Principal Component Analysis EDA on Telecom Churn Data

By *Alex Vaillant*

Research Question:

"As an exploratory data analysis step in our journey to reducing customer churn rates, to what extent can we explain the most variance of our customer data based on PCA reductions of customer age, income, monthly charges, tenure, usage, and survey results? These results will be used in the logistic regression analysis in the future to focus on a smaller set of 13 continuous variables plus various other categorical variables."

Import Libraries/Packages

In [1]:

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from numpy import random
from matplotlib import pyplot as plt
```

Set Random Seed to obtain the same results every run

In [2]:

```
random.seed(1234)
```

Import my dataset by using pandas::read_csv

In [3]:

```
url = "C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Unsupervised Lear
churn_data = pd.read_csv(url)
churn_data.head()
```

Out[3]:

	CaseOrder	Customer_id	Interaction	UID	City	State	County
0	1	K409198	aa90260b-4141-4a24-8e36-b04ce1f4f77b	e885b299883d4f9fb18e39c75155d990	Point Baker	AK	Prince o Wales Hyde
1	2	S120509	fb76459f-c047-4a9d-8af9-e0f7d4ac2524	f2de8bef964785f41a2959829830fb8a	West Branch	MI	Ogemav
2	3	K191035	344d114c-3736-4be5-98f7-c72c281e2d35	f1784cfa9f6d92ae816197eb175d3c71	Yamhill	OR	Yamhi

	CaseOrder	Customer_id	Interaction	UID	City	State	Count
3	4	D90850	abfa2b40-2d43-4994-b15a-989b8c79e311	dc8a365077241bb5cd5ccd305136b05e	Del Mar	CA	Sa Dieg
4	5	K662701	68a861fd-0d20-4e51-a587-8a90407ee574	aabb64a116e83fdc4befc1fbab1663f9	Needville	TX	For Ben

5 rows × 50 columns

Index the data to remove categorical variables and unnecessary columns from our analysis

In [4]:

```
churn_indexed = churn_data[['Age','Income','Tenure','MonthlyCharge','Bandwidth_GB_Year'
                          ,'Item5', 'Item6','Item7','Item8']]
```

Standardize the dataset to use with PCA

In [5]:

```
cols = set(churn_indexed.columns)
scaler = StandardScaler()
churn_std = scaler.fit_transform(churn_indexed)
churn_std = pd.DataFrame(churn_std, columns = cols)
```

In [6]:

```
churn_std.to_csv("C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Unsupe
```

Fit the PCA model to the standardized data & print the explained variance

In [7]:

```
pca = PCA()
pca.fit_transform(churn_std)
print('Each Principal Components Explained Variance Matrix:\n', pca.explained_variance_
print('Cumulative Explained Variance Matrix:\n', pca.explained_variance_ratio_.cumsum()
```

```
Each Principal Components Explained Variance Matrix:
 [0.22673235 0.15352983 0.12581859 0.07789078 0.07689261 0.0762398
 0.06018131 0.05314031 0.04559372 0.04140589 0.03711805 0.02500108
 0.00045566]
Cumulative Explained Variance Matrix:
 [0.22673235 0.38026218 0.50608077 0.58397155 0.66086416 0.73710396
 0.79728528 0.85042559 0.89601931 0.9374252  0.97454325 0.99954434
 1.        ]
```

Create a variance variable with the explained variance data from the PCA model

In [8]:

```
variance = pca.explained_variance_ratio_
```

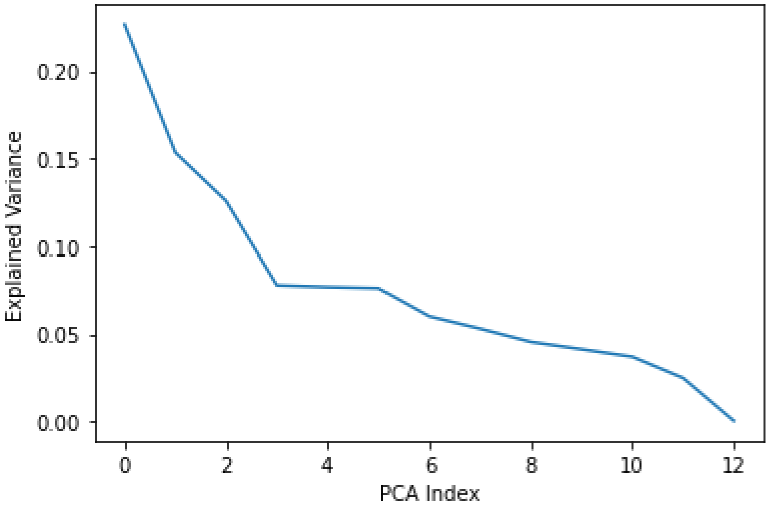
Create an elbow plot by plotting the explained variance ratios above

In [9]:

```
plt.xlabel('PCA Index')
plt.ylabel('Explained Variance')
plt.plot(variance)
```

Out[9]:

```
[<matplotlib.lines.Line2D at 0x1cc722e1610>]
```



Identify the Explained Variance of Each Principal Component (1-4)

In [10]:

```
print('Each Principal Components 1-4s Explained Variances:\n',pca.explained_variance_ra
```

```
Each Principal Components 1-4s Explained Variances:
 [0.22673235 0.15352983 0.12581859 0.07789078]
```

Identify the Cumulative Explained Variance of the 4 Principal Components combined

In [11]:

```
print('Cumulative Explained Variance by Principal Components 1-4:\n',pca.explained_vari
```

```
Cumulative Explained Variance by Principal Components 1-4:
 [0.22673235 0.38026218 0.50608077 0.58397155]
```