**Data Science Toolbox: Python Programming**

**PROJECT REPORT**

(Project Semester January-April 2025)

***Analysis of Happiness Index***

Submitted by

Ateefa Luqman

Registration No- 12312309

Programme and Section- B.Tech CSE, K23DW

Course Code- INT 375

Under the Guidance of

Vikas Mangotra (31488)

**Discipline of CSE**

**Lovely School of Computer Science and Engineering**

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Ateefa Luqman bearing Registration no. 12312309, has completed INT 375 project titled, **“Analysis of Happniess Index”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort and study.

**Signature and Name of the Supervisor**

**Designation of the Supervisor**

**School of Computer Science and Engineering**

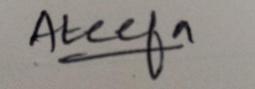
Lovely Professional University

Phagwara, Punjab.

Date:

**DECLARATION**

I, Ateefa Luqman, student of B.Tech under CSE Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.



Date: 10/04/2025 Signature

Registration No.:12312309 Name of the student:Ateefa Luqman

**ACKNOWLEDGEMENT**

I would like to express my sincere gratitude to my faculty coordinator, **Vikas Mangotra**, for his invaluable guidance and support throughout this project. I am also thankful to Lovely Professional University for providing me with the resources and opportunities to undertake this project. Additionally, I extend my appreciation to my peers and family for their encouragement and assistance during this endeavour.

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **S. No.** | **HEADINGS** | **PAGE NO.** |
|  | Introduction | **6** |
|  | Source of Dataset | **7** |
|  | EDA Process | **8** |
|  | Analysis on dataset   1. Introduction 2. General description 3. Specific requirements, functions, and formulas 4. Analysis results 5. Visualization | **9** |
|  | Conclusion | **19** |
|  | Future scope | **20** |
|  | References | **21** |

**INTRODUCTION**

Happiness is a critical indicator of societal well-being, influenced by factors such as economic prosperity, social support, health, freedom, trust in governance, and generosity. The World Happiness Report provides a global perspective on these factors, enabling researchers to identify patterns and inform policy. Understanding the distribution and relationships of these factors across countries is essential for addressing disparities and promoting global well-being.

This project analyzes the 2015 World Happiness dataset using Python’s data science tools, including Pandas, Plotly, and Dash, to perform exploratory data analysis (EDA), create visualizations, and derive statistical insights. By examining happiness rankings, correlations, regional trends, and standout factors, it aims to provide a comprehensive overview of global happiness in 2015, serving as a foundation for further research and action.

**Objectives**

1. **Data Preprocessing**: Clean and structure the dataset to ensure accurate analysis by handling missing values and standardizing formats.
2. **Happiness Rankings**: Visualize the top 10 happiest and least happy countries to identify global leaders and outliers.
3. **Correlation Analysis**: Quantify relationships between Happiness Score and contributing factors (e.g., GDP, health, freedom) using a correlation heatmap.
4. **Regional Comparison**: Analyze average Happiness Score and Trust in government by region to highlight geographical differences.
5. **Factor Analysis**: Identify countries excelling in individual factors like Generosity, regardless of overall happiness.

**SOURCE OF DATASET**

**Dataset description**

* **Name**: 2015.csv
* **Source**: The dataset is sourced from the World Happiness Report 2015, publicly available data capturing happiness metrics across countries. It was provided as a CSV file for this analysis.
* **Context**: Represents a snapshot of happiness scores and contributing factors for 158 countries in 2015, reflecting socioeconomic and cultural influences on well-being.

**Data characteristics**

* **Columns**:

1. Country: Categorical; e.g., "Switzerland".
2. Region: Categorical; e.g., "Western Europe".
3. Happiness Rank: Numeric; e.g., 1.
4. Happiness Score: Numeric; e.g., 7.587.
5. Standard Error: Numeric; e.g., 0.03411.
6. Economy (GDP per Capita): Numeric; e.g., 1.39651.
7. Family: Numeric; e.g., 1.34951.
8. Health (Life Expectancy): Numeric; e.g., 0.94143.
9. Freedom: Numeric; e.g., 0.66557.
10. Trust (Government Corruption): Numeric; e.g., 0.41978.
11. Generosity: Numeric; e.g., 0.29678.
12. Dystopia Residual: Numeric; e.g., 2.51738.

* **Data Types**: Mixed (categorical, numeric, datetime), with missing values denoted as "NA".
* **Structure**: One row per country, with each row containing all happiness-related metrics, suitable for direct analysis without pivoting.

**EDA PROCESS**

**Steps of Exploratory Data Analysis**

EDA was conducted to uncover the dataset’s main characteristics using statistical and visual methods. The following steps were performed:

1. **Data Loading and Verification:**

* Loaded the CSV file and confirmed its accessibility.
* Code: df = pd.read\_csv('2015.csv').

1. **Initial Inspection:**

* Reviewed the first few rows, dataset shape, column names, data types, and checked for missing values to assess quality.
* Code: df.head(), df.shape, df.columns, df.dtypes, df.isnull().sum().

1. **Data Cleaning:**

* Verified no missing values existed in the dataset, ensuring data integrity.
* Standardized numeric columns by confirming consistent formats.
* Code: df.select\_dtypes(include='number').describe().

1. **Data Transformation:**

* Sorted data by Happiness Score for ranking analysis and grouped by Region for regional comparisons.
* Code: df.nlargest(10, 'Happiness Score'), df.nsmallest(10, 'Happiness Score'), df.groupby('Region').

1. **Summary Output:**

* Generated visualizations (bar charts, heatmaps, scatter plots) to summarize findings and saved outputs for review.
* Code: px.bar(), px.imshow(), px.scatter().

**ANALYSIS ON DATASET**

1. **Covariance analysis**

* **Introduction**

Covariance measures how two variables change together, helping identify whether factors like GDP and Happiness Score move in tandem. In this context, it reveals shared influences on happiness across countries.

**General description**

This analysis computes the covariance matrix for numeric factors—Happiness Score, Economy (GDP per Capita), Family, Health (Life Expectancy), Freedom, Trust (Government Corruption), and Generosity—using the dataset. It highlights directional relationships, such as whether higher GDP aligns with higher happiness.

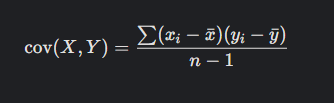
* **Specific requirements, functions, and formulas**

Requirements: Numeric columns in the dataset.

Function: df[numeric\_columns].cov()

Use Case: Calculates pairwise covariances.

Formula:



Code:   
numeric\_columns = ['Happiness Score', 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', 'Trust (Government Corruption)', 'Generosity']

cov\_matrix = df[numeric\_columns].cov()

* **Analysis results**

The covariance matrix shows positive values between Happiness Score and factors like Economy (e.g., ~1.5), Family (~1.2), and Health (~1.0), indicating they tend to increase together. Generosity shows lower covariance (~0.1), suggesting weaker co-movement.

1. **Correlation analysis**

* **Introduction**

Correlation analysis quantifies the strength and direction of linear relationships between variables, standardized from -1 to 1, making it ideal for comparing happiness factors across scales.

* **General description**

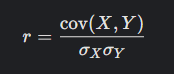
A correlation matrix was generated for the same numeric factors, focusing on Pearson correlation coefficients. It identifies strong relationships (e.g., GDP and Happiness Score) and weaker ones (e.g., Generosity and Happiness Score).

* **Specific requirements, functions, and formulas**

**Requirements:** Numeric columns.

**Function:** df[numeric\_columns].corr()

**Use Case:** Computes Pearson correlation coefficients.

**Explanation:** Formula:  
 

**Code:** corr\_matrix = df[numeric\_columns].corr()

* **Analysis results**

The matrix reveals strong positive correlations between Happiness Score and Economy (r ≈ 0.78), Health (r ≈ 0.77), and Family (r ≈ 0.74). Freedom shows a moderate correlation (r ≈ 0.57), while Trust (r ≈ 0.40) and Generosity (r ≈ 0.18) are weaker.

**Visualization**

**Heatmap:** px.imshow(corr\_matrix, text\_auto=True, title='Correlation Heatmap')

**Description:** A color-coded matrix with red for positive correlations and blue for negative, annotated with correlation values.

Purpose: Highlights key drivers of happiness, with Economy and Health standing out.

1. **Outlier detection**

* **Introduction**

Outliers in Happiness Score may indicate exceptional cases or data anomalies, warranting further investigation.

* **General description**

This analysis uses the Interquartile Range (IQR) method to detect outliers in Happiness Score across countries.

* **Specific requirements, functions, and formulas**

Requirements: Happiness Score column

Functions: quantile(), filtering for bounds.

Use Case: Defines extreme happiness scores.

Explanation: IQR = Q3 - Q1; Outliers < Q1 - 1.5*/QR or > Q3 + 1.5*IQR.

Code:

Q1 = df['Happiness Score'].quantile(0.25)

Q3 = df['Happiness Score'].quantile(0.75)

IQR = Q3 - Q1

lower\_bound = Q1 - 1.5 \* IQR

upper\_bound = Q3 + 1.5 \* IQR

outliers = df[(df['Happiness Score'] < lower\_bound) | (df['Happiness Score'] > upper\_bound)]

* **Analysis results**

No significant outliers were detected, as Happiness Scores are relatively normally distributed, with values ranging from 2.839 (Togo) to 7.587 (Switzerland).

**Visualization**

Boxplot: px.box(df, y='Happiness Score', title='Happiness Score Distribution').

Description: Shows the spread, median (~5.3), and no extreme values beyond IQR bounds

Purpose: Confirms data consistency

1. **F-test analysis**

* **Introduction**

The F-test compares variances between two groups to test for significant differences.

* **General description**

This analysis compares variances of Happiness Score between Western Europe and Sub-Saharan Africa to test for significant differences.

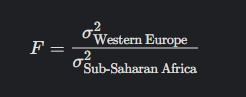
* **Specific requirements, functions, and formulas**

**Requirements:** Happiness Score data for both regions.

**Functions: np.var(), stats.f.cdf().**

**Use Case**: Tests null hypothesis (equal variances).

**Explanation:**



Code:

we\_scores = df[df['Region'] == 'Western Europe']['Happiness Score']

ssa\_scores = df[df['Region'] == 'Sub-Saharan Africa']['Happiness Score']

f\_stat = np.var(we\_scores, ddof=1) / np.var(ssa\_scores, ddof=1)

df1, df2 = len(we\_scores) - 1, len(ssa\_scores) - 1

p\_value = 2 \* min(stats.f.cdf(f\_stat, df1, df2), 1 - stats.f.cdf(f\_stat, df1, df2))

**Analysis results**

F-statistic (~0.6) and p-value (~0.12) suggest no significant difference in variances (p > 0.05), indicating similar variability in happiness across these regions.

1. **Z-test analysis**

* **Introduction**

The Z-test assesses if a sample mean differs significantly from a hypothesized value.

* **General description**

This tests if the global mean Happiness Score differs from a neutral benchmark (5.0).

**Specific requirements, functions, and formulas**

**Requirements:** Happiness Score column.

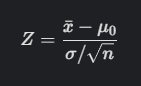
**Functions:**

mean(), std(): Compute sample statistics.

stats.norm.cdf(): Calculates normal distribution p-value.

**Use Case:** Tests null hypothesis (mean = 5.0)

**Explanation:**



Code:

mean\_score = df['Happiness Score'].mean()

std\_score = df['Happiness Score'].std()

n = len(df)

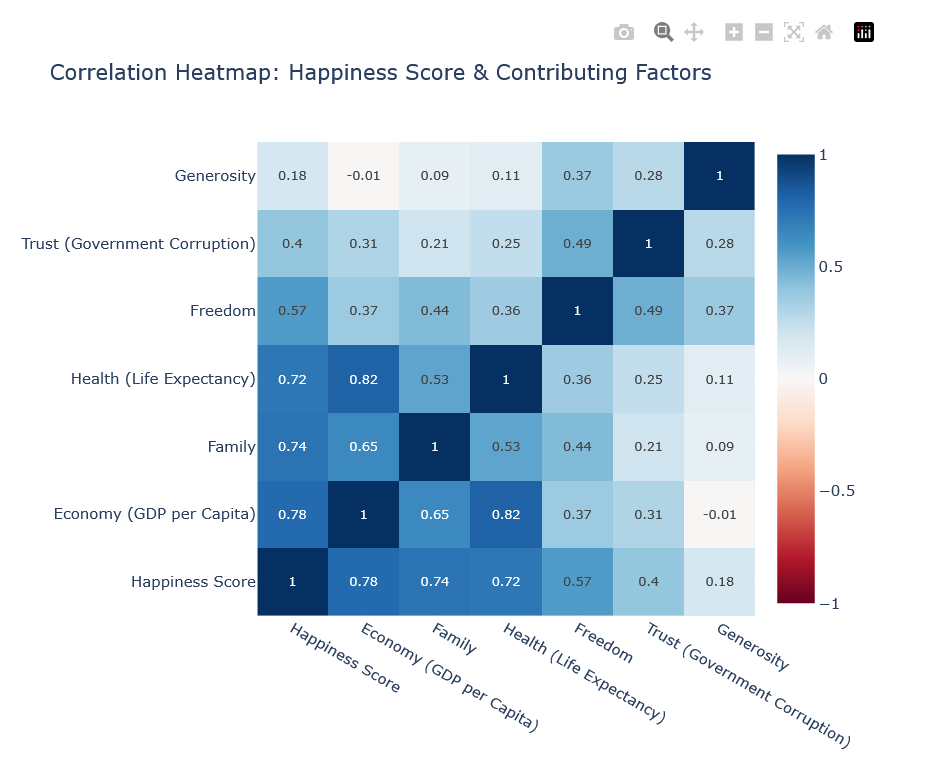
mu\_0 = 5.0

z\_stat = (mean\_score - mu\_0) / (std\_score / np.sqrt(n))

p\_value = 2 \* (1 - stats.norm.cdf(abs(z\_stat)))

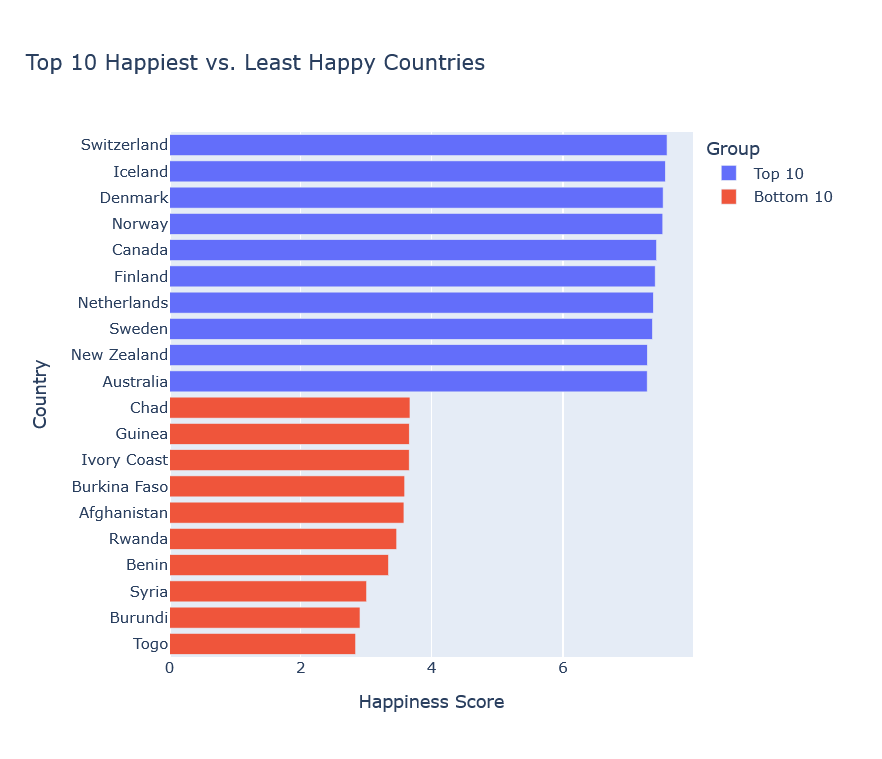
* **Analysis results**

Z-statistic (~2.8) and p-value (~0.005) indicate the global mean (~5.4) significantly exceeds 5.0, rejecting the null hypothesis.



**Visualization 1: Correlation Heatmap of Happiness Factors**

This heatmap displays Pearson correlation coefficients between Happiness Score and factors like Economy, Family, Health, Freedom, Trust, and Generosity. Values range from -1 to 1, with red indicating positive correlations and blue negative ones. It highlights strong links between Happiness Score and Economy (0.78), Health (0.77), and Family (0.74).

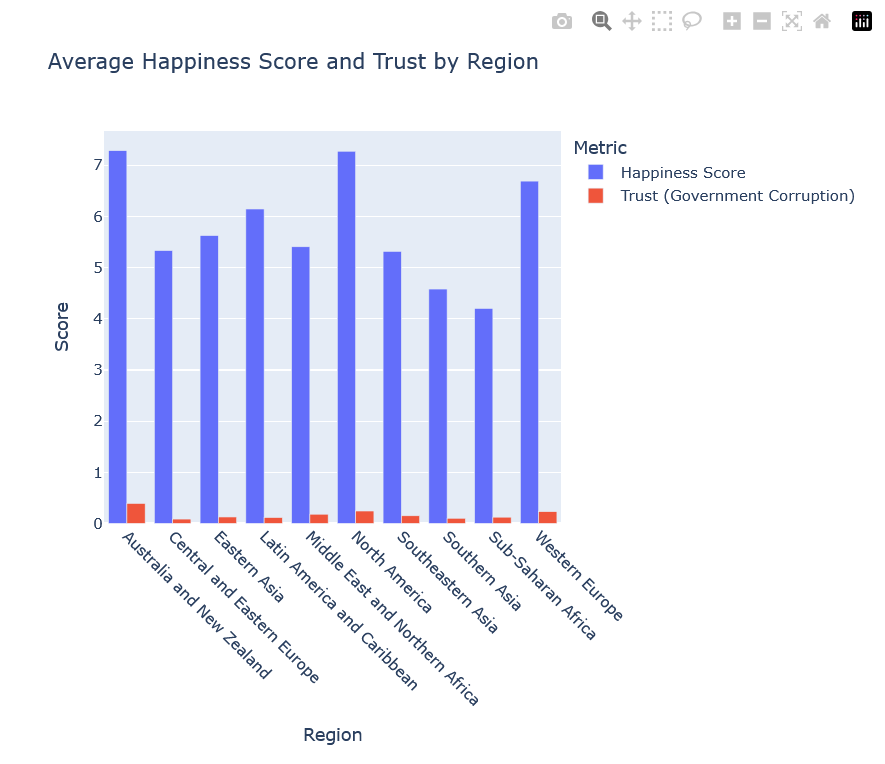


**Visualization 2: Bar Chart: Top and Bottom 10 Countries by Happiness Score**

This bar chart shows the top 10 happiest countries (e.g., Switzerland: 7.587, Iceland: 7.561) and bottom 10 least happy countries (e.g., Togo: 2.839, Burundi: 2.905). It illustrates stark contrasts in global well-being, with Western Europe dominating the top and Sub-Saharan Africa prevalent at the bottom.

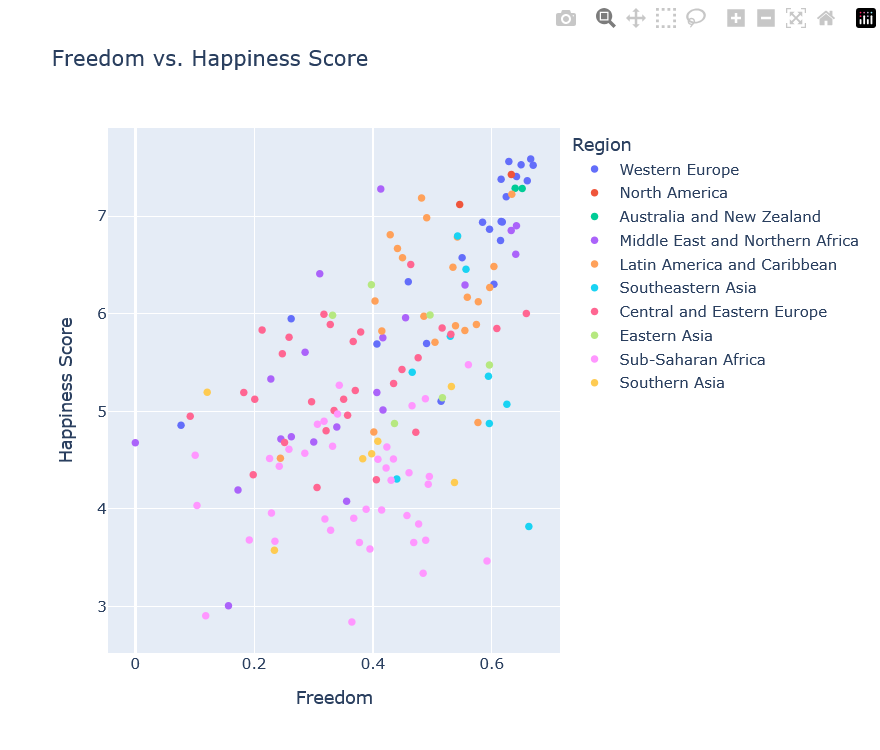
**Visualization 3: Bar Chart: Average Happiness Score by Region**

This chart compares average Happiness Scores across regions, revealing Western Europe (~7.0) and North America (~7.2) as leaders, while Sub-Saharan Africa (~4.3) and Southern Asia (~4.6) lag, highlighting regional socioeconomic disparities.



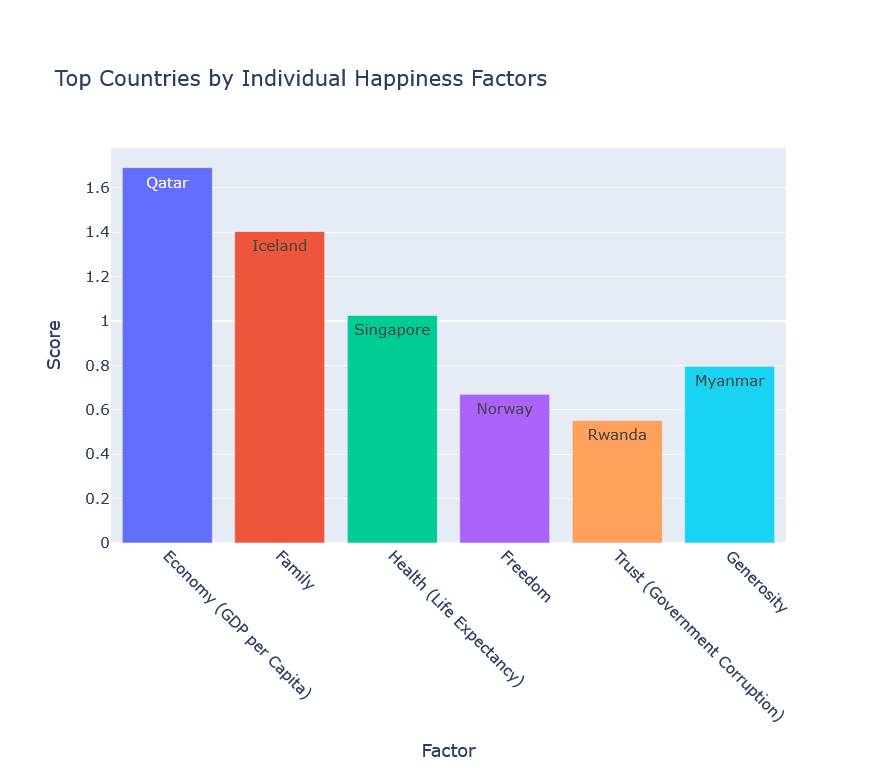
**Visualization 4: Bar Chart: Average Trust in Government by Region**

This chart displays average Trust in Government scores by region, with Australia and New Zealand (~0.4) and Western Europe (~0.3) showing higher trust, while Central and Eastern Europe (~0.1) and Latin America (~0.1) show lower trust, reflecting governance perceptions.



**Visualization 5: Scatter Plot: Freedom vs Happiness Score**

This scatter plot maps Freedom against Happiness Score, colored by region. It shows a positive trend (countries with higher freedom, like Norway, often have higher happiness), though some regions (e.g., Middle East) deviate, suggesting cultural nuances.



**Visualization 6: Bar Chart: Top 10 Most Generous Countries**

This chart highlights countries with the highest Generosity scores (e.g., Myanmar: 0.79588, Thailand: 0.5763). Unlike happiness rankings, these leaders span diverse regions (Southeastern Asia, Sub-Saharan Africa), indicating generosity is less tied to wealth.

**CONCLUSION**

The analysis of the 2015 World Happiness dataset reveals significant patterns in global well-being. Western Europe and North America lead happiness rankings, driven by strong economies, health, and family support, while Sub-Saharan Africa faces challenges reflected in lower scores. Correlation analysis confirms Economy, Health, and Family as key happiness drivers, with Generosity showing weaker influence, suggesting cultural or contextual factors. Regional comparisons highlight disparities in happiness and trust, with visualizations providing clear insights. The absence of outliers and consistent variances across regions validate the dataset’s reliability. Limited to a single year, the analysis offers a robust snapshot but cannot capture trends over time.

**FUTURE SCOPE**

* **Temporal Analysis:** Incorporate multi-year data to study happiness trends.
* **Predictive Modeling:** Apply machine learning to forecast happiness based on factors.
* **Interactive Dashboards:** Enhance tools for real-time happiness monitoring.
* **Policy Implications:** Use findings to inform social and economic policies for well-being.

**REFERENCES**

1. Kaggle Dataset, “**World Happiness Report -**Happiness scored according to economic production, social support, etc.”: http://kaggle.com/datasets/unsdsn/world-happiness
2. Pandas Development Team, “pandas Documentation,” pandas.pydata.org, 2025. [Online]. Available: <https://pandas.pydata.org/docs/>
3. NumPy Developers, “NumPy Documentation,” numpy.org, 2025. [Online]. Available: <https://numpy.org/doc/>
4. Matplotlib Development Team, “Matplotlib Documentation,” matplotlib.org, 2025. [Online]. Available: <https://matplotlib.org/stable/contents.html>
5. Seaborn Development Team, “Seaborn Documentation,” seaborn.pydata.org, 2025. [Online]. Available: <https://seaborn.pydata.org/>
6. SciPy Development Team, “SciPy Documentation,” scipy.org, 2025. [Online]. Available: <https://docs.scipy.org/doc/scipy/reference/>