Ex.no:12 **DHAMODARAN.B**

Date:10.10.2020 **1832016**

FACTOR ANALYSIS

Problem Statement:

The inner beauty is always cherished rather than the outer appearance. These lines show that the character of a person is more important rather than his appearance. Business firms these days take these into considerations and aim at selecting a person rather than a talent employee. Thus, deciding the personality of a person becomes necessary for firms to increase their productivity. Here, we are given the scores for various personalities of a person we try to reduce them and bring the unobserved feature or behavior into consideration. This can be done with the help of dimension reduction techniques such as the Factor Analysis.

Problem Description:

. The dataset had scores for various personalities for a person ranging from 1 t0 10. The various personalities given are "distant", "talkatv", "carelss", "hardwrk", "anxious","agreebl", "te nse", "kind", "opposng", "relaxed”,"disorgn", "outgoin", "approvn", "shy", "discipl","harsh", "per sevr", "friendl", "worryin", "respnsi","contrar", "sociabl", "lazy", "coopera", "quiet","organiz", "c riticl", "lax", "laidbck", "withdrw","givinup", "easygon”.The dataset had about 292 instances.

Initially, the data is checked for any null values. Later, the data are scaled using the standard scaling technique. Then, the scaled data are passed through various tests such as the Bartlett’s test of sphericity and the KMO test to determine whether the dimensionality reduction techniques such as the Factor Analysis can be applied on this dataset. With the help of Scree plot, the optimal number of factors are determined. Then the Factor Analysis is implemented using the Factor Analysis Module.

Code:

Data Preprocessing:

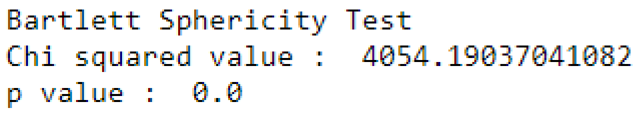
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
with open('Standford.txt','r') as file:
   header = file.readline()
   data=[]
   for row in file.readlines()[1:]:
      row = row.split()[1:]
```

```
        data.append(row)
    data = np.array(data,dtype='int')
dataset = pd.DataFrame(data,columns=np.array(header.split(),dtype=object))
dataset.head()
np.array(list(dataset.columns),dtype=object)
dataset.isnull().sum()
from sklearn.preprocessing import StandardScaler
scaler =  StandardScaler()
dataframe = scaler.fit_transform(dataset)
dataframe = pd.DataFrame(data=dataframe,columns=dataset.columns)
dataframe.head(10)
```

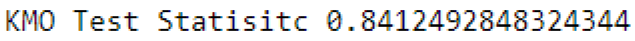
Bartlett and KMO Test:

```
from factor_analyzer.factor_analyzer import calculate_bartlett_sphericity,calculate_kmo
chi2,p = calculate_bartlett_sphericity(dataframe)
print("Chi squared value : ",chi2)
print("p value : ",p)
kmo_all,kmo_model = calculate_kmo(dataset)
print(kmo_model)
```

OUTPUT:



```
Bartlett Sphericity Test
Chi squared value :  4054.19037041082
p value :  0.0
```



```
KMO Test Statisitc 0.8412492848324344
```

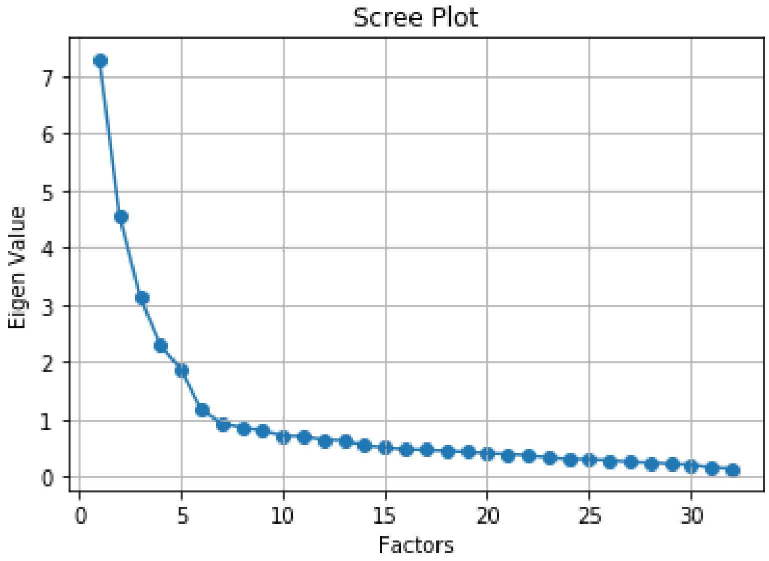
INFERENCE:

The Bartlett Test of Sphericity tests the hypothesis that the correlation is present among the features. The p-statistic for the test is 0.0 with 95% confidence. This implies that the correlation matrix is not an identity matrix i.e. correlation is present among the features with 95% confidence. The KMO test can be used to determine the sampling adequacy. The KMO score of 0.841 indicates that the dataset satisfies the sampling adequacy. Since both the tests showed green signals, factor analysis can be done on the dataset.

DETERMIANTINING NUMBER OF FACTORS AND SCREE PLOT:

```
fa = FactorAnalyzer(rotation = None,impute = "drop",n_factors=dataframe.shape[1])

fa.fit(dataframe)

ev,_ = fa.get_eigenvalues()


plt.scatter(range(1,dataframe.shape[1]+1),ev)

plt.plot(range(1,dataframe.shape[1]+1),ev)

plt.title('Scree Plot')

plt.xlabel('Factors')

plt.ylabel('Eigen Value')

plt.grid()
```

OUTPUT:



INFERENCE:

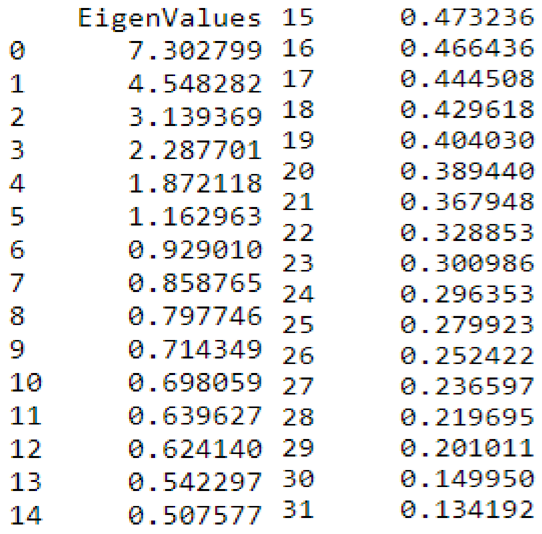
The scree plot shows the effect on eigen values with increase in number of factors. It can be inferred from the graph that the eigen values is above 1 until the 6 factors. So, the optimal number of Factors is 6 as the eigen value drops below 1 after 6 factors.

Computing Eigen Values, Factor Loadings, Uniqueness through Factor analysis:

```
fa = FactorAnalyzer(n_factors=6,rotation='varimax')

fa.fit(dataset)

with np.printoptions(suppress=True,precision=6):
    print(pd.DataFrame(fa.loadings_,index=dataframe.columns))
with np.printoptions(suppress=True,precision=6):
    print(pd.DataFrame(fa.get_factor_variance(),index=['Eigen Values','Proportional Var','Cumulative Var']))
with np.printoptions(suppress=True,precision=6):
    print(pd.DataFrame(fa.get_uniquenesses(),index=dataframe.columns,columns=['Uniqueness']))
with np.printoptions(precision=4,suppress=True):
    print(pd.DataFrame(fa.get_communalities(),index=dataframe.columns,columns=['Communalities']))
```

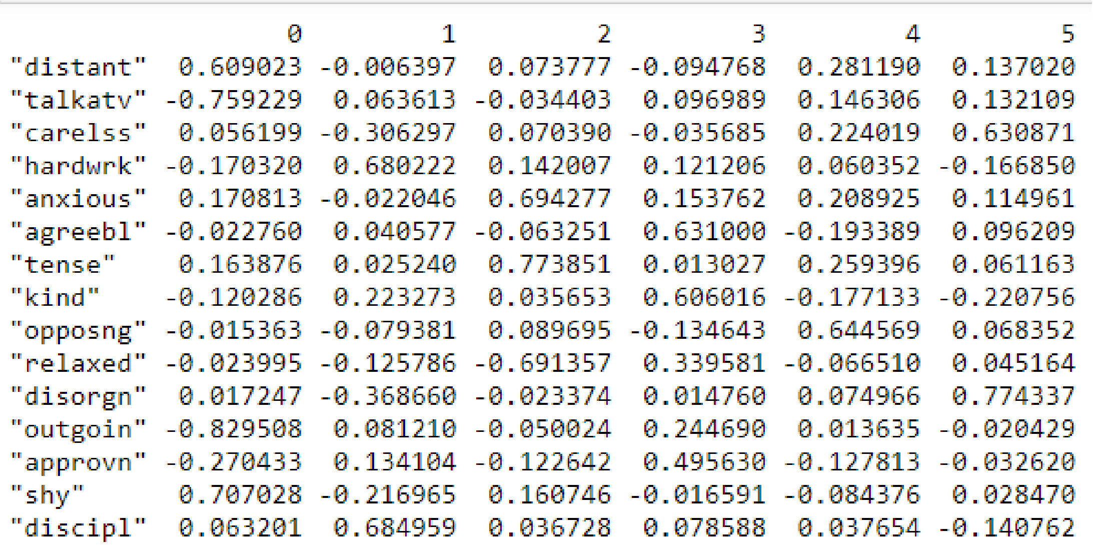
Output:

EIGEN VALUES:

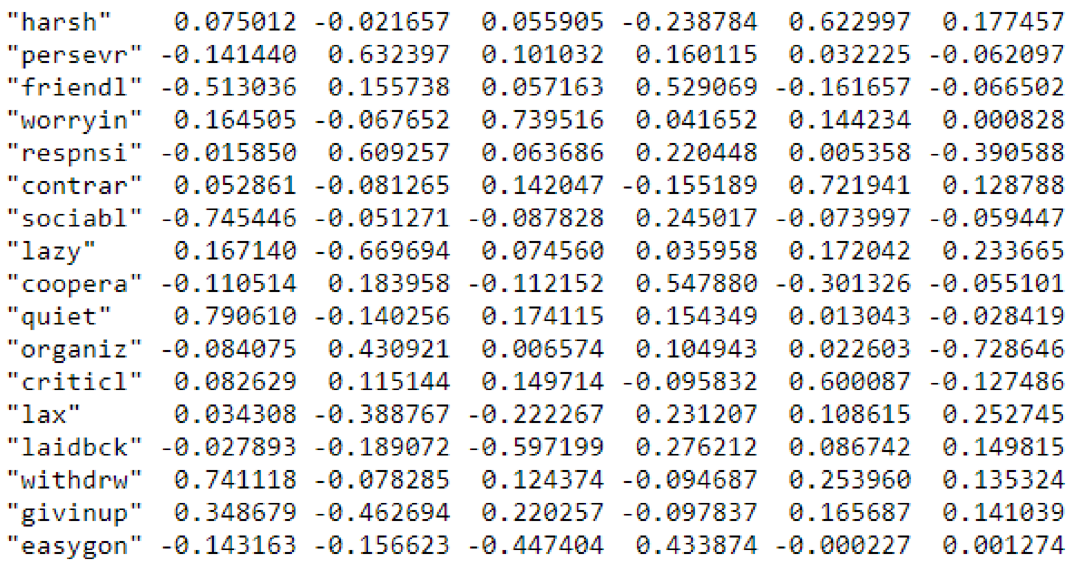


```
        EigenValues 15      0.473236
0          7.302799 16      0.466436
1          4.548282 17      0.444508
2          3.139369 18      0.429618
3          2.287701 19      0.404030
4          1.872118 20      0.389440
5          1.162963 21      0.367948
6          0.929010 22      0.328853
7          0.858765 23      0.300986
8          0.797746 24      0.296353
9          0.714349 25      0.279923
10         0.698059 26      0.252422
11         0.639627 27      0.236597
12         0.624140 28      0.219695
13         0.542297 29      0.201011
14         0.507577 30      0.149950
                    31      0.134192
```

FACTOR LOADINGS:

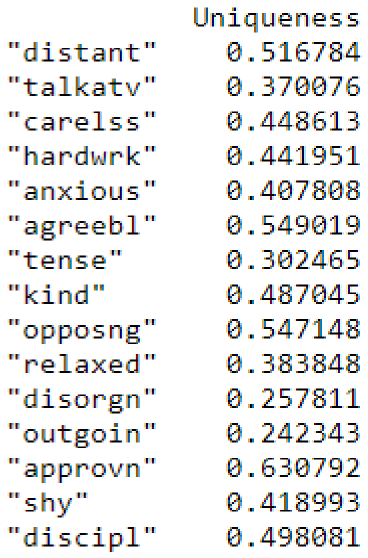


```
                  0         1         2         3         4         5
"distant"  0.609023 -0.006397  0.073777 -0.094768  0.281190  0.137020
"talkatv" -0.759229  0.063613 -0.034403  0.096989  0.146306  0.132109
"carelss"  0.056199 -0.306297  0.070390 -0.035685  0.224019  0.630871
"hardwrk" -0.170320  0.680222  0.142007  0.121206  0.060352 -0.166850
"anxious"  0.170813 -0.022046  0.694277  0.153762  0.208925  0.114961
"agreebl" -0.022760  0.040577 -0.063251  0.631000 -0.193389  0.096209
"tense"    0.163876  0.025240  0.773851  0.013027  0.259396  0.061163
"kind"    -0.120286  0.223273  0.035653  0.606016 -0.177133 -0.220756
"opposng" -0.015363 -0.079381  0.089695 -0.134643  0.644569  0.068352
"relaxed" -0.023995 -0.125786 -0.691357  0.339581 -0.066510  0.045164
"disorgn"  0.017247 -0.368660 -0.023374  0.014760  0.074966  0.774337
"outgoin" -0.829508  0.081210 -0.050024  0.244690  0.013635 -0.020429
"approvn" -0.270433  0.134104 -0.122642  0.495630 -0.127813 -0.032620
"shy"      0.707028 -0.216965  0.160746 -0.016591 -0.084376  0.028470
"discipl"  0.063201  0.684959  0.036728  0.078588  0.037654 -0.140762
```

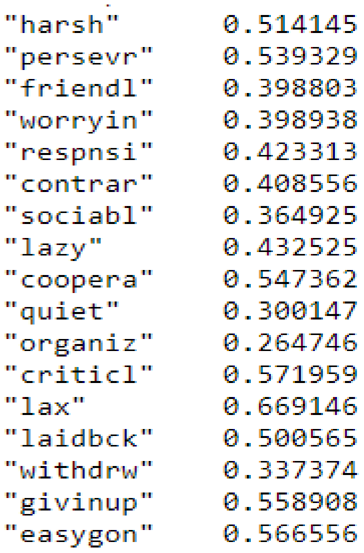


```
"harsh"    0.075012 -0.021657  0.055905 -0.238784  0.622997  0.177457
"persevr" -0.141440  0.632397  0.101032  0.160115  0.032225 -0.062097
"friendl" -0.513036  0.155738  0.057163  0.529069 -0.161657 -0.066502
"worryin"  0.164505 -0.067652  0.739516  0.041652  0.144234  0.000828
"respnsi" -0.015850  0.609257  0.063686  0.220448  0.005358 -0.390588
"contrar"  0.052861 -0.081265  0.142047 -0.155189  0.721941  0.128788
"sociabl" -0.745446 -0.051271 -0.087828  0.245017 -0.073997 -0.059447
"lazy"     0.167140 -0.669694  0.074560  0.035958  0.172042  0.233665
"coopera" -0.110514  0.183958 -0.112152  0.547880 -0.301326 -0.055101
"quiet"    0.790610 -0.140256  0.174115  0.154349  0.013043 -0.028419
"organiz" -0.084075  0.430921  0.006574  0.104943  0.022603 -0.728646
"criticl"  0.082629  0.115144  0.149714 -0.095832  0.600087 -0.127486
"lax"      0.034308 -0.388767 -0.222267  0.231207  0.108615  0.252745
"laidbck" -0.027893 -0.189072 -0.597199  0.276212  0.086742  0.149815
"withdrw"  0.741118 -0.078285  0.124374 -0.094687  0.253960  0.135324
"givinup"  0.348679 -0.462694  0.220257 -0.097837  0.165687  0.141039
"easygon" -0.143163 -0.156623 -0.447404  0.433874 -0.000227  0.001274
```

UNIQUENESS:

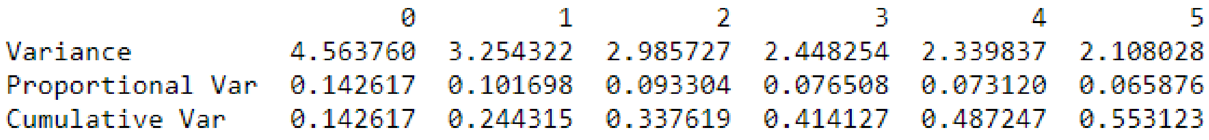


```
           Uniqueness
"distant"    0.516784
"talkatv"    0.370076
"carelss"    0.448613
"hardwrk"    0.441951
"anxious"    0.407808
"agreebl"    0.549019
"tense"      0.302465
"kind"       0.487045
"opposng"    0.547148
"relaxed"    0.383848
"disorgn"    0.257811
"outgoin"    0.242343
"approvn"    0.630792
"shy"        0.418993
"discipl"    0.498081
```



```
"harsh"      0.514145
"persevr"    0.539329
"friendl"    0.398803
"worryin"    0.398938
"respnsi"    0.423313
"contrar"    0.408556
"sociabl"    0.364925
"lazy"       0.432525
"coopera"    0.547362
"quiet"      0.300147
"organiz"    0.264746
"criticl"    0.571959
"lax"        0.669146
"laidbck"    0.500565
"withdrw"    0.337374
"givinup"    0.558908
"easygon"    0.566556
```

VARIANCE:



```
                         0         1         2         3         4         5
Variance          4.563760  3.254322  2.985727  2.448254  2.339837  2.108028
Proportional Var  0.142617  0.101698  0.093304  0.076508  0.073120  0.065876
Cumulative Var    0.142617  0.244315  0.337619  0.414127  0.487247  0.553123
```

Inference:

The weight of the variables in the factor can be obtained from the loadings table. Eigen value is the amount of variance of observed variables explained by the factor. It can be seen that the Factor 1 explains more variance than all other factors i.e. about 14% of the common variance is explained by the factor 1. The 6 factors cumulatively explain about 55% of the common variance i.e. the variance due the correlation among the observed variables. Communality is the amount of commonness the variables share and uniqueness is the exact opposite to the communality. Uniqueness measures the uniqueness of the variables i.e. the amount of contradiction or independence.

Conclusion:

The Bartlett's test of sphericity had a p test statistic of 0.0 at 95% confidence which states that the correlation matrix is not an identity matrix i.e. correlation is present among the variables. The overall KMO statistic value is 0.88 which states the sampling is adequate and thus providing the way for applying factor analysis. The optimal number of factors is 6 as their eigen values are above 1 which can also be inferred from the scree plot. Factor explains creates factor which can explain the amount of variance due to correlation among the variables. The 6 factors cumulatively explain about 55% of the common variance where the factor 1 leads with explaining about 14% of the common variance. Thus, factor analysis has helped us reducing the dimensions by introducing factors where each factor has helped in explaining the variance due to the correlation among the variables.