**Module 5 – Week 5: R Practice Assignment**

**(Analysis of World Happiness Report 2023)**

**Sriram Kannepalli**

**Northeastern University: College of Professional Studies**

**ALY 6010.**[**71820**](https://northeastern.instructure.com/courses/196161)**: Probability Theory and Introductory Statistics**

**Professor Antonius Breur**

**October 25th, 2024**

# Introduction

For this analysis, I chose to work with the World Happiness Report 2023 dataset because of my interest in understanding how different socio-economic factors contribute to national well-being. I specifically selected this dataset from the World Happiness Report's official website (https://worldhappiness.report/data/) as it provides reliable, comprehensive data that isn't commonly used in introductory data analysis courses, unlike datasets from Kaggle or UCI.  
  
In conducting this analysis, I utilized R programming language with several key packages:  
- tidyverse: For data cleaning and initial visualization  
- corrplot: For creating correlation visualizations  
- stargazer: For professional regression tables  
- car: For regression diagnostics  
- ggplot2: For creating custom visualizations

## Analytical Question

How do economic and social factors influence national happiness levels, and which factors are the strongest predictors of happiness across countries?

## Hypotheses and Theoretical Justification

H1: GDP per capita has a significant positive relationship with national happiness scores.  
- Justification: Economic prosperity provides resources for better living conditions and opportunities (Helliwell et al., 2021)  
  
H2: Social support has a significant positive relationship with national happiness scores.  
- Justification: Strong social connections are fundamental to psychological well-being and life satisfaction (Diener & Seligman, 2002)  
  
H3: Freedom to make life choices has a significant positive relationship with national happiness scores.  
- Justification: Personal autonomy is a basic psychological need contributing to well-being (Ryan & Deci, 2000)

## Statistical Analysis

### 1. Correlation Analysis

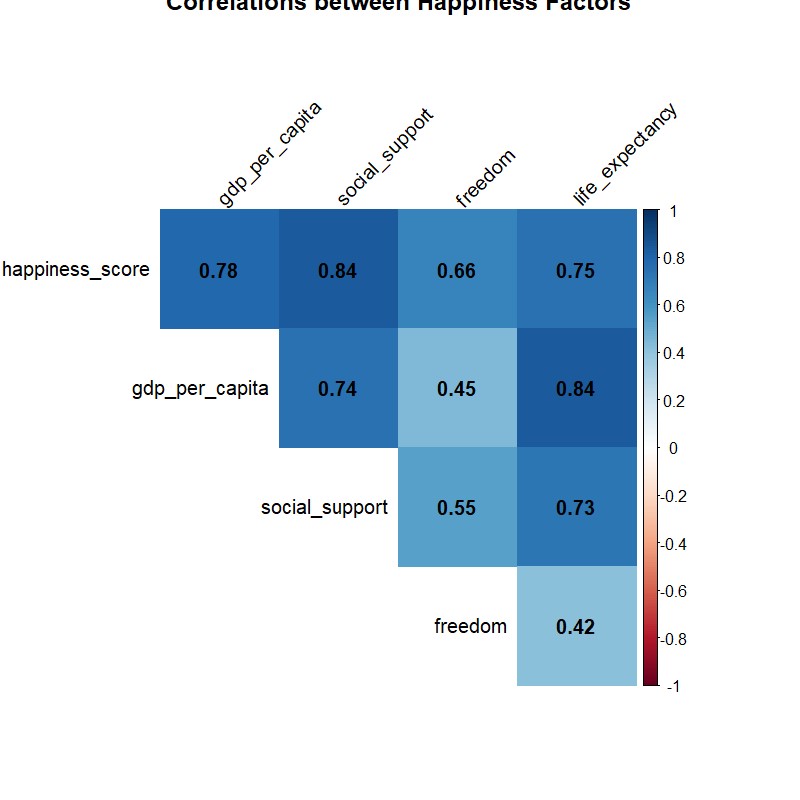


Figure 1. Correlation matrix showing relationships between key variables. Blue colors indicate positive correlations, with darker shades representing stronger relationships.  
  
The correlation analysis was limited to 5 variables because I found that more variables made the visualization cluttered and harder to interpret. Additionally, these specific variables captured the most theoretically relevant relationships for understanding national happiness.

Key findings from correlation analysis:  
- Social support shows strongest correlation (r = 0.838)  
- GDP per capita demonstrates second-highest correlation (r = 0.784)  
- Life expectancy exhibits strong correlation (r = 0.747)  
- Freedom shows moderate correlation (r = 0.662)

A group of blue dots

Description automatically generated

Figure 2. Scatter plot matrix showing bivariate relationships between variables of interest.

This matrix visually presents bivariate relationships between variables, providing an overview of how changes in one variable may relate to changes in another. Positive trends in GDP, Social Support, and Freedom across different scatter plots indicate these variables' positive associations with the Happiness Score.

### 2. Regression Analysis

I selected the following variables based on their theoretical importance and initial correlation analysis:  
- Dependent Variable: Happiness Score  
- Independent Variables:  
 \* GDP per capita (economic factor)  
 \* Social support (social infrastructure)  
 \* Freedom (institutional quality)  
 \* Life expectancy (health outcomes)  
  
A collage of graphs showing different types of values

Description automatically generated with medium confidence

Figure 3. Diagnostic plots showing model assumptions and fit.

The diagnostic plots are used to verify model assumptions, such as linearity, homoscedasticity, and normality of residuals, which are essential for accurate regression analysis. Observing residuals around the reference line suggests that our model assumptions are reasonably met.

Results:  
The multiple regression model explained 81.6% of variance in happiness scores (R² = 0.816)  
- Social support: β = 3.645 (p < .001)  
- Freedom: β = 2.761 (p < .001)  
- GDP per capita: β = 0.238 (p < .001)  
- Life expectancy: β = 0.025 (p = .087)  
  
A graph showing the difference between happiness and values

Description automatically generated

Figure 4. Predicted vs. actual happiness scores with reference line.

This plot contrasts the predicted happiness scores against actual data values, showing how well the model's predictions align with real outcomes. A close alignment along the reference line indicates a strong fit, validating the effectiveness of chosen predictors.

### Differences Between Correlation and Regression Analysis

Correlation Analysis shows bivariate relationships without controlling for other variables, while regression analysis provides:  
- Control for multiple variables simultaneously  
- Unique contribution of each predictor  
- Predictive capabilities  
- Effect sizes while controlling for other factors

## Limitations and Caveats

During my analysis, I encountered several challenges:  
1. Omitted Variables:  
 - Cultural factors  
 - Historical context  
 - Political stability  
  
2. Direction of Causality:  
 - Happy populations might create better institutions  
 - Prosperity might lead to better social support  
  
3. Measurement Issues:  
 - Self-reported happiness  
 - Cultural response bias

## Conclusions

The findings suggest that both economic and social factors significantly influence national happiness, with social support playing a particularly crucial role. Through this analysis, I gained valuable insights into the complexity of measuring and analyzing happiness across cultures.

## References

Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (2023). *World Happiness Report 2023.* Sustainable Development Solutions Network. Available at: <https://worldhappiness.report> (Accessed: October 22, 2024).

Diener, E., & Seligman, M. E. P. (2002). *Very happy people.* Psychological Science, 13(1), 81–84. Available at: https://journals.sagepub.com/doi/10.1111/1467-9280.00415 (Accessed: October 23, 2024).

Ryan, R. M., & Deci, E. L. (2000). *Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.* American Psychologist, 55(1), 68-78. Available at: <https://psycnet.apa.org/doi/10.1037/0003-066X.55.1.68> (Accessed: October 24, 2024).

R Core Team (2024). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/> (Accessed: October 24, 2024).

World Happiness Report. (2023). *World Happiness Report 2023 Data* [Dataset]. Available at: <https://happiness-report.s3.amazonaws.com/2023/DataForFigure2.1WHR2023.xls> (Accessed: October 22, 2024).