6.1 Sourcing Open Data

By Claudia Lan Yee Chiu

Table of Contents

World Happiness Report 2015 – 2023

Data Source

Data Description

The *World Happiness Report up to 2023* dataset offers a comprehensive and updated examination of happiness metrics and the factors influencing well-being on a global scale.

Summary of data source

The *World Happiness Report* is a publication of the Sustainable Development Solutions Network, powered by the Gallup World Poll data. Life evaluations from the Gallup World Poll provide the basis for the annual happiness rankings. The *World Happiness Report* is released annually around March 20th as part of the International Day of Happiness celebration.

The dataset used for the analysis, has been merged from 9 CSV files, each listing the same items for each different years and countries starting 2015 until 2023.

Initial Data Sets Used

- WHR_2015.csv
- WHR_2016.csv
- WHR_2017.csv
- WHR_2018.csv
- WHR_2019.csv
- WHR_2020.csv
- WHR_2021.csv
- WHR_2022.csv
- WHR_2023.csv

Data Source: https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors/

Transformed Data Set: WHR_2015_2023.csv

Data Collection

The annual happiness rankings, derived from the Gallup World Poll, rely on responses to the Cantril ladder question. Respondents rate their current lives on a scale from 0 to 10. The rankings use nationally representative samples over a year. Six variables, including GDP per capita, social support, healthy life expectancy, freedom, generosity, and corruption, are analyzed to explain variations in life evaluations across countries. Rankings are solely based on individuals' self-assessments, not an index of these factors. Typically, around 1,000 responses are gathered anually for each country Weights are used to construct population-representative national averages for each year in each country.

Reason for choosing this data set

The World Happiness Report highlights a global call for prioritizing happiness and well-being in governmental policies. This analysis seeks to explore the key factors shaping a nation's happiness and identify the happiest countries. Understanding these factors can significantly improve a country's citizens' quality of life and culture, especially in the wake of the transformative impact of the COVID-19 pandemic in 2020, which reshaped the world's outlook on life and the concept of happiness.

Data Limitations

The information is sourced from a well-established global yearly survey, indicating the reliability of this data.
Nonetheless, it is important to note that the data utilized was created and modified by Kaggle community members, which could introduce the possibility of human errors or biases.
Furthermore, it is worth mentioning that we lack consistent data for every country in each year due to the fact that some countries do not participate in the survey on an annual basis.

Ethics

The dataset does not include any personal information or sensitive variables, so there is no need for Personal Data Protection or Privacy Act (PLA) security measures.

Data Profiling

Data Set Name	Number of Rows	Number of Columns	Amendments done
WHR_2015	158	9	None
WHR_2016	157	9	None
WHR_2017	155	9	None
WHR_2018	156	9	United Arab Emirates, ´perceptions_of_corruption´ is NaN, whilst all other years have an update. Replaced with the mean of other years available
WHR_2019	156	9	None
WHR_2020	153	9	None
WHR_2021	149	9	None
WHR_2022	146	9	None
WHR_2023	137	9	'State of Palestine', healthy_life_expectancy is NaN. Dropped this line since Palestine only appears in report of 2023
WHR_2015_2023	1366	10	Changed Year from Object to Type Added Overall Rank column Removed Index for Country

Data Cleaning & Consistency of new merged dataset: WHR_2015_2023

Variables	Data Type	Explanation
Country	Character	Country taking part of the survey
Region	Character	Geographic region of a country is part of
Happiness_score	Numeric	Happiness score or subjective well-being, also known as ´ladder score´ or (´Cantril life ladder´). Respondents are asked to think of a ladder, with the best possible life for them being a 10 and the worst possible life being a 0.
GDP_per_Capita	Numeric	Gross Domestic Product measurement per its population
Social_support	Numeric	Support of friends and family assisting in times of need or crisis. Social support improves the quality of life and provides a buffer against adverse life events.
Healthy_life_expectancy	Numeric	Healthy life expectancy is the average timespan in years that a newborn can expect to live when enjoying a good health
Freedom_to_make_life_choices	Numeric	Freedom to make life choices is the national average of binary responses to the *GWP question ´Are you satisfied or dissatisfied with your freedom to choose what you do with your life?`
Generosity	Numeric	Is the residual of regressing the national average of *GWP responses to the donation question ´Have you donated money to a charity in the past month?´
Perceptions_of_corruption	Numeric	Is the average of binary answers to two *GWP questions: ´Is corruption widespread within business or not?` and ´Is corruption widespread throughout the government or not? `
Newly added: Year	Numeric	Changed from object to int type
Newly added: Overall_rank	Numeric	
Removed Index and replaced with ´Country´ above		

** GWP: Gallup World Poll, which remains the principal source of data for the reports, provide the basis for the annual happiness rankings.*

Data Profiling of merged dataset

	Data Types			
Variables	**Time Component: Invariant or Variant?**	**Structured or Unstructured?**	**Qualitative or Quantitative?**	**Nominal or Ordinal? Discrete or Continuous?**
Country	Invariant	Structured	Qualitative	Nominal
Region	Invariant	Structured	Qualitative	Nominal
Happiness_score	Variant	Structured	Quantitative	Continuous
GDP_per_Capita	Variant	Structured	Quantitative	Continuous
Social_support	Variant	Structured	Quantitative	Continuous
Healthy_life_expectancy	Variant	Structured	Quantitative	Continuous
Freedom_to_make_life_choices	Variant	Structured	Quantitative	Continuous
Generosity	Variant	Structured	Quantitative	Continuous
Perceptions_of_corruption	Variant	Structured	Quantitative	Continuous
Year	Invariant	Structured	Qualitative	Nominal
Overall_rank	Variant	Structured	Qualitative	Discrete

Descriptive Statistical Analysis

	happiness_score	**gdp_per_capita**	**social_support**	**healthy_life_expectancy**	**freedom_to_make_life_choices**	**generosity**	**perceptions_of_corruption**	**year**	**overall_rank**
count	1366	1366	1366	1366	1366	1366	1366	1366	1366
mean	5.441476	1.019331	1.045141	0.584043	0.450797	0.196356	0.132426	2018.900439	76.564422
std	1.118226	0.453856	0.331207	0.245117	0.156787	0.113287	0.112606	2.559541	44.100799
min	1.859	0	0	0	0	0	0	2015	1
25%	4.59775	0.696163	0.832011	0.402301	0.356	0.115002	0.056826	2017	38.25
50%	5.4481	1.042	1.08284	0.61283	0.467672	0.182825	0.097095	2019	76
75%	6.25695	1.33871	1.299028	0.777614	0.568842	0.252909	0.16675	2021	114
max	7.842	2.209	1.644	1.141	0.772	0.838075	0.587	2023	158

Defining Questions to Explore

- What factors contribute most to a country´s happiness score?
- What regions have the happiest countries? Lowest happiness scores?
- How has COVID-19 altered different aspects of life in the happiness scores during the pandemy in 2020 to 2021?
- What is the difference from the top 10 countries with the highest happiness score compared to the top 10 lowest happiness scores?