# CS 579: Project 2: Final Report


Department of Computer Science Illinois Institute of Technology

May 4th, 2024

Pranav Saji (A20518289)
Mithilesh Bade (A20514157)


## Project Outline

# **1.Introduction**

Our World Happiness Analysis Project aims to explore and understand the varied levels of happiness across countries by delving into the factors that significantly impact national happiness. Utilizing data from the World Happiness Report, which includes metrics like GDP per capita, social support, healthy life expectancy, freedom, trust, and generosity, my teammate and I are analyzing how these factors interact and influence overall life satisfaction. We employ advanced statistical techniques to uncover patterns and trends, providing actionable insights into the socio-economic and psychological drivers of happiness.

Motivated by a desire to quantify the subjective experience of happiness, our goals include identifying key factors affecting happiness, discovering their interdependencies, and predicting changes based on shifts in these influential factors. We aim to provide data-driven insights that can inform policymakers and educators, helping them to enhance well-being in their communities. By the project's conclusion, we aspire to contribute to the broader discussion on global well-being, offering nuanced perspectives supported by comprehensive data analysis.

However, we face challenges such as variability in data quality, representativeness of survey samples, and the inherent biases in self-reported happiness measures. We are addressing these issues through meticulous data management and sophisticated analytical methods to ensure the reliability of our findings. Through this rigorous approach, our project seeks not just to explore what makes people happy but to understand how to foster environments that improve life quality globally.

# **2.Data**

## 2.1) Dataset Description and Acquisition

2.1.a) Dataset Overview
For our project, we utilized the World Happiness Report dataset, an annual publication by the Sustainable Development Solutions Network. This dataset ranks countries based on their citizens' reported happiness levels. The key metrics include life satisfaction ratings and contributing factors such as GDP per capita, social support, healthy life expectancy, freedom, trust, and generosity.

| Country name | year | Life Ladder | Log GDP per capita | Social support | Healthy life expectancy at birth | Freedom to make life choices | Generosity | Perceptions of corruption | Positive affect | Negative affect |
|---|---|---|---|---|---|---|---|---|---|---|
| Afghanistan | 2008 | 3.724 | 7.350 | 0.451 | 50.500 | 0.718 | 0.164 | 0.882 | 0.414 | 0.258 |
| Afghanistan | 2009 | 4.402 | 7.509 | 0.552 | 50.800 | 0.679 | 0.187 | 0.850 | 0.481 | 0.237 |
| Afghanistan | 2010 | 4.758 | 7.614 | 0.539 | 51.100 | 0.600 | 0.118 | 0.707 | 0.517 | 0.275 |
| Afghanistan | 2011 | 3.832 | 7.581 | 0.521 | 51.400 | 0.496 | 0.160 | 0.731 | 0.480 | 0.267 |
| Afghanistan | 2012 | 3.783 | 7.661 | 0.521 | 51.700 | 0.531 | 0.234 | 0.776 | 0.614 | 0.268 |
| Afghanistan | 2013 | 3.572 | 7.680 | 0.484 | 52.000 | 0.578 | 0.059 | 0.823 | 0.547 | 0.273 |
| Afghanistan | 2014 | 3.131 | 7.671 | 0.526 | 52.300 | 0.509 | 0.102 | 0.871 | 0.492 | 0.375 |
| Afghanistan | 2015 | 3.983 | 7.654 | 0.529 | 52.600 | 0.389 | 0.078 | 0.881 | 0.491 | 0.339 |
| Afghanistan | 2016 | 4.220 | 7.650 | 0.559 | 52.925 | 0.523 | 0.040 | 0.793 | 0.501 | 0.348 |
| Afghanistan | 2017 | 2.662 | 7.648 | 0.491 | 53.250 | 0.427 | -0.123 | 0.954 | 0.435 | 0.371 |
| Afghanistan | 2018 | 2.694 | 7.631 | 0.508 | 53.575 | 0.374 | -0.095 | 0.928 | 0.385 | 0.405 |
| Afghanistan | 2019 | 2.375 | 7.640 | 0.420 | 53.900 | 0.394 | -0.109 | 0.924 | 0.324 | 0.502 |
| Afghanistan | 2021 | 2.436 | 7.325 | 0.454 | 54.550 | 0.394 | -0.085 | 0.946 | 0.179 | 0.607 |
| Afghanistan | 2022 | 1.281 | | 0.228 | 54.875 | 0.368 | | 0.733 | 0.206 | 0.576 |
| Afghanistan | 2023 | 1.446 | | 0.368 | 55.200 | 0.228 | | 0.738 | 0.261 | 0.460 |
| Albania | 2007 | 4.634 | 9.122 | 0.821 | 66.760 | 0.529 | -0.013 | 0.875 | 0.489 | 0.246 |
| Albania | 2009 | 5.485 | 9.241 | 0.833 | 67.320 | 0.525 | -0.162 | 0.864 | 0.564 | 0.279 |
| Albania | 2010 | 5.269 | 9.283 | 0.733 | 67.600 | 0.569 | -0.176 | 0.726 | 0.576 | 0.300 |
| Albania | 2011 | 5.867 | 9.310 | 0.759 | 67.880 | 0.487 | -0.209 | 0.877 | 0.566 | 0.257 |
| Albania | 2012 | 5.510 | 9.326 | 0.785 | 68.160 | 0.602 | -0.173 | 0.848 | 0.553 | 0.271 |
| Albania | 2013 | 4.551 | 9.338 | 0.759 | 68.440 | 0.632 | -0.131 | 0.863 | 0.541 | 0.338 |
| Albania | 2014 | 4.814 | 9.358 | 0.626 | 68.720 | 0.735 | -0.029 | 0.883 | 0.573 | 0.335 |
| Albania | 2015 | 4.607 | 9.382 | 0.639 | 69.000 | 0.704 | -0.085 | 0.885 | 0.579 | 0.350 |
| Albania | 2016 | 4.511 | 9.417 | 0.638 | 69.025 | 0.730 | -0.021 | 0.901 | 0.567 | 0.322 |
| Albania | 2017 | 4.640 | 9.455 | 0.638 | 69.050 | 0.750 | -0.033 | 0.876 | 0.547 | 0.334 |
| Albania | 2018 | 5.004 | 9.497 | 0.684 | 69.075 | 0.824 | 0.005 | 0.899 | 0.592 | 0.319 |
| Albania | 2019 | 4.995 | 9.522 | 0.686 | 69.100 | 0.777 | -0.103 | 0.914 | 0.548 | 0.274 |
| Albania | 2020 | 5.365 | 9.494 | 0.710 | 69.125 | 0.754 | 0.002 | 0.891 | 0.563 | 0.265 |
| Albania | 2021 | 5.255 | 9.588 | 0.702 | 69.150 | 0.827 | 0.039 | 0.896 | 0.554 | 0.254 |
| Albania | 2022 | 5.212 | 9.649 | 0.724 | 69.175 | 0.802 | -0.070 | 0.846 | 0.547 | 0.255 |
| Albania | 2023 | 5.445 | 9.689 | 0.691 | 69.200 | 0.872 | 0.068 | 0.855 | 0.597 | 0.314 |
| Algeria | 2010 | 5.464 | 9.306 | | 65.500 | 0.593 | -0.212 | 0.618 | | |
| Algeria | 2011 | 5.317 | 9.316 | 0.810 | 65.600 | 0.530 | -0.188 | 0.638 | 0.503 | 0.255 |
| Algeria | 2012 | 5.605 | 9.330 | 0.839 | 65.700 | 0.587 | -0.179 | 0.690 | 0.540 | 0.230 |
| Algeria | 2014 | 6.355 | 9.355 | 0.818 | 65.900 | | | | 0.558 | 0.177 |
| Algeria | 2016 | 5.341 | 9.383 | 0.749 | 66.100 | | | | 0.565 | 0.377 |
| Algeria | 2017 | 5.249 | 9.377 | 0.807 | 66.200 | 0.437 | -0.174 | 0.700 | 0.555 | 0.289 |
| Algeria | 2018 | 5.043 | 9.370 | 0.799 | 66.300 | 0.583 | -0.153 | 0.759 | 0.534 | 0.293 |
| Algeria | 2019 | 4.745 | 9.361 | 0.803 | 66.400 | 0.385 | -0.002 | 0.741 | 0.544 | 0.215 |
| Algeria | 2020 | 5.438 | 9.291 | 0.868 | 66.500 | 0.574 | -0.124 | 0.724 | 0.524 | 0.311 |
| Algeria | 2021 | 5.217 | 9.308 | 0.841 | 66.600 | 0.558 | -0.116 | 0.712 | 0.498 | 0.258 |
| Algeria | 2022 | 5.538 | 9.323 | 0.783 | 66.700 | 0.440 | -0.045 | 0.611 | 0.583 | 0.259 |
| Angola | 2011 | 5.589 | 8.944 | 0.723 | 51.220 | 0.584 | 0.050 | 0.911 | 0.667 | 0.361 |

Fig - Representation of the dataset

2.1.b) Data Source and Acquisition
The dataset was directly downloaded from the World Happiness Report website. Accessing the most recent data ensured the relevance and timeliness of our analysis. We specifically downloaded the data for the year 2024 from their official data repository, available in Excel format. This approach ensured that we were working with comprehensive and up-to-date information.

## 2.2) Data Sampling Methodology

Given the dataset's nature as a comprehensive global survey, there was no need for sampling; we used the complete dataset provided. This dataset includes data from various countries, enabling a full-scale analysis of global happiness without the need for further sampling, which might compromise the breadth of insights obtained.Our decision to use the entire dataset was driven by the objective to maintain the integrity and representativeness of the data. Using the full dataset avoids potential biases and errors that might arise from sampling methods. This allows for an accurate representation of global happiness trends and factors.

## 2.3) Data Preparation and Adjustments

2.3.a) Initial Data Checking
Upon downloading, we reviewed the dataset for completeness and consistency. This step involved checking for any missing data or strange entries, ensuring that the data was clean and ready for analysis.

2.3.b) Changes in Data Handling
Originally, our data description deliverable outlined a straightforward use of the dataset. However, during the initial data preparation phase, we decided to adjust our approach slightly by incorporating more detailed checks for data integrity and handling missing

values more systematically. This adjustment was necessary to enhance the reliability of our subsequent analyses.

### 2.4) Reflection and Modifications

As we delved deeper into the dataset and its characteristics, our understanding of the data's complexity increased. This enhanced understanding led us to adopt more sophisticated data handling and analysis techniques than initially planned.The changes were primarily driven by the realization that the data's depth could provide more nuanced insights if analyzed with advanced statistical methods and machine learning models. These techniques require a highly rigorous approach to data preparation and analysis, prompting us to refine our methods to suit these sophisticated tools.

## **3.Project Process**

### 3.1) File Format and Missing Data

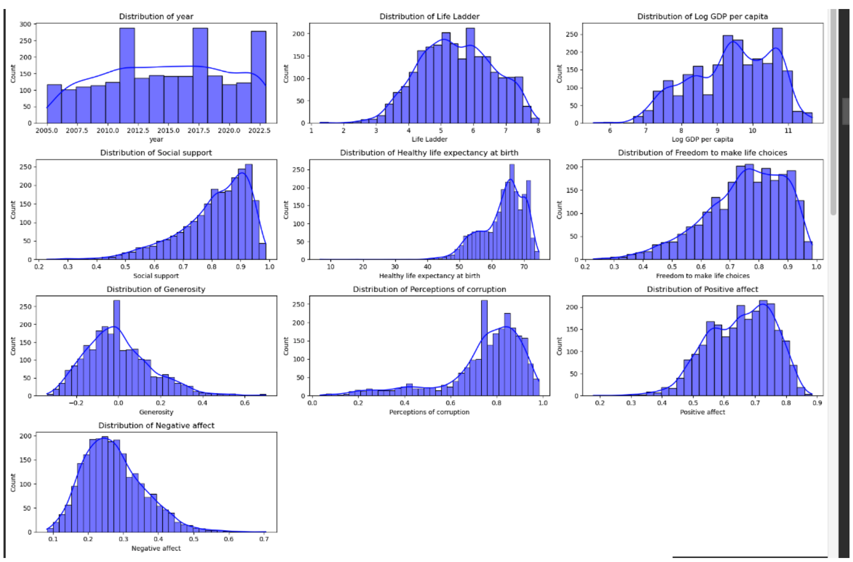
One challenge we encountered was the file format issue when uploading the dataset in Excel, which couldn't be read due to missing libraries and compatibility problems. To resolve this, we converted the file to CSV format, which works better with our Python environment. Another challenge was missing data in various columns like Log GDP per capita and Social support. We tackled this by applying mean imputation to fill in the missing values with the average of each column, ensuring our dataset is complete for analysis.

```python
df = pd.read_csv('/content/DataForTable2.1 (1).csv')

# Handling missing data by imputation with the mean for each column
df.fillna(df.mean(), inplace=True)

#  'year' to integer
df['year'] = df['year'].astype(int)
print(df.head())
```

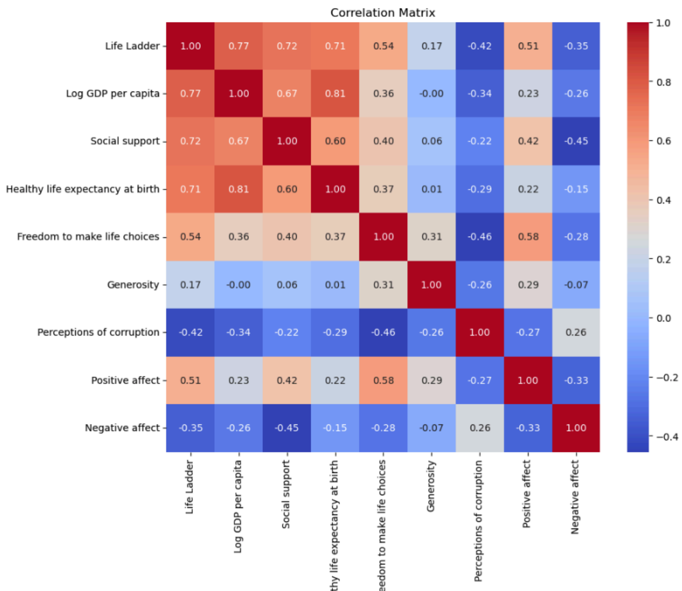
```
  Country name  year  Life Ladder  Log GDP per capita  Social support  \
0  Afghanistan  2008        3.724               7.350           0.451
1  Afghanistan  2009        4.402               7.509           0.552
2  Afghanistan  2010        4.758               7.614           0.539
3  Afghanistan  2011        3.832               7.581           0.521
4  Afghanistan  2012        3.783               7.661           0.521

   Healthy life expectancy at birth  Freedom to make life choices  Generosity  \
0                              50.5                         0.718       0.164
1                              50.8                         0.679       0.187
2                              51.1                         0.600       0.118
3                              51.4                         0.496       0.160
4                              51.7                         0.531       0.234

   Perceptions of corruption  Positive affect  Negative affect
0                      0.882            0.414            0.258
1                      0.850            0.481            0.237
2                      0.707            0.517            0.275
3                      0.731            0.480            0.267
4                      0.776            0.614            0.268
/var/folders/z5/36fz2y8j5dqdzn04zypyxy8h0000gn/T/ipykernel_57859/1399924553.py:5: FutureWarning: Dropping 
  df.fillna(df.mean(), inplace=True)
```

## 3.2) Data Visualization:

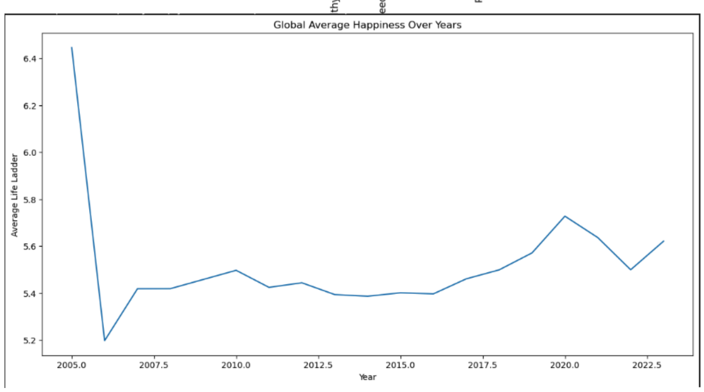
We aimed to visually explore the World Happiness Report dataset to comprehend the distributions, relationships, and potential correlations among variables. This phase was crucial for identifying patterns, outliers, and insights that informed subsequent analyses and model development. Our initial visualizations included histograms for each numerical attribute, serving to reveal distribution characteristics and outlier presence. Leveraging Python's Matplotlib and Seaborn libraries, we constructed these histograms. Additionally, we employed a scatter plot to examine the relationship between a country's GDP per capita and its happiness score, integrating factors like social support and healthy life expectancy. Further, we utilized Seaborn's heatmap function to depict the correlation strength among key numerical variables, aiding in understanding inter-variable relationships. To analyze global happiness trends over time, we employed Seaborn's line plot function to plot the average happiness score across years. Finally, we employed a cluster map to visualize variable groupings based on correlations, facilitated by Seaborn's clustermap function, providing hierarchical clustering for organizing variables and countries by similarity.



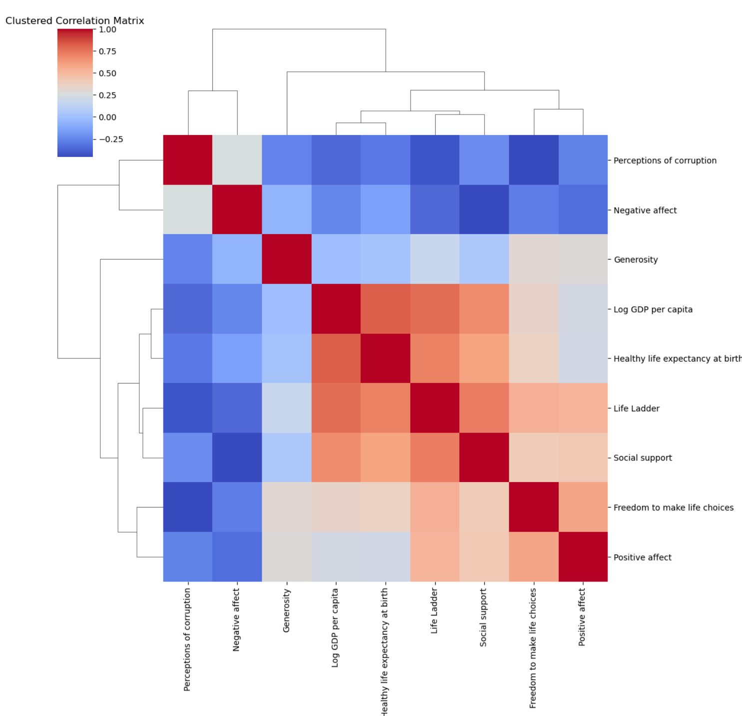

### Scatter Plot: Happiness Score vs. GDP per Capita
- **Insight**: Positive trend between GDP per capita and happiness, more pronounced with better social support and health.




### Correlation Matrix
- **Positive Correlations**: Life Ladder scores strongly correlate with GDP per capita, social support, and healthy life expectancy.
- **Negative Correlations**: Happiness and positive factors correlate negatively with perceptions of corruption.
- **Positive Affects and Freedom**: Significant positive correlation between positive affects and freedom.




### Time Series Plot: Global Average Happiness Over Years
- **Trend Insight**: Reveals trends over time in global happiness, reflecting economic, social, and political changes.




These visualizations provide a comprehensive understanding of happiness factors, informing further analysis and policymaking efforts.

### 3.3) Transforming Data :

In Step 3 we were focused on transforming and preparing the data for advanced analysis, aiming to optimize it for predictive modeling and deeper insights. We faced a few challenges, like we had to mitigate distortions caused by variables on different scales, we applied standard scaling using StandardScaler from sklearn.preprocessing. Additionally, we tackled the need for capturing complex interactions between variables by creating a new metric, the "Economic Impact Index." To deal with high dimensionality, we employed PCA and Factor Analysis for dimensionality reduction while retaining data variance and structure. Finally, to adapt the dataset for network-based analyses, particularly for exploring bipartite relationships, we developed a bipartite matrix representation, enhancing insights into country-year relationships. These solutions ensured that our dataset was well-prepared for advanced analysis and modeling.

Output :

```
Data Head with PCA Components:
     Country name    year  Life Ladder  Log GDP per capita  Social support  \
0     Afghanistan  2008.0        3.724           -1.790140       -2.965353
1     Afghanistan  2009.0        4.402           -1.651273       -2.129622
2     Afghanistan  2010.0        4.758           -1.559568       -2.237191
3     Afghanistan  2011.0        3.832           -1.588389       -2.386133
4     Afghanistan  2012.0        3.783           -1.518519       -2.386133
...           ...     ...          ...                 ...             ...
2358     Zimbabwe  2019.0        2.694           -1.486204       -0.416785
2359     Zimbabwe  2020.0        3.160           -1.575289       -0.764317
2360     Zimbabwe  2021.0        3.155           -1.522012       -1.029103
2361     Zimbabwe  2022.0        3.296           -1.510659       -1.186320
2362     Zimbabwe  2023.0        3.572           -1.502798       -0.954632

     Healthy life expectancy at birth  Freedom to make life choices  \
0                           -1.911568                     -0.233484
1                           -1.867119                     -0.515558
2                           -1.822670                     -1.086938
3                           -1.778221                     -1.839135
4                           -1.733773                     -1.585992
...                               ...                           ...
2358                        -1.526345                     -0.855493
2359                        -1.455968                     -0.775934
2360                        -1.385591                     -0.595117
2361                        -1.315213                     -0.710840
2362                        -1.244836                     -0.110529

     Generosity  Perceptions of corruption  Positive affect  Negative affect  \
0      1.033675                   0.767385        -2.251046        -0.174514
```

| | | | | |
|---|---|---|---|---|
| 1 | 1.178728 | 0.589478 | -1.617034 | -0.416402 |
| 2 | 0.743569 | -0.205543 | -1.276370 | 0.021300 |
| 3 | 1.008448 | -0.072113 | -1.626497 | -0.070848 |
| 4 | 1.475141 | 0.178068 | -0.358472 | -0.059330 |
| ... | ... | ... | ... | ... |
| 2358 | -0.322256 | 0.483846 | 0.057894 | -0.439439 |
| 2359 | 0.018304 | 0.250343 | 0.086282 | 0.839111 |
| 2360 | -0.498842 | 0.072436 | -0.396324 | -0.358810 |
| 2361 | -0.461002 | 0.050198 | -0.102975 | -0.946252 |
| 2362 | -0.435775 | 0.072436 | -0.396324 | -1.084474 |

| | Economic Impact Index |
|---|---|
| 0 | 3.421973 |
| 1 | 3.083122 |
| 2 | 2.842577 |
| 3 | 2.824508 |
| 4 | 2.632766 |
| ... | ... |
| 2358 | 2.268460 |
| 2359 | 2.293569 |
| 2360 | 2.108886 |
| 2361 | 1.986838 |
| 2362 | 1.870737 |

[2363 rows x 12 columns]

PCA Components:

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| 0 | 3.706228 | 1.698998 | 0.312060 | 1.073711 | 0.903008 |
| 1 | 3.015411 | 1.727014 | -0.105606 | 0.863598 | 1.130774 |
| 2 | 3.056076 | 1.543760 | 0.231256 | 1.305795 | 0.668188 |
| 3 | 3.506154 | 1.337524 | 0.263943 | 1.452374 | 1.331214 |
| 4 | 2.936568 | 2.042726 | 0.094913 | 0.601696 | 1.150837 |
| ... | ... | ... | ... | ... | ... |
| 2358 | 1.813197 | 0.816890 | -1.330552 | 0.180034 | -0.004025 |
| 2359 | 2.172140 | 1.066350 | -0.125752 | -0.177189 | -0.306897 |
| 2360 | 2.011988 | 0.816917 | -0.875142 | 0.763024 | -0.301501 |
| 2361 | 1.812840 | 0.975078 | -1.299758 | 0.920601 | -0.185741 |
| 2362 | 1.508948 | 0.990098 | -1.297277 | 0.955001 | -0.229787 |

[2363 rows x 5 columns]

Factor Analysis Components:

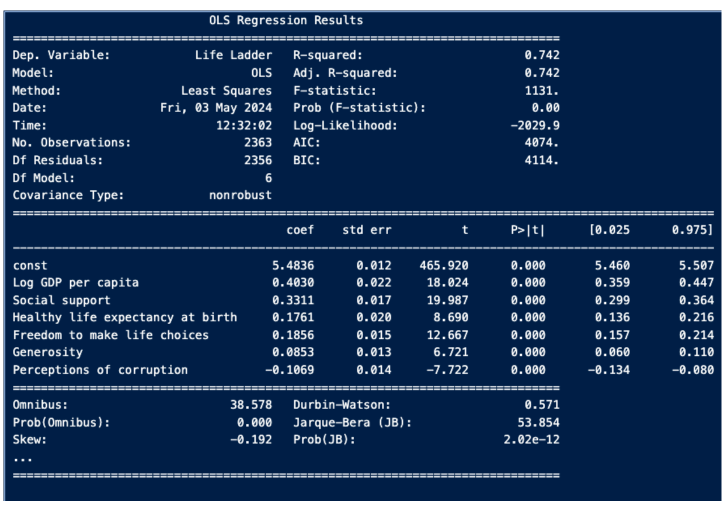
| | Factor1 | Factor2 | Factor3 | Factor4 | Factor5 |
|---|---|---|---|---|---|
| 0 | 2.166949 | -0.309601 | 1.039728 | 0.443278 | -0.068289 |
| 1 | 1.867604 | -0.456422 | 0.440683 | 0.580104 | 0.217986 |
| 2 | 1.852274 | -0.323958 | 0.799763 | 0.927294 | 0.249920 |
| 3 | 2.004880 | 0.077545 | 0.756166 | 1.175804 | 0.553113 |
| 4 | 1.860268 | -0.394208 | 0.580241 | 0.613662 | 0.667619 |

```
...        ...        ...        ...        ...        ...
2358  1.248088 -0.539804 -0.902411  0.410880 -0.007387
2359  1.378441 -0.502734 -0.171348 -0.077790  0.170809
2360  1.337815 -0.511718 -0.205109  0.565749 -0.264576
2361  1.310523 -0.621174 -0.311795  0.744310 -0.300596
2362  1.176537 -0.784435 -0.328139  0.661160 -0.447347

[2363 rows x 5 columns]

Cross Product Matrix:
 [[ 2.36300000e+03  1.59195981e+03  1.90886002e+03  8.44895480e+02
  -1.78784754e+00 -8.08543827e+02  5.37244647e+02 -6.03312110e+02]
 [ 1.59195981e+03  2.36300000e+03  1.40746566e+03  9.51558861e+02
   1.50575175e+02 -5.16825301e+02  1.00008268e+03 -1.07446777e+03]
 [ 1.90886002e+03  1.40746566e+03  2.36300000e+03  8.70648182e+02
   3.51577280e+01 -6.94182628e+02  5.08628298e+02 -3.44923810e+02]
 [ 8.44895480e+02  9.51558861e+02  8.70648182e+02  2.36300000e+03
   7.41170640e+02 -1.07522399e+03  1.36180723e+03 -6.52915919e+02]
 [-1.78784754e+00  1.50575175e+02  3.51577280e+01  7.41170640e+02
   2.36300000e+03 -6.14861595e+02  6.96427995e+02 -1.65170576e+02]
 [-8.08543827e+02 -5.16825301e+02 -6.94182628e+02 -1.07522399e+03
  -6.14861595e+02  2.36300000e+03 -6.36891931e+02  6.13670214e+02]
 [ 5.37244647e+02  1.00008268e+03  5.08628298e+02  1.36180723e+03
   6.96427995e+02 -6.36891931e+02  2.36300000e+03 -7.87569143e+02]
 [-6.03312110e+02 -1.07446777e+03 -3.44923810e+02 -6.52915919e+02
  -1.65170576e+02  6.13670214e+02 -7.87569143e+02  2.36300000e+03]]

Bipartite Matrix:
             2008.0  2009.0  2010.0  2011.0  2012.0  2013.0  2014.0  2015.0  \
Afghanistan       1       1       1       1       1       1       1       1
Albania           0       1       1       1       1       1       1       1
Algeria           0       0       1       1       1       0       1       0
Angola            0       0       0       1       1       1       1       0
Argentina         1       1       1       1       1       1       1       1
...             ...     ...     ...     ...     ...     ...     ...     ...
Venezuela         1       1       1       1       1       1       1       1
Vietnam           1       1       1       1       1       1       1       1
Yemen             0       1       1       1       1       1       1       1
Zambia            1       1       0       1       1       1       1       1
Zimbabwe          1       1       1       1       1       1       1       1

             2016.0  2017.0  2018.0  2019.0  2021.0  2022.0  2023.0  2007.0  \
Afghanistan       1       1       1       1       1       1       1       0
Albania           1       1       1       1       1       1       1       1
Algeria           1       1       1       1       1       1       0       0
Angola            0       0       0       0       0       0       0       0
Argentina         1       1       1       1       1       1       1       1
...             ...     ...     ...     ...     ...     ...     ...     ...
Venezuela         1       1       1       1       1       1       1       0
```

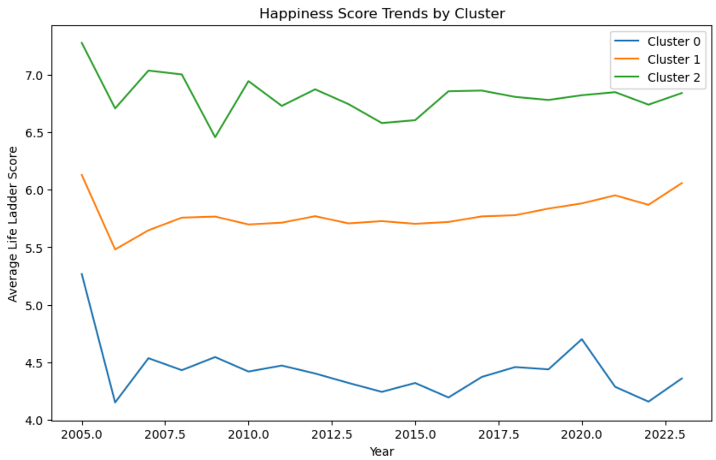
```
Vietnam      1    1    1    1    1    1    1    1
Yemen        1    1    1    1    0    1    1    1
Zambia       1    1    1    1    1    1    1    1
Zimbabwe     1    1    1    1    1    1    1    1

            2020.0  2006.0  2005.0
Afghanistan      0       0       0
Albania          1       0       0
Algeria          1       0       0
Angola           0       0       0
Argentina        1       1       0
...            ...     ...     ...
Venezuela        1       1       1
Vietnam          1       1       0
Yemen            0       0       0
Zambia           1       1       0
Zimbabwe         1       1       0

[165 rows x 19 columns]
```

## 3.4) Quantitative Analysis :

In Step 4 of the analysis using the World Happiness Report dataset, various advanced data analysis techniques were applied to delve deeper into the data. An Ordinary Least Squares (OLS) regression model was fitted using principal components, enabling the understanding of factors impacting happiness scores. K-Means clustering segmented countries into distinct groups based on their happiness-related characteristics, aiding in identifying similar country clusters. Visualization of these clusters, along with trends in happiness scores over years, provided insights into countries' happiness evolution. Additionally, a two-dimensional PCA facilitated visual representation of clusters. The analysis also involved identifying countries within each cluster, offering a clear understanding of their composition. This step was pivotal for extracting profound insights, blending statistical modeling with clustering to unveil patterns among countries and enhance comprehension of the dataset's nuances.



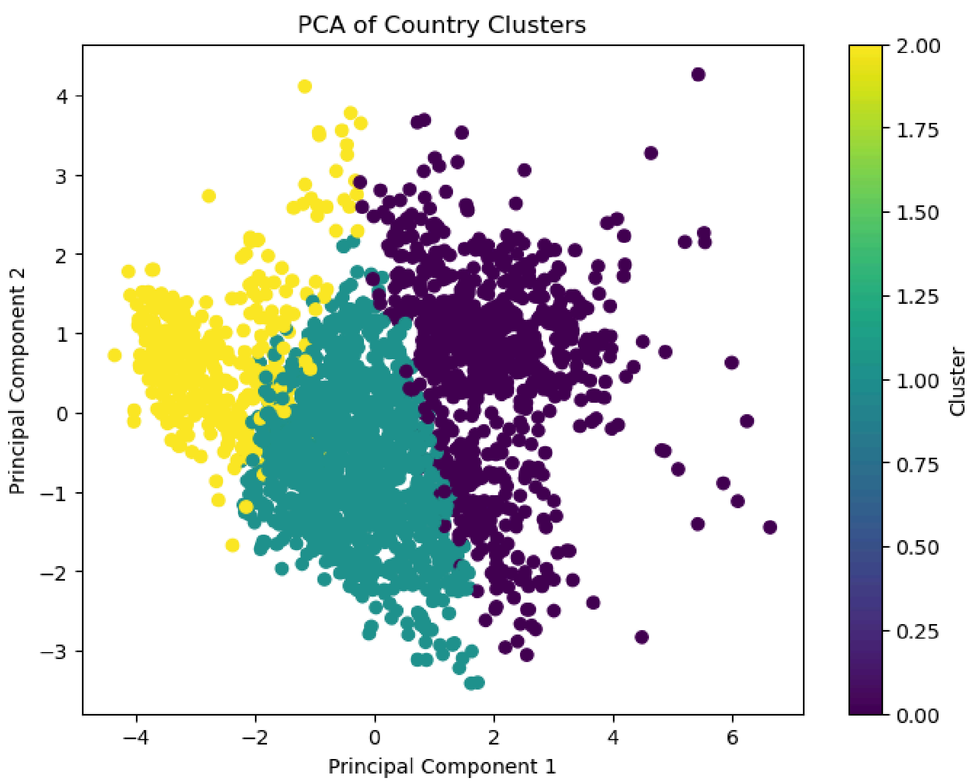
```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Life Ladder   R-squared:                       0.742
Model:                            OLS   Adj. R-squared:                  0.742
Method:                 Least Squares   F-statistic:                     1131.
Date:                Fri, 03 May 2024   Prob (F-statistic):               0.00
Time:                        12:32:02   Log-Likelihood:                -2029.9
No. Observations:                2363   AIC:                             4074.
Df Residuals:                    2356   BIC:                             4114.
Df Model:                           6
Covariance Type:            nonrobust
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                                5.4836      0.012    465.920      0.000       5.460       5.507
Log GDP per capita                   0.4030      0.022     18.024      0.000       0.359       0.447
Social support                       0.3311      0.017     19.987      0.000       0.299       0.364
Healthy life expectancy at birth     0.1761      0.020      8.690      0.000       0.136       0.216
Freedom to make life choices         0.1856      0.015     12.667      0.000       0.157       0.214
Generosity                           0.0853      0.013      6.721      0.000       0.060       0.110
Perceptions of corruption           -0.1069      0.014     -7.722      0.000      -0.134      -0.080
==============================================================================
Omnibus:                       38.578   Durbin-Watson:                   0.571
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               53.854
Skew:                          -0.192   Prob(JB):                     2.02e-12
...
==============================================================================
```



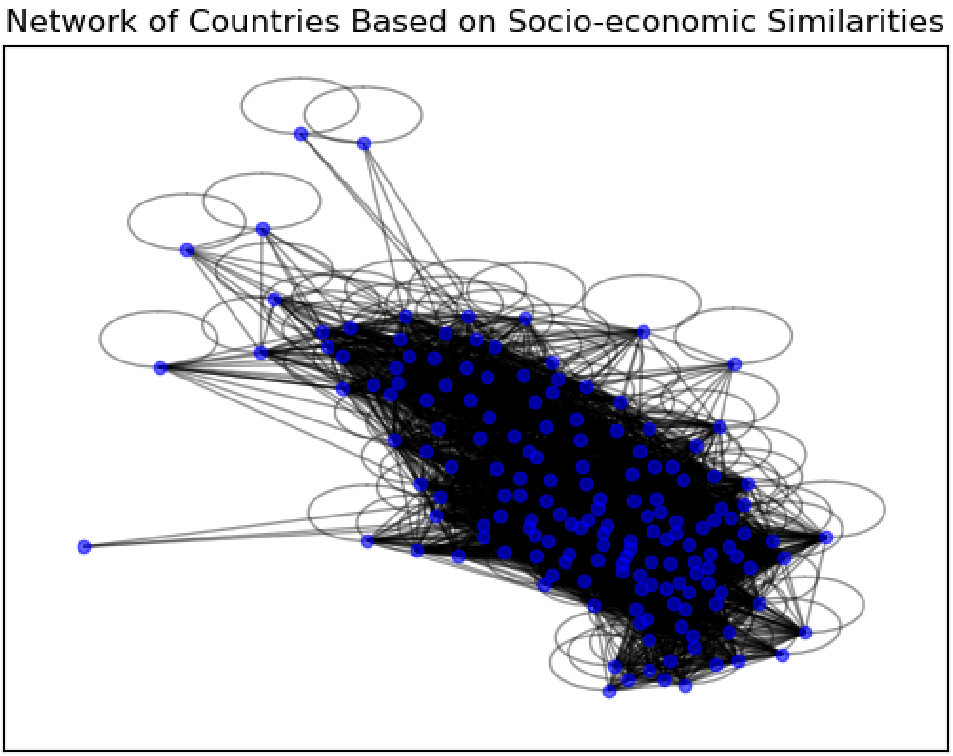

```
Countries in each cluster:
Cluster 0: ['Afghanistan' 'Angola' 'Bangladesh' 'Benin' 'Bolivia' 'Botswana'
 'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon' 'Central African Republic'
 'Chad' 'Comoros' 'Congo (Brazzaville)' 'Congo (Kinshasa)' 'Djibouti'
 'Eswatini' 'Ethiopia' 'Gabon' 'Gambia' 'Ghana' 'Guatemala' 'Guinea'
 'Guyana' 'Haiti' 'Honduras' 'India' 'Indonesia' 'Iran' 'Iraq'
 'Ivory Coast' 'Kenya' 'Kosovo' 'Laos' 'Lesotho' 'Liberia' 'Madagascar'
 'Malawi' 'Mali' 'Mauritania' 'Morocco' 'Mozambique' 'Myanmar' 'Namibia'
 'Nepal' 'Niger' 'Nigeria' 'Pakistan' 'Philippines' 'Rwanda' 'Senegal'
 'Sierra Leone' 'Somalia' 'South Africa' 'South Sudan'
 'State of Palestine' 'Sudan' 'Syria' 'Tajikistan' 'Tanzania' 'Togo'
 'Uganda' 'Venezuela' 'Yemen' 'Zambia' 'Zimbabwe']
```

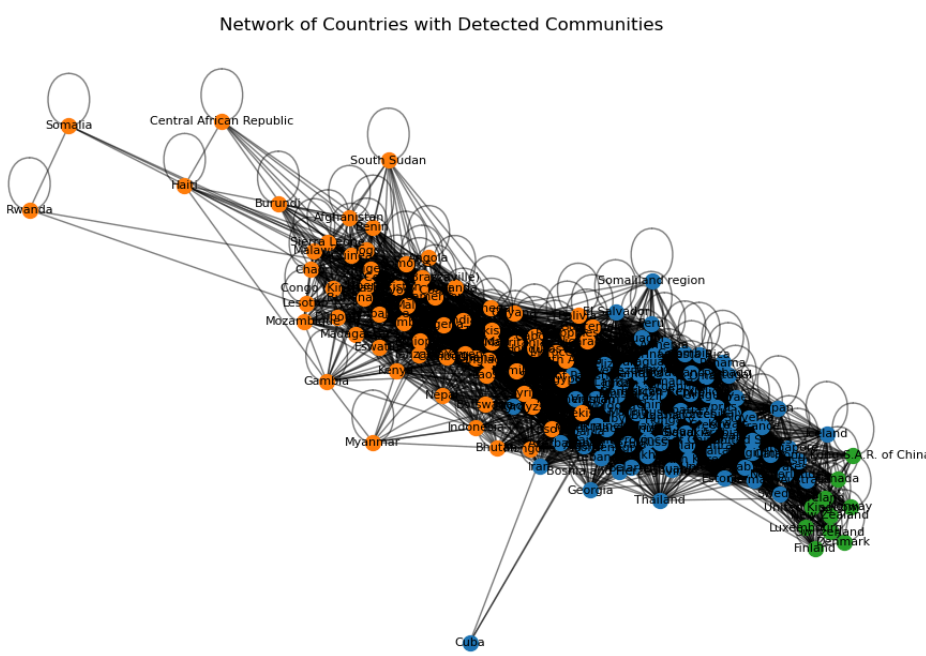
```
Cluster 1: ['Argentina' 'Australia' 'Austria' 'Azerbaijan' 'Bahrain' 'Belgium'
 'Belize' 'Bhutan' 'Bolivia' 'Bosnia and Herzegovina' 'Brazil' 'Canada'
 'Chile' 'China' 'Colombia' 'Costa Rica' 'Cyprus' 'Czechia' 'Denmark'
 'Dominican Republic' 'Ecuador' 'El Salvador' 'Estonia' 'Finland' 'France'
 'Germany' 'Guatemala' 'Honduras' 'Hong Kong S.A.R. of China' 'Hungary'
 'Iceland' 'Indonesia' 'Ireland' 'Israel' 'Italy' 'Jamaica' 'Japan'
 'Jordan' 'Kazakhstan' 'Kosovo' 'Kuwait' 'Kyrgyzstan' 'Laos' 'Latvia'
 'Luxembourg' 'Malaysia' 'Maldives' 'Malta' 'Mauritius' 'Mexico' 'Myanmar'
 'Netherlands' 'New Zealand' 'Nicaragua' 'Norway' 'Oman' 'Panama'
 'Paraguay' 'Philippines' 'Poland' 'Qatar' 'Russia' 'Rwanda'
 'Saudi Arabia' 'Serbia' 'Singapore' 'Slovakia' 'Slovenia'
 'Somaliland region' 'South Africa' 'Spain' 'Sri Lanka' 'Suriname'
 'Sweden' 'Switzerland' 'Taiwan Province of China' 'Tajikistan' 'Thailand'
 'Trinidad and Tobago' 'Turkmenistan' 'United Arab Emirates'
 'United Kingdom' 'United States' 'Uruguay' 'Uzbekistan' 'Venezuela'
 'Vietnam']
```

```
Cluster 2: ['Albania' 'Algeria' 'Argentina' 'Armenia' 'Azerbaijan' 'Bahrain'
 'Bangladesh' 'Belarus' 'Belgium' 'Belize' 'Bolivia'
 'Bosnia and Herzegovina' 'Botswana' 'Brazil' 'Bulgaria' 'Chile' 'China'
 'Colombia' 'Croatia' 'Cuba' 'Cyprus' 'Czechia' 'Dominican Republic'
 'Ecuador' 'Egypt' 'El Salvador' 'Estonia' 'Gabon' 'Georgia' 'Greece'
 'Honduras' 'Hong Kong S.A.R. of China' 'Hungary' 'Iran' 'Iraq' 'Israel'
 'Italy' 'Jamaica' 'Japan' 'Jordan' 'Kazakhstan' 'Kosovo' 'Kuwait'
 'Kyrgyzstan' 'Latvia' 'Lebanon' 'Libya' 'Lithuania' 'Malta' 'Mauritania'
 'Mexico' 'Moldova' 'Mongolia' 'Montenegro' 'Morocco' 'Namibia'
 'Nicaragua' 'North Macedonia' 'Paraguay' 'Peru' 'Poland' 'Portugal'
 'Qatar' 'Romania' 'Russia' 'Saudi Arabia' 'Serbia' 'Slovakia' 'Slovenia'
 'South Africa' 'South Korea' 'Spain' 'Sri Lanka' 'State of Palestine'
 'Sudan' 'Syria' 'Taiwan Province of China' 'Tajikistan' 'Tunisia'
 'Turkmenistan' 'Türkiye' 'Ukraine' 'United Arab Emirates' 'Uzbekistan'
 'Venezuela' 'Vietnam' 'Yemen']
```




The process involved constructing a network graph from country data, focusing on socio-economic features to analyze relationships and community structures based on similarity. Initially, a graph `G` was initialized using `networkx`, where nodes represented countries and node attributes captured country-specific features like GDP and Social Support.

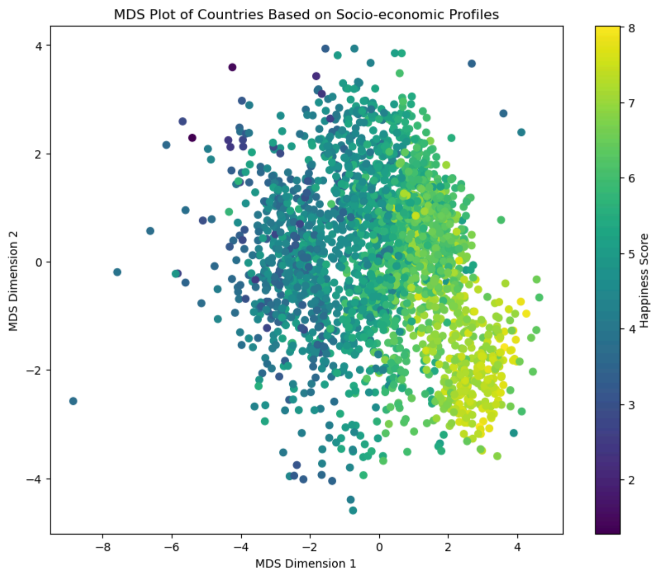



Challenges included ensuring data integrity and handling complex data structures, addressed through thorough pre-processing using `pandas`. Nodes were then added to the graph for each country, incorporating socio-economic features as node attributes. Edges were added based on Euclidean distances between feature vectors, with efficient distance calculations using `numpy` and threshold adjustments to balance connectivity and clarity. Centrality measures were computed to identify influential nodes, while community detection using the greedy modularity method grouped densely interconnected countries. Visualization was achieved using `networkx` drawing functions, optimizing clarity and aesthetics with force-directed layouts and adjustments to node sizes and transparency. This process facilitated a deeper understanding of complex international socio-economic relationships and structures, supported by the NetworkX, NumPy, and Pandas documentation for implementation guidance.

## 3.5) Advance Data Analysis Techniques:

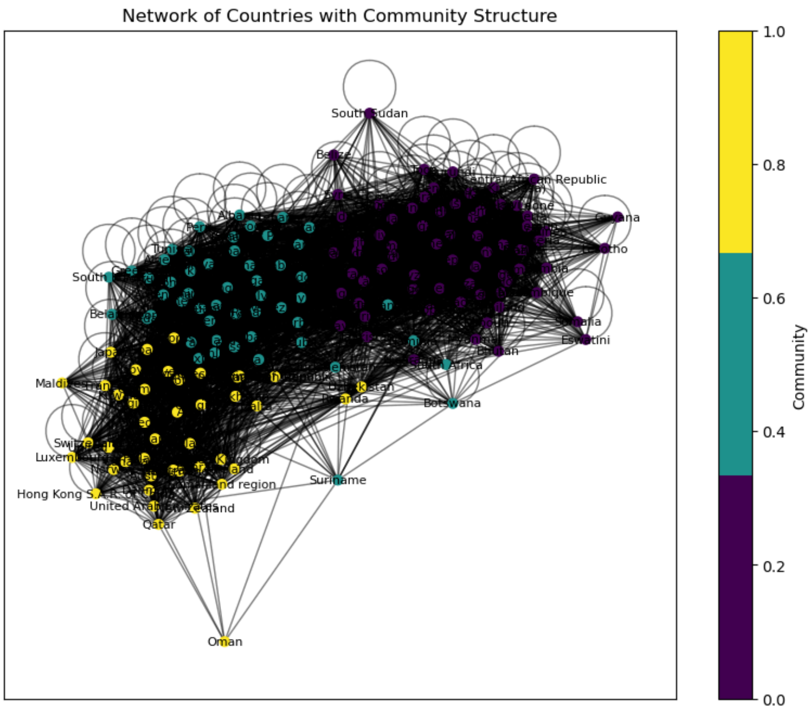
Within the step 5 of the code, a series of advanced data analysis techniques were implemented to understand the relationships and importance of various factors influencing happiness across different countries. First, a Random Forest Regressor was employed to handle nonlinear relationships and interactions between variables, providing insights into feature importance for predicting happiness. Following model fitting, feature importance ranking was conducted to identify the most influential variables. Multidimensional Scaling (MDS) was then applied to visualize country clusters based on socio-economic profiles and happiness scores. Additionally, Singular Value Decomposition (SVD) was utilized to decompose the dataset, revealing underlying patterns in a reduced two-dimensional space. These steps collectively facilitated a comprehensive understanding of the data structure, aiding in predictive modeling and strategic decision-making for improving national happiness.




Feature importances ranked:
Social support: 0.6154
Freedom to make life choices: 0.1335
Healthy life expectancy at birth: 0.1047
Generosity: 0.0684
Positive affect: 0.0401
Perceptions of corruption: 0.0378
Log GDP per capita: 0.0000
Top 5 Countries by Degree Centrality:
Moldova 0.6585365853658537
Sri Lanka 0.6585365853658537
Vietnam 0.6585365853658537
Brazil 0.6463414634146342

## 3.6) Block Modeling and Network Analysis:

In Step 6 of the project,Using the `networkx` library, a network graph was constructed wherein nodes represented countries and edges depicted strong correlations between them. Challenges encountered included determining an optimal correlation threshold to define edges, ultimately set at 0.8 to retain only significant relationships. Additionally, computational efficiency was managed through the strategic use of data structures and algorithms from the `networkx` library, focusing on relevant features to mitigate computational load. The analysis delved further into community detection using the Louvain method, unveiling clusters of countries sharing similar happiness profiles, and identified a core-periphery structure through the Girvan-Newman algorithm, distinguishing closely connected countries from more isolated ones.




```
Top central countries based
on connectivity:
Serbia: 0.6890
Azerbaijan: 0.6768
Sri Lanka: 0.6402
Ukraine: 0.6402
Malaysia: 0.6280
```

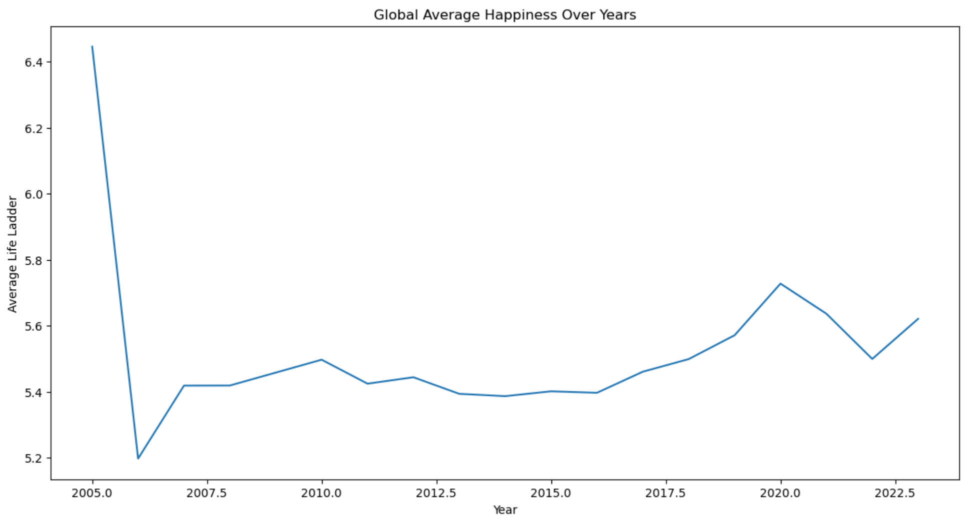
The interpretation of community structures and influential countries within the network was facilitated by visualizations such as network plots and centrality measures, aiding in making actionable insights. The visualization of the network's community structure using the Louvain method provided a clear depiction of countries grouped based on shared happiness determinants. Throughout this phase, external resources including the `networkx` documentation for guidance on construction and visualization techniques, along with references from the Scikit-learn user guide for standardization and clustering algorithms, and documentation from NumPy and Pandas for data manipulation and correlation calculations, were instrumental. This comprehensive analysis not only demonstrated the efficacy of network analysis in uncovering complex dataset relationships but also underscored the significance of preprocessing and algorithm selection in deriving meaningful insights, effectively communicated through visualizations to stakeholders.

## 3.7) Bipartite Graph:

In Step 7 of our project, we focused on constructing and analyzing a bipartite graph using network analysis techniques. We categorized nodes into two distinct sets representing countries and years, respectively, based on the dataset's structure. Edges were added between countries and years where data was available, illustrating occurrences or events. We utilized `networkx` for graph construction and `matplotlib` for visualization, adjusting node size, color, and edge transparency to enhance clarity. Challenges included addressing data sparsity and inconsistency, managed through thorough data cleaning and preprocessing, and optimizing visualization clarity and computational efficiency. The resulting bipartite graph provided insights into the temporal distribution of data, facilitating the identification of patterns or trends over time in global happiness indices. This step was crucial for analyzing the dataset's structural properties, highlighting the temporal aspect of data collection, and potentially uncovering meaningful temporal trends in happiness metrics. Resources such as the NetworkX and Matplotlib documentation were instrumental in guiding graph construction, visualization, and customization techniques.

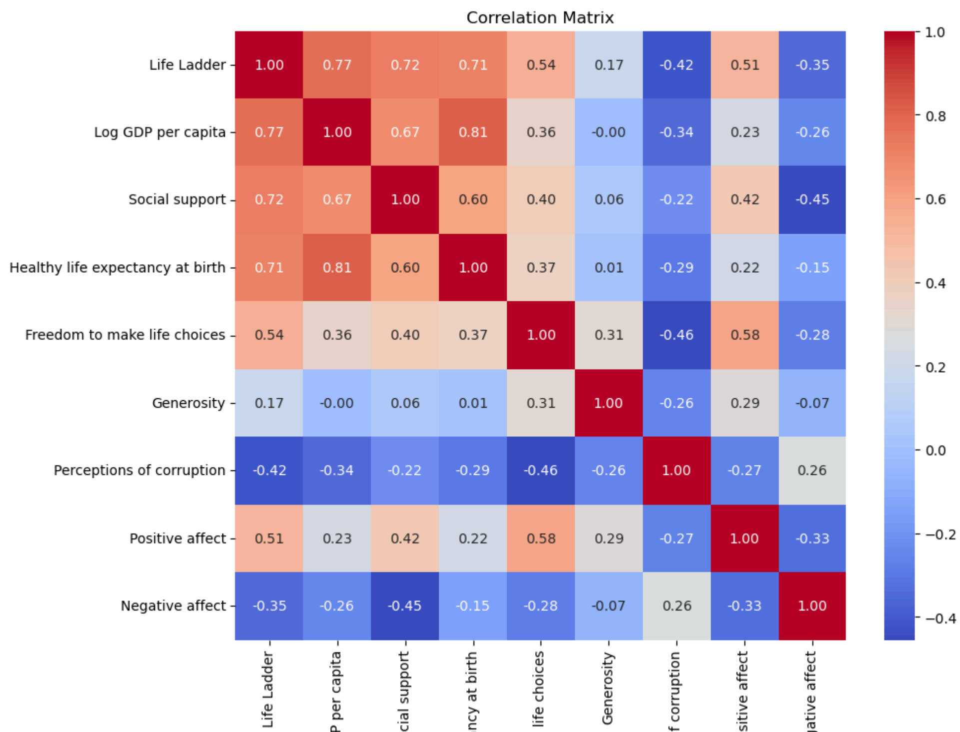
# **4.Project Results**

## 4.1) Global Happiness Report



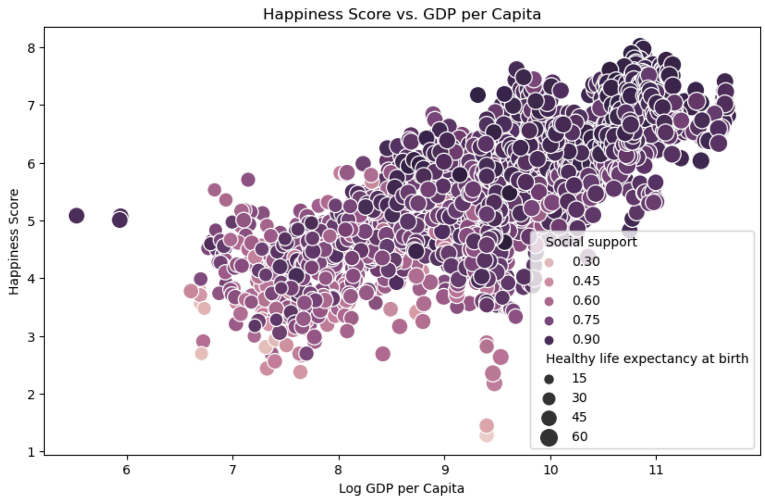

Here We get to know how the happiness meter has changed over the years peaking at 2005 and seeing a dip in between 2005 - 2007 as the second peak would be at 2020 right before covid. There could also be a possibility of covid impacting happiness indicating from the sudden decline in happiness meter after 2020 till mid 2021.

## 4.2) Correlation Matrix:



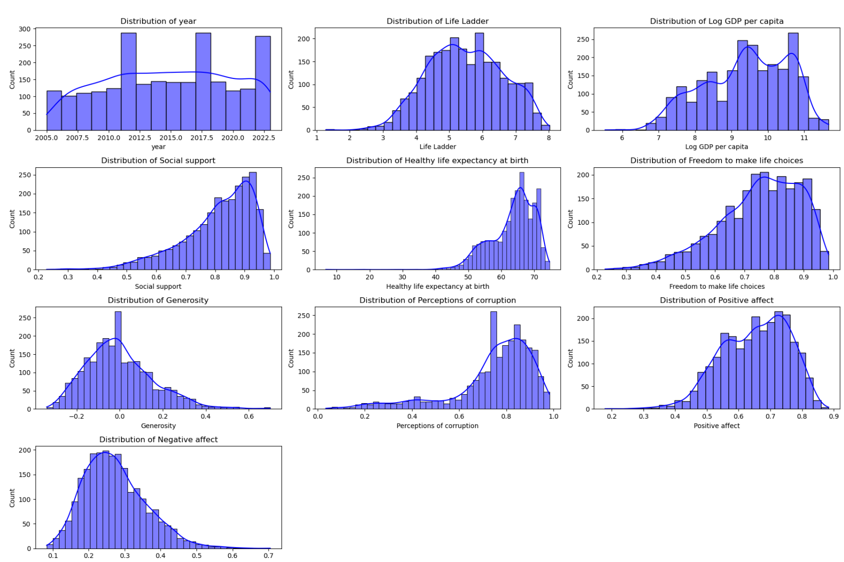

From the correlation matrix we can say that there is a high correlation between happiness indicator (Life Ladder) and Log GDP suggesting that it is the highest contributor towards happiness.

Happiness Score vs GDP per Capita:




The scatter plot visualizes the relationship between GDP per capita (logged) and happiness scores, with additional insights provided by social support (represented as hue) and healthy life expectancy at birth (indicated by point size). It demonstrates a positive correlation where higher GDP per capita is associated with higher happiness scores, suggesting that wealthier countries often have happier populations. The hue and size dimensions further enrich the analysis by showing that strong social support networks and longer healthy life expectancies are more prevalent in these wealthier countries. This visualization thus underscores the importance of economic, social, and health factors in influencing national happiness levels and offers nuanced insights for policymakers focusing on holistic development rather than solely economic growth.

## 4.3) Histograms:




Distribution of Year: This histogram shows the count of data entries over different years. The non-uniform distribution suggests varying numbers of data points collected in different years, with peaks around 2010, 2015, and 2020. This might indicate either increased focus on data collection during these periods or a larger availability of data.

Distribution of Life Ladder : The "Life Ladder" represents the happiness score on a scale from 1 to 8. The distribution is roughly bell-shaped, centered around 5, indicating a normal distribution where most countries report moderate levels of happiness.
Distribution of Log GDP per Capita: This histogram reveals a right-skewed distribution, suggesting that while most countries have lower GDP per capita, there is a tail of countries with significantly higher GDP.

Distribution of Social Support: Social support is shown to be left-skewed, with most countries reporting high levels of social support, indicating strong social networks or family support structures in these regions.

Distribution of Healthy Life Expectancy at Birth:The distribution of healthy life expectancy displays a right-skew. Most countries have a lower expectancy, with fewer countries achieving higher values, reflecting disparities in global health outcomes.

Distribution of Freedom to Make Life Choices:This metric also shows a left-skewed distribution with most countries having high freedom scores, suggesting that many countries provide substantial freedoms in life choices to their citizens.

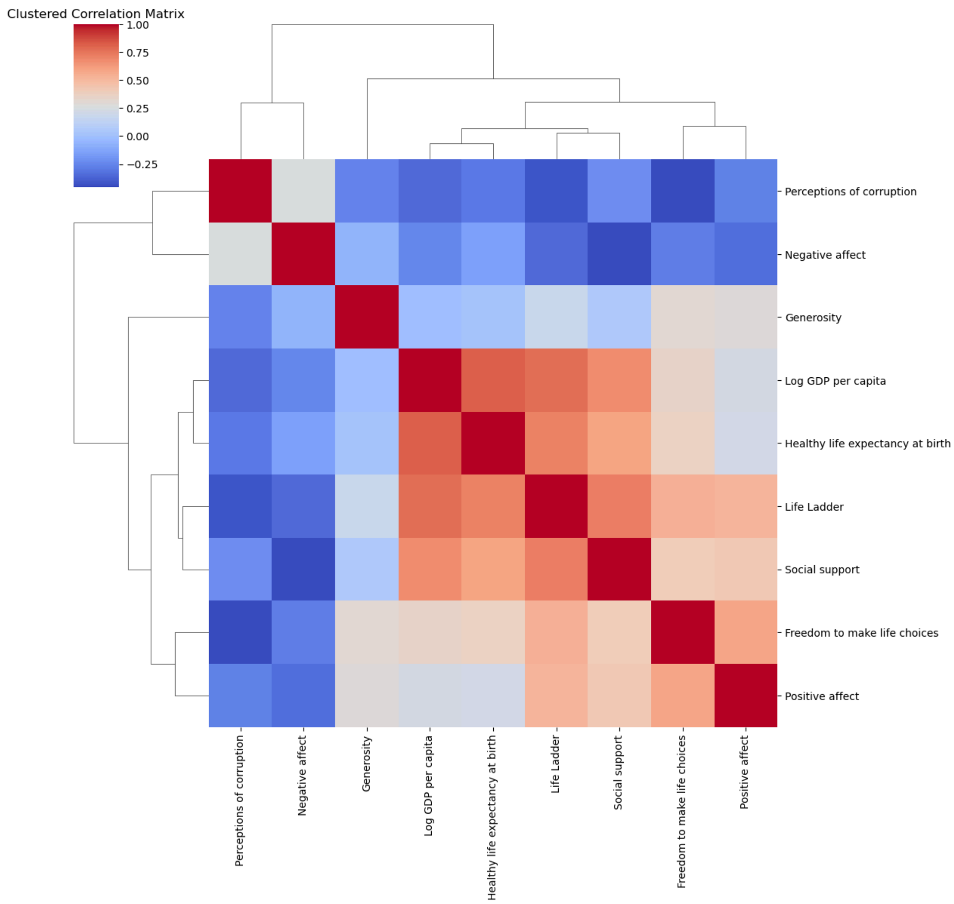
Distribution of Generosity:Generosity scores are centered around a central value with a normal distribution. This suggests a varied but balanced level of generosity across surveyed nations.

Distribution of Perceptions of Corruption:The distribution of perceptions of corruption is somewhat bimodal, indicating two groups of countries: one where corruption is perceived as low and another where it is perceived as high, with fewer countries in the moderate perception range.

Distribution of Positive Affect:Positive affect is moderately skewed to the right, implying that while many countries report higher positive emotions, there's still a significant number experiencing lower positive effects.

Distribution of Negative Affect:Negative affect has a normal distribution, indicating a balanced view of negative emotions across the globe, with most countries having moderate levels of negative emotions.

### 4.4) Clustered Correlation Matrix




The clustered correlation matrix provides a detailed visualization of the relationships between socio-economic and well-being variables. Analyzing the matrix reveals several key clusters and insights:

Economic and Health Well-being Cluster:A strong positive correlation exists between Life Ladder (happiness index), Log GDP per capita, Healthy life expectancy at birth, and Social support. This cluster indicates that higher GDP per capita is associated with longer healthy life expectancy, stronger social support, and increased happiness.

Freedom and Emotional Well-being Cluster:Freedom to make life choices and Positive affect are closely correlated, suggesting that greater personal freedoms are linked to higher positive emotional experiences.

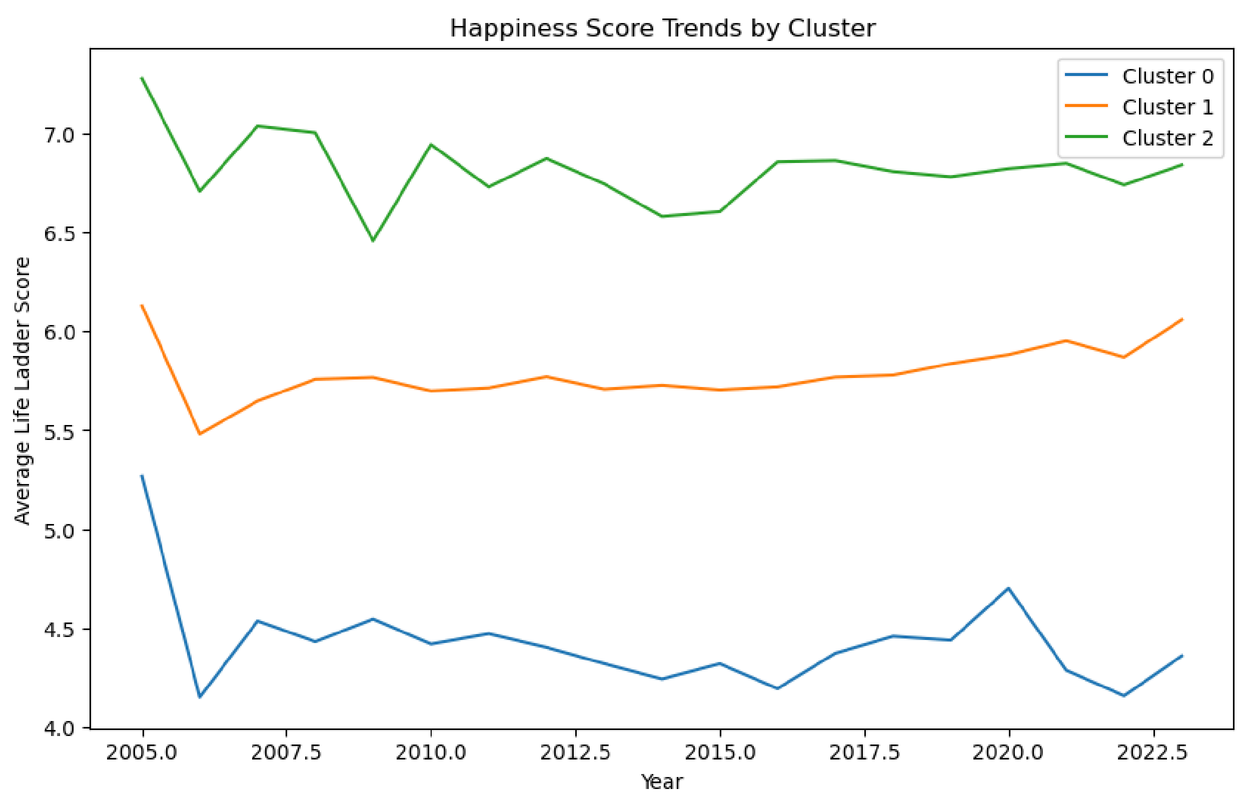
Impact of Governance:Perceptions of corruption show strong negative correlations with Life Ladder, Log GDP per capita, and Social support, highlighting the detrimental impact of corruption on economic prosperity and subjective well-being.

Isolated Impact of Generosity:Generosity exhibits weaker correlations with well-being metrics, though it mildly correlates with Positive affect, indicating its limited but positive influence on emotional well-being.

Dendrogram Insights: Hierarchical clustering groups similar variables, underscoring synergistic areas for policy interventions. For instance, improving economic conditions and healthcare simultaneously can have compounded positive effects on national happiness levels.

This analysis underscores the multifaceted nature of societal well-being and highlights the importance of holistic policy approaches that consider economic, health, and governance factors in enhancing life quality and happiness.

## 4.5) Happiness Scores by Clusters :




```
Countries in each cluster:
Cluster 0: ['Afghanistan' 'Angola' 'Bangladesh' 'Benin' 'Bolivia' 'Botswana'
 'Burkina Faso' 'Burundi' 'Cambodia' 'Cameroon' 'Central African Republic'
 'Chad' 'Comoros' 'Congo (Brazzaville)' 'Congo (Kinshasa)' 'Djibouti'
 'Eswatini' 'Ethiopia' 'Gabon' 'Gambia' 'Ghana' 'Guatemala' 'Guinea'
 'Guyana' 'Haiti' 'Honduras' 'India' 'Indonesia' 'Iran' 'Iraq'
 'Ivory Coast' 'Kenya' 'Kosovo' 'Laos' 'Lesotho' 'Liberia' 'Madagascar'
 'Malawi' 'Mali' 'Mauritania' 'Morocco' 'Mozambique' 'Myanmar' 'Namibia'
 'Nepal' 'Niger' 'Nigeria' 'Pakistan' 'Philippines' 'Rwanda' 'Senegal'
 'Sierra Leone' 'Somalia' 'South Africa' 'South Sudan'
 'State of Palestine' 'Sudan' 'Syria' 'Tajikistan' 'Tanzania' 'Togo'
 'Uganda' 'Venezuela' 'Yemen' 'Zambia' 'Zimbabwe']
```

```
Cluster 1: ['Argentina' 'Australia' 'Austria' 'Azerbaijan' 'Bahrain' 'Belgium'
 'Belize' 'Bhutan' 'Bolivia' 'Bosnia and Herzegovina' 'Brazil' 'Canada'
 'Chile' 'China' 'Colombia' 'Costa Rica' 'Cyprus' 'Czechia' 'Denmark'
 'Dominican Republic' 'Ecuador' 'El Salvador' 'Estonia' 'Finland' 'France'
 'Germany' 'Guatemala' 'Honduras' 'Hong Kong S.A.R. of China' 'Hungary'
 'Iceland' 'Indonesia' 'Ireland' 'Israel' 'Italy' 'Jamaica' 'Japan'
 'Jordan' 'Kazakhstan' 'Kosovo' 'Kuwait' 'Kyrgyzstan' 'Laos' 'Latvia'
 'Luxembourg' 'Malaysia' 'Maldives' 'Malta' 'Mauritius' 'Mexico' 'Myanmar'
 'Netherlands' 'New Zealand' 'Nicaragua' 'Norway' 'Oman' 'Panama'
 'Paraguay' 'Philippines' 'Poland' 'Qatar' 'Russia' 'Rwanda'
 'Saudi Arabia' 'Serbia' 'Singapore' 'Slovakia' 'Slovenia'
 'Somaliland region' 'South Africa' 'Spain' 'Sri Lanka' 'Suriname'
 'Sweden' 'Switzerland' 'Taiwan Province of China' 'Tajikistan' 'Thailand'
 'Trinidad and Tobago' 'Turkmenistan' 'United Arab Emirates'
 'United Kingdom' 'United States' 'Uruguay' 'Uzbekistan' 'Venezuela'
 'Vietnam']
```

The time-series plot depicting "Happiness Score Trends by Cluster" illustrates how average happiness scores (measured by the Life Ladder) have evolved over the years from 2005 to approximately 2022 for three distinct clusters of countries or regions. Here's a detailed analysis of each cluster's trend:

```
Cluster 2: ['Albania' 'Algeria' 'Argentina' 'Armenia' 'Azerbaijan' 'Bahrain'
 'Bangladesh' 'Belarus' 'Belgium' 'Belize' 'Bolivia'
 'Bosnia and Herzegovina' 'Botswana' 'Brazil' 'Bulgaria' 'Chile' 'China'
 'Colombia' 'Croatia' 'Cuba' 'Cyprus' 'Czechia' 'Dominican Republic'
 'Ecuador' 'Egypt' 'El Salvador' 'Estonia' 'Gabon' 'Georgia' 'Greece'
 'Honduras' 'Hong Kong S.A.R. of China' 'Hungary' 'Iran' 'Iraq' 'Israel'
 'Italy' 'Jamaica' 'Japan' 'Jordan' 'Kazakhstan' 'Kosovo' 'Kuwait'
 'Kyrgyzstan' 'Latvia' 'Lebanon' 'Libya' 'Lithuania' 'Malta' 'Mauritania'
 'Mexico' 'Moldova' 'Mongolia' 'Montenegro' 'Morocco' 'Namibia'
 'Nicaragua' 'North Macedonia' 'Paraguay' 'Peru' 'Poland' 'Portugal'
 'Qatar' 'Romania' 'Russia' 'Saudi Arabia' 'Serbia' 'Slovakia' 'Slovenia'
 'South Africa' 'South Korea' 'Spain' 'Sri Lanka' 'State of Palestine'
 'Sudan' 'Syria' 'Taiwan Province of China' 'Tajikistan' 'Tunisia'
 'Turkmenistan' 'Türkiye' 'Ukraine' 'United Arab Emirates' 'Uzbekistan'
 'Venezuela' 'Vietnam' 'Yemen']
```

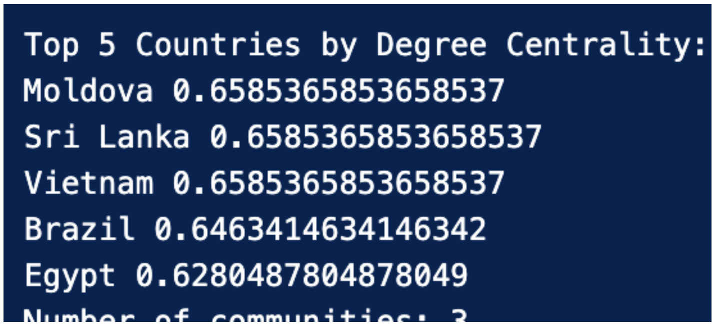
Insights:
- The distinct patterns across clusters indicate varying underlying conditions that influence happiness. Cluster 0's recovery could be driven by economic recovery or successful policy interventions. In contrast, Cluster 1's steady rise suggests incremental improvements over a long period.
- Cluster 2's consistency at a high level of happiness underscores the importance of sustained quality in life factors. This could serve as a benchmark or model for other clusters aiming to improve their happiness scores.
- The variability and trends observed can inform policymakers about the effectiveness of their strategies over time and guide future decisions to foster greater well-being in their respective populations.

Overall, this analysis helps understand the temporal dynamics of happiness across different global segments, providing a macroscopic view of how well-being evolves in response to various factors and interventions.

## 4.6) Random Forest Model and Feature Importance for happiness:

1. Social Support (0.6154) Social support is the most influential factor, with an importance score of 0.6154. This indicates that the perception of support from friends, family, or

community has the strongest relationship with the target variable (possibly happiness or life satisfaction). This high value suggests that social connectivity and support systems are critical to the perceived quality of life.



```
Top 5 Countries by Degree Centrality:
Moldova 0.6585365853658537
Sri Lanka 0.6585365853658537
Vietnam 0.6585365853658537
Brazil 0.6463414634146342
Egypt 0.6280487804878049
```

2. Freedom to Make Life Choices (0.1335) This factor has the second-highest importance, suggesting that the autonomy to make personal life decisions significantly impacts happiness. It demonstrates that freedom and personal agency are valued aspects of well-being.

3. Healthy Life Expectancy at Birth (0.1047) This feature reflects the expected number of years of healthy living from birth and has a notable influence on the target variable. It implies that health expectancy is a key component of overall life satisfaction.

4. Generosity (0.0684) Generosity's importance suggests that altruistic behavior, measured by how much people report giving to others, has a moderate impact on how people rate their happiness or satisfaction with life.

5. Positive Affect (0.0401) This factor represents the prevalence of positive emotions (like joy and pride) and has a smaller, yet significant, contribution to happiness. This underscores that frequent positive experiences contribute to overall life satisfaction, though less so than factors like social support or personal freedom.

6. Perceptions of Corruption (0.0378) The influence of perceptions of corruption on happiness is relatively low compared to social support and freedom but still noteworthy. It suggests that how people perceive the integrity of businesses and government can slightly impact their overall happiness.

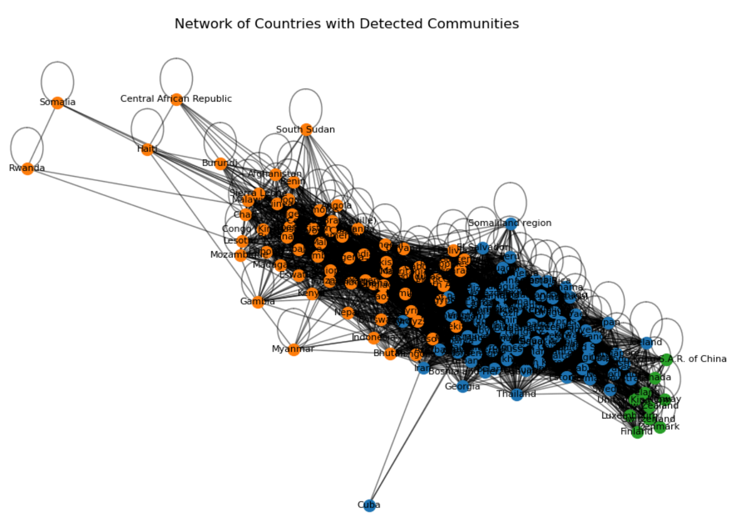
7. Log GDP per Capita (0.0000) Surprisingly, GDP per capita has no discernible effect on happiness in this model, which is represented by a feature importance score of 0.0000. This indicates that, within this model's context, economic output per person is not a determining factor of life satisfaction. This might suggest that once basic economic conditions are met, other factors like social relationships and health take precedence in influencing happiness.

These insights are crucial for understanding the key drivers of happiness and can guide policies focusing on improving life satisfaction through enhancing social support systems, personal freedoms, and health rather than focusing solely on economic growth.

## 4.7) Community Detection:

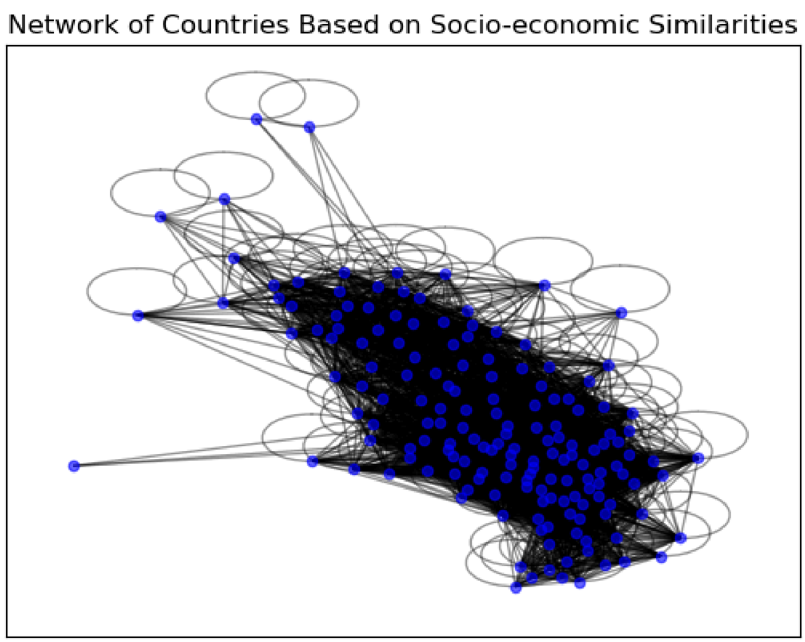
Number of communities: 3
Community 1: ['Chile', 'Peru', 'Uruguay', 'Greece', 'Portugal', 'Malta', 'Mauritius', 'Lebanon', 'Turkmenistan', 'Latvia', 'Serbia', 'Sri Lanka', 'Bahrain', 'United Arab Emirates', 'Italy', 'Ecuador', 'Slovenia', 'Belgium', 'Belarus', 'Somaliland region', 'Oman', 'Poland', 'Japan', 'Israel', 'Slovakia', 'Venezuela', 'Sweden', 'Belize', 'Argentina', 'Taiwan Province of China', 'Dominican Republic', 'Trinidad and Tobago', 'Netherlands', 'Ukraine', 'Jamaica', 'Colombia', 'State of Palestine', 'Libya', 'Austria', 'Bulgaria', 'Montenegro', 'Germany', 'Armenia', 'Bosnia and Herzegovina', 'Cyprus', 'Qatar', 'Singapore', 'Brazil', 'Iran', 'Russia', 'Tunisia', 'South Korea', 'Australia', 'El Salvador', 'Georgia', 'Suriname', 'France', 'Algeria', 'Kazakhstan', 'North Macedonia', 'Jordan', 'Hungary', 'Maldives', 'Paraguay', 'Spain', 'Türkiye', 'United States', 'Romania', 'Croatia', 'Vietnam', 'Panama', 'Cuba', 'Moldova', 'Thailand', 'Iraq', 'Mexico', 'Albania', 'Lithuania', 'Costa Rica', 'Estonia', 'Iceland', 'Kuwait', 'Azerbaijan', 'Saudi Arabia', 'Malaysia', 'Czechia', 'China']
Community 2: ['Lesotho', 'Zambia', 'Kyrgyzstan', 'Yemen', 'Senegal', 'Honduras', 'Mongolia', 'Congo (Brazzaville)', 'Nicaragua', 'Myanmar', 'Comoros', 'Bolivia', 'Togo', 'Kosovo', 'Ivory Coast', 'Eswatini', 'South Africa', 'Gabon', 'Nepal', 'Guatemala', 'Uganda', 'Mauritania', 'Sierra Leone', 'Guyana', 'Congo (Kinshasa)', 'Egypt', 'Angola', 'Cameroon', 'Malawi', 'Cambodia', 'Niger', 'Bangladesh', 'Djibouti', 'Sudan', 'Chad', 'Haiti', 'India', 'Laos', 'Tajikistan', 'Benin', 'Syria', 'Pakistan', 'Morocco', 'Kenya', 'Botswana', 'Afghanistan', 'Tanzania', 'Indonesia', 'Ethiopia', 'Ghana', 'Madagascar', 'Bhutan', 'Uzbekistan', 'Philippines', 'Liberia', 'Gambia', 'Namibia', 'Burkina Faso', 'South Sudan', 'Rwanda', 'Zimbabwe', 'Mozambique', 'Burundi', 'Central African Republic', 'Somalia', 'Mali', 'Nigeria', 'Guinea']
Community 3: ['United Kingdom', 'Luxembourg', 'Finland', 'Canada', 'Ireland', 'Denmark', 'Switzerland', 'Hong Kong S.A.R. of China', 'Norway', 'New Zealand']

The graph represents a network of countries visualized based on socio-economic similarities, with countries clustered into communities indicating groups with closely related socio-economic characteristics. Each node represents a country, and the edges denote the similarity between these countries based on various socio-economic factors like GDP, social support, and health indices. The detection of different communities within this network helps to highlight how countries can be grouped based on their socio-economic profiles. For example, the color-coded clusters (orange and blue) suggest distinct groups with shared traits which might imply similar challenges and potential for targeted policy approaches. Peripheral countries with fewer



connections might face unique conditions or challenges not widely shared with others in the network. This visualization provides valuable insights into global socio-economic patterns and can guide international collaborations and policy-making by identifying which countries share significant socio-economic ties and common developmental stages.

This graph visualizes a network of countries based on socio-economic similarities where each node represents a country, and edges represent the strength of similarity between countries in terms of socio-economic factors. The dense core of the network, with a multitude of closely interconnected nodes, suggests that many countries share strong socio-economic ties, possibly due to globalization or similar development stages.



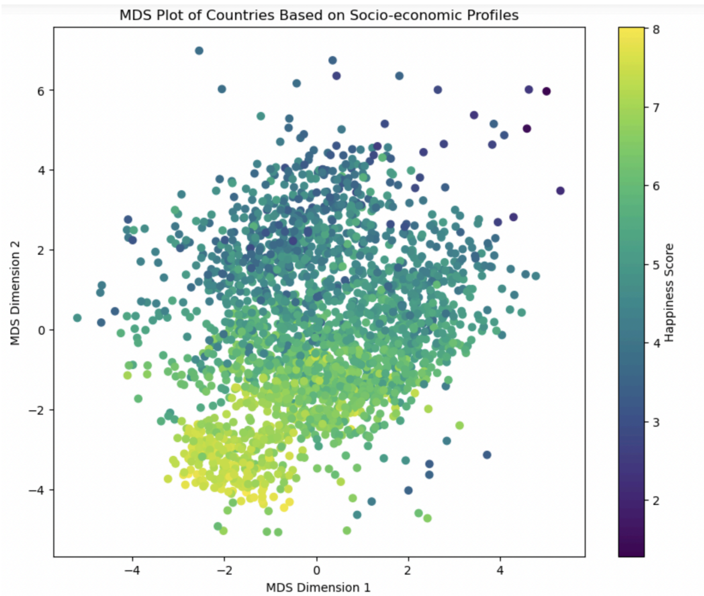

The sparser connections at the periphery might indicate countries with unique socio-economic profiles or those less integrated into the global economy. Community detection in this network could reveal groups of countries with common socio-economic challenges or advantages, which can be crucial for understanding global patterns and for policy-making. The visualization emphasizes the interconnected nature of the global economy and the significant variance in how deeply countries are embedded within this network.

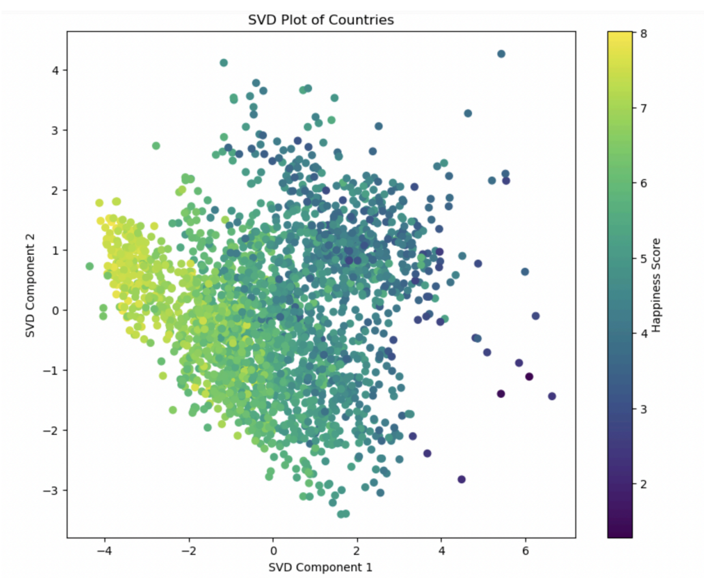
## 4.8) Feature Importance and Multidimensional Analysis:

Feature Importances Ranked
Through the Random Forest model, the importance of various features in predicting happiness scores was quantified. The features are ranked as follows:

Social Support: 0.7725
Healthy Life Expectancy at Birth: 0.0889

Generosity: 0.0500
Freedom to Make Life Choices: 0.0462
Perceptions of Corruption: 0.0425
Log GDP per Capita: 0.0000




Insights from Feature Importance
Social Support: Predominantly the most critical factor, with the highest importance score of 0.7725. This underlines the vital role of strong social networks and support mechanisms in enhancing societal happiness.
Economic Status: Interestingly, 'Log GDP per Capita' scored 0.0000, suggesting that beyond meeting basic economic needs, factors like social support and health are more pivotal in influencing happiness.

Multidimensional Scaling (MDS)
The MDS plot clusters countries based on their socio-economic profiles related to happiness:
Countries are visually grouped by similarities in happiness-related features.
The color coding indicates happiness scores, showing that countries with similar scores cluster closely, emphasizing the role of social and health conditions over economic status.

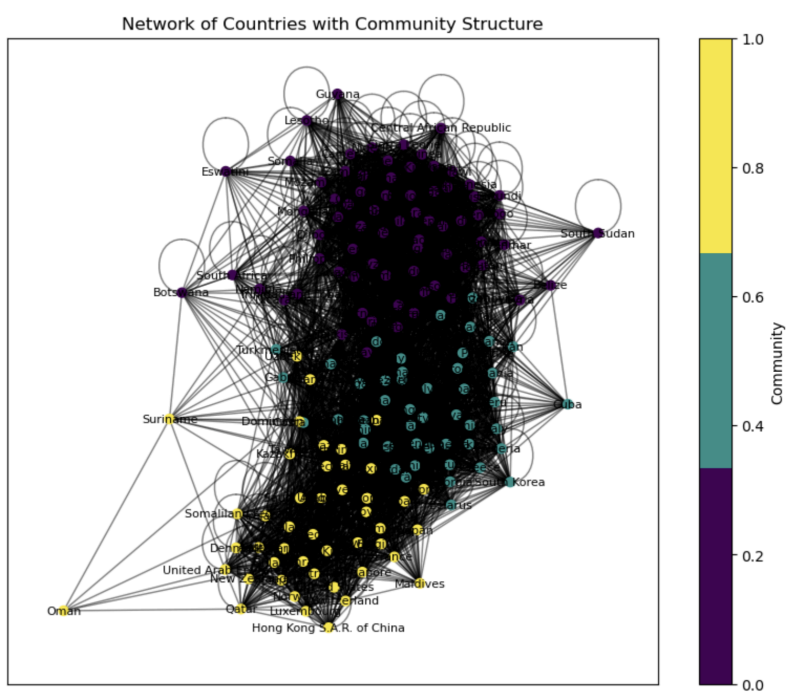
Implications for Policy Making and Social Planning
Policy Making: These insights can direct policymakers to prioritize enhancements in social support systems and healthcare to foster happiness.
Social Planning: Implementing programs that encourage community interactions and strengthen support networks could serve as effective strategies for boosting national happiness levels.
Health Initiatives: Given health's significant influence on happiness, public health initiatives should aim not only to extend life expectancy but also to enhance life quality, thus contributing positively to societal happiness levels.

## 4.9) Core-Periphery and Factions Analysis:

Core-Periphery Structure:
Core countries: The extensive list of core countries, including nations like the United States, Germany, and Japan, suggests that these countries share strong similarities in the features considered (e.g., GDP, social support). Their interconnectedness in the network highlights a global similarity in how these factors influence happiness across diverse regions.
Periphery countries: Oman stands out as a periphery country, indicating it might have unique characteristics or weaker correlations in the considered features with other countries. This could suggest unique policy needs or different socio-economic dynamics affecting happiness.
Factions Analysis Using Louvain Method:

Louvain method for community detection
The visualization clearly delineates distinct communities within the network, each representing a cluster of countries with similar socio-economic and happiness profiles. This is pivotal for understanding regional similarities and differences in happiness drivers. The gradient in the plot from teal to yellow represents varying levels of happiness metrics used, with closely clustered countries sharing similar levels.

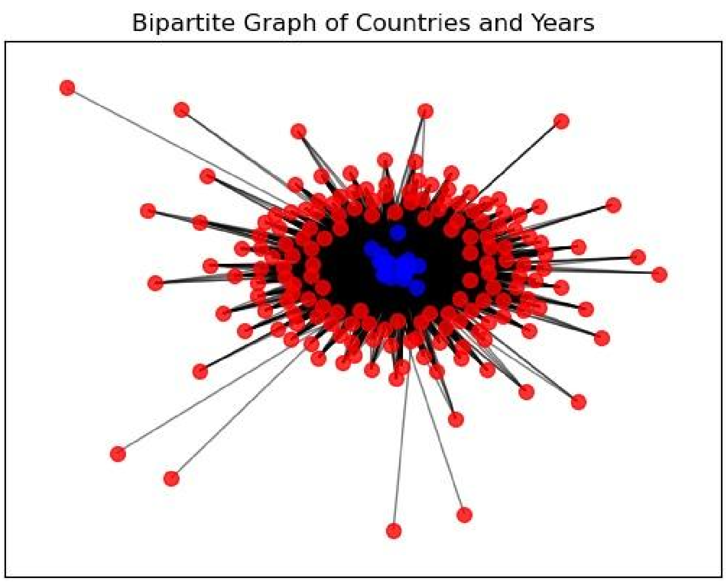
Key Results:
Connectivity and Centrality:
Serbia, Azerbaijan, Sri Lanka, Ukraine, and Malaysia are identified as the top central countries based on connectivity. This high centrality indicates these countries' features are very representative or closely correlated with many other countries in the dataset. These countries could potentially influence or be influenced by changes in the global socio-economic environment.



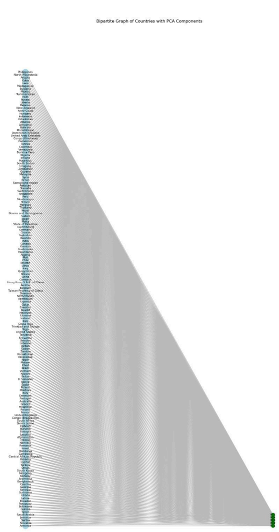

Community Detection:
The community detection algorithm has effectively grouped countries into different factions, each sharing specific characteristics. Such grouping can aid in targeted policy-making, where strategies can be customized

for groups of countries sharing similar socio-economic statuses or happiness scores.
Strategic Implications:
For Policy Makers: Understanding core versus periphery countries can help in crafting global versus targeted policies. For instance, strategies that work for core countries might not be as effective for periphery countries like Oman.
For Researchers and Economists: The network analysis provides a foundation for further exploration into the causal relationships between economic indicators and happiness, potentially leading to more effective interventions.

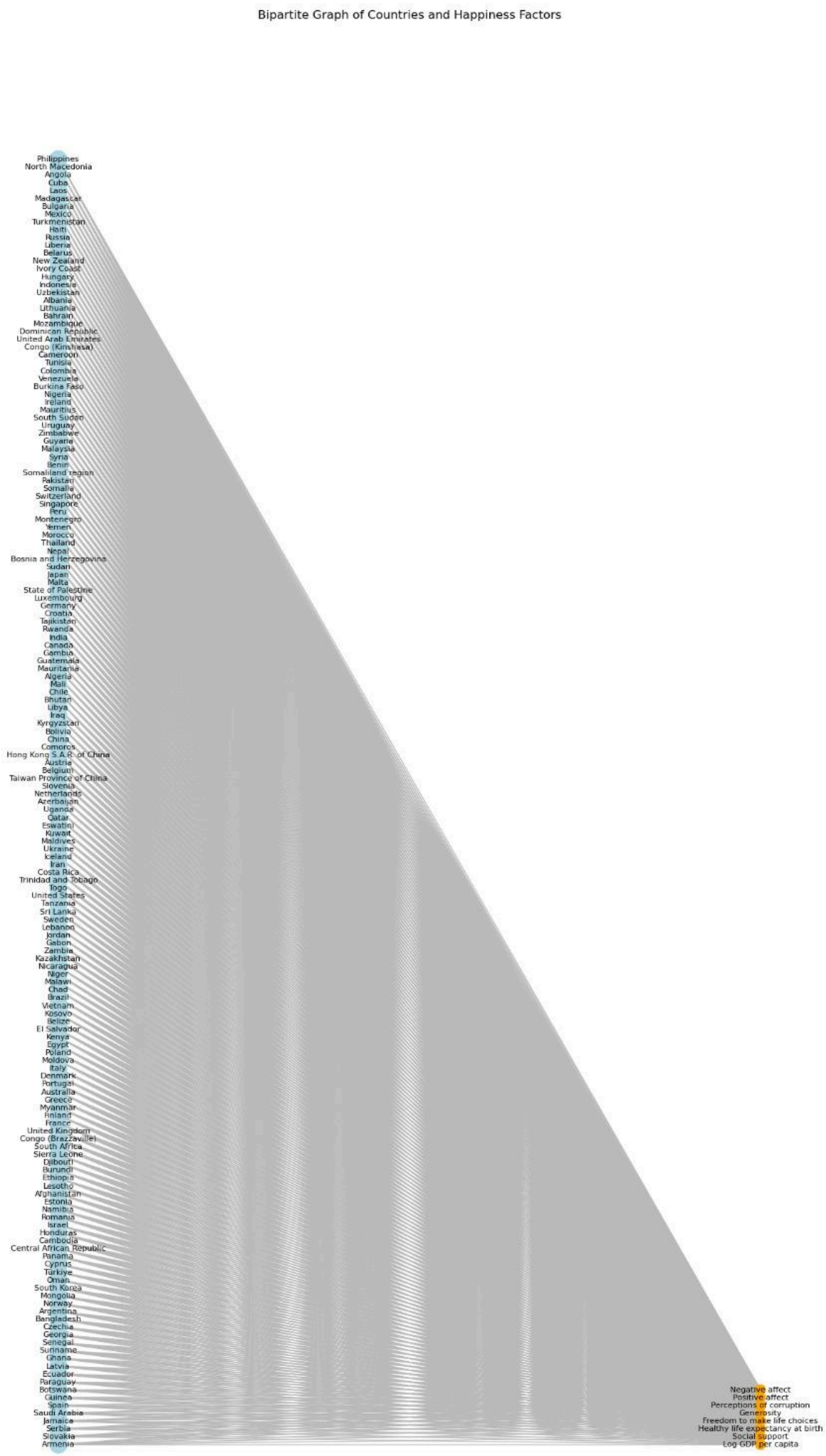
## 4.10) Bipartite Modeling:




Countries and Years
Year-Centric Connectivity: The years with high degree centrality (like 2017, 2011, etc.) have more comprehensive data available or represent significant years in terms of data collection across multiple countries. These might be years of particular interest for further analysis on how certain global events or conditions influenced multiple countries.



Countries and Happiness Factors
Universal Influence of Factors: Factors like 'Log GDP per capita', 'Social support', etc., showing high centrality indicate their universal presence across countries in the dataset. This suggests these factors are common metrics of happiness across different nations and might be critical levers for policy-making.
Country Specific Analysis: Lower centrality scores for some countries indicate missing data or lesser relevance of certain happiness factors in those regions. It might be useful to investigate if these countries are underrepresented or have unique characteristics affecting data availability or factor impacts.




Countries and PCA Components
Dominance of PCA Components: High centrality of PCA components suggests that these components are highly representative of variability across countries. Each principal component captures significant unique variance across the dataset, making them potentially powerful for predictive modeling or clustering countries.
Applications of Insights
Policy Making: Understanding which happiness factors are most central can help policymakers focus on what might truly impact happiness scores across different demographics.
Data Imputation Strategies: For countries with low centrality and hence, likely missing data, targeted surveys or data imputation methods could be considered to improve data reliability.

Predictive Analysis: Using PCA components with high centrality in predictive models could enhance the accuracy of predictions regarding country-specific outcomes related to happiness.

# **5.Discussion of Project**

Reflecting on the project, it was quite a challenging and enlightening experience for us. We struggled initially with organizing the implementation and structuring our code efficiently, which proved difficult as we navigated through the complexities of a large dataset. Employing advanced statistical methods and machine learning to uncover the drivers of happiness across countries also presented its own set of challenges, particularly in deriving random insights that were not immediately apparent.

If there were aspects to improve for next time, placing a stronger emphasis on addressing the challenges associated with data quality and representativeness from the start could have significantly enhanced the reliability of our findings. Additionally, incorporating more diverse data sources or expanding the dataset to cover a broader time frame might have allowed us to provide deeper insights into long-term trends and the effects of global events on happiness.

From this project, we learned the critical importance of thorough data preparation and the benefits of sophisticated analytical techniques. We gained practical experience in handling and visualizing complex datasets, understanding the nuances of feature importance, and the value of community detection in network analysis to comprehend socio-economic connections between countries. This experience not only honed our technical skills but also highlighted the profound impact of socio-economic factors on global happiness, offering valuable lessons for future research and policy applications.

# **6.Future Work**

In advancing this project, the following areas of exploration are outlined for future work, each promising to enrich the understanding and implications of the findings already established

Expanding the variety of data sources to include regional economic reports, psychological well-being surveys, and environmental quality indices could enrich the analysis and potentially uncover new determinants of happiness. Additionally, extending the dataset to cover a longer period would allow for a more comprehensive analysis of trends over time and help understand the long-term effects of socio-economic policies on happiness.

Employing deep learning techniques could more effectively capture non-linear relationships within the data, and neural networks might be used to predict happiness scores with greater accuracy by considering complex interactions between features. Furthermore, integrating geospatial data could facilitate the exploration of the geographical distribution of happiness scores and identify regional clusters or disparities within countries.

Incorporating psychological factors such as personality traits and cultural influences could provide deeper insights into the determinants of happiness. Similarly, examining sociological aspects like community engagement or social capital might enhance the understanding of the social support variable. Integrating macroeconomic models would also allow a better understanding of how economic policies impact happiness.

Investigating the interaction effects between different socio-economic factors could reveal how combined influences impact happiness differently than individual factors. Developing new features that capture the complexities of happiness, such as an index combining health, economic, and social data, could also prove beneficial.

Developing simulation models to forecast the impact of potential policy changes on happiness scores could aid policymakers in visualizing the potential effects of their decisions before implementation. Using system dynamics modeling could help understand the feedback loops and time delays in the influence of socio-economic factors on happiness, aiding in predicting long-term trends

# **7.References**


- Hanneman, Robert A., and Mark Riddle. "Two-Mode Networks." Introduction to Social Network Methods. University of California, Riverside. Accessed April 2024. https://faculty.ucr.edu/~hanneman/nettext/C17_Two_mode.html.
- Hanneman, Robert A., and Mark Riddle. "Multiplex Networks." Introduction to Social Network Methods. University of California, Riverside. Accessed April 2024. https://faculty.ucr.edu/~hanneman/nettext/C16_Multi_Plex.html.
- Newman, M.E.J. Networks: An Introduction. Oxford University Press, 2010.
- Fortunato, Santo. "Community detection in graphs." Physics Reports 486, no. 3-5 (2010): 75-174.

- Breiman, Leo. "Random Forests." Machine Learning 45, no. 1 (2001): 5-32.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics, 2009.
- Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." Journal of Statistical Mechanics: Theory and Experiment 2008.10 (2008): P10008.
- Newman, M.E.J., and M. Girvan. "Finding and evaluating community structure in networks." Physical review E 69, no. 2 (2004): 026113.
- Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008.
- Jolliffe, I.T. Principal Component Analysis. Springer-Verlag, 2002.
- Fabrigar, Leandre R., and Duane T. Wegener. Exploratory Factor Analysis. Oxford University Press, 2011.
- OpenAI. "ChatGPT: Optimizing Language Models for Dialogue." Accessed April 2024. https://www.openai.com.