World Happiness Report

Data source

Data Sourcing

This is an external data source. The data World Happiness Report provided by Kaggle originates from the Gallup World Poll, which is conducted by the Gallup Institute, a reputable organization known for its global public opinion polls. We can verify this as a trustworthy data source.

Data Collection

Smapling: Organization uses a representative sampling method to select participants for their surveys. Approximately 1000 participants from more than 150 countries are surveyed for happiness levels.
Data Collection Frequency: Gallup conducts these surveys on an annual basis, which allows to track changes in happiness and well-being over time.
Method: surveys are conducted in the form of face-to-face or telephone interviews.
Survey Questions: Gallup includes a series of questions in their surveys to measure various aspects that influence happiness level.
Metrics: Gallup uses standardized metrics and scales to assess happiness and well-being. This is known as the Cantril ladder. They ask respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale.

Data Contents

The dataset consists of 5 tables, each for one year, from 2015 to 2019.
Each of tables include country name, region, country's respective positions in the happiness ranking, and the overall happiness scores for each country. The columns following the happiness score estimate the extent to which each of six factors – GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption – contribute to making life evaluations higher in each country than they are in Dystopia (a hypothetical country that has values equal to the world's lowest national averages for each of the six factors).

Limitations and Ethical Considerations

Sampling bias: The survey was conducted on a sample of about a thousand people from each country, we need to ensure that this sample reflect whole population.
Exclusion Bias: The dataset relies on responses from the Gallup World Poll, which may not always be fully representative of all demographic groups within each country. There may be concerns about underrepresented or marginalized groups whose experiences might not be adequately captured.
Temporal Coverage: The dataset covers the years 2015-2019. While it provides valuable insights for those specific years, it may not reflect more recent changes in happiness levels or factors influencing happiness in later years.

Exercise 6.1 Sourcing Open Data

Limited Geographical Coverage: The dataset includes a specific set of countries, and not all countries in the world are represented, not exactly same countries are represented every year.
Limited Factors: The dataset focuses on a limited set of factors that contribute to happiness. Other important variables, such as environmental factors, political stability, and cultural aspects, are not included, which may provide a more comprehensive understanding of happiness.
Contextual bias: The dataset lacks detailed information about the contextual factors that might influence individual and national happiness, such as cultural norms, historical conditions, or current political situation.
Subjectivity bias: The happiness scores are based on respondents' self-assessment using the Cantril ladder, which asks individuals to rate their own happiness on a scale from 0 to 10. This can be influenced by cultural, social, and individual biases even health condition or current mood. Also different people around the word may have different interpretations of happiness
Cultural bias: Cultural differences can significantly impact how people perceive and report happiness.

Relevancy

Since this dataset was recommended in the Project Brief, I assume it is relevant to the project. I have ensured that it meets all the required conditions:
- is open-source
- comes from an authentic source
- includes non-anonymized column names
- is not more than 10 years old
- Contain at least 2-3 continuous variables and 2-3 categorical variables
- contains appropriate number of rows
- includes a geographical object.

Why I've chosen this data set

I selected the World Happiness Report dataset for my project due to its relevance to the project requirements, but there are deeper reasons behind this choice.
I have often heard about rankings of the happiest countries in the world, but I was curious about the methodology behind these rankings and how happy the countries outside the top places in the rankings were. This project provides an excellent opportunity to delve deeper into this topic.
The World Happiness Report is highly respected and widely recognized as a tool for measuring global happiness levels, and as a result, there is a wealth of information available on the Internet about the Gallup Happiness Report. I'm excited about the prospect of doing more comprehensive research in the coming days (or weeks).
Moreover, I think it will be particularly interesting to later extend the analysis to 2020-2023 and compare it with pre-2020 data. As we know, significant global events have been taking place since 2020 that are likely to have a significant impact on people's happiness levels.

Exercise 6.1 Sourcing Open Data

Data Profile

Variable	Description	Time-variant / invariant	Structured / Unstructured	Qualitative / Quantitative	Nominal / Ordinal, Discrete/ Continuous
Country	Country name	Invariant	Structured	Qualitative	Nominal
Region	The region to which a country belongs.	Invariant	Structured	Qualitative	Nominal
Happiness Rank	Rank of the country based on the Happiness Score	Variant	Structured	Quantitative	Discrete
Happiness Score	The Happiness Index calculated by averaging the survey results of (happiness rate on a scale from 0 to 10)	Variant	Structured	Quantitative	Continuous
Standard Error	The standard error of the happiness score	Variant	Structured	Quantitative	Continuous
Economy (GDP per Capita)	The extent to which GDP contributes to the calculation of the Happiness Score.	Variant	Structured	Quantitative	Continuous
Family	The extent to which Family contributes to the calculation of the Happiness Score	Variant	Structured	Quantitative	Continuous
Health (Life Expectancy)	The extent to which Life expectancy contributed to the calculation of the Happiness Score	Variant	Structured	Quantitative	Continuous
Freedom	The extent to which Freedom contributed to the calculation of the Happiness Score.	Variant	Structured	Quantitative	Continuous
Trust (Government Corruption)	The extent to which Perception of Corruption contributes to Happiness Score	Variant	Structured	Quantitative	Continuous
Generosity	The extent to which Generosity contributed to the calculation of the Happiness Score	Variant	Structured	Quantitative	Continuous
Dystopia Residual	The extent to which Dystopia Residual contributed to the calculation of the Happiness Score	Variant	Structured	Quantitative	Continuous

Wrangling steps

Data set	Columns dropped	Column renamed	Data type changed	Column created
df_2015	'Standard Error', 'Dystopia Residual'	'Happiness Rank' : 'Happiness_Rank', 'Happiness Score' : 'Happiness_Score', 'Economy (GDP per Capita)' : 'GDP per Capita', 'Health (Life Expectancy)' : 'Health', 'Trust (Government Corruption)' : 'Corruption'		Year
df_2016	'Lower Confidence Interval', 'Upper Confidence Interval', 'Dystopia Residual'	'Happiness Rank' : 'Happiness_Rank', 'Happiness Score' : 'Happiness_Score', 'Economy (GDP per Capita)' : 'GDP per Capita', 'Health (Life Expectancy)' : 'Health', 'Trust (Government Corruption)' : 'Corruption'		Year
df_2017	'Whisker.high', 'Whisker.low', 'Dystopia.Residual'	'Happiness.Rank' : 'Happiness_Rank', 'Happiness.Score' : 'Happiness_Score', 'Economy..GDP.per.Capita.' : 'GDP per Capita', 'Health..Life.Expectancy.' : 'Health', 'Trust..Government.Corruption.' : 'Corruption'		Year Region
df_2018		'Overall rank' : 'Happiness_Rank', 'Country or region' : 'Country' 'Score' : 'Happiness_Score', 'GDP per capita' : 'GDP_per_Capita', 'Social support' : 'Family' 'Healthy life expectancy' : 'Health', 'Freedom to make life choices' : 'Freedom' 'Perceptions of corruption' : 'Corruption'		Year Region
df_2019		'Overall rank' : 'Happiness_Rank', 'Country or region' : 'Country' 'Score' : 'Happiness_Score', 'GDP per capita' : 'GDP_per_Capita', 'Social support' : 'Family' 'Healthy life expectancy' : 'Health', 'Freedom to make life choices' : 'Freedom' 'Perceptions of corruption' : 'Corruption'		Year Region
df	‚Gambia‘ (row) as it only appears in 2019	‘Country’ names: 'Taiwan Province of China' : 'Taiwan', d'Hong Kong S.A.R., China' : 'Hong Kong', 'Trinidad & Tobago' : 'Trinidad and Tobago', 'Northern Cyprus' : 'North Cyprus', 'North Macedonia' : 'Macedonia'	‚Region‘-Category,	

Other: Setting column ‚Country‘ as an Index

Consistency checks

Data set	Missing values	Addressing missing values	Duplicates
df	Region value for : 'Hong Kong S.A.R., China', 'Trinidad & Tobago', 'Northern Cyprus', 'North Macedonia', 'Gambia'. One missing value in 'corruption' column – replaced with mean.	County names replaced with correct ones, and regions mapped with regions_dic, 'Gambia' - removed	No duplicates

Questions to explore

1. What are the most and the least happy countries each year?
2. Considering the average for all years, how does the ranking change?
3. How happy is my home country according to Gallup Happines Report?
4. How indicators contribute to the happiness scores in my country?
5. What are the most and least happy regions of world according to the World Happiness Report in last years?
6. Did any countries experience significant changes in their happiness scores or rankings between the 2015 and 2019?
7. How do the indicators affect the happines level?
8. How are the indicators correlated?