Final Project - STA 6707

Multivariate Data Analysis on World Happiness Report Dataset

Mina Akhondzadeh

December 4, 2024

About Data

The **World Happiness Report** is an annual publication that ranks countries based on happiness levels, using various socio-economic and psychological factors. Drawing on survey data and statistical analysis, it assesses global well-being and highlights the key drivers of happiness, offering valuable insights for policymakers and researchers worldwide.

Dataset Description

- **country**: The name of the country.
- **region**: The geographical region to which the country belongs.
- **score**: The country's overall happiness score.
- **gdp_per_capita**: A measure of the country's economic output per person.
- **social_support**: A measure of the social network and support available to individuals in the country.
- **healthy_life_expectancy**: The average number of years a person is expected to live in good health.
- **freedom_to_make_life_choices**: A measure of how free individuals feel to make their own life decisions.
- **generosity**: The level of generosity, often measured by charitable donations.
- **perceptions_of_corruption**: A measure of how corrupt a country's institutions are perceived to be.
- **year**: The year of the report.

Data Exploration

The correlation matrix reveals the relationships between the variables.

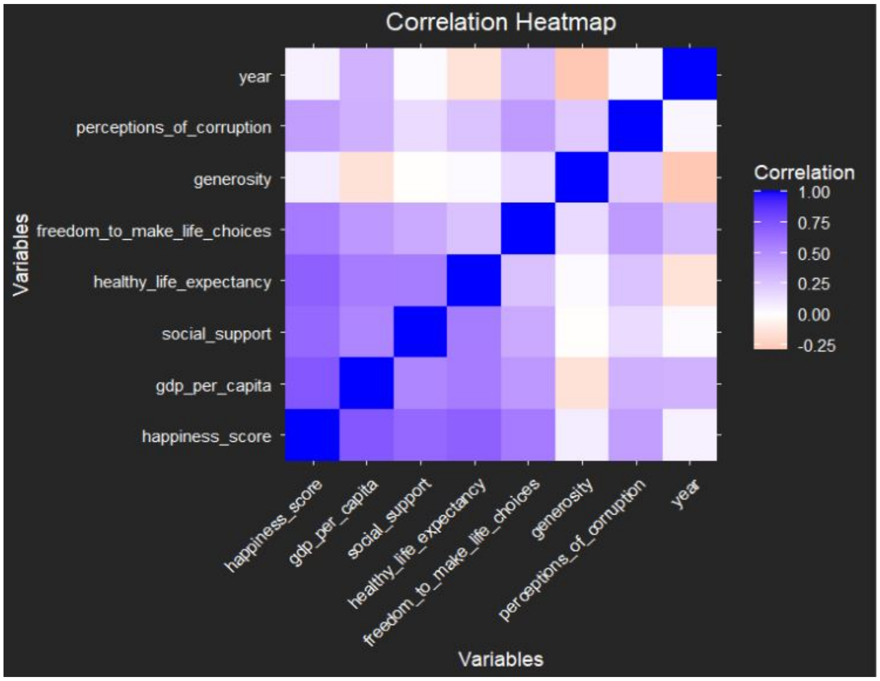


Figure 1: Correlation Matrix

- **Personal Freedom** correlates moderately with happiness.
- **Generosity** has a very weak positive correlation with happiness, suggesting that generosity alone may not strongly impact overall happiness. Over time, generosity has decreased slightly on average. Generosity shows a negative correlation with GDP per capita, possibly reflecting that generosity is not necessarily higher in wealthier countries.
- **Perceptions of corruption** shows positive correlations with the rest of the variables, suggesting a relationship between trust in governance and various underlying patterns in the data. This may indicate that governance perceptions shift depending on broader societal or economic conditions.
- **Year** has a negligible correlation with happiness, indicating minimal variation in happiness trends over time.
- **GDP per capita** is moderately correlated with:
  - **Healthy life expectancy**, showing that wealthier nations tend to have better health outcomes.
  - **Social Support**, indicating stronger social networks in wealthier nations.

- **Freedom to make life choices** and perceptions of corruption are moderately correlated, indicating a connection between perceived freedom and trust in institutions.

Although the happiness score is not included in the PCA, observing its spread in Figure 2 helps us understand the range and variability of the target outcome, which is influenced by the predictors used in PCA.

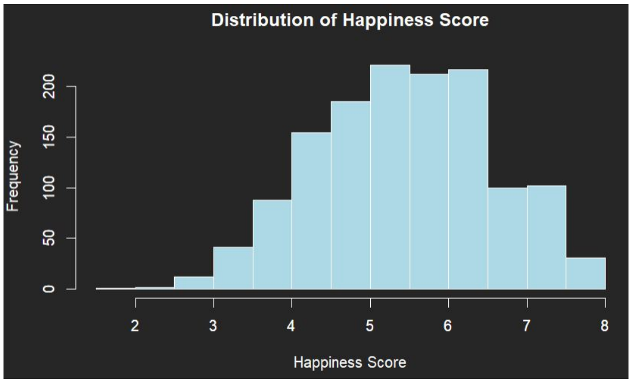


Figure 2: Distribution of Happiness Scores

Figure 3 tells us that economic prosperity has the highest correlation with happiness, indicating that wealthier nations tend to have higher happiness scores.

Western Europe and North America are clustered at the higher end of both GDP and happiness, reflecting their strong economic performance and higher levels of well-being.

In contrast, regions such as Sub-Saharan Africa and South Asia are concentrated at the lower end of GDP and happiness, highlighting the disparities in economic development and their impact on happiness.

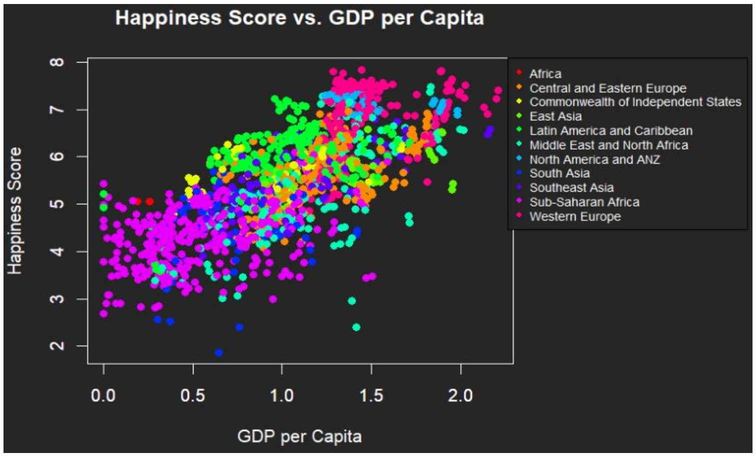


Figure 3: Happiness score distribution

Figure 4 shows that North America and Western Europe have the highest median happiness score and relatively small variability. Sub-Saharan Africa and South Asia have some of the lowest happiness scores, with significant variability in their distributions. Regions like

Latin America and the Caribbean, and Central and Eastern Europe display moderate scores with varying spreads. The Middle East and North Africa show significant variability, with happiness scores ranging from the minimum to the higher end of the scale.

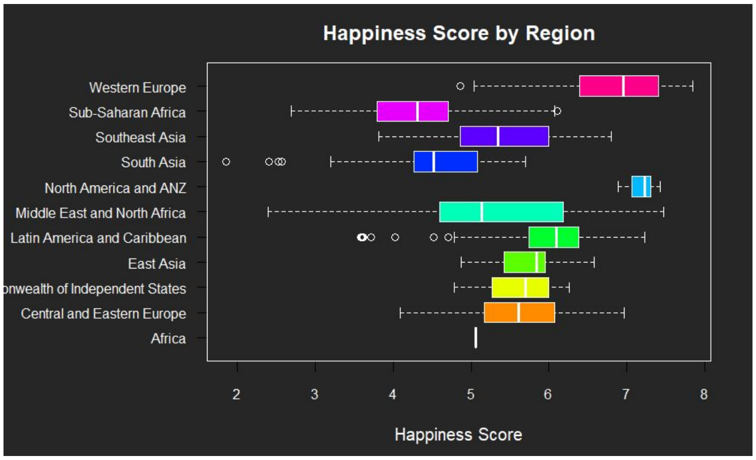


Figure 4: Happiness Score distributions

Principal Component Analysis

After exploring the data using various plots to understand its distributions and relationships, we now perform PCA to uncover underlying patterns and reduce dimensionality for further analysis.

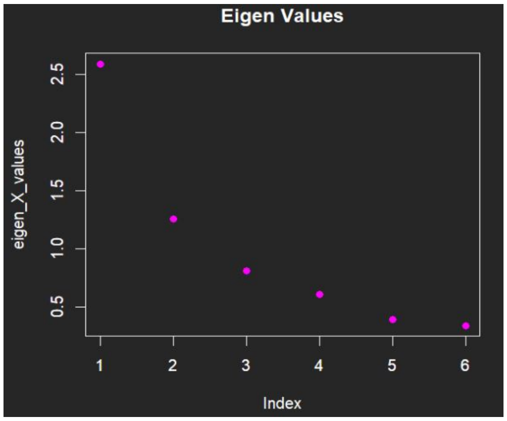
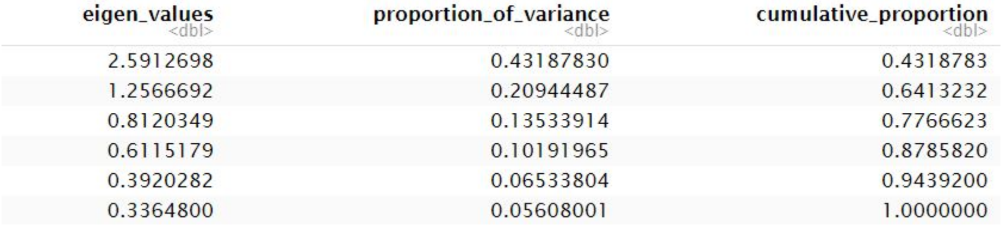


Figure 5: Eigenvalues

The sharp drop after the first eigenvalue suggests that the first principal component captures most of the variance in the data, but we also have relatively sharp drops after 2nd, 3rd, and 4th eigenvalues. We can say the top 4 eigenvalues are explaining the majority of the variance of the data.

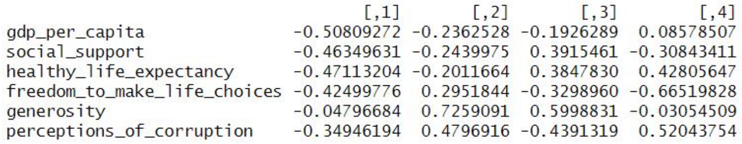
The variance explained by each PC and their cumulative sum confirm our findings above. Figure 6 shows that the first four PCs account for over 87% of the total variance in the data, with PC1 alone explaining approximately 43% of the variability.



eigen_values <dbl>	proportion_of_variance <dbl>	cumulative_proportion <dbl>
2.5912698	0.43187830	0.4318783
1.2566692	0.20944487	0.6413232
0.8120349	0.13533914	0.7766623
0.6115179	0.10191965	0.8785820
0.3920282	0.06533804	0.9439200
0.3364800	0.05608001	1.0000000

Figure 6: Proportion of Variance

Figure 7 illustrates the weights of each variable in the first four PCs. PC1 is primarily influenced by GDP. PC2 and PC3 are mainly influenced by generosity, PC4 is driven by freedom to make life choices.



```
                                   [,1]       [,2]       [,3]        [,4]
gdp_per_capita              -0.50809272 -0.2362528 -0.1926289  0.08578507
social_support              -0.46349631 -0.2439975  0.3915461 -0.30843411
healthy_life_expectancy     -0.47113204 -0.2011664  0.3847830  0.42805647
freedom_to_make_life_choices -0.42499776  0.2951844 -0.3298960 -0.66519828
generosity                  -0.04796684  0.7259091  0.5998831 -0.03054509
perceptions_of_corruption   -0.34946194  0.4796916 -0.4391319  0.52043754
```

Figure 7: PCs

Figure 8 displays the transformed data along different PCs, colored by various variables from the original dataset. Figure 9 is the two dimensional format of it. This visualization reveals how the PCs are connected to the original data.

PC1 is the most influential component, capturing significant trends related to happiness and GDP. PC2 and PC4 show moderate relationships with corruption.

These plots effectively demonstrate how the principal components capture distinct variations within the dataset, particularly in economic and governance-related indicators.
Now we can plot the correlation between the first 4 PCs and the variables.

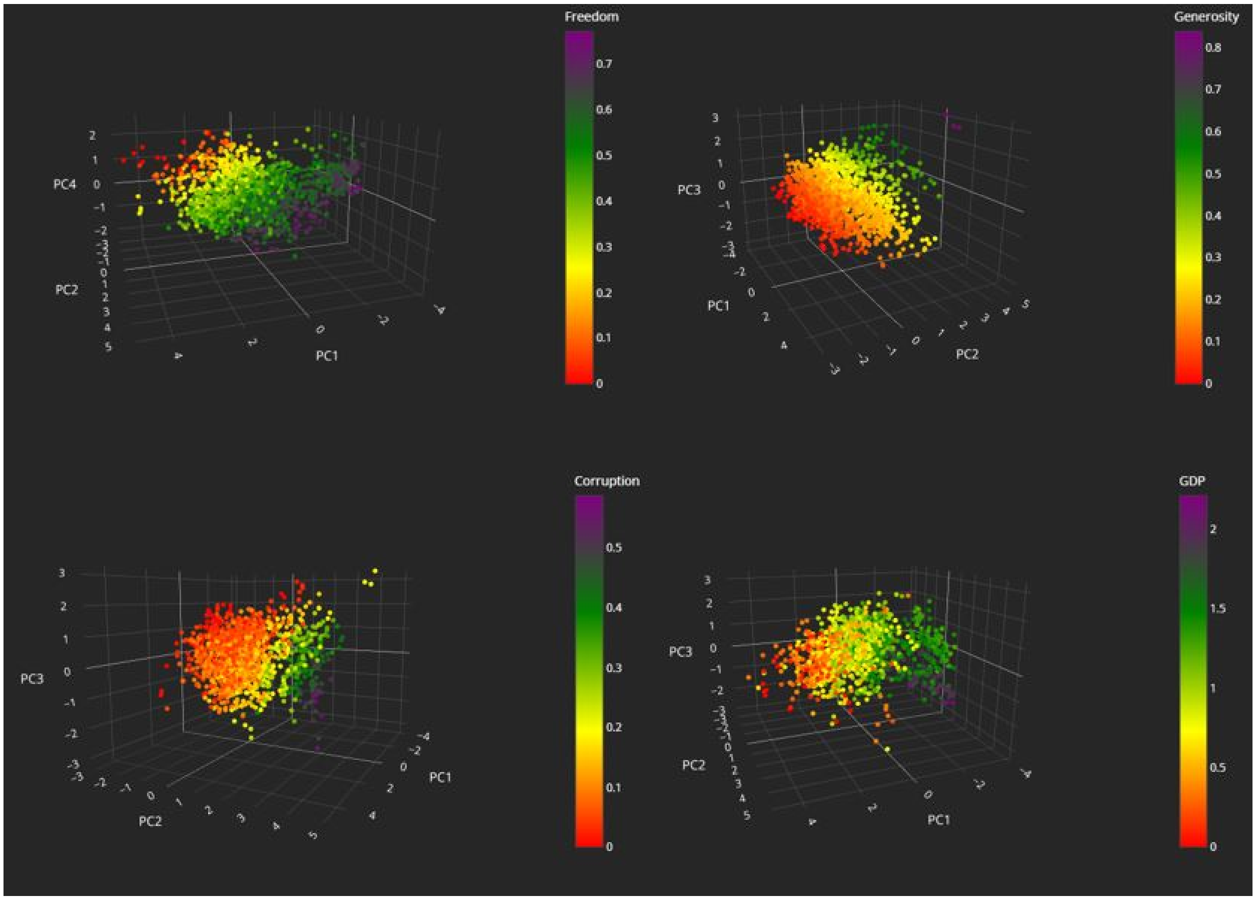


Figure 8: Transformed Data Across PCs in 3D

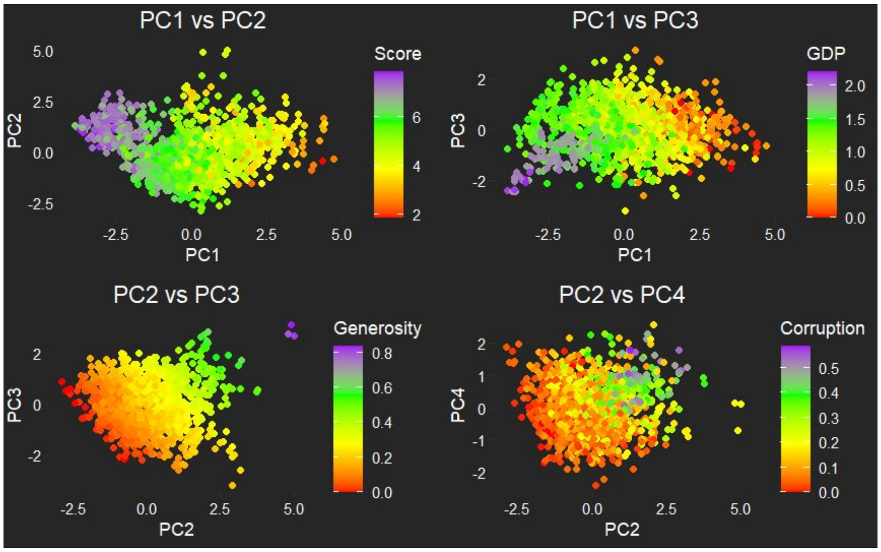


Figure 9: Transformed Data Across PCs in 2D

From Figures 10, 11, and 12, we conclude that: PC1 is influenced the most by economic and social factors such as GDP per capita, social support, and healthy life expectancy, indi-

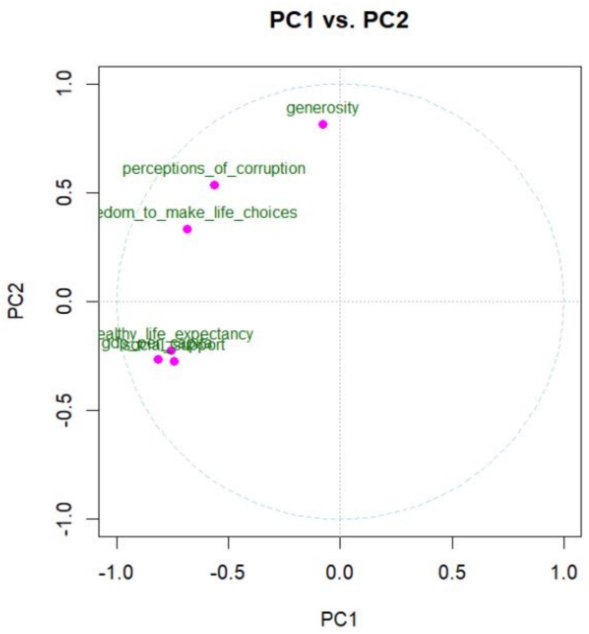


Figure 10: PC1 vs. PC2

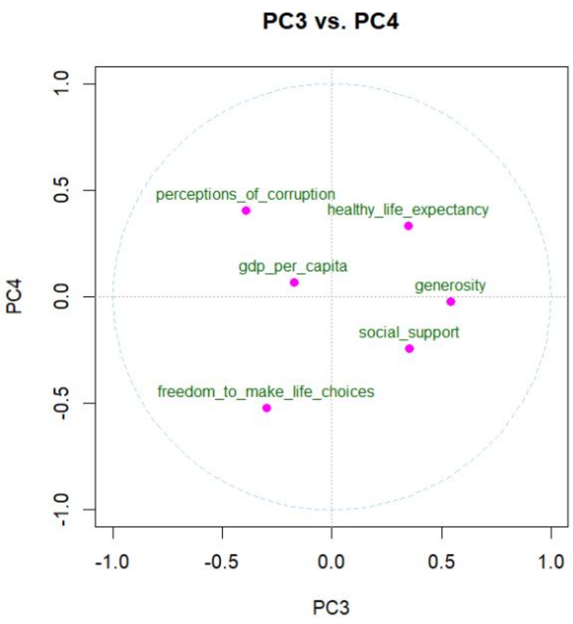


Figure 11: PC3 vs. PC4

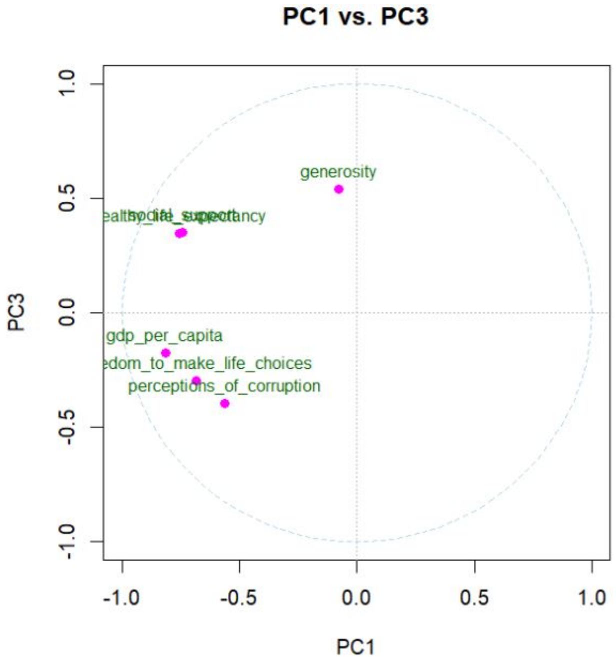


Figure 12: PC1 vs. PC3

cating that it captures economic well-being and social support.
PC2 is primarily influenced by generosity and perceptions_of_corruption, suggesting it captures aspects related to social trust and selflessness. Similar to PC2, PC3 highlights social trust and selflessness factors, driven by generosity and perception of corruption. PC4 is primarily influenced by individual freedom, while GDP, social support, and generosity have

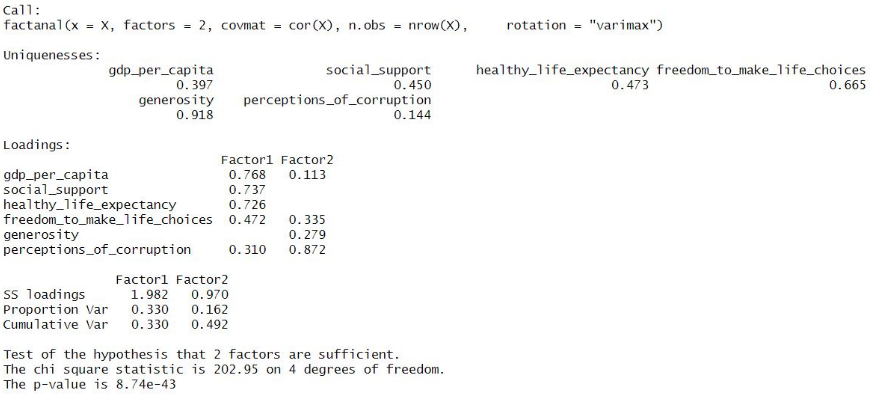
minimal influence on it.

Factor Analysis

In the factor analysis, we are looking for unobserved variables (factors) that can explain the variability of our data. As previously discussed, the variables in the dataset can be clustered into different categories based on different factors such as economic well-being, social trust and selflessness.

With factor analysis, we can explore the correlations between the variables and see if we can explain them with fewer underlying factors.

The first step in factor analysis is determining the appropriate number of factors. For $p = 6$ observed variables, the degrees of freedom formula limits the maximum number of factors to $k = 2$. When using only one factor, the uniqueness of `generosity` and `perceptions_of_corruption` is very large. To minimize uniqueness while ensuring valid degrees of freedom, we use two factors.



```
Call:
factanal(x = X, factors = 2, covmat = cor(X), n.obs = nrow(X),     rotation = "varimax")

Uniquenesses:
              gdp_per_capita               social_support      healthy_life_expectancy freedom_to_make_life_choices
                       0.397                        0.450                        0.473                        0.665
                  generosity    perceptions_of_corruption
                       0.918                        0.144

Loadings:
                             Factor1 Factor2
gdp_per_capita                0.768   0.113
social_support                0.737
healthy_life_expectancy       0.726
freedom_to_make_life_choices  0.472   0.335
generosity                            0.279
perceptions_of_corruption     0.310   0.872

               Factor1 Factor2
SS loadings      1.982   0.970
Proportion Var   0.330   0.162
Cumulative Var   0.330   0.492

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 202.95 on 4 degrees of freedom.
The p-value is 8.74e-43
```

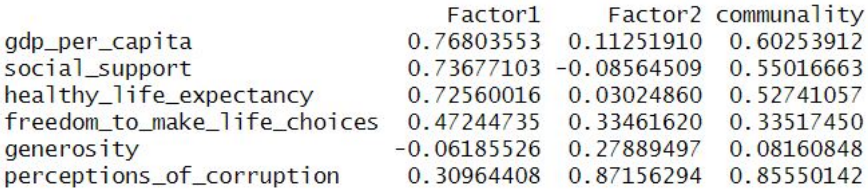
Figure 13: Factor Analysis Solution for 2 factors

Based on Figure 13, the uniqueness of “generosity” and “freedom_to_make_life_choices” are relatively high, meaning these two variables are not fully explained by the factors identified. This suggests there might be additional dimensions that these variables represent, which are not captured by the current factors.

Factor 1: `gdp_per_capita`, `social_support`, and `healthy_life_expectancy` have high loadings, indicating that this factor primarily explains economic well-being and social support.

Factor 2: The `perceptions_of_corruption` variable is the only variable with a high loading on Factor 2. It is also the only variable that significantly influences all three principal components: PC2, PC3, and PC4. Therefore, we can interpret Factor 2 as reflecting these three principal components, suggesting that this factor represents trust in governance, while the other loadings remain relatively low.

The cumulative variance is 0.492, which means these two factors explain about 50% of the variance in the data. Also, the p-value is 8.74e-43 which is too small and we reject the null and we conclude that 2 factors are sufficient for Factor Analysis.



```
                                Factor1     Factor2 communality
gdp_per_capita               0.76803553  0.11251910  0.60253912
social_support               0.73677103 -0.08564509  0.55016663
healthy_life_expectancy      0.72560016  0.03024860  0.52741057
freedom_to_make_life_choices 0.47244735  0.33461620  0.33517450
generosity                  -0.06185526  0.27889497  0.08160848
perceptions_of_corruption    0.30964408  0.87156294  0.85550142
```

Figure 14: Loadings of 2 factors

Based on Figure 14, `perceptions_of_corruption` and `gdp_per_capita` are well explained by the two factors. However, `generosity` has the lowest communality, with the two factors explaining only about 8% of its variance.

The results indicate that the factors effectively capture trust in governance and society, economic well-being, and social support but fail to sufficiently represent generosity and selflessness.

We can visualize the loadings of Factor 1 against Factor 2. Figure 15 confirms our previous conclusions. Factor 1 is strongly influenced by `gdp_per_capita`, `healthy_life_expectancy`, and `social_support`, while Factor 2 has near-zero loadings for these variables. Factor 2 is strongly influenced by `perceptions_of_corruption`.

`Generosity` is slightly explained only by Factor 2, whereas `freedom_to_make_life_choices` is slightly and equally explained by both factors.

Conclusion: Both Principal Component Analysis (PCA) and Factor Analysis (FA) aim to simplify complex data by identifying patterns and reducing the number of variables, but they achieve this in slightly different ways. PCA focuses on finding combinations of variables (PCs) that explain the most variation in the data. FA, on the other hand, seeks underlying factors that explain the relationships between variables by modeling their shared patterns.

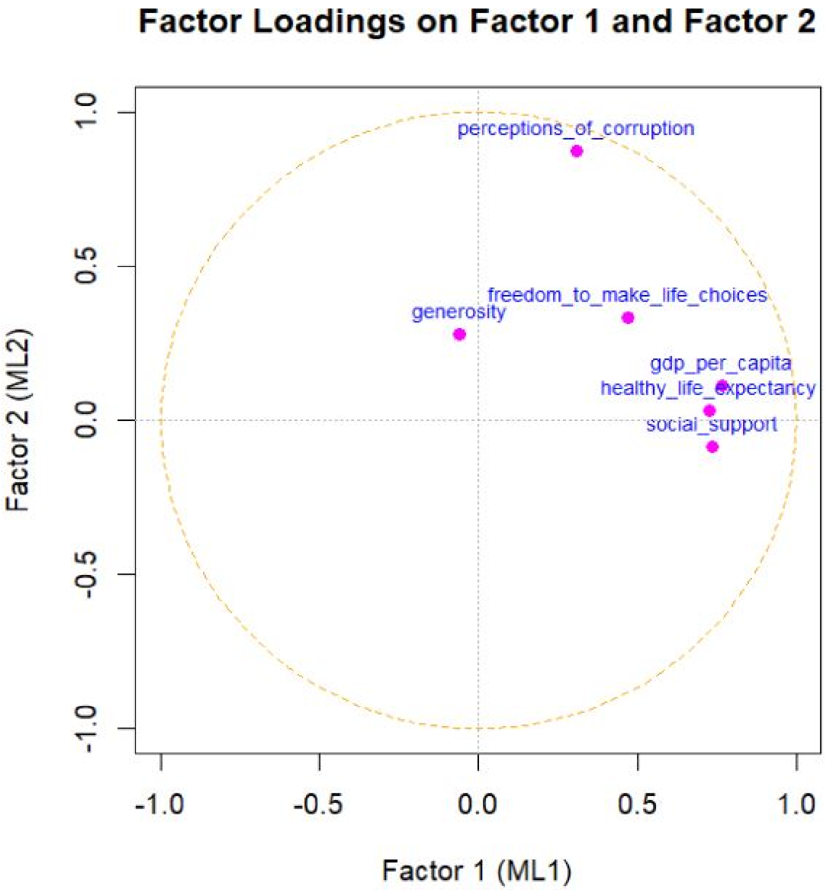


Figure 15: Factor Loadings Factor1 vs. Factor2

In the World Happiness Report data, both PCA and FA identify the key drivers of happiness, reflecting their shared goal of summarizing the data. For instance, PC1 in PCA closely aligns with Factor 1 in FA, as both emphasize economic well-being and social support. Similarly, PC2 in PCA highlights generosity and perceptions of corruption, which is comparable to Factor 2 in FA, representing perceptions of corruption.

However, unlike PC2 in PCA, generosity remains largely unexplained by the two factors in FA, with a high uniqueness value. This suggests the possibility of additional underlying factors influencing generosity that were not captured in the surveys.

`https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors?select=WHR_2023.csv`