Giada Griso
May 6, 2024
Data Immersion Course

Exercise 6.1: Sourcing Open Data

Section 1: Data Source

The datasets I have chosen for Achievement 6 are part of a collection called World Happiness Reports 2013 – 2023 on Kaggle. I will analyse the reports from 2019 to 2023.

The data was manually downloaded by the datasets owner from https://worldhappiness.report/ and is available under the Community Data License Agreement - Permissive - Version 1.0.

Given its global significance, I decided to delve into the current state of global happiness and investigate the factors influencing it, aiming to uncover parallels and distinctions in well-being across different cultures and regions.

I used ChatGPT to help me create this document and get ideas on how to approach the analysis.

Data Source Summary

The World Happiness Report is a partnership of Gallup, the Oxford Wellbeing Research Centre, the UN Sustainable Development Solutions Network, and the WHR's Editorial Board. The report is produced under the editorial control of the WHR Editorial Board.
These organizations are well-established in their respective fields and are known for their rigorous research methodologies and standards, suggesting a high level of credibility and trustworthiness in the data.

Data Collection Summary

The Gallup World Poll (GWP) is a comprehensive global survey that tracks key issues such as food access, employment, leadership performance, and well-being across more than 160 countries. It employs telephone surveys in countries with high telephone coverage and face-to-face interviews in the developing world. The surveys are conducted using probability-based sampling, ensuring national representativeness. Interviews typically last around one hour for face-to-face and 30 minutes for telephone surveys. Each survey includes at least 1,000 individuals, with oversampling in some cases. The frequency of surveys varies by country. Samples are weighted to correct for unequal selection probability, nonresponse, and demographic imbalances, ensuring the reliability of the data collected.
Source: https://www.gallup.com/178667/gallup-world-poll-work.aspx

Data Contents

I will analyse 5 different datasets, each one being a World Happiness Report spanning from 2019 to 2023. Each dataset contains different variables, but the following appear in all reports:

- **Happiness Score (also called *Ladder score* or just *Score*)**

Unless stated otherwise, it is the national average response to the question of life evaluations. Respondents rate their life satisfaction on a scale from 0 (worst possible life) to 10 (best possible life) using the *Cantril life ladder* question.

- **GDP per Capita**

  GDP figures are sourced from either the World Development Indicators (WDI) or the Penn World Table. Some GDP per capita values are estimated using country-specific forecasts of real GDP growth from OECD Economic Outlook and World Bank's Global Economic Prospects.

- **Healthy Life Expectancy (HLE)**

  Healthy life expectancies at birth are extracted from the World Health Organization's (WHO) Global Health Observatory data repository. Interpolation and extrapolation are used to match the sample periods of the reports.

- **Social Support**

  National average of binary responses (0 or 1) to the GWP question: "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"

- **Freedom to Make Life Choices**

  National average of responses to the GWP question: "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"

- **Generosity**

  Residual of regressing the national average of response to the GWP question: "Have you donated money to a charity in the past month?" on GDP per capita.

- **Corruption Perception**

  National average of survey responses to questions in the GWP regarding the perception of corruption in government and businesses. Overall perception is the average of responses, with business corruption perception used if government corruption perception is missing.

Note on "Dystopia + residual" columns

Dystopia serves as an imaginary benchmark representing the world's least-happy population. It allows for favorable comparisons across countries, ensuring that no nation performs worse than Dystopia in terms of the six key variables: income, life expectancy, generosity, corruption perception, freedom, and social support. Life in Dystopia would be characterized by low incomes, short life expectancy, minimal generosity, high corruption, limited freedom, and little social support, making it a symbol of extreme unhappiness.

Residuals represent the unexplained components of happiness evaluations for each country, indicating the extent to which the six key variables either over- or under-explain the average happiness evaluations in each country. The column "Dystopia + residual" combines the residuals with the estimate for happiness level in Dystopia. This makes sure that the combined result is always positive. Essentially, this column accounts for the unexplained portion of happiness evaluations in each country and adjusts it relative to the benchmark set by Dystopia.

Limitations and ethical considerations

- **Sampling bias:** the data collection methods, such as telephone surveys and face-to-face interviews, may introduce sampling bias. For example, certain populations, such as those

without access to phones or living in remote areas, may be underrepresented in the surveys, leading to a skewed perspective of happiness levels.

- **Data availability bias:** some countries may lack comprehensive data for all the factors included in the report, leading to gaps in the analysis and potentially skewing the overall findings. The absence of comprehensive data for certain countries can create biases by disproportionately representing countries with more robust data collection systems or those more willing to participate in surveys. This can result in an incomplete or skewed picture of global happiness trends, potentially underestimating or overestimating the well-being of specific populations.
- **Subjectivity of responses:** responses to survey questions about happiness and life satisfaction are inherently subjective and may be influenced by various factors such as mood, context, and cultural norms. This subjectivity can introduce measurement error and affect the reliability of the data.
- **Social desirability bias:** the data relies heavily on self-reported measures of happiness and life satisfaction, which may be influenced by social desirability bias. Respondents may provide answers that they believe are socially acceptable rather than reflecting their true feelings.
- **Limited variables:** while the World Happiness Reports consider multiple factors contributing to happiness, they may not capture all relevant variables. For example, factors like environmental quality, access to healthcare, and political stability may also influence well-being but are not included in the analysis.
- **Privacy concerns:** it is essential to ensure that respondents understand the purpose of the survey, how their data will be used, and that their participation is voluntary.

While the reports offer a comprehensive overview of happiness levels worldwide, users should interpret the data cautiously and consider the context of each country's data collection methods and limitations. With proper acknowledgment of these limitations, the World Happiness Reports can still provide valuable insights into global well-being trends over time.

Section 2: Data Profile

Data Shape

	Rows	Columns
World Happiness Report 2019	156	9
World Happiness Report 2020	153	20
World Happiness Report 2021	149	20
World Happiness Report 2022	146	12
World Happiness Report 2023	137	19

***Note on Data Shape:** the number of countries included in each dataset diminishes over time (156 countries in 2019 versus 137 countries in 2023). This reduction in representation could lead to **biases in the analysis**, as certain regions or populations may be underrepresented or excluded altogether. Changes in the composition of the sample may affect the interpretation of trends and

make it challenging to assess whether observed changes are reflective of true shifts in happiness levels or artifacts of sample composition.

Data Profile

The first 8 variables in the table below are common to all five datasets, while the remaining variables appear in different datasets. The exact names may change, but the variables evaluate the same factors.

Variable	Time-variant/invariant	Structured/ Unstructured	Qualitative/ Quantitative	Qualitative: Nominal/Ordinal Quantitative: Discrete/Continuous
Country	Time-invariant	Structured	Qualitative	Nominal
Ladder/Happiness score	Time-variant	Structured	Quantitative	Continuous
GDP per capita	Time-variant	Structured	Quantitative	Continuous
Healthy life expectancy	Time-variant	Structured	Quantitative	Continuous
Social support	Time-variant	Structured	Quantitative	Continuous
Freedom to make life choices	Time-variant	Structured	Quantitative	Continuous
Generosity	Time-variant	Structured	Quantitative	Continuous
Perceptions of corruption	Time-variant	Structured	Quantitative	Continuous
Rank	Time-variant	Structured	Quantitative	Discrete
Regional indicator	Time-invariant	Structured	Qualitative	Nominal
Standard error of ladder score	Time-invariant	Structured	Quantitative	Continuous
upperwhisker	Time-invariant	Structured	Quantitative	Continuous
lowerwhisker	Time-invariant	Structured	Quantitative	Continuous
Ladder score in Dystopia	Time-variant	Structured	Quantitative	Continuous
Dystopia + residual	Time-variant	Structured	Quantitative	Continuous
Explained by	Time-variant	Structured	Quantitative	Continuous

Consistency Checks

	Missing Values	Duplicates	Mixed-type Data
World Happiness Report 2019	N/A	N/A	N/A
World Happiness Report 2020	N/A	N/A	N/A
World Happiness Report 2021	N/A	N/A	N/A
World Happiness Report 2022	N/A	N/A	N/A
World Happiness Report 2023	3 missing values	N/A	N/A

The datasets appear to be quite clean, with the exception of 3 missing values in the World Happiness Report 2023: 1 missing value in the “Healthy life expectancy” column, a second one in the “Explained by: Healthy life expectancy” column and a third one in the “Dystopia + residual”

column. As the missing values are only 3 in total, I will keep them in the dataset because their presence will not create any additional bias.

Descriptive Statistics

World Happiness Report 2019

	Overall rank	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000	156.000000
mean	78.500000	5.407096	0.905147	1.208814	0.725244	0.392571	0.184846	0.110603
std	45.177428	1.113120	0.398389	0.299191	0.242124	0.143289	0.095254	0.094538
min	1.000000	2.853000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	39.750000	4.544500	0.602750	1.055750	0.547750	0.308000	0.108750	0.047000
50%	78.500000	5.379500	0.960000	1.271500	0.789000	0.417000	0.177500	0.085500
75%	117.250000	6.184500	1.232500	1.452500	0.881750	0.507250	0.248250	0.141250
max	156.000000	7.769000	1.684000	1.624000	1.141000	0.631000	0.566000	0.453000

World Happiness Report 2020

	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	1
count	153.00000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	1.530000e+02	
mean	5.47324	0.053538	5.578175	5.368304	9.295706	0.808721	64.445529	0.783360	-0.014568	0.733120	1.972317e+00	
std	1.11227	0.018183	1.096823	1.128631	1.201588	0.121453	7.057848	0.117786	0.151809	0.175172	1.559417e-15	
min	2.56690	0.025902	2.628270	2.505530	6.492642	0.319460	45.200001	0.396573	-0.300907	0.109784	1.972317e+00	
25%	4.72410	0.040698	4.826248	4.603149	8.350645	0.737217	58.961712	0.714839	-0.127015	0.683019	1.972317e+00	
50%	5.51500	0.050606	5.607728	5.430644	9.456313	0.829204	66.305145	0.799805	-0.033665	0.783122	1.972317e+00	
75%	6.22850	0.060677	6.363886	6.138881	10.265124	0.906747	69.289192	0.877709	0.085429	0.849151	1.972317e+00	
max	7.80870	0.120590	7.869766	7.747634	11.450681	0.974670	76.804581	0.974998	0.560664	0.935585	1.972317e+00	

Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
153.000000	153.000000	153.000000	153.000000	153.000000	153.000000	153.000000
0.868771	1.155607	0.692869	0.463583	0.189375	0.130718	1.972317
0.372416	0.286866	0.254094	0.141172	0.100401	0.113097	0.563638
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.257241
0.575862	0.986718	0.495443	0.381457	0.115006	0.055805	1.629928
0.918549	1.203987	0.759818	0.483293	0.176745	0.098435	2.046272
1.169229	1.387139	0.867249	0.576665	0.255510	0.163064	2.350267
1.536676	1.547567	1.137814	0.693270	0.569814	0.533162	3.440810

In this dataset we can observe the presence of the usual 6 metrics contributing to happiness scores, but in two different versions. The "Explained by" column, from a glance, seems to be the proportion of the ladder score that the corresponding metric contributes. The columns circled with blue have very different and even negative minimum and maximum values, while the ones circled with yellow have all the same minimum (=0,00) and maximum values in the range 0,5–1,6. The variables circled with yellow in this dataset have more similar statistics to the ones in the 2019 and 2022 datasets when compared to the variables circled in blue, therefore I think they

are the ones that should be used for further analysis. The columns circled with blue can be dropped, as they seem to be the result of different calculations.

World Happiness Report 2021

	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia
count	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	1.490000e+02
mean	5.532839	0.058752	5.648007	5.417631	9.432208	0.814745	64.992799	0.791597	-0.015134	0.727450	2.430000e+00
std	1.073924	0.022001	1.054330	1.094879	1.158601	0.114889	6.762043	0.113332	0.150657	0.179226	5.347044e-15
min	2.523000	0.026000	2.596000	2.449000	6.635000	0.463000	48.478000	0.382000	-0.288000	0.082000	2.430000e+00
25%	4.852000	0.043000	4.991000	4.706000	8.541000	0.750000	59.802000	0.718000	-0.126000	0.667000	2.430000e+00
50%	5.534000	0.054000	5.625000	5.413000	9.569000	0.832000	66.603000	0.804000	-0.036000	0.781000	2.430000e+00
75%	6.255000	0.070000	6.344000	6.128000	10.421000	0.905000	69.600000	0.877000	0.079000	0.845000	2.430000e+00
max	7.842000	0.173000	7.904000	7.780000	11.647000	0.983000	76.953000	0.970000	0.542000	0.939000	2.430000e+00

Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000
0.977161	0.793315	0.520161	0.498711	0.178047	0.135141	2.430329
0.404740	0.258871	0.213019	0.137888	0.098270	0.114361	0.537645
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.648000
0.666000	0.647000	0.357000	0.409000	0.105000	0.060000	2.138000
1.025000	0.832000	0.571000	0.514000	0.164000	0.101000	2.509000
1.323000	0.996000	0.665000	0.603000	0.239000	0.174000	2.794000
1.751000	1.172000	0.897000	0.716000	0.541000	0.547000	3.482000

The situation with the yellow-circled and blue-circled columns is the same described above.

World Happiness Report 2022

	RANK	Happiness score	Whisker-high	Whisker-low	Dystopia (1.83) + residual	Explained by: GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption
count	146.000000	146.000000	146.000000	146.000000	146.000000	146.000000	146.000000	146.000000	146.000000	146.000000	146.000000
mean	73.500000	5.553575	5.673589	5.433568	1.831808	1.410445	0.905863	0.586171	0.517226	0.147377	0.154781
std	42.290661	1.086843	1.065621	1.109380	0.534994	0.421663	0.280122	0.176336	0.145859	0.082799	0.127514
min	1.000000	2.404000	2.469000	2.339000	0.187000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	37.250000	4.888750	5.006250	4.754750	1.555250	1.095500	0.732000	0.463250	0.440500	0.089000	0.068250
50%	73.500000	5.568500	5.680000	5.453000	1.894500	1.445500	0.957500	0.621500	0.543500	0.132500	0.119500
75%	109.750000	6.305000	6.448750	6.190000	2.153000	1.784750	1.114250	0.719750	0.626000	0.197750	0.198500
max	146.000000	7.821000	7.886000	7.756000	2.844000	2.209000	1.320000	0.942000	0.740000	0.468000	0.587000

World Happiness Report 2023

	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia
count	137.000000	137.000000	137.000000	137.000000	137.000000	137.000000	136.000000	137.000000	137.000000	137.000000	1.370000e+02
mean	5.539796	0.064715	5.666526	5.412971	9.449796	0.799073	64.967632	0.787394	0.022431	0.725401	1.778000e+00
std	1.139929	0.023031	1.117421	1.163724	1.207302	0.129222	5.750390	0.112371	0.141707	0.176956	2.897173e-15
min	1.859000	0.029000	1.923000	1.795000	5.527000	0.341000	51.530000	0.382000	-0.254000	0.146000	1.778000e+00
25%	4.724000	0.047000	4.980000	4.496000	8.591000	0.722000	60.648500	0.724000	-0.074000	0.668000	1.778000e+00
50%	5.684000	0.060000	5.797000	5.529000	9.567000	0.827000	65.837500	0.801000	0.001000	0.774000	1.778000e+00
75%	6.334000	0.077000	6.441000	6.243000	10.540000	0.896000	69.412500	0.874000	0.117000	0.846000	1.778000e+00
max	7.804000	0.147000	7.875000	7.733000	11.660000	0.983000	77.280000	0.961000	0.531000	0.929000	1.778000e+00

Explained by: Log GDP per capita	Explained by: Social support	Explained by: Healthy life expectancy	Explained by: Freedom to make life choices	Explained by: Generosity	Explained by: Perceptions of corruption	Dystopia + residual
137.000000	137.000000	136.000000	137.000000	137.000000	137.000000	136.000000
1.406985	1.156212	0.366176	0.540000	0.148474	0.145898	1.777838
0.432963	0.326322	0.156691	0.149501	0.076053	0.126723	0.504390
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-0.110000
1.099000	0.962000	0.248500	0.455000	0.097000	0.060000	1.555250
1.449000	1.227000	0.389500	0.557000	0.137000	0.111000	1.848500
1.798000	1.401000	0.487500	0.656000	0.199000	0.187000	2.078750
2.200000	1.620000	0.702000	0.772000	0.422000	0.561000	2.955000

The situation with the yellow-circled and blue-circled columns is the same described above.

Wrangling Steps

First of all, I removed the blue-circled columns from their respective datasets, as they are not useful for the purpose of this analysis.

I then removed other columns from each dataset that are not shared with the others, except for the "Rank" and the "Regional indicator" columns, which could provide valuable insights for analysis. Specifically, I dropped the standard error, upper- and lower-whisker, ladder score in Dystopia and Dystopia residual columns from their respective datasets.

As a third step, I renamed the remaining columns so there is consistency across the datasets. The new names for the remaining columns are the following:

Country

Region

Year

Happiness_rank

Happiness_score

GDP_capita

Social_support

Life_expectancy

Freedom

Generosity

Corruption_perception

As a fourth step, I re-orderd the columns following a specif preferencial order and created a new column called “Year” with the corresponding value for each dataset.

After making sure the datasets are uniform and have the same columns, I combined them using the pd.concat function in Python and exported the combined_df as a csv file.

Section 3: Questions to Explore

1. How does the distribution of happiness scores vary across different regions and years?
2. Is there a correlation between GDP per capita and happiness scores?
3. How does social support correlate to happiness scores?
4. What are the differences in average life expectancy between the happiest and least happy countries?
5. What is the relationship between a country's level of freedom and its happiness score?
6. How does the degree of generosity in a country relate to its happiness score?
7. Is there a correlation between the level of corruption perception and happiness score?
8. Which factors have the strongest impact on happiness scores across different regions and years?
9. Are there any outliers or anomalies in the data that require further investigation?
10. Which countries consistently maintain high or low happiness rankings over multiple years?
11. Do significant differences exist in happiness scores between developed and developing countries?
12. Have there been notable fluctuations in happiness scores following major political or economic events?
13. How do societal factors, such as corruption perception and generosity, contribute to variations in happiness scores across regions?
14. What recommendations can be made to policymakers to improve overall well-being in different parts of the world?