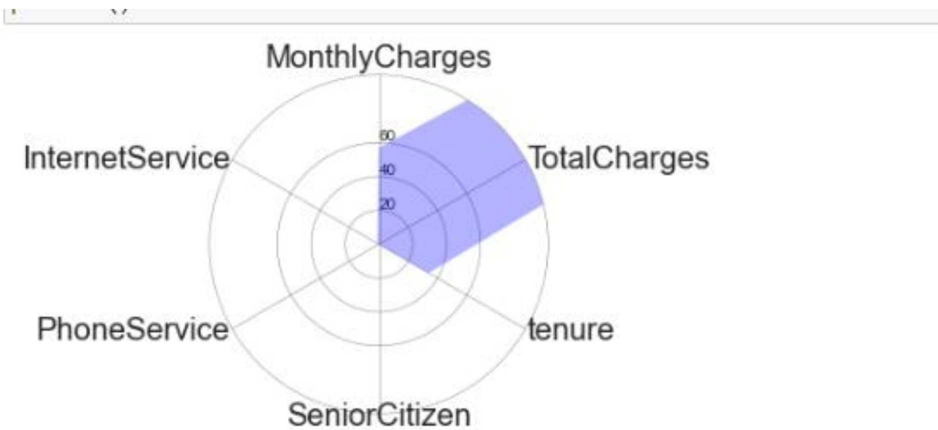
Part 2

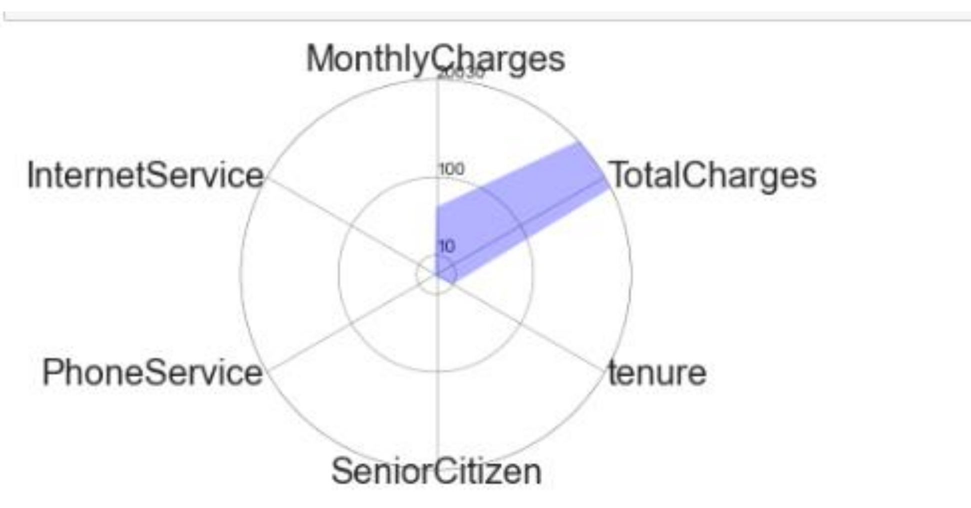
1. Are there types of people who churn at higher rates?

R) Those with a monthly contract are more likely to churn than those with an annual or bi-annual contract, as well as those who use online backup are more likely to stay than those who pay by e-check.

Come up with 2-3 profiles to give executives an idea of who churns often. Try to look several factors deep for example: people with no internet service and no phone service, or women who are senior citizens

R)





2. Create a case study for one of your customer profiles. Show how much additional revenue you could make by increasing sales by 10% in that profile.

Let's take, customer # with customerID '4881-JVQOD, if we increase his purchases by 10%, he will earn for the month 38,005 and for TotalCharges will become 398.86.

3. Do you have any recommendations on how to reach groups of people who churn at high rates?
R) First of all, if there is an internal problem, the company has to manage it, then if it is not the case, it has to give more or less attractive advantages to its customers, especially those who have a high probability of unemployment, such as improving the quality of service, reducing the price of some services, or maybe keeping the same price for the product but offering gifts to these customers to force them to stay, let's say, not to be attracted elsewhere to other competing companies.