Report : Wrangle and Profile Stages

Mahtab Nahayati

1. Introduction

This project investigates the interplay between happiness scores and socio-economic as well as environmental factors across countries from 2015 to 2021. The analysis leverages two datasets: a primary dataset, the World Happiness Reports, which provides happiness scores and ranks along with various associated factors, and a secondary dataset containing global socio-economic indicators such as CO2 production and the Human Development Index (HDI). By combining these datasets, this project aims to uncover meaningful patterns and relationships that influence happiness on a global scale.

2. Wrangle Stage

2.1 Joining the Datasets

To integrate the primary dataset with the socio-economic indicators from the secondary dataset, the columns "**Country"** and "**Year"** were used as the join keys. These columns served as common identifiers across both datasets, facilitating their merging without requiring the introduction of new keys. However, discrepancies in country names between the datasets posed challenges, as some countries were labeled differently in the primary and secondary datasets. For example, "Côte d'Ivoire" in the secondary dataset corresponded to "Ivory Coast" in the primary dataset. These inconsistencies were systematically resolved through a mapping process, ensuring a smooth and accurate merge of the datasets.

2.2 Data Cleaning Steps

Cleaning the data was a multi-step process. First, temporal alignment was necessary, as the secondary dataset contained data up to 2021, while the primary dataset included entries for 2022 and 2023. To maintain consistency, data points for 2022 and 2023 were excluded, as were years prior to 2015 from the secondary dataset. This resulted in a final dataset covering the years 2015 to 2021.

For the primary dataset, inconsistencies in column names across yearly files were addressed. Each year's data was stored in a separate CSV file with slight variations in naming conventions, which were standardized during preprocessing. Unnecessary columns were removed to focus solely on those relevant to the analysis, such as "Happiness Score", "Happiness Rank", and socio-economic indicators.

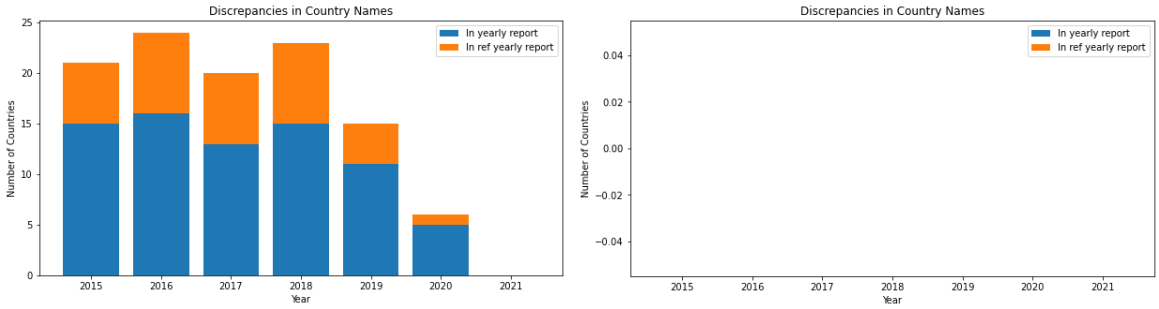
A critical step involved resolving country name discrepancies within the yearly files of the primary dataset. Using the year 2021 as a reference, country names from earlier years were mapped and corrected to align with those in the reference year. For example, "Macedonia" was updated to "North Macedonia" to ensure uniformity. Additionally, anomalies in the data were identified, such as the absence of the "Happiness Rank" column for certain years. This column was reconstructed based on the sorted "Happiness Score". Furthermore, the columns "Happiness Score" and "Happiness Rank" were mistakenly interchanged for 2017, causing erroneous values in that year's dataset. This anomaly was rectified during the cleaning process.

In the secondary dataset, column names representing yearly data (e.g., co2_prod_2015) were reshaped into a long format with a single "Year" column to facilitate merging. Relevant columns were renamed to "CO2 Production" and "HDI" for clarity and consistency.

Finally, the two datasets were merged based on “Country” and “Year”, producing a clean, integrated dataset suitable for analysis. Rows with missing or unmatched data were dropped, ensuring the final dataset was comprehensive and accurate.

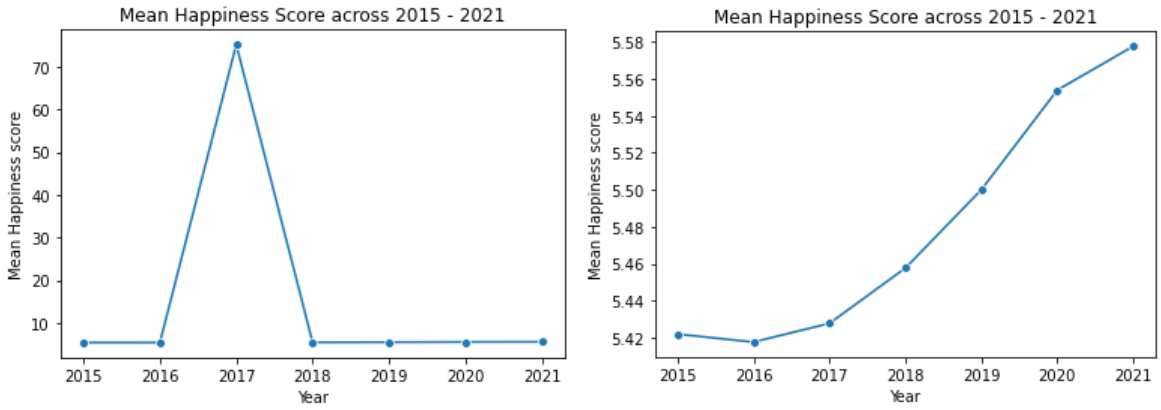
2.3 Visualizations to Illustrate Data Quality

Several visualizations were used to assess and communicate the quality of the data. For instance, bar charts highlighted discrepancies in country names before and after alignment, demonstrating the effectiveness of the mapping process. Additionally, a line graph revealed an anomaly in happiness scores for 2017, where scores unrealistically exceeded 80. After rectification, the visual confirmed the consistency and reliability of the corrected dataset.



(a)Before Mapping (b) After Mapping

Figure1: The number of discrepancies before and after mapping the country names



(a) Before the Fixing the mix-up (b) After Fixing the Mix-up

Figure2: The Mix-up of the column names Happiness Rank for the year 2017

3. Profile Stage

3.1 Insight 1: Distribution of Happiness Scores

A boxplot was created to analyze the distribution of happiness scores across countries from 2015 to 2021. The boxplot revealed that the median happiness score remained relatively stable at approximately 5.5 during this period. However, there was noticeable variability in the interquartile range across years, indicating differences in the spread of scores. Some years exhibited outliers, representing countries with exceptionally high or low happiness scores. This visualization provided an overview of global happiness

trends and highlighted potential areas for further exploration, such as the factors contributing to outliers in specific years.

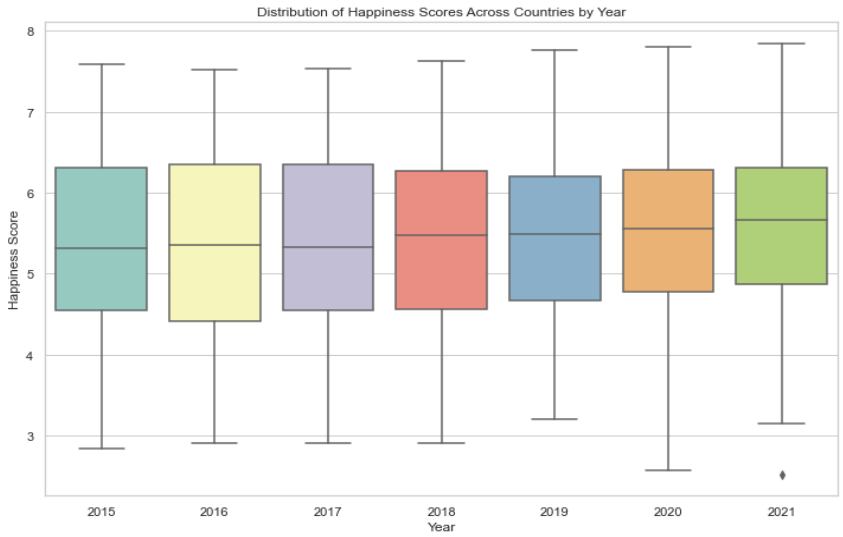
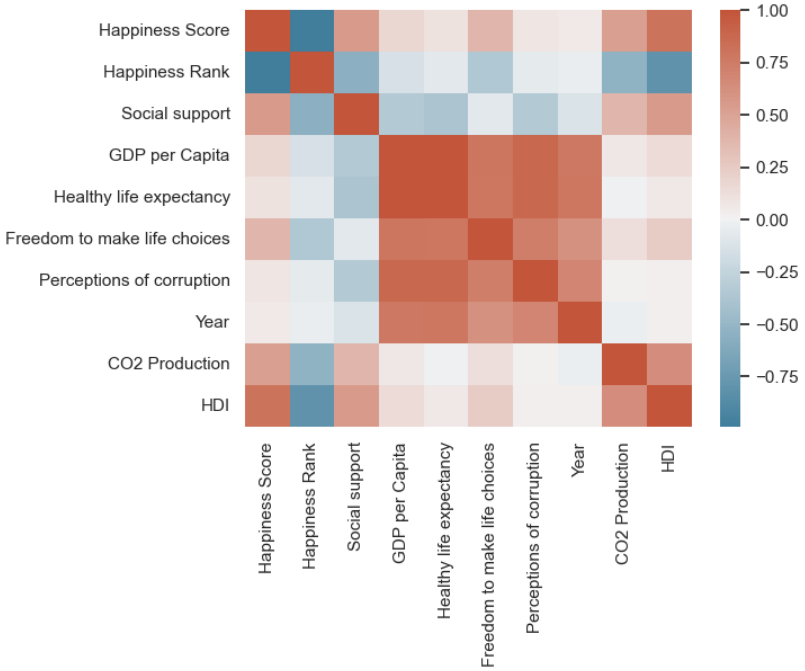


Figure3: Distribution of Happiness Across Countries

3.2 Insight 2: Correlation Between Factors

A heatmap was generated to explore the correlations between happiness scores and various socio-economic and environmental factors. The analysis showed strong positive correlations between happiness scores and factors such as GDP per capita, social support, healthy life expectancy, freedom, and low corruption. These findings align with theoretical understanding that well-being is closely linked to economic stability, social connections, and governance quality. Conversely, CO2 production exhibited negative correlations with other factors, suggesting its potential adverse impact on societal well-being. The heatmap not only reinforced existing knowledge but also provided a quantitative basis for examining these relationships further.



3.3 Insight 3: Trends in Happiness Scores for Selected Countries

A line graph was used to track happiness scores for a selection of countries over time. This analysis revealed distinct trends. For instance, Canada maintained a relatively stable score of around 7.0, while France showed a gradual improvement. Conversely, Argentina experienced a sharp decline in happiness scores, reflecting potential socio-economic challenges. Thailand's score peaked in 2018 before slightly declining in subsequent years, whereas Malaysia experienced a sharp drop between 2018 and 2019 but partially recovered by 2021. These country-specific insights underline the complexity of happiness trends and the need to consider local contexts when analyzing global data.

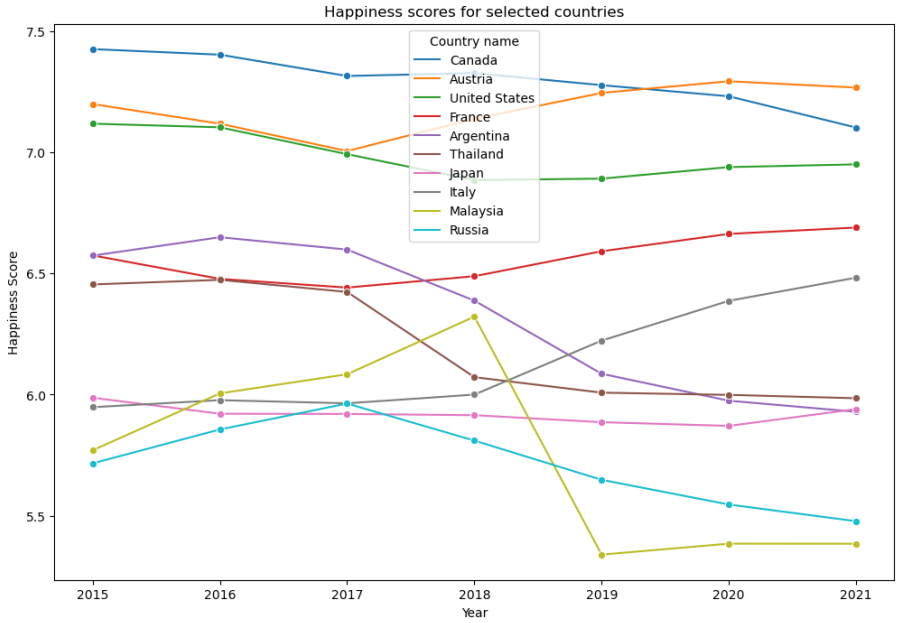


Figure5: Evolution of Happiness Scores across selected countries