Alie Bakytbekova

Ms. Adhikari

COMP 4353 Data Mining

April 7th 2025

Data Preprocessing Report

World Happiness Report (2020–2024)

1. Introduction

I chose the World Happiness Report because it includes global data on happiness and well-being based on various life factors like GDP, health, and freedom. It covers multiple years and countries, making it a rich dataset for analysis.

2. Data Cleaning

I cleaned the dataset by removing rows with missing values and renaming the columns for easier access. Each year from 2020 to 2024 was combined into one dataframe, and duplicates were checked.

3. Data Normalization

To standardize the data, I applied Min-Max normalization on all numerical features like economy, social support, and health to bring them into a [0, 1] range.

4. PCA (Principal Component Analysis)

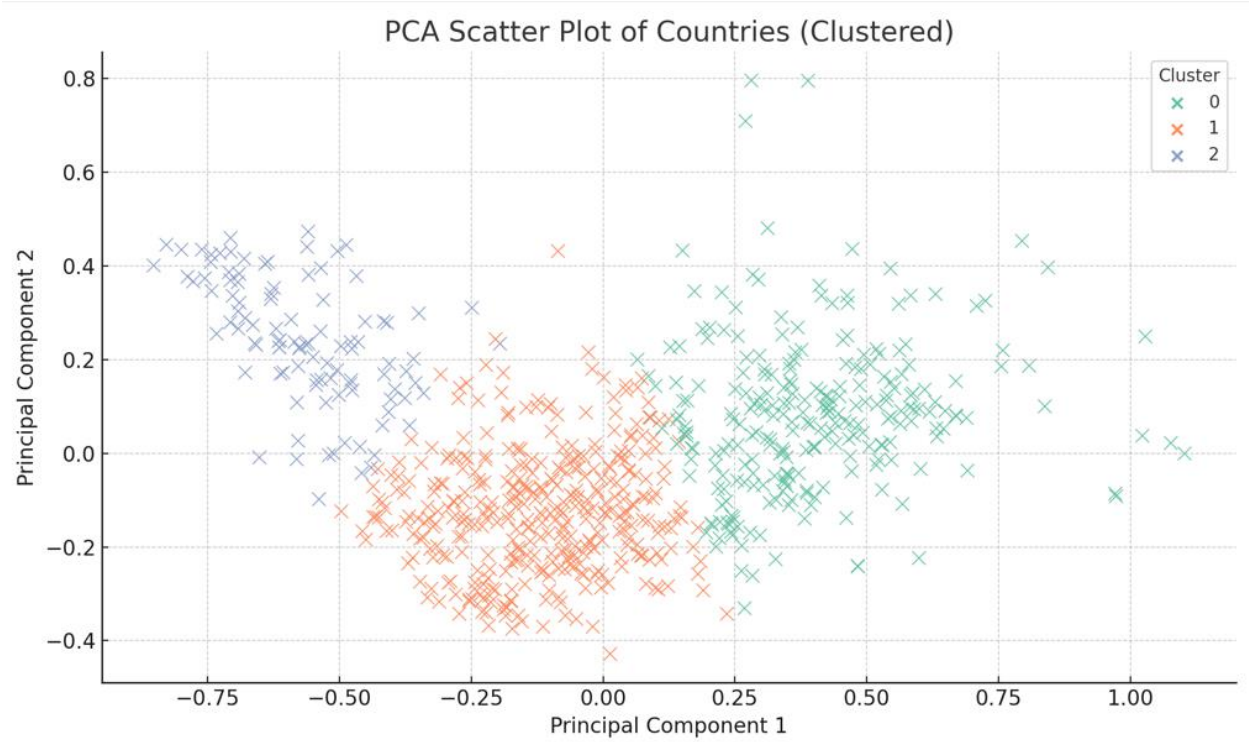
I used PCA to reduce the dimensions of the dataset from 7 features to 2 principal components. This made it easier to visualize patterns and groupings among countries.

5. Linear Regression

I performed linear regression to predict the happiness score based on GDP, social support, and health. The model explained 0.66 of the variance in happiness scores, which shows a strong relationship between these features.

6. Clustering

I used K-Means clustering to group countries into 3 clusters based on their normalized happiness-related metrics. These clusters help to identify patterns across different regions and years. The PCA scatter plot below shows how countries are grouped into three clusters based on their happiness-related features. Each point represents a country, and the colors indicate the cluster it belongs to.



7. Conclusion

Overall, data preprocessing helped clean, simplify, and extract useful patterns from the World Happiness Report. PCA and clustering revealed meaningful groups, and regression showed key contributors to happiness like GDP and health.