1/5/2024

Project 6.1 – Sourcing Open Data

Nathan Evans

World Happiness Report 2015-2019

Source: This data set was sourced from Kaggle.com. Kaggle is made up of data analysts, data scientists, and developers. It is the largest data science community online with over a million registered users. Users can participate in discussions and share their work on data sets, models and network with others.

Collection: The happiness scores and rankings use data from the Gallup World Poll.

Contents: The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks 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 cleaning & wrangling was performed in Python. Columns were renamed, unused columns were dropped, Year column was added, Missing and Zero Values were imputed through various methods, a Data Dictionary was utilized to map Countries to Regions, and the 5 CSV files were combined into a single CSV file in preparation for Exploratory Data Analysis.

Relevance: This data does meet the requirements for the project as the data was open source, contains geospatial components and meets the data requirements. The data is a few years old; however, it was suggested and chosen from the project brief.

Data Profile

Variable	Description	Time Variant/Invariant	Quantitative/Qualitative	Nominal/Ordinal/Discrete/Continuous
Country	Country Name	Invariant	Qualitative	Nominal
Happiness_Rank	Ranking of Happiness	Invariant	Quantitative	Ordinal
Happiness_Score	Score of Happiness	Invariant	Quantitative	Continuous
GDP_Per_Capita	GDP per Capita	Invariant	Quantitative	Continuous
Social_Support	Level of Social Support	Invariant	Quantitative	Continuous
Health_Life_Expectancy	Health Life Expectancy	Invariant	Quantitative	Continuous

Freedom_Life_Choices	Freedom in Life Choices	Invariant	Quantitative	Continuous
Generosity	Generosity Level	Invariant	Quantitative	Continuous
Perceptions_of_Corruption	Perception of Corruption Level	Invariant	Quantitative	Continuous
Year	Year	Variant	Quantitative	Discrete
Region	Region of the Country	Invariant	Qualitative	Nominal

Limitations

Sampling Bias: The data might suffer from sampling bias if it primarily includes responses from certain demographics, regions, or age groups. This could limit the generalizability of findings to the entire population.

Self-Reporting: Happiness, perceptions of corruption, and generosity are subjective and self-reported metrics. They can be influenced by cultural norms, individual biases, or social desirability bias, affecting the accuracy of the data.

Missing Variables: There might be important variables not included in the dataset that could influence happiness scores, like political stability, access to education, or environmental factors. Omitting such variables may limit the depth of analysis.

Data Quality: Data quality issues like missing values, outliers, or inconsistencies can affect the reliability of analysis and conclusions drawn from the dataset.

Ethical Considerations

Privacy: Ensuring the anonymity and confidentiality of respondents is crucial, especially when dealing with personal perceptions or sensitive data like perceptions of corruption.

Informed Consent: It's important to ensure that participants have given informed consent for their data to be used in research or analysis, particularly in surveys or studies related to happiness and personal beliefs.

Fair Representation: Ensure fair representation of diverse groups within the dataset to avoid marginalization or underrepresentation of certain demographics, which could lead to biased results.

Responsible Use: Utilize the data responsibly, ensuring that findings are not misinterpreted or used to perpetuate stereotypes, discrimination, or misinformation about particular regions or demographics.

Avoiding Harm: Analyzing and presenting the data in a manner that avoids potential harm or stigma towards specific countries, cultures, or groups is crucial.

Addressing these limitations and ethical considerations is essential to maintain the integrity of the research and ensure that the data is used responsibly and ethically.

Questions to Explore

General Analysis Questions:

Trend Analysis: How have happiness scores changed over the years globally and within different regions?

Correlation: What is the relationship between GDP per capita, social support, health life expectancy, and happiness scores?

Inequality: Are there disparities in happiness scores within and between regions? What factors contribute to these differences?

Impact of Corruption: How does the perception of corruption affect happiness scores across different countries?

Generosity and Freedom: Is there a correlation between generosity, freedom in life choices, and overall happiness?

Geospatial Analysis Questions:

Regional Happiness Variation: What are the spatial patterns of happiness scores globally? Are there geographical clusters of high or low happiness?

Regional Disparities: How do happiness scores vary geographically within specific regions? Are there spatial trends in happiness within continents?

Correlation with Geographic Features: Is there any correlation between happiness scores and geographical factors like proximity to water bodies, altitude, or climatic conditions?

Temporal Changes in Happiness: How have happiness scores changed spatially over time? Are there spatial patterns in the change of happiness scores?

These questions can guide the analysis to understand global trends in happiness, explore factors contributing to happiness disparities, and delve into spatial patterns and variations in happiness scores across different regions. They offer a mix of general insights into happiness determinants as well as specific geospatial exploration to understand how happiness varies across different geographic locations.