WORLD HAPPINESS ANALYSIS

Project Summary

Oksana Stepanova
CareerFoundry – Data Analytics Course

Summary

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th.

The World Happiness Report reflects a worldwide demand for more attention to happiness and well-being as criteria for government policy. It reviews the state of happiness in the world today and shows how the science of happiness explains personal and national variations in happiness.

We will explore the geographical and temporal trends in the happiest countries, impact of Covid-19 outbreak on the happiness score to have a better understanding and to make a prediction for the following year.

Data sourcing

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.

Data is available here: https://www.kaggle.com/datasets/unsdsn/world-happiness

Merged dataset is based on country results for 9 years since 2015 to 2023 in a prepared datasets and allows to conduct spatial and time series analysis.

Data Collection

The typical annual sample for each country is 1,000 people. If a typical country had surveys each year, the sample size would be 3,000. It uses responses from the three most recent years to provide an up-to-date and robust estimate of life evaluations.

The number of people and countries surveyed varies year to year, but by and large more than 100,000 people in 130 countries participate in the Gallup World Poll each year. They are based entirely on the survey scores, using the Gallup weights to make the estimates representative.

Data contents

The data is categorized by country for 2015-2023 and contains the following variables:

- Happiness rank
- Happiness score
- GDP per capita
- Family / social support
- Healthy life expectancy
- Freedom to make life choices
- Generosity (willingness to donate for charity)
- Perceptions of corruption

The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – 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. They have no impact

on the total score reported for each country, but they do explain why some countries rank higher than others.

Data limitations and Ethics

The data is collected via survey based on random sample of respondents. Therefore, it might lead to sample errors.

There was a special chapter on social media in *World Happiness Report 2019*, emphasizing the damaging effects of social media use on the happiness and self-image of adolescents, mainly based on data from the United States. This runs parallel to evidence from earlier Reports showing that in-person friendships support happiness while online connections do not. But COVID-19 and its limitations on in-person meetings offered a chance for electronic connections to develop their potential for creating and maintaining the social bonds that support happiness. Social media have, in consequence, become much more social in the uses to which they have been put, as virtual hugs have been used to fill in for the real thing.

There is no PII in the dataset and therefore it is considered low risk in terms of breaching ethics.

Data relevance

The data is relevant because we can identify happiness rank by country and by time period 2015-2023, and conduct spatial and time series analysis to make a prediction of happiness level.

Data Cleaning

The main cleaning and data wrangling steps:

- Dropping the columns that are irrelevant for the analysis
- Renaming columns to unify the DataFrames
- Adding the column 'year' for further merging
- Converting data type to get the variables suitable for descriptive analysis

The conducted wrangling procedure is available in the scripts 6.1 Data Wrangling - Part 1 & Part 2.

Year	# rows	# columns	Droping columns	renaming columns	Adding columns	# columns after wrangling	# rows after wrangling	DF name after wrangling
2015	158	12	Standard Error' 'region'	10	year'	11	158	wh_2015_2
2016	157	13	Lower Confidence Interval' 'Upper Confidence Interval' 'region'	10	year'	11	157	wh_2016_2
2017	155	12	Whisker.high' 'Whisker.low'	9	year'	11	155	wh_2017_1
2018	156	9	n/a	9	dystopia_residual' 'year'	11	156	wh_2018
2019	156	9	n/a	9	dystopia_residual' 'year'	11	156	wh_2019
2020	153	20	Standard error of ladder score' 'upperwhisker' 'lowerwhisker' 'Logged GDP per capita' 'Social support' 'Healthy life expectancy' 'Freedom to make life choices' 'region' 'Generosity' 'Perceptions of corruption' 'Ladder score in Dystopia'	10	happiness_rank' 'year'	11	153	wh_2020_2
2021	149	20	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' 'region'	10	happiness_rank'	11	149	wh_2021_2
2022	147	12	Whisker-high' 'Whisker-low'	10	year'	11	147	wh_2022_1
2023	137	9	region'	5	happiness_rank' 'dystopia_residual' 'year'	11	137	wh_2023_1

Data merging and consistency check

The main steps of consistency check are

- Merging 9 DataFrames into one
- Checking the country names and renaming the countries to have consistent names of the countries in the merged DataFrame.
- No mixed-type data has been found
- Replacing missing values with the mean value
- No duplicates have been found

The conducted consistency check procedure is available in the scripts 6.1 Consistency Check - Part 3

Data profile

Column name	Description	Data type	Time Variant
country	Country of measurement	Qualitative, nominal	No
year	Year of measurement	Quantitative, discrete	Yes
happiness_rank	Country rank based on the happiness score	Quantitative, discrete	Yes
happiness_score	The national average response to the question of life evaluations	Quantitative, continuous	Yes
economy_GDP_per_capita	The score of purchasing power	Quantitative, continuous	Yes
family	The score of social support (or having someone to count on in times of trouble) is the national average.	Quantitative, continuous	Yes
health_life_expectancy	The score of healthy life expectancies at birth	Quantitative, continuous	Yes
freedom	The score of freedom to make life choices is the national average.	Quantitative, continuous	Yes
trust_government_corruption	The score of corruption perception measured as the national average.	Quantitative, continuous	Yes
generosity	The score of generosity (donations to charity)	Quantitative, continuous	Yes
dystopia_residual	The metric of happiness of the imaginary country Dystopia	Quantitative, continuous	Yes

Questions to explore

- Which region/top 10 countries are the happiest?
- What geographical and temporal trends could be observed?
- Could we predict the following year the happiest countries/region?
- How was the score affected by COVID-19 outbreak?