Wrangle and Profile

Sherlyne Ndiwa

December 20, 2023

1 Introduction

Analysing happiness report dataset can help identify correlations and trends between socio-economic and environmental factors and the happiness scores for each country. The primary dataset can be found here World Happiness Reports , and the secondary dataset is Global Socio-economic Indicators .

2 Wrangle Stage

2.1 How I joined the datasets

The Primary dataset contains the happiness score, rank and several factors influencing happiness score for each country The second dataset contains socio-economic and environmental indicators for countries. From the second dataset I only needed the Co2 production and the human development index columns to add onto the primary dataset as the influencing factors and analyse how it contributes to the happiness score. I used the Country and Year columns as keys to join the two datasets. I did not have to introduce new keys to join the two datasets, as using the Country name and Year was sufficient.

2.2 Necessary Data Cleaning Steps

I first had to clean the two datasets separately before joining them. The second set of dataset has datapoints until 2021, therefore the years 2022 and 2023 from the happiness report were excluded. In relation, the years prior to 2015 in the second dataset were excluded. Therefore, the year range that was considered for this exercise is from 2015 to 2021. The first step was to drop all the data points for the years not in the consideration for the exercise.

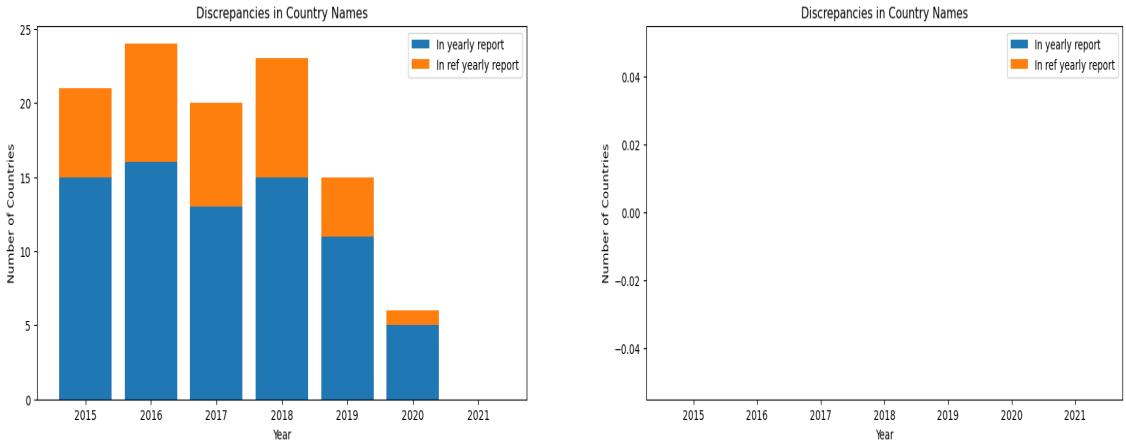
Both datasets did not have any missing values, so I did not have to fix any missing values. From the first dataset, there was inconsistencies with how the column names were named through the years. Since the data for each year was in a different CSV file, I had to align the column names across the yearly datasets. At this point I also dropped the column names that I did not require for my final dataset.

To merge the yearly datasets together, I had to use countries and year as the index. Therefore, I had to check for discrepancies of the country names across the yearly datasets. I decided to use the year 2021 as the reference year to align the country names across the datasets. I therefore had to rectify the discrepancies in the earlier reports to match up with the year 2021. There were a few years completely missing from either the reference year or the other years and I had to drop them entirely.

For the second dataset, I had to extract the full year from the column names for the indicators. I also had to rename the two considered column names: CO2 Production and HDI. The primary and the secondary dataset were now in a better condition to be merged on the Country name and Year index.

2.3 Visuals to Explain the Quality of the Dataset

The was a number of discrepancies in the country names for yearly reports. To fix this, I selected the year 2021 as the reference year since it was the latest year considered for my analysis. It was then possible to ensure that all the country names from 2015 to 2020 matched the ones in 2021.

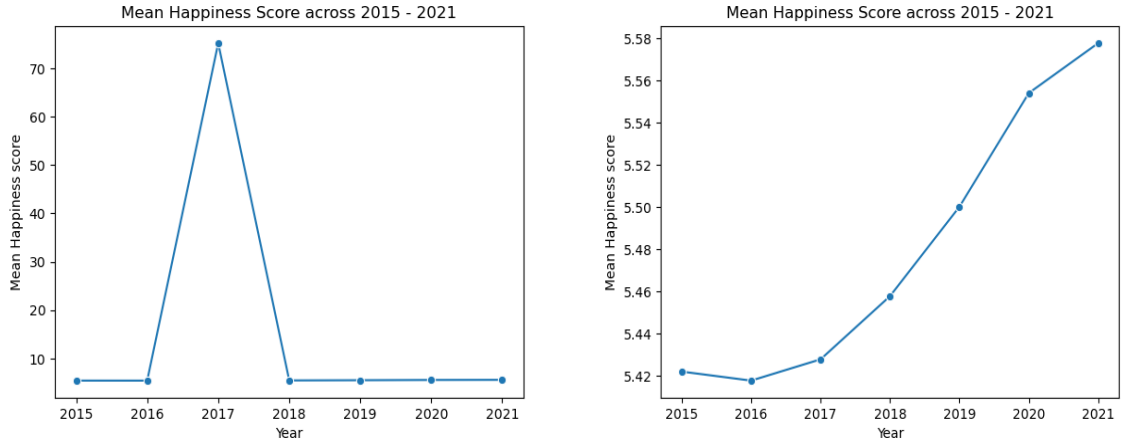


(a) Before Mapping

(b) After Mapping

Figure 1: The number of discrepancies before and after mapping the Country Names

During the process of cleaning the dataset, the column Happiness Rank was missing and had to be introduced for some years basing of the index since it was further based on the Happiness Score. For some reason, for the year 2017, the column names Happiness Score and Happiness Rank were interchanged. This can be seen as the anomaly when visualizing the happiness scores shooting up to around 80 for the year 2017, when in general it should not be more than 10.



(a) Before Fixing the Mix-up

(b) After Fixing the Mix-up

Figure 2: The Mix-up of the column names Happiness Rank for the year 2017

3 Profile Stage

3.1 Informative Insight 1

The box plot graph that displays the distribution of happiness scores across different countries for each year from 2015 to 2021. The y-axis represents the happiness score, which ranges from approximately 3 to 8. Each year is represented by a colored box plot that shows the spread of happiness scores for that year. The median happiness scores have remained relatively stable over these years, with around a score of approximately 5.5. However, there are variations in the spread of scores each year, with some years having wider interquartile ranges. This indicates more variability in reported happiness levels from 2015 to 2021.

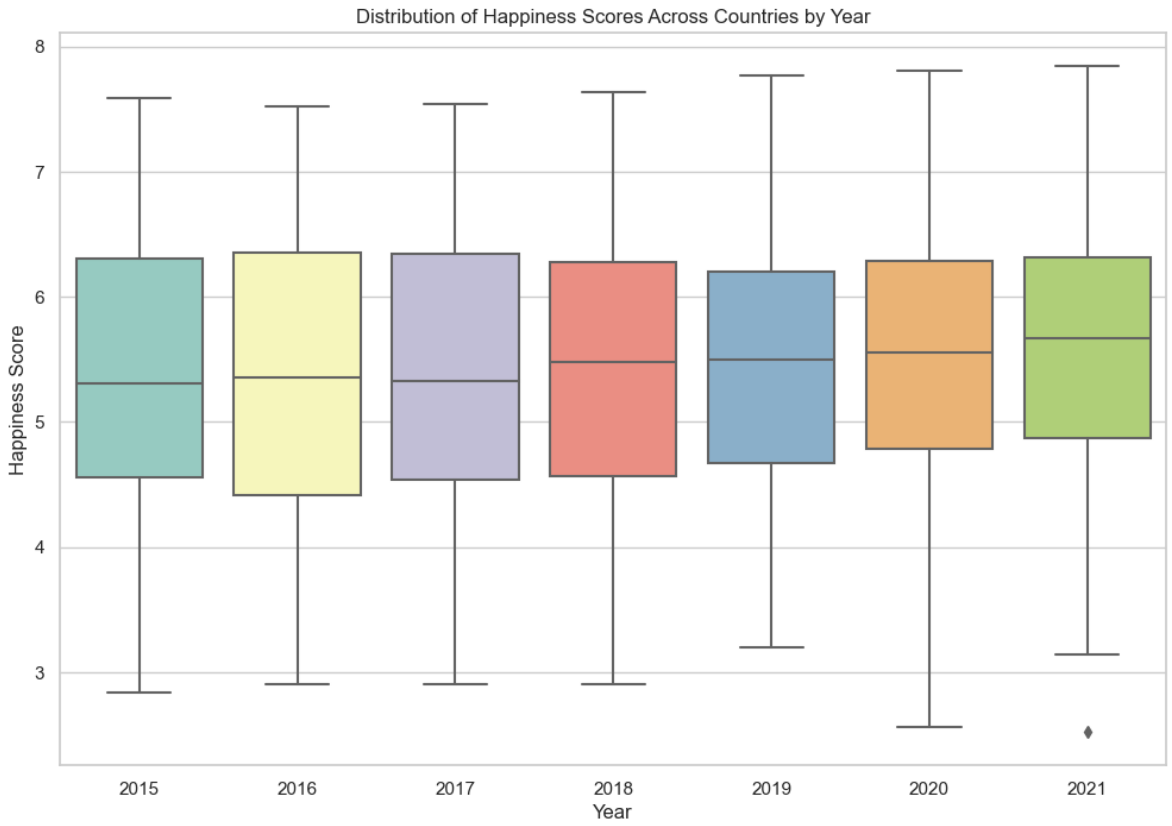


Figure 3: Distribution of Happiness Across Countries

3.2 Informative Insight 2

This is a heatmap that represents the correlation between various factors related to happiness and development.

Figure 4: Correlation Heatmap of the factors related to Happiness Score

There are prominent red squares indicating positive correlations among many factors. The graph shows that the happiness score of a country is generally correlated with the factors that are considered to be important for well-being. Countries with higher GDP per capita, healthier life expectancies, more social support, more freedom, and less corruption tend to have higher happiness scores.

On the other hand, CO2 production mostly shows blue squares, indicating negative correlations with many of the other parameters.

3.3 Informative Insight 3

The line graph is a visual representation of the evolution of happiness scores in selected countries from 2015 to 2021. Canada's happiness score has remained relatively stable around 7.0 throughout the years. Austria has seen a slight increase in its happiness score over time. The United States experienced a slight decrease until 2019, followed by an increase in subsequent years. France's score has fluctuated but shows an overall increase by 2021. Argentina's happiness score has significantly decreased over time. Thailand's score increased until it peaked in 2018 and then slightly decreased in the following years. Both Japan and Italy's scores have remained relatively stable with minor fluctuations. Malaysia experienced a significant drop in happiness score between 2018 and 2019 but partially recovered by 2021. Lastly, Russia's happiness score shows an overall increase over time. These trends in happiness scores could be influenced by a variety of factors and might require further investigation to understand the underlying causes.

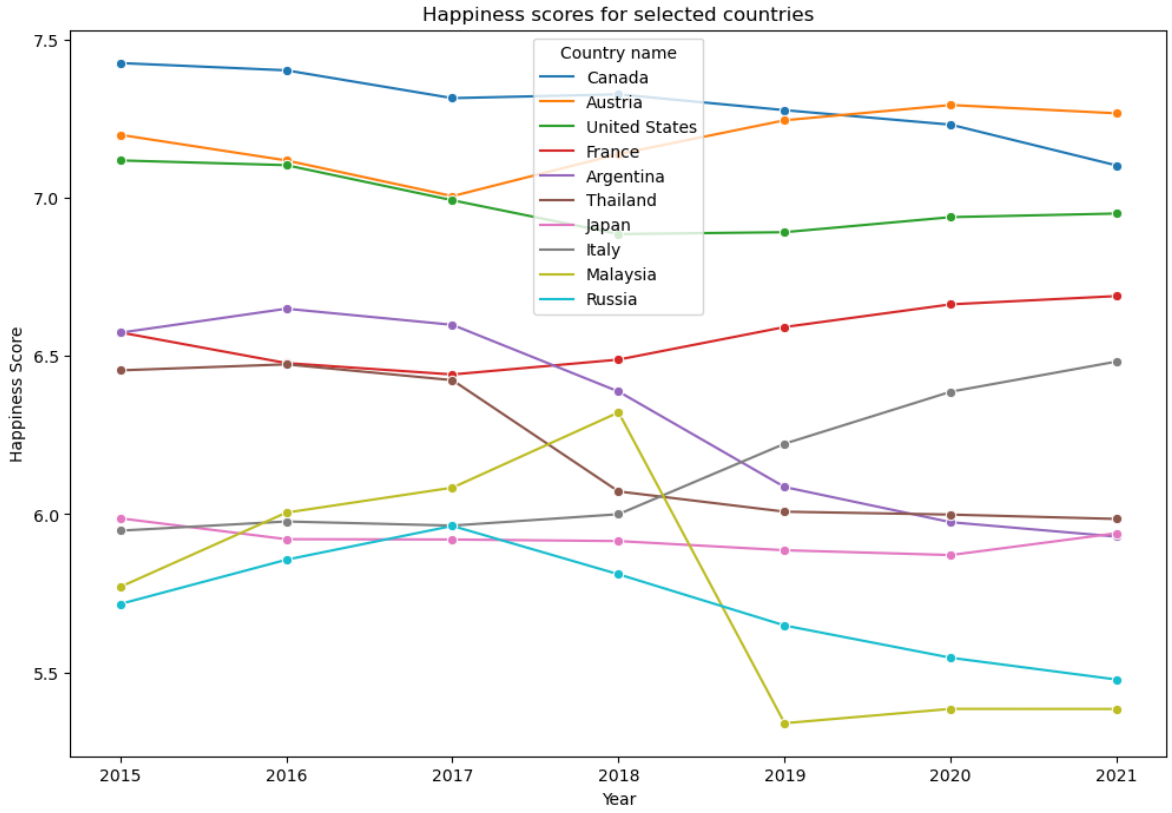


Figure 5: Evolution of Happiness Scores across select countries