Name- Prakhar Khanduri
Registration no. 19BCE0486

In [11]:
```
import pandas as pd
import numpy as np
import matplotlib

from matplotlib import pyplot as plt
%matplotlib inline
```

In [3]:
```
data = pd.read_csv('Churn.csv')
```

In [4]:
```
print('Data shape - ', data.shape)
data.head()
```

```
Data shape -  (50, 24)
```

Out[4]:

	customerID	gender	CreditScore	Geography	SeniorCitizen	NumOfProducts	Partner	Depe
0	7590-VHVEG	Female	619.0	France	0.0	1.0	Yes	
1	5575-GNVDE	Male	608.0	Spain	0.0	1.0	No	
2	3668-QPYBK	Male	502.0	France	0.0	3.0	No	
3	7795-CFOCW	Male	699.0	France	0.0	2.0	No	
4	9237-HQITU	Female	850.0	Spain	0.0	1.0	No	

5 rows × 24 columns

In [5]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50 entries, 0 to 49
Data columns (total 24 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   customerID        48 non-null     object 
 1   gender            48 non-null     object 
 2   CreditScore       48 non-null     float64
 3   Geography         48 non-null     object 
 4   SeniorCitizen     48 non-null     float64
 5   NumOfProducts     48 non-null     float64
 6   Partner           48 non-null     object 
 7   Dependents        48 non-null     object 
 8   Tenure            50 non-null     int64  
 9   Balance           50 non-null     float64
 10  NumOfProducts.1   50 non-null     int64  
 11  HasCrCard         50 non-null     int64  
 12  IsActiveMember    50 non-null     int64  
 13  EstimatedSalary   50 non-null     float64
 14  Exited            50 non-null     int64  
 15  PhoneService      48 non-null     object 
 16  InternetService   48 non-null     object 
 17  Contract          48 non-null     object 
 18  PaperlessBilling  48 non-null     object 
 19  PaymentMethod     48 non-null     object 
 20  MonthlyCharges    48 non-null     float64
 21  yearly Charges    48 non-null     float64
 22  TotalCharges      48 non-null     float64
 23  Churn             48 non-null     object 
dtypes: float64(8), int64(5), object(11)
memory usage: 9.5+ KB
```

Q1.1 - Using python script do density plot any three continuous variables with respect to categorical variables.

In [12]:
```
# Continuous vars.
data.select_dtypes('float').columns
```

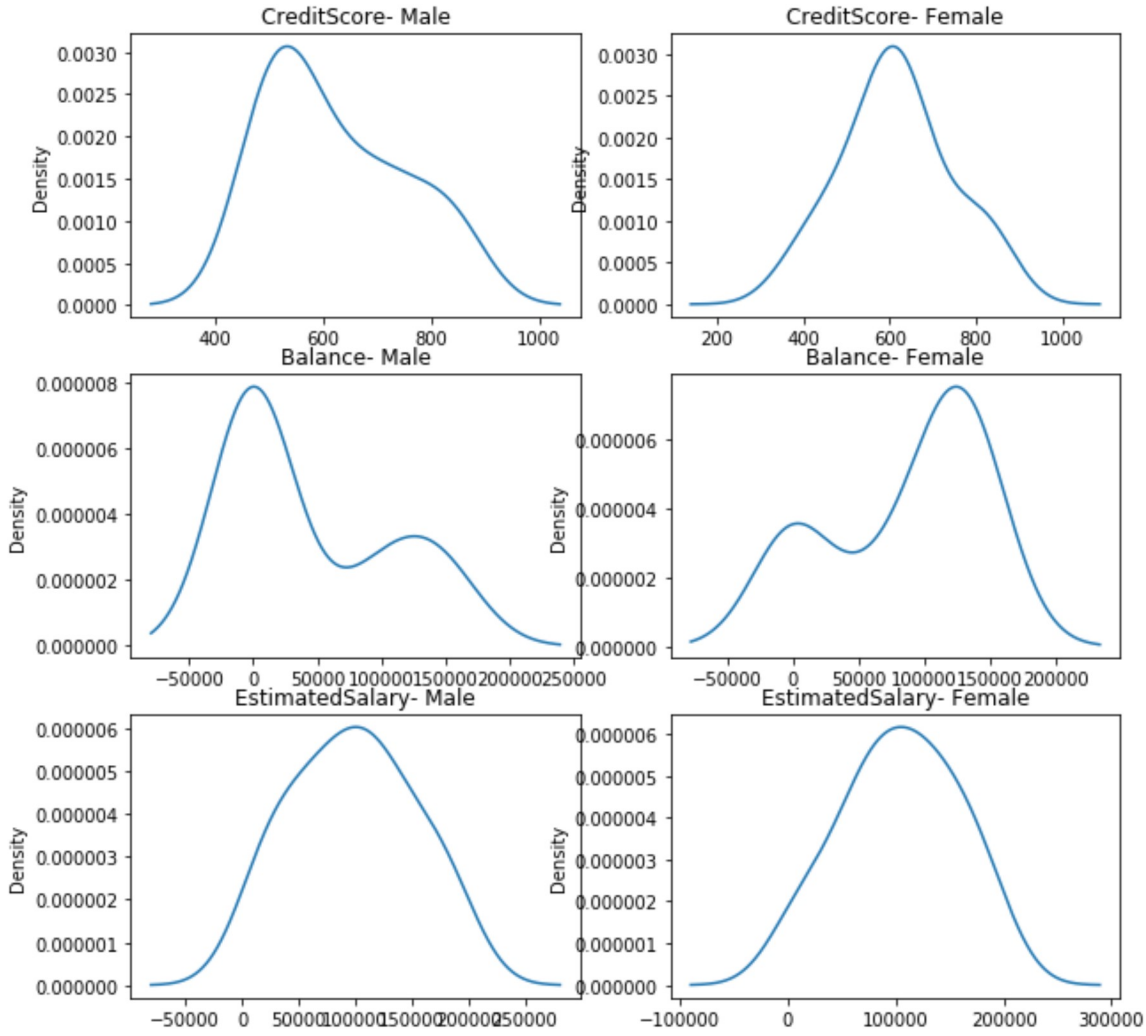
Out[12]:
```
Index(['CreditScore', 'SeniorCitizen', 'NumOfProducts', 'Balance',
       'EstimatedSalary', 'MonthlyCharges', 'yearly Charges', 'TotalC
harges'],
      dtype='object')
```

In [13]:
```
# Categorical vars.
data.select_dtypes('object').columns
```

Out[13]:
```
Index(['customerID', 'gender', 'Geography', 'Partner', 'Dependents',
       'PhoneService', 'InternetService', 'Contract', 'PaperlessBilli
ng',
       'PaymentMethod', 'Churn'],
      dtype='object')
```

In [ ]:
```
# Selecting the below continuous against categorical variable
Continuous - CreditScore, Balance, EstimatedSalary
Categorical - gender
```

In [29]:
```
fig, ((ax1,ax2),(ax3,ax4),(ax5,ax6)) = plt.subplots(3,2, figsize=(10,10))
i=1
for s in ['CreditScore', 'Balance', 'EstimatedSalary']:
    data.loc[data.gender=='Male',s].plot(kind='density', ax=eval('ax'+str(i)), title=s+'- Male')
    i=i+1
    data.loc[data.gender=='Female',s].plot(kind='density', ax=eval('ax'+str(i)), title=s+'- Female')
    i=i+1
```



Q1.2 - Find the IQR of any three continuous variables using python script. First mention which variables are you considering. Then find the IQR. Paste the code along with result.

In [33]:
```
cont_var = ['CreditScore', 'Balance', 'EstimatedSalary']

for i in cont_var:
    q75, q25 = np.nanpercentile(data[i], [75 ,25])
    iqr = q75 - q25
    print('The IQR of ', i, 'is - ', iqr)
```

```
The IQR of  CreditScore is -  178.75
The IQR of  Balance is -  124763.695
The IQR of  EstimatedSalary is -  74639.85500000003
```

Q1.3 - Using ggplot library, do box plot of the selected continuous variables with respect to each of categorical variables. Paste code and graphics.

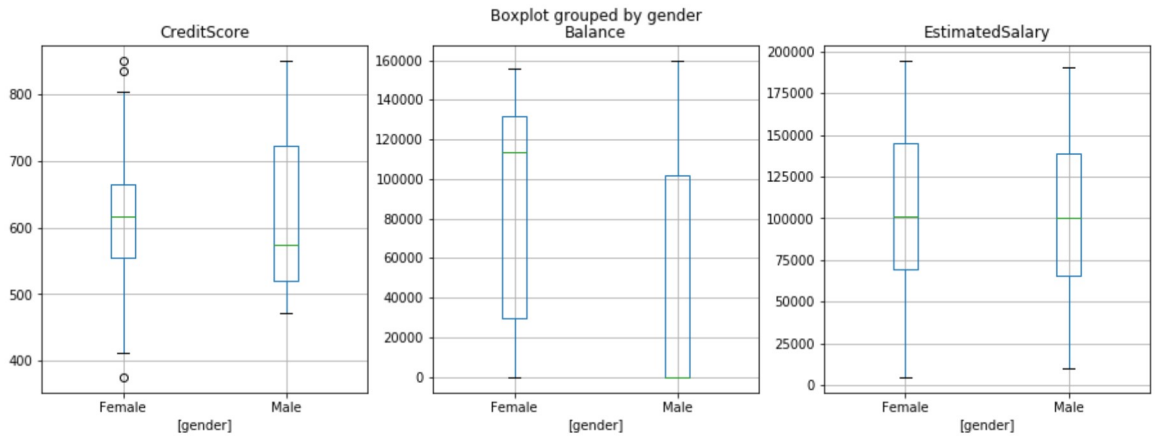
In [36]:
```
# Since having issue with ggplot package using pandas

cont_var = ['CreditScore', 'Balance', 'EstimatedSalary']
cat_vars = list(data.select_dtypes('object').columns)
```

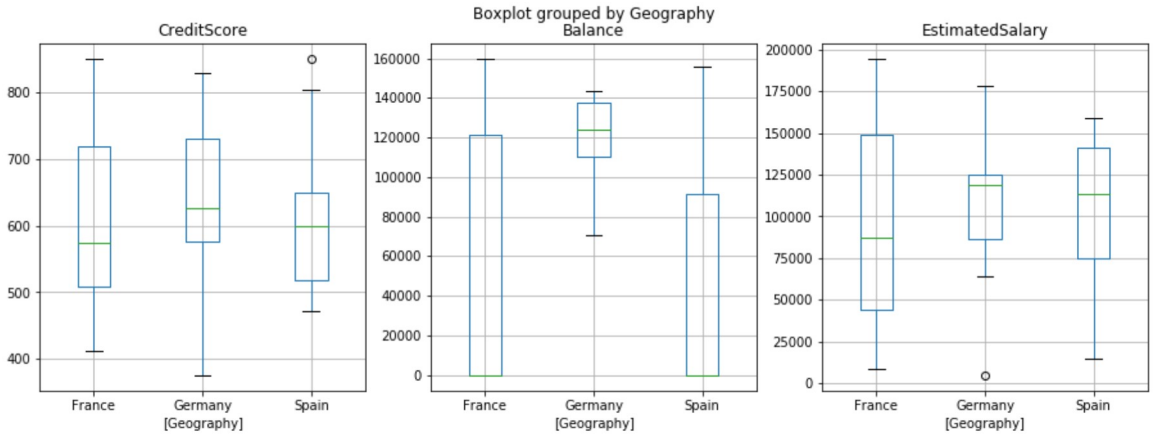
In [37]:
```
len(cat_vars)
```

Out[37]: 11

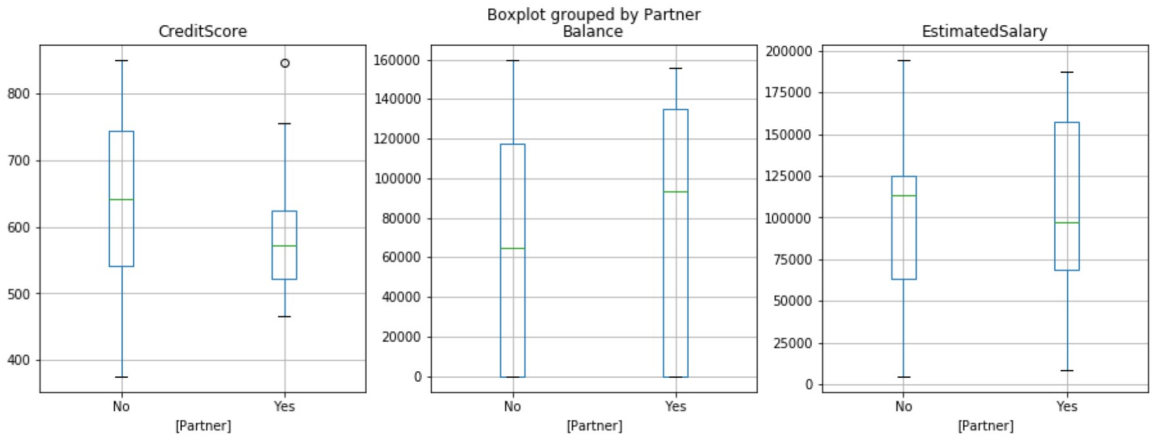
In [57]:
```
fig, ax = plt.subplots(1,3, figsize=(15,5))
j = cat_vars[1]
i=0
for i,s in enumerate(cont_var):
    data[[s,j]].boxplot(by=j, ax=ax[i]);
    i=i+1
```



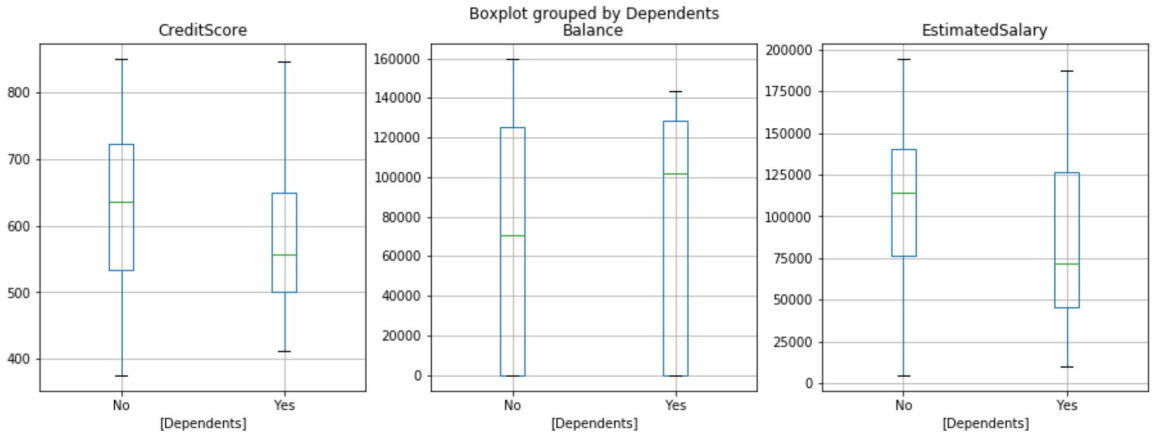
In [58]:
```
fig, ax = plt.subplots(1,3, figsize=(15,5))
j = cat_vars[2]
i=0
for i,s in enumerate(cont_var):
    data[[s,j]].boxplot(by=j, ax=ax[i]);
    i=i+1
```



In [59]:
```
fig, ax = plt.subplots(1,3, figsize=(15,5))
j = cat_vars[3]
i=0
for i,s in enumerate(cont_var):
    data[[s,j]].boxplot(by=j, ax=ax[i]);
    i=i+1
```



In [60]:
```
fig, ax = plt.subplots(1,3, figsize=(15,5))
j = cat_vars[4]
i=0
for i,s in enumerate(cont_var):
    data[[s,j]].boxplot(by=j, ax=ax[i]);
    i=i+1
```



In [ ]:

Q2.1 - Analysis the churn dataset and answer the following question using python /R

How different user behavior, subscription, and demographic features correlate with churn in Internet service

In [79]:
```
data.corr(method='pearson')
```

Out[79]:

	CreditScore	SeniorCitizen	NumOfProducts	Tenure	Balance	NumOfPro
CreditScore	1.000000	6.015097e-02	-0.326858	0.070195	-0.056935	-0
SeniorCitizen	0.060151	1.000000e+00	0.169932	0.188525	-0.208957	0
NumOfProducts	-0.326858	1.699324e-01	1.000000	0.201372	-0.211632	1
Tenure	0.070195	1.885248e-01	0.201372	1.000000	-0.243118	0
Balance	-0.056935	-2.089568e-01	-0.211632	-0.243118	1.000000	-0
NumOfProducts.1	-0.326858	1.699324e-01	1.000000	0.202267	-0.211939	1
HasCrCard	0.131001	6.567020e-02	0.040918	0.042193	-0.085158	0
IsActiveMember	0.399406	8.991371e-02	-0.264113	-0.049492	0.064372	-0
EstimatedSalary	0.131434	2.937220e-01	-0.098617	0.041568	0.164051	-0
Exited	-0.117712	4.831626e-18	0.036155	-0.009351	0.278141	0
MonthlyCharges	0.147293	2.243296e-02	-0.015230	0.115975	0.194043	-0
yearly Charges	0.147293	2.243296e-02	-0.015230	0.115975	0.194043	-0
TotalCharges	0.147293	2.243296e-02	-0.015230	0.115975	0.194043	-0

Q2.2 - The proportion of Churn to Non-Churn.

In [63]:
```
pd.crosstab(pd.data.Churn, columns='N', normalize=True)
```

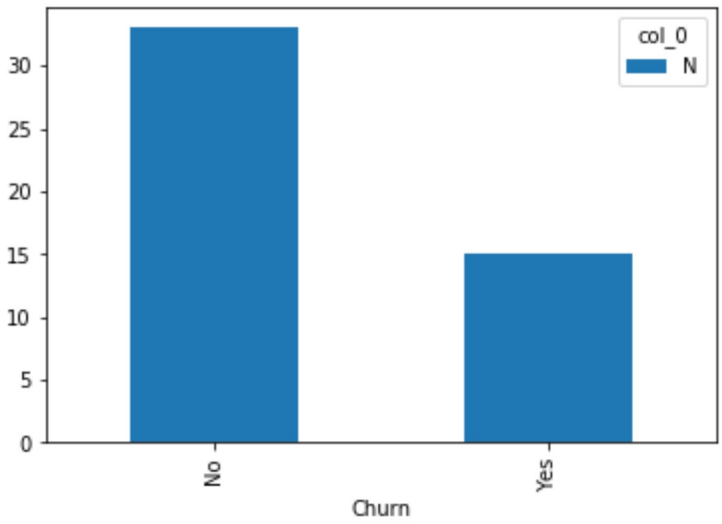
Out[63]:

col_0	N
Churn	
No	0.6875
Yes	0.3125

Q2.3 - Analysis the churn distribution.

In [66]:
```
pd.crosstab(data.Churn, columns='N').plot(kind='bar')
```

Out[66]: <matplotlib.axes._subplots.AxesSubplot at 0x1c0a42e2828>



In [ ]:

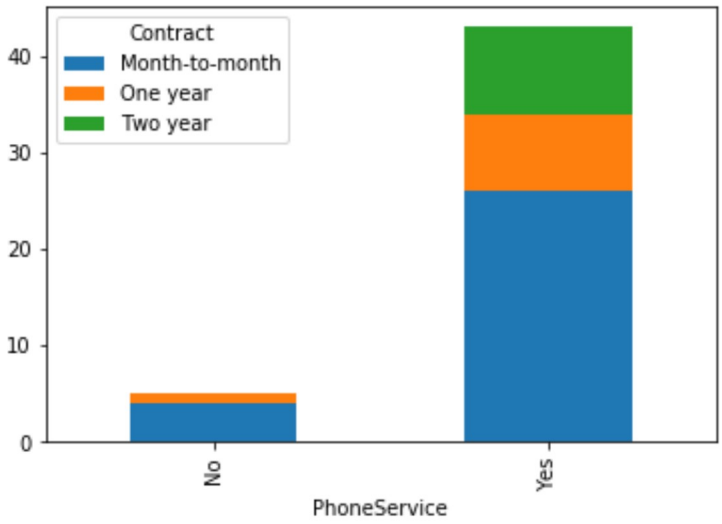
In [ ]:

Q2.4 - Analysis the service purchased by contract .

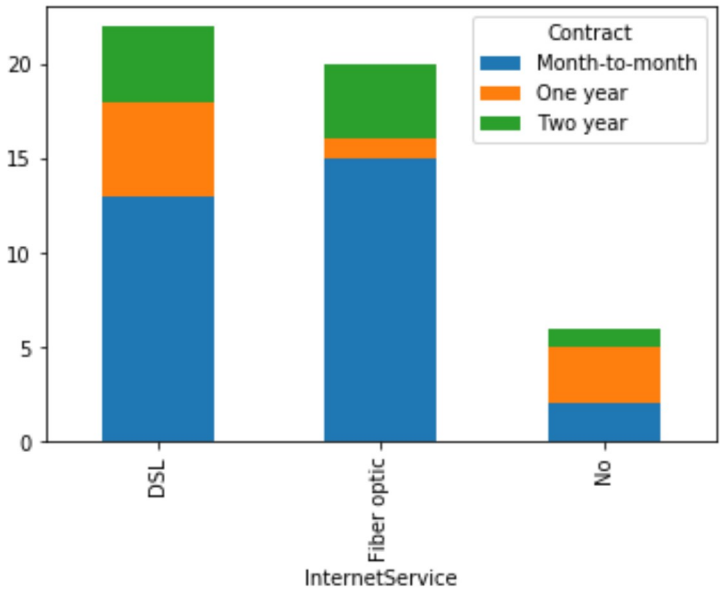
In [67]:
```
cat_vars
```

Out[67]:
```
['customerID',
 'gender',
 'Geography',
 'Partner',
 'Dependents',
 'PhoneService',
 'InternetService',
 'Contract',
 'PaperlessBilling',
 'PaymentMethod',
 'Churn']
```

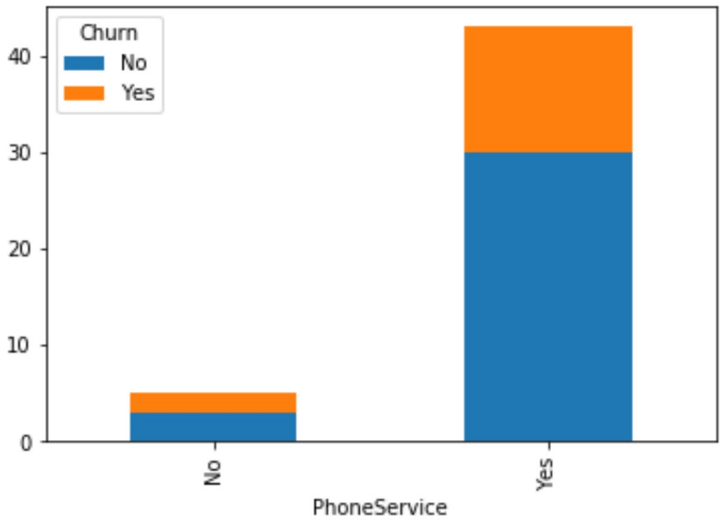
In [77]:
```
pd.crosstab(data['PhoneService'],data['Contract']).plot(kind='bar', stacked=True);
```



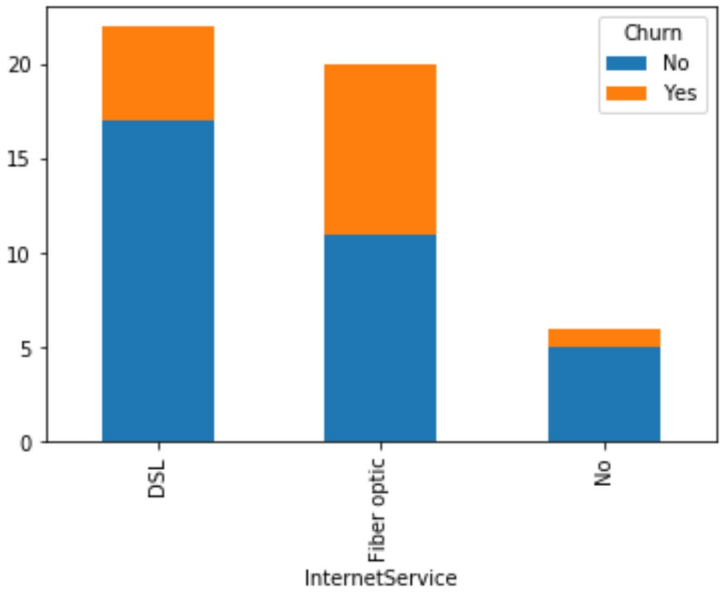
In [78]:
```
pd.crosstab(data['InternetService'],data['Contract']).plot(kind='bar', stacked=True);
```



In [70]:
```
pd.crosstab(data['PhoneService'],data['Churn']).plot(kind='bar', stacked=True);
```



In [74]:
```
pd.crosstab(data['InternetService'],data['Churn']).plot(kind='bar', stacked=True);
```



In [ ]:
```

```