

Sourcing Open Data

Exercise 6.1

Vineeta Sinha

07/03/2024

Data Source - <https://www.kaggle.com/datasets/unsdsn/world-happiness>

Data Source Summary -

This dataset provides a comprehensive view of the factors contributing to happiness across different countries, allowing for detailed analysis and insights into global well-being. World Happiness Report dataset is a survey based data and is generally viewed as a reliable and valuable resource for understanding global happiness trends. The dataset might not capture all dimensions of well-being and may be influenced by economic and social reporting biases.

Data Collection Method -

The data collection method for the World Happiness Report involves

Gallup World Poll: The primary source of data is the Gallup World Poll, which conducts surveys in over 150 countries. The poll uses a standardized questionnaire to ensure consistency and comparability across different nations.

Representative Samples: The Gallup World Poll uses probability sampling to ensure that the sample is representative of the entire population. This often involves stratified sampling, where the population is divided into subgroups (strata) and random samples are taken from each subgroup.

Survey Administration: Surveys are conducted face-to-face or via telephone, depending on the country's infrastructure and accessibility. In-person interviews are typically used in regions where telephone penetration is low.

Questionnaire: The questionnaire includes various questions that measure different aspects of well-being. One of the key questions is the Cantril ladder, where respondents rate their current lives on a scale from 0 to 10, with 0 being the worst possible life and 10 being the best possible life.

Data Contents -

The variables of the dataset are -

- **Country:** The name of the country.
- **Overall Rank:** The rank of the country based on the Happiness Score.
- **Happiness Score:** The average score of happiness, based on survey responses.
- **Economy (GDP per Capita):** The extent to which GDP per capita contributes to the calculation of the Happiness Score.
- **Family:** The extent to which social support (measured by having someone to count on) contributes to the calculation of the Happiness Score.
- **Life Expectancy:** The extent to which life expectancy contributes to the calculation of the Happiness Score.
- **Freedom:** The extent to which freedom to make life choices contributes to the calculation of the Happiness Score.
- **Absence of Corruption:** The extent to which perceptions of corruption in government and business contribute to the calculation of the Happiness Score.
- **Generosity:** The extent to which generosity (measured by donations) contributes to the calculation of the Happiness Score.

Why I chose this data -

As a surgical scheduler aspiring to become a data analyst, I chose this dataset to analyze because my background in healthcare will enable me to interpret the data accurately and draw meaningful insight. Using the World Happiness Report dataset for my analysis allows me to develop and demonstrate a wide range of data analysis skills on a topic with significant real-world implications. It can also enhance my portfolio and prepare me for a career in data analysis by providing hands-on experience with a rich and multifaceted dataset.

Descriptive Analysis -

	Overall Rank	Happiness Score	Economy (GDP per Capita)	Family	Life Expectancy	Freedom to make life choices	Absence of Corruption	Generosity
count	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000	158.000000
mean	79.493671	5.375734	0.846137	0.991046	0.630259	0.428615	0.143422	0.237296
std	45.754363	1.145010	0.403121	0.272369	0.247078	0.150693	0.120034	0.126685
min	1.000000	2.839000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	40.250000	4.526000	0.545808	0.856823	0.439185	0.328330	0.061675	0.150553
50%	79.500000	5.232500	0.910245	1.029510	0.696705	0.435515	0.107220	0.216130
75%	118.750000	6.243750	1.158448	1.214405	0.811013	0.549092	0.180255	0.309883
max	158.000000	7.587000	1.690420	1.402230	1.025250	0.669730	0.551910	0.795880

Data Cleaning - Basic data cleaning and consistency checks

Checked the dataset for basic structure information

Missing values - No missing values were found

Duplicate records - No duplicate records were found

Mixed Data Types - No mixed data types were found

Rename Columns - (Columns were renamed for more clarity)

1. Trust(Government Corruption) was renamed as Absence of Corruption
2. Freedom was renamed as Freedom to make life choices
3. Happiness Rank was renamed as Overall Rank

Filtering & Dropping Data

Three columns - Region, Dystopia Residual and standard error were dropped as they were irrelevant to the analysis.

Data Profile of Final Dataset

General Information -

Total records - 158

Columns - 9

Continuous Variable - 8

Categorical Variable - 1

Column	Type	Description
Country	Categorical	Name of the Country
Overall Rank	Continuous	Rank of the country based on the happiness score.
Happiness Score	Continuous	A Metric measured in 2015 through feedback from individuals.
Economy(GDP per Capita)	Continuous	The extent to which GDP contributes to the calculation of happiness score.
Family	Continuous	The extent to which family contributes to the happiness score
Life expectancy	Continuous	The extent to which life expectancy contributes to the happiness score
Freedom to make choices	Continuous	The extent to which freedom contributes to the happiness score
Absence of corruption	Continuous	The extent to which absence of corruption contributes to the happiness score
Generosity	Continuous	The extent to which generosity contributes to the happiness score

Limitation , Ethical considerations & Potential Bias

While the World Happiness dataset provides valuable insights into global well-being, it is essential to be mindful of its limitations and ethical considerations.

Limitations -

- Happiness is inherently subjective and may be influenced by cultural, social, and personal factors that are difficult to quantify accurately.

- Different cultures may have varying definitions and understandings of happiness, leading to inconsistencies in how happiness is reported across countries.
- Happiness levels can fluctuate over time due to economic, political, social, and environmental changes. A snapshot in time may not reflect long-term trends.

Ethical Considerations & Potential Bias-

Privacy and Confidentiality -

- Ensuring that individual responses are anonymized to protect the privacy of survey participants.
- Safeguarding the data against unauthorized access and misuse.

Misinterpretation and Misuse

- Avoiding oversimplification of the data, such as using happiness rankings to make sweeping generalizations about countries.
- Being cautious about using the data to support political agendas or policies without considering the broader context.

Equity and Inclusion:

- Ensuring that the data collection methods are inclusive and representative of all segments of the population, including marginalized and vulnerable groups.
- Being aware of potential biases in the survey design that might exclude or under-represent certain demographics.

Impact on Policy:

- Using the data responsibly to inform policy decisions, considering the potential consequences for different populations.
- Ensuring that policies based on happiness data are aimed at improving well-being in an equitable and just manner.

Define questions to explore -

- What countries rank the highest in overall happiness and each of the factors contributing to happiness?
 - Did any country experience a significant increase or decrease in happiness?
 - How do life expectancy and access to healthcare correlate with happiness scores?
 - What is the impact of mental health services on happiness scores?
 - Is there a correlation between perceived freedom to make life choices and happiness scores?
 - How does GDP per capita correlate with happiness scores?
 - What is the relationship between unemployment rates and happiness scores?
-