```
from google.colab import drive
drive.mount('/content/gdrive')
path='gdrive/My Drive/Data Mining (Nguyen Duc Vu Duy)'
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive"
```

Bài thực hành môn Khai Thác Dữ Liệu - Tuần 1 Nguyễn Đức Vũ Duy - 18110004

```
import os
path='gdrive/MyDrive/Data Mining (Nguyen Duc Vu Duy)'
print(os.path.isdir(path))
print(os.path.isfile(path+'/Dataset/CustomerChurn.csv'))
```

```
True
True
```

```
import pandas as pd
ChurnData=pd.read_csv(path+'/Dataset/CustomerChurn.csv')
print(ChurnData.head(10).to_string())
```

```
  State  Account length  Area code International plan Voice mail plan  Number vmail messages  Total day minu
0    KS             128        415                 No             Yes                     25             26
1    OH             107        415                 No             Yes                     26             16
2    NJ             137        415                 No              No                      0             24
3    OH              84        408                Yes              No                      0             29
4    OK              75        415                Yes              No                      0             16
5    AL             118        510                Yes              No                      0             22
6    MA             121        510                 No             Yes                     24             21
7    MO             147        415                Yes              No                      0             15
8    LA             117        408                 No              No                      0             18
9    WV             141        415                Yes             Yes                     37             25
```

```
from IPython.display import Image
Image(filename=path+'/Dataset/CustomerChurn.png')
```

Variable Description

Name	Description	Value Type	Statistical Type
State	State Abbreviation (like KS – Kansas)	String	Categorical
Account length	How long the client has been with the company	Numerical	Quantitative
Area code	Phone number prefix	Numerical	Categorical
International plan	International Plan (on/off)	String. "Yes"/"No"	Categorical/Binary
Voice mail plan	Voice mail (on/off)	String. "Yes"/"No"	Categorical/Binary
Number vmail messages	Number of Voice mail messages	Numerical	Quantitative
Total day minutes	Total duration of daytime calls	Numerical	Quantitative
Total day calls	Total number of daytime calls	Numerical	Quantitative
Total day charge	Total charges for daytime services	Numerical	Quantitative
Total eve minutes	Total duration of evening calls	Numerical	Quantitative
Total eve calls	Total number of evening calls	Numerical	Quantitative
Total eve charge	Total charges for evening services	Numerical	Quantitative
Total night minutes	Total duration of nighttime calls	Numerical	Quantitative
Total night calls	Total number of nighttime calls	Numerical	Quantitative
Total night charge	Total charges for nighttime services	Numerical	Quantitative
Total intl minutes	Total duration of international calls	Numerical	Quantitative
Total intl calls	Total number of international calls	Numerical	Quantitative
Total intl charge	Total charges for international services	Numerical	Quantitative
Customer service calls	Number of calls to customer service	Numerical	Categorical/Ordinal

```
1 import warnings
2 warnings.filterwarnings('ignore')
3
4 print('Churn Data Dimensionality: '+str(ChurnData.shape))
```

```
Churn Data Dimensionality: (3333, 20)
```

```
1 FeatureList=ChurnData.columns
2 print("Number of features: "+str(len(FeatureList)))
3 print("Feature List: "+str(FeatureList))
```

```
Number of features: 20
Feature List: Index(['State', 'Account length', 'Area code', 'International plan',
       'Voice mail plan', 'Number vmail messages', 'Total day minutes',
       'Total day calls', 'Total day charge', 'Total eve minutes',
       'Total eve calls', 'Total eve charge', 'Total night minutes',
       'Total night calls', 'Total night charge', 'Total intl minutes',
       'Total intl calls', 'Total intl charge', 'Customer service calls',
       'Churn'],
      dtype='object')
```

```
1 ChurnData.dtypes
```

```
State                    object
Account length            int64
Area code                 int64
International plan       object
Voice mail plan          object
Number vmail messages     int64
Total day minutes       float64
Total day calls           int64
Total day charge        float64
Total eve minutes       float64
Total eve calls           int64
Total eve charge        float64
Total night minutes     float64
Total night calls         int64
Total night charge      float64
Total intl minutes      float64
Total intl calls          int64
Total intl charge       float64
Customer service calls    int64
Churn                      bool
dtype: object
```

Xuất danh sách các tiêu chí định tính và kèm theo số lượng có bao nhiêu định tính theo mỗi tiêu chí

```
dt=ChurnData.columns[ChurnData.dtypes=='object']
dt
```

```
Index(['State', 'International plan', 'Voice mail plan'], dtype='object')
```

Có 3 cột có dữ liệu định tính là State, International plan và Voice mail plan

Số lượng có bao nhiêu định tính theo mỗi tiêu chí

```
print(' Number of unique values in State: '+str(len(ChurnData['State'].unique())))
print(' List of unique values in State: ')
print(ChurnData['State'].unique())
print(' The number of value in each State: ')
print(ChurnData['State'].value_counts())
```

```
 Number of unique values in State: 51
 List of unique values in State:
['KS' 'OH' 'NJ' 'OK' 'AL' 'MA' 'MO' 'LA' 'WV' 'IN' 'RI' 'IA' 'MT' 'NY'
 'ID' 'VT' 'VA' 'TX' 'FL' 'CO' 'AZ' 'SC' 'NE' 'WY' 'HI' 'IL' 'NH' 'GA'
 'AK' 'MD' 'AR' 'WI' 'OR' 'MI' 'DE' 'UT' 'CA' 'MN' 'SD' 'NC' 'WA' 'NM'
 'NV' 'DC' 'KY' 'ME' 'MS' 'TN' 'PA' 'CT' 'ND']
 The number of value in each State:
WV    106
MN     84
NY     83
AL     80
WI     78
OH     78
OR     78
WY     77
VA     77
CT     74
VT     73
MI     73
ID     73
```

```
UT     72
TX     72
IN     71
MD     70
KS     70
MT     68
NJ     68
NC     68
CO     66
NV     66
WA     66
RI     65
MS     65
MA     65
AZ     64
MO     63
FL     63
NM     62
ME     62
ND     62
OK     61
DE     61
NE     61
SD     60
SC     60
KY     59
IL     58
NH     56
AR     55
GA     54
DC     54
HI     53
TN     53
AK     52
LA     51
PA     45
IA     44
CA     34
Name: State, dtype: int64
```

```
1 print(' Number of unique values in International plan: '+str(len(ChurnData['International plan'].unique())))
2 print(' List of unique values in International plan: ')
3 print(ChurnData['International plan'].unique())
```

```
print(' The number of value in each International plan: ')
print(ChchurnData['International plan'].value_counts())
```

```
 Number of unique values in International plan: 2
 List of unique values in International plan:
['No' 'Yes']
 The number of value in each International plan:
No     3010
Yes     323
Name: International plan, dtype: int64
```

```
print(' Number of unique values in Voice mail plan: '+str(len(ChurnData['Voice mail plan'].unique())))
print(' List of unique values in Voice mail plan: ')
print(ChurnData['Voice mail plan'].unique())
print(' The number of value in each Voice mail plan: ')
print(ChurnData['Voice mail plan'].value_counts())
```

```
 Number of unique values in Voice mail plan: 2
 List of unique values in Voice mail plan:
['Yes' 'No']
 The number of value in each Voice mail plan:
No     2411
Yes     922
Name: Voice mail plan, dtype: int64
```

Xuất các giá trị định tính có tần xuất xuất hiện cao nhất trong mỗi đặc trưng định tính

```
print(' The top 5 value in State: ')
print(ChurnData['State'].value_counts()[:5])
```

```
 The top 5 value in State:
WV    106
MN     84
NY     83
AL     80
WI     78
Name: State, dtype: int64
```

5 State WV, MN, NY, AL và OH có tần xuất xuất hiện cao nhất

```
print(' The number of value in each International plan: ')
print(ChurnData['International plan'].value_counts())
```

```
 The number of value in each International plan:
No     3010
Yes     323
Name: International plan, dtype: int64
```

Yes có tần xuất cao nhất ở International plan

```
print(' The number of value in each Voice mail plan: ')
print(ChurnData['Voice mail plan'].value_counts())
```

```
 The number of value in each Voice mail plan:
No     2411
Yes     922
Name: Voice mail plan, dtype: int64
```

No có tần xuất xuất hiện cao nhất ở Voice mail plan

```
ChurnData.describe(include=['object'])[2:4]
```

	State	International plan	Voice mail plan
top	WV	No	No
freq	106	3010	2411

✓ 0s completed at 11:43 AM