ST405_Project1

W.M.C.C.M.Wijesingha S/18/836

4/6/2024

1. INTRODUCTION

The purpose of this study is to investigate the impact of these variables on the Happiness Index Score in 2019 and to identify the factors that significantly influence individuals' happiness levels.

2. METHODOLOGY

2.1 Dataset

The dataset provides comprehensive information regarding the World Happiness Report. This dataset, compiled by the United Nations Sustainable Development Solutions Network (UNSDSN), offers valuable insights into global happiness levels and the factors contributing to well-being across different countries and regions.
The pursuit of happiness has been a central theme in human societies throughout history. The Happiness Index Score is calculated based on several variables, including GDP per capita, social support, healthy life expectancy, freedom to make life choices, perceptions of corruption, and generosity.

- Score: List of happiness scores of different countries.
- GDP per capita: The GDP per capita score of different
- countries. Social support: The social support of different countries.
- Healthy life expectancy: The healthy life expectancy of different countries.
- Freedom to make life choices: The score of perception of freedom of different countries.
- Generosity: Generosity (the quality of being kind and generous) score of different countries.
- Perceptions of corruption: The score of the perception of corruption in different countries.

2.2 Statistical Methods

Here I am going to perform the factor analysis technique using principal components method. Since different variables have different measurement units, I standardized the whole data set. Also I have used only numerical variables of the dataset for my analysis purposes.

3. Results and Discussion

As we need only numerical data first we remove unwanted variables from the dataset.

<u>dimention of dataset</u>

```
dim(HappinessData)

## [1] 156   9
```

Remove unwanted variables

```
Happiness_data <- HappinessData[ ,-which(names(HappinessData) %in% c("Overall rank", "Country or regio
n"))]
```

After normalizing the data first of all I check the dataset for whether we can apply the factor analysis techniques. For that I use KMO test and Bartlett's Test.

Kaiser-Mayer-Olkin results

```
KMO(r=cor_matrix)

## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = cor_matrix)
## Overall MSA =  0.82
## MSA for each item =
##                        Score             GDP per capita
##                         0.85                       0.81
##               Social support      Healthy life expectancy
##                         0.87                       0.85
## Freedom to make life choices                 Generosity
##                         0.80                       0.62
##    Perceptions of corruption
##                         0.65
```

Bartlett's Test

```
cortest.bartlett(Happiness_data)

## R was not square, finding R from data

## $chisq
## [1] 657.8741
##
## $p.value
## [1] 1.036055e-125
##
## $df
## [1] 21
```

Since the p-value of Bartlett's Test is less than 0.05 and MSA value of KMO test is 0.82 we proceed to factor analysis.
Then I obtain the correlation matrix and eigen values. From that we can see there is only 2 eigen values greater than 1.
After that I perform principal component analysis. Then construct scree plot and parallel analysis scree plots. By that we can determine two factors are needed.

Correlation Matrix for the data

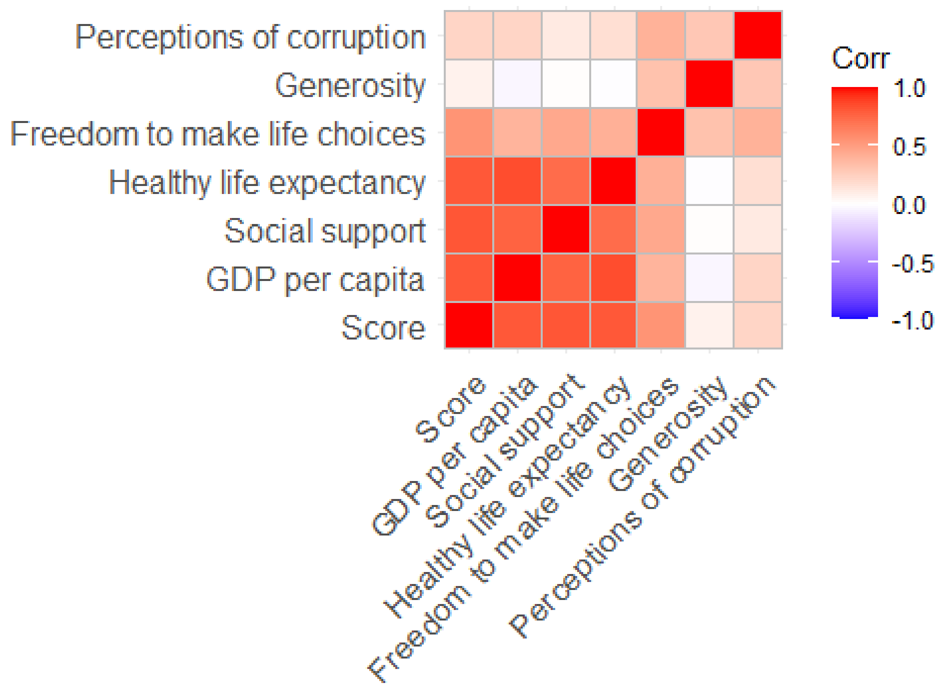
```
cor_matrix<-cor(Happiness_data,method = "spearman")
cor_matrix

##                              Score GDP per capita Social support
## Score                    1.0000000     0.81448340    0.816180736
## GDP per capita           0.8144834     1.00000000    0.768539468
```

```
## Social support             0.8161807     0.76853947    1.000000000
## Healthy life expectancy      0.8072746     0.85151281    0.725997312
## Freedom to make life choices 0.5519742     0.39342614    0.449491258
## Generosity                   0.0704826    -0.04130395    0.007695104
## Perceptions of corruption    0.2173484     0.22320256    0.114017347
##                              Healthy life expectancy
## Score                                   0.807274574
## GDP per capita                          0.851512811
## Social support                          0.725997312
## Healthy life expectancy                 1.000000000
## Freedom to make life choices            0.412905791
## Generosity                             -0.006773188
## Perceptions of corruption               0.165308577
##                              Freedom to make life choices   Generosity
## Score                                           0.5519742  0.070482597
## GDP per capita                                  0.3934261 -0.041303954
## Social support                                  0.4494913  0.007695104
## Healthy life expectancy                         0.4129058 -0.006773188
## Freedom to make life choices                    1.0000000  0.324146014
## Generosity                                      0.3241460  1.000000000
## Perceptions of corruption                       0.4035862  0.286171214
##                              Perceptions of corruption
## Score                                        0.2173484
## GDP per capita                               0.2232026
## Social support                               0.1140173
## Healthy life expectancy                      0.1653086
## Freedom to make life choices                 0.4035862
## Generosity                                   0.2861712
## Perceptions of corruption                    1.0000000
```

Graphical view of Correlation Matrix

```
ggcorrplot(cor_matrix)
```



Eigen values

```
eigen_val<-eigen(cor_matrix)
eigen_val

## eigen() decomposition
## $values
## [1] 3.7876146 1.4449619 0.7055176 0.5019651 0.2752231 0.1500510 0.1346667
```

There is only 2 eigen values greater than 1. There fore 2 factors are needed.

```
## $vectors
##            [,1]        [,2]        [,3]        [,4]         [,5]        [,6]        [,7]
## [1,] -0.48004729 -0.07284003  0.08288148  0.03003839  0.101960891  0.84027354 0.20005411
## [2,] -0.46343526 -0.17930747 -0.09244679 -0.26557198 -0.195744374 -0.40188333 0.68860412
## [3,] -0.44828455 -0.16617529  0.15008022  0.02659439  0.756552945 -0.31114229 -0.28109344
## [4,] -0.45683775 -0.17471236  0.01438337 -0.22705477 -0.558666473 -0.05217022  -0.62784087
## [5,] -0.33289137  0.40420264  0.06668338  0.80396943 -0.207171321 -0.17469724 0.03938235
## [6,] -0.06202367  0.65703881  0.62323044 -0.41703962 -0.002685312 -0.03244880  0.03247637
## [7,] -0.17452506  0.55616972 -0.75431455 -0.23668376  0.154532727  0.03642614   -0.09999685
```

Find PCA

```
pca <- princomp(cor_matrix)
pca

## Call:
## princomp(x = cor_matrix)
```

```
##
## Standard deviations:
##     Comp.1     Comp.2     Comp.3              Comp.4                 Comp.5    Comp.6           Comp.7
## 0.75901758 0.26960842 0.21783734 0.10536095 0.06586333 0.05176683 0.00000000
```

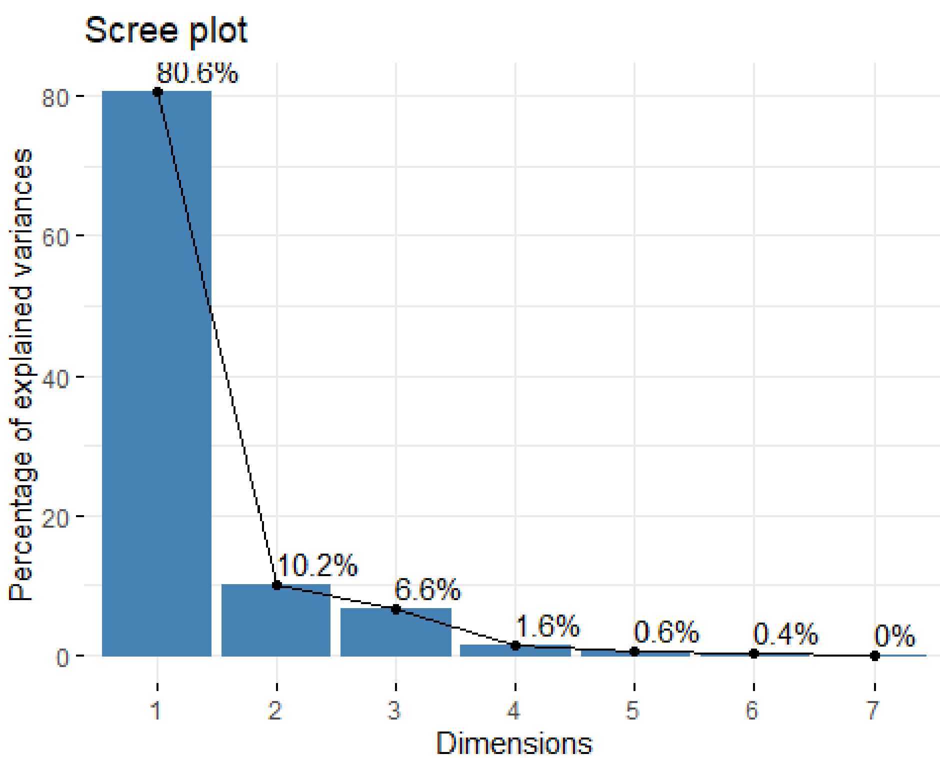
```
summary(pca)
```

```
## Importance of components:
##                        Comp.1    Comp.2     Comp.3     Comp.4      Comp.5
## Standard deviation     0.7590176 0.2696084 0.21783734 0.10536095 0.065863333
## Proportion of Variance 0.8064576 0.1017524 0.06642668 0.01553951 0.006072469
## Cumulative Proportion  0.8064576 0.9082101 0.97463673 0.99017624 0.996248707
##                        Comp.6 Comp.7
## Standard deviation     0.051766828      0
## Proportion of Variance 0.003751293      0
## Cumulative Proportion  1.000000000      1
```

Scree plot

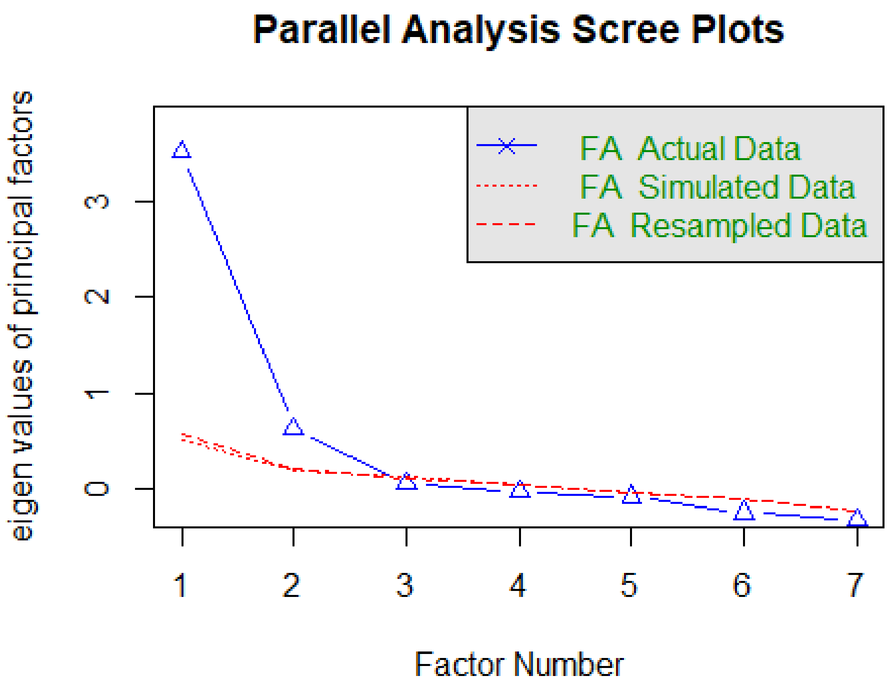
```
library(factoextra)

fviz_eig(pca,addlabels=TRUE)
```



Parallel analysis

```
fa.parallel(Happiness_data,fm = "pa",fa = "fa")
```




```
## Parallel analysis suggests that the number of factors =  2  and the number of components =  NA
```

Covariance matix

```
cov_matrix<-cov(Happiness_data)
cov_matrix

##                              Score        GDP per capita Social support
## Score                         1.239035842  0.352051508  0.258788134
## GDP per capita                0.352051508  0.158714165  0.089980763
## Social support                0.258788134  0.089980763  0.089515494
## Healthy life expectancy       0.210188673  0.080588383  0.052086065
## Freedom to make life choices  0.090394390  0.021639735  0.019177610
## Generosity                    0.008039557 -0.003023055 -0.001371571
## Perceptions of corruption     0.040578813  0.011258182  0.005145010
##                              Healthy life expectancy
## Score                         0.2101886732
## GDP per capita                0.0805883832
## Social support                0.0520860649
## Healthy life expectancy       0.0586240306
## Freedom to make life choices  0.0135442859
## Generosity                   -0.0006806203
## Perceptions of corruption     0.0067589878
##                      Freedom to make life choices    Generosity
## Score                         0.090394390  0.0080395568
## GDP per capita                0.021639735 -0.0030230546
## Social support                0.019177610 -0.0013715707
## Healthy life expectancy       0.013544286 -0.0006806203
## Freedom to make life choices  0.020531872  0.0036816948
## Generosity                    0.003681695  0.0090734084
```

```
## Perceptions of corruption                0.005944693  0.0029405191
##                         Perceptions of corruption
## Score                                0.040578813
## GDP per capita                       0.011258182
## Social support                       0.005145010
## Healthy life expectancy              0.006758988
## Freedom to make life choices         0.005944693
## Generosity                           0.002940519
## Perceptions of corruption            0.008937402
```

Proportion of variation explained by each PCs.

```
PVE <- eigen_val$values / sum(eigen_val$values) * 100
round(PVE,3)
```

```
## [1] 54.109    20.642      10.079     7.171     3.932     2.144      1.924
```

By the 1st component 54% of population variability is described. Likewise 20.6% ,10%, 7%, 3.9%, 2% , 1.9% total variability is described by the other components respectively.

```
pca$loadings[, 1:2]
```

```
##                                   Comp.1          Comp.2
## Score                              0.41387523      0.03176005
## GDP per capita                     0.46137354     -0.08919941
## Social support                     0.44323640      0.12632486
## Healthy life expectancy            0.45327365      0.01078677
## Freedom to make life choices       0.01489168     -0.11257009
## Generosity                        -0.40340432      0.53682349
## Perceptions of corruption         -0.22576412     -0.82104139
```

#Factor analysis

With 2 factors

```
Happiness_data.fa<-factanal(Happiness_data,factors = 2)
Happiness_data.fa
```

```
## Call:
## factanal(x = Happiness_data, factors = 2)
##
## Uniquenesses:
##                   Score            GDP per capita
##                   0.152                 0.149
##            Social support     Healthy life expectancy
##                   0.304                 0.209
## Freedom to make life choices              Generosity
##                   0.425                 0.711
##    Perceptions of corruption
##                   0.658
##
## Loadings:
##                                    Factor1                     Factor2
```

```
## Score                          0.852                0.350
## GDP per capita                  0.921
## Social support                 0.822                0.143
## Healthy life expectancy        0.885
## Freedom to make life choices   0.390                0.650
## Generosity                     -0.116               0.525
## Perceptions of corruption      0.252                0.528
##
##                                Factor1              Factor2
## SS loadings                    3.263                1.130
## Proportion Var                 0.466                0.161
## Cumulative Var                 0.466                0.628
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 20.63 on 8 degrees of freedom.
## The p-value is 0.00821
```

<u>With 3 factors</u>

```
Happiness_data.fa<-factanal(Happiness_data,factors = 3)

Happiness_data.fa

##Test of the hypothesis that 3 factors are sufficient.

##The chi square statistic is 0.82 on 3 degrees of freedom.

##The p-value is 0.844
```

Here I perform factor analysis. From the output we can see that 0.628 proportion of the variation is explained by the factor model. Since p-value is less than 0.05 we reject the null hypothesis and can conclude two factors are not sufficient.
Then I perform factor analysis using three factors obtain a p-value of 0.844 then I conclude that 3 factor model is sufficient.

4. Conclusion and Recommendation

- By the above results we can conclude that factor 1 is correlated with Score, GDP per capita, Social support, Healthy life expectancy, Freedom to make life choices , Perceptions of corruption.
- Factor 2 is correlated with Score, Freedom to make life choices, Generosity, Perceptions of corruption and factor 3 correlated with Score, GDP per capita, Social support, Freedom to make life choices, Perceptions of corruption

5. References

Thurstone, L. L. (1947). Multiple factor analysis. University of Chicago Press.
Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

6. Appendices

6.1 Dataset

https://www.kaggle.com/datasets/unsdsn/world-happiness/data

Overall r	Country or region	Score	GDP per capita	Social support	Healthy life e	Freedom to m	Generosity	Perception
1	Finland	7.769	1.34	1.587	0.986	0.596	0.153	0.393
2	Denmark	7.6	1.383	1.573	0.996	0.592	0.252	0.41
3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
4	Iceland	7.494	1.38	1.624	1.026	0.591	0.354	0.118
5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
6	Switzerland	7.48	1.452	1.526	1.052	0.572	0.263	0.343
7	Sweden	7.343	1.387	1.487	1.009	0.574	0.267	0.373
8	New Zealand	7.307	1.303	1.557	1.026	0.585	0.33	0.38
9	Canada	7.278	1.365	1.505	1.039	0.584	0.285	0.308
10	Austria	7.246	1.376	1.475	1.016	0.532	0.244	0.226

6.2 R codes

- ```
  library(psych)
  library(tidyverse)
  library(ggplot2)
  library(ggcorrplot)
  library(gridExtra)
  ```

- ```
  #Read dataset
  HappinessData <- read_csv("../data/Happiness_data_2019.csv")
  head(HappinessData)
  #Dimentions of dataset
  dim(HappinessData)
  ```

- ```
  #Keep numerical variables
  Happiness_data <- HappinessData[ ,-which(names(HappinessData) %in% c("Overall rank", "Country
  or region"))]
  ```

- ```
  apply(Happiness_data,2,mean)
  ```
- ```
  apply(Happiness_data,2,var)
  ```

- ```
  #Correlation matrix
  cor_matrix<-cor(Happiness_data,method = "spearman")
  cor_matrix
  ggcorrplot(cor_matrix)
  ```

- ```
  #Eigen values and Eigen vector

  eigen_val<-eigen(cor_matrix)
  eigen_val
  ```

- ```
  #Principle component Analysis
  pca <- princomp(cor_matrix)
  pca

  summary(pca)
  ```

- **#Scree plot**
  ```
  library(factoextra)
  fviz_eig(pca,addlabels=TRUE)
  ```

- **#Parallel analysis**
  ```
  fa.parallel(Happiness_data,fm = "pa",fa = "fa")
  ```

- **#Covariance matix**
  ```
  cov_matrix<-cov(Happiness_data)
  cov_matrix
  ```

- **#Kaiser-Mayer-Olkin results**
  ```
  KMO(r=cor_matrix)
  ```

- **#Bartlett's Test**
  ```
   cortest.bartlett(Happiness_data)
  ```

- **#determinant of covariance matrix**
  ```
  det(cov_matrix)
  ```

- **#Proportion of variation explained by each PCs.**
  ```
  PVE <- eigen_val$values / sum(eigen_val$values) * 100
  round(PVE,3)
  sum(eigen_val$values)
  sum(PVE)
  ```

- **#Factor analysis**
  ```
  pca$loadings[, 1:2]
  Happiness_data.fa<-factanal(Happiness_data,factors = 2)
  Happiness_data.fa
  ```