Guided Assignment On Feature Engineering

Name: **Shreya Shrivastava**

Course: **Artificial Intelligence And Machine Learning**

Batch Four

Duration: **12 Months**

Problem statement: Factor analysis is a useful technique to find latent factors that can potentially describe multiple attributes, which is sometimes very useful for dimensionality reduction. Use the **Airline Passenger Satisfaction dataset** to perform factor analysis. (Use only the columns that represent the ratings given by the passengers, only 14 columns). Choose the best features possible that helps in dimensionality reduction, without much loss in information.

Prerequisites:

The libraries as well as things required in order for the program to work:

I. **Python 3.6** : The following url https://www.python.org/downloads/ can be referred to download python. Once you have python downloaded and installed, you will need to setup PATH variables (if you want to run python program directly, detail instructions are below in how to run software section). To do that check this: https://www.pythoncentral.io/add-python-to-path-python-is-not-recognized-as-an-internal-or-external-command/. Setting up PATH variable is optional as you can also run program without it and more instruction are given below on this topic. Second option is to download anaconda and use its anaconda prompt to run the commands. To install anaconda check this url : https://www.anaconda.com/download/

II. **ADDITIONAL PACKAGES :** You will also need to download and install below 3 packages- numpy,pandas and seaborn after you install either python or anaconda from the steps above. If you have chosen to install python 3.6,then run the following commands in command prompt/terminal to install these packages :

NUMPY: pip install -U numpy

SEABORN: pip install -U seaborn

PANDAS: pip install -U pandas

If using Anaconda then run the following commands in anaconda prompt to install these packages:

NUMPY: conda install -c anaconda numpy

SEABORN: conda install -c anaconda seaborn

PANDAS: conda install -c anaconda pandas

III. The dataset used can be accessed from here: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

IV. **METHOD USED:**

A. EXPLORATORY FACTOR ANALYSIS

THE PROJECT :

1. Importing the libraries and loading the **Airline Passenger Satisfaction dataset**. We then proceed to load the training as well as the test dataset and the target variable.

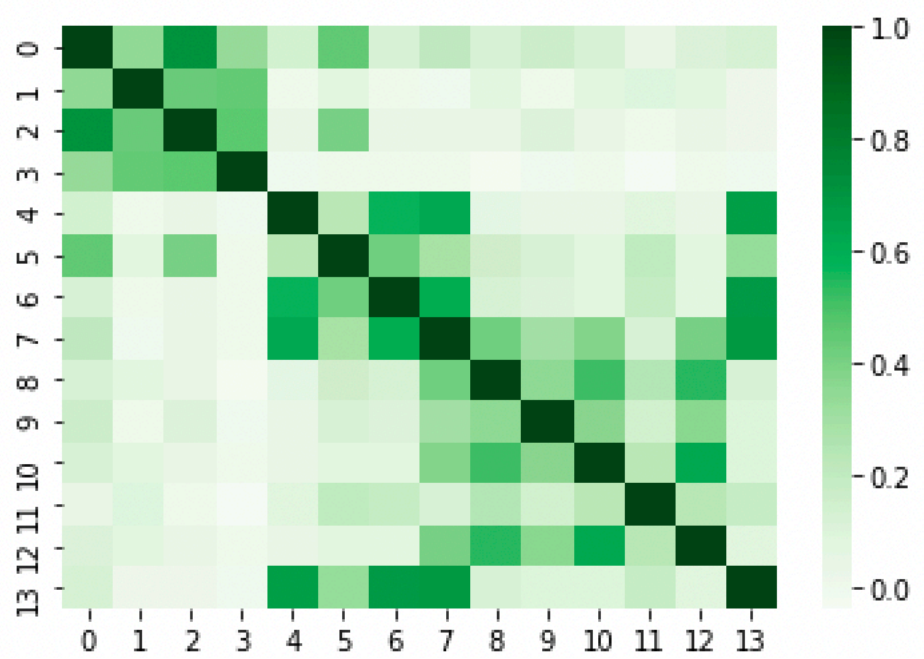
```
import pandas as pd
import numpy as np
import seaborn as sns

train=pd.read_csv("train.csv")
test=pd.read_csv("test.csv")
X_train = train.iloc[:,8:-3]
X_test=test.iloc[:8:-3]
target_tr= train.iloc[:,-1:]
```

2.Normalising the Data.

```
x=X_train.values
x_mean=np.mean(x,axis=0)
x_norm=x-np.matrix(x_mean)
x_norm=x_norm.T #important otherwise we'll end up having a large dimension of the covariance matrix
```

3. Calculating the covariance and correlation and plotting a heat map to speculate the number of latent factors

```
c=np.cov(x_norm)
co=np.corrcoef(x_norm)
ax=sns.heatmap(co,cmap='Greens')
```



4. Calculating the eigenvectors and eigenvalues of the covariance matrix and appending the favourable no of them to their respective lists.

```
eig_val,eig_vec=np.linalg.eig(c)
eig_sort= np.sort(eig_val)[::-1]
arg_sort=np.argsort(eig_val)[::-1]
eig_vec_ls=[]
eig_val_ls=[]
i_vec=arg_sort[:5]
for i in i_vec:
    eig_vec_ls.append(eig_vec[:,i])
    eig_val_ls.append(eig_val[i])
```

5.Estimation of V,S and W.

```
#estimation of parameter V
eig_val_arr= np.array(eig_val_ls)
lm=np.diag(eig_val_arr)
eig_vec_mat= np.matrix(eig_vec_ls).T
V=eig_vec_mat@np.sqrt(lm)

#Estimation of S
var_ls=[]
var_x=np.var(x_norm,axis=1)
var_x=np.ravel(var_x)
for i in range(V.shape[0]):
    s=np.sum(np.square(np.ravel(V[i,:])))
    sig_sq=var_x[i]-s
    var_ls.append(sig_sq)
var_ls=np.array(var_ls)
s=np.diag(var_ls)
c_inv=np.linalg.inv(c)
w=V.T@c_inv
```

6.Estimation of Z.

```
z=w@x_norm
z_=z.T

print(z_.shape)
```