|  |  |  |
| --- | --- | --- |
| Checkpoint I | Checkpoint I: Project Proposal | |
| Group: | G06 |
| Date: | 2024/09/23 |
|  |  |

# Problem Domain

Our project aims to explore different relationships between happiness and various interesting factors, such as GDP per capita, geographical location, and average temperature by looking at it from a country’s level perspective. By integrating the *World Happiness Report* (WHR) with other datasets, our goal is to visualize that data in an interactive form, in order to have a deeper understanding of what it represents.

# Task Abstraction

We can ask ourselves interesting questions in an attempt to get some insight into our data:

## Example Questions

* How has world happiness varied over time? (Trend)
* Is there a positive relation between average temperature and happiness? (Correlation)
* What are the top 10 happiest countries in the world, ranked by year? (Ranked by Year)
* How do different factors that contribute to happiness (e.g., GDP, social support) compare across the happiest and least happy countries? (Comparison)
* What proportion of countries report a happiness score above 7, by region? (Proportion)

# Data

Our primary data source for this project is Kaggle. We use these 2 datasets

* World Happiness Report (2015-2023): [World Happiness Report (till 2023) (kaggle.com)](https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors?select=WHR_2019.csv)
* Mean temperature (1901-2022): [Mean temperature for countries by year 1901-2022 (kaggle.com)](https://www.kaggle.com/datasets/palinatx/mean-temperature-for-countries-by-year-2014-2022)

## Initial Dataset

The initial dataset for the WHR was composed of multiple .csv files, one for each year (2015-2023) with a combined 1368 entries and was made up of 10 columns:

* country: Name of the country in each row
* region: Name of the country’s region
* happiness\_score: Happiness score assigned to each country from 0-10
* generosity: Factor to represent how much generosity contributes to happiness
* freedom\_to\_make\_life\_choices: Factor to represent how much freedom contributes to happiness
* perception\_of\_corruption: Factor to represent how much perception of corruption contributed to happiness
* healthy\_life\_expectancy: Factor to represent how much healthy life expectancy contributed to happiness
* social\_support: Factor to represent how much social support contributed to happiness
* gdp\_per\_capita: Factor to represent how much GDP per capita contributed to happiness

*Note*: It is important to consider that the happiness score is not the sum of the factors from generosity to GDP per capita, these factors simply exist in an attempt to explain the score based on those metrics.

The Mean Temperature dataset used was composed of a single .csv file with 5 columns and a total of 20740 entries. Its columns are:

* country: Name of the country in each row
* year: Year of each entry
* annual\_Mean: Average temperature in celsius for a country in a specific year
* 5-yr smooth: Smoothing average for temperature over 5 years
* code: Country’s 3-letter code

## Selected/Derived Data

The relevant columns chosen from both datasets were:

* year
* country
* region
* happiness\_score
* freedom\_to\_make\_life\_choices
* healthy\_life\_expectancy
* perception\_of\_corruption
* generosity
* gdp\_per\_capita
* social\_support
* temperature

## Data Processing

1. Initially, on the WHR dataset, all the yearly .csv files were merged into a single .csv (a new column year was created when merging).
2. Then, all the missing values in the WHR dataset were replaced by their average on other years for their specific country.
3. The values for the year of 2023 were removed from the WHR dataset, because that year is not present in the Mean temperature dataset.
4. Finally both datasets were merged where a new column in the final dataset was created for the mean temperature for each country and for each year. Some countries had different names in the 2 datasets, so when merging their names had to be changed.

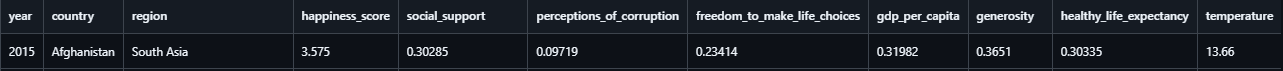
Note: All data processing was done using python with the [pandas library](https://pandas.pydata.org/)

## Data Abstraction

The happiness score will serve as the independent variable, while the independent variables will be used to explore trends and answer our questions.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Semantics** |
| Country | Nominal | Name of the country |
| Region | Nominal | Geographical region |
| Year | Ordinal (Linear, Non-Hierarchichal) | Year of data (2015-2022) |
| Happiness\_score | Quantitative (Ratio, Sequential) | Overall happiness score (0-10) |
| GDP\_per\_capita | Quantitative (Ratio, Sequential) | Natural log of GDP per capita |
| Social\_support | Quantitative (Ratio, Sequential) | National average of binary responses to having someone to count on in times of trouble |
| Healthy\_life\_expectancy | Quantitative (Ratio, Sequential) | Healthy life expectancy at birth (in years) |
| Freedom\_to\_make\_life\_choices | Quantitative (Ratio, Sequential) | National average of binary responses to satisfaction with freedom of life choices |
| Generosity | Quantitative (Ratio, Sequential) | Residual of regressing national average of charitable donations on GDP per capita |
| Perceptions\_of\_corruption | Quantitative (Ratio, Sequential) | Average of binary responses to corruption perception in government and businesses |
| Temperature | Quantitative (Interval, Sequential) | Average yearly temperature (°C) |

## Mapping (Data sample/Questions)



How has world happiness varied over time?

* year
* happiness\_score

Is there a positive relation between Average temperature and happiness?

* happiness\_score
* temperature
* country

What are the top 10 happiest countries in the world, ranked by year?

* year
* country
* happiness\_score

How do different factors that contribute to happiness (e.g., GDP, social support) compare across the happiest and least happy countries?

* happiness\_score
* country
* social\_support
* perceptions\_of\_corruption
* freedom\_to\_make\_life\_choices
* gdp\_per\_capita,generosity
* healthy\_life\_expectancy

What proportion of countries report a happiness score above 7, by region?

* happiness\_score
* country
* region