Happiness and Life Expectancy

Andre Kleber and Trevor McCalmont
SIADS 591-592
University of Michigan, School of Information

Motivation

Many governments use country-level data points like GDP, happiness indices, and life expectancy to direct and inform their policy making decisions across several aspects of society[1]. The motivation for this project is to examine how happiness, health and economic factors interact and how those decisions are impacting health outcomes for various countries across the world.

In this project we will analyze the links between human happiness and life expectancy using a country happiness dataset from the Gallup World Poll and a country life expectancy dataset from the World Bank. Using these data sources, we believe we can answer questions like 'Is there a relationship between health and happiness?'

Questions

- What is the most important factor on a country's happiness score?
- Which country has the lowest life expectancy and why?
- Does life expectancy have a positive or negative correlation with a country's happiness?

[1]Horton & El-Ganainey. Accessed September 15, 2021.

Data Sources

Our primary dataset is the World Happiness Report from the Gallup World Poll which was released at the United Nations at an event celebrating the International Day of Happiness. This dataset contains happiness scores for 158 countries from 2015-2019 and six additional factors that affect the happiness score: economic production, social support, life expectancy, freedom, absence of corruption, and generosity.

The happiness score is calculated by adding together the scores of each of the six categories. A dystopia residual is added and takes the unexplained value of the happiness score. For example, if the happiness score is 7.5 and each of the six factors had a value of one after normalization, the dystopia residual would be 1.5.

The dataset is split into five separate .csv files. Each file contains the data for one year and each country has one row of data per year. In total, the dataset is 79 kb.

World Happiness Data

Location: https://www.kaggle.com/unsdsn/world-happiness

Size: 79 kb

Format: CSV, one file per year

Variables: Country, Country Code (added), Happiness, GDP, Factors that affect happiness

Date Range: 2015 - 2019

Data Sources

Our secondary dataset is the World Development Indicators dataset from the World Bank. This dataset includes observations from 193 countries between 1960-2020. The factors included range from community health (i.e. infant mortality rate, alcohol consumption, % of GDP spent on health) to economic factors (GDP, % of exports) to education and also include data points like infrastructure and access to electricity. We chose to focus on the community health related factors as those answer our questions most directly, and we filtered this dataset to only include the years 2015-2019 to match the happiness dataset. Because the World Bank dataset has raw data and the happiness data set normalizes values like life expectancy and GDP per capita, we will use the raw data from the World Bank data set for our upcoming analyses.

The CSV file is roughly 194 Mb with approximately 384,000 rows and 65 columns. This dataset is unique in that the years are the columns and there are indicator names and codes in each row. Using the pandas.melt function, we were able to manipulate the data into a more usable format.

World Bank Data

Location: https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators (we used Download CSV from here)

Size: 194 Mb

Format: CSV within zip file

Variables: Life expectancy, suicide rate, alcohol consumption, health expenditure (% of GDP), infant mortality rate, incidence of HIV, education expenditure (% of GDP)

Date Range: 1960 - 2020

Data Cleaning and Manipulation

Country Names

Both our datasets are CSV files, so the importing of the data was relatively straightforward. However, in the description on the previous two slides, you'll notice that there was no column to easily merge the two datasets. We could use country name to merge the data sets but we found several issues of consistency (i.e. Trinidad and Tobago vs. Trinidad & Tobago, Saint Lucia vs. St. Lucia, Taiwan Province of China vs. Taiwan).

We did proceed with making the country name uniform across all years (see examples above). This step is not necessary for the subsequent merging step, but we decided to keep it for the sake of completeness and consistency.

Country Codes

The World Bank dataset includes the 3-letter country codes, so using the dataset at the github link below, we appended the appropriate 3-letter country code to the happiness dataset which enabled us to merge the two datasets: https://raw.githubusercontent.com/plotly/datasets/master/2014_world_gdp_with_codes.csv

Regions

Another challenge we faced with the happiness dataset was that the years 2017 - 2019 did not have information about the region in the happiness dataset. In order to map the missing regions to the countries for 2015 and 2016, we used a function which looks up the missing values and returns the happiness dataframe with the respective filled in regions for all countries. We saw that one country (Gambia) was still missing a region. Gambia only has data available for 2019, so we mapped the region manually.

Data Cleaning and Manipulation

Happiness Dataset

To clean the happiness dataset, we added the year as a column to each dataframe. We cleaned up the column names by shortening and clarifying the names and concatenated each year dataframe into one happiness dataframe (i.e. pd.concat(happy_2015, happy_2016, etc.))

World Bank Dataset

As mentioned before, the World Bank dataset contains indicator names and codes in each row and the values for the corresponding years are contained as columns. First, we selected the variables we wanted to use in our analysis from the indicator name column. Together with the corresponding indicator code, we created a dictionary to map the indicator codes to the indicator names. Then we used the original World Bank dataset and the dictionary as input to a function which created a dataframe which has the same structure as the happiness dataset. The pandas.melt function helped us to unpivot the dataframe from wide to long format.

Merging Dataframes

Once both data frames had been cleaned, merging on the 3-letter country code was straightforward.

Missing Data

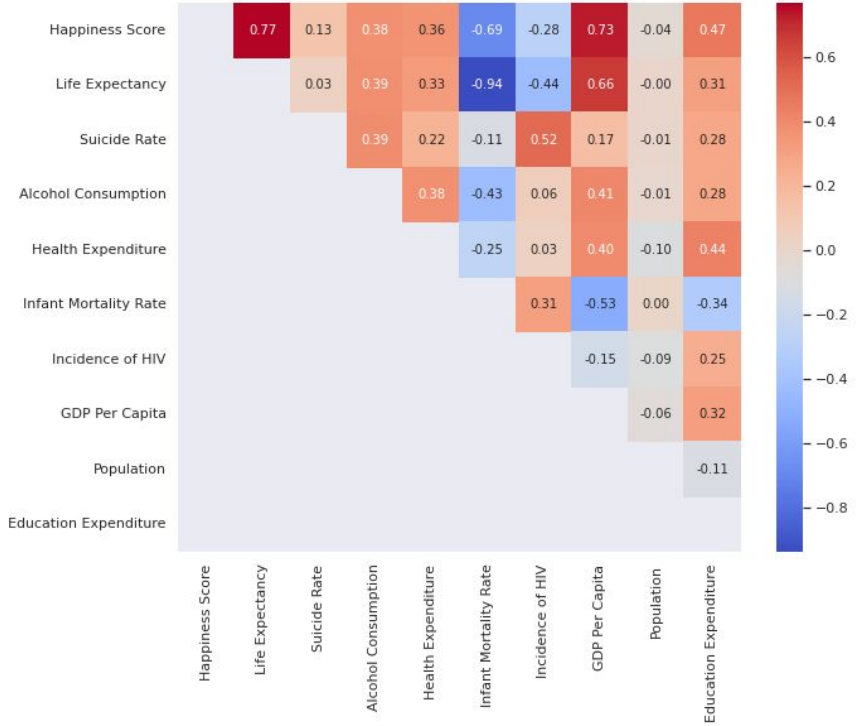
Lastly, we filled NaN and missing values with the column means for that given country. There are a few countries that are missing all values for a given factor for all years and therefore, estimates could not be imputed (i.e. Incidence of HIV per 1,000 people). Those countries were not dropped, but could not be included in all analyses.

Analysis: Most Important Factors in Happiness

After cleaning the merged dataset, the first thing we wanted to do was examine which variables had the strongest relationship with a country's happiness. To do this, we looked at the correlation between happiness score and the other numerical variables, see the top row in the table to the right.

We can see that life expectancy (0.77) and GDP per capita (0.73) have the strongest positive correlation with happiness score. Infant mortality rate (-0.69) has a strong negative correlation with happiness score which makes sense given that infant mortality rate has a near perfect negative correlation with life expectancy. Education expenditure (0.47) has the fourth highest correlation with happiness score.

These findings suggest governments should work to increase GDP per capita and invest in healthcare and education to increase a country's happiness.

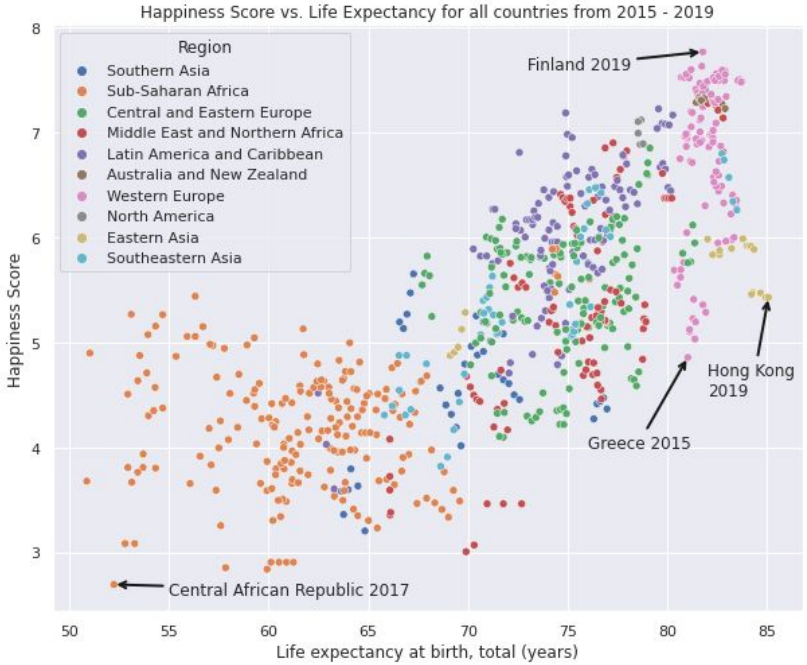


Analysis: Happiness Score by Life Expectancy

To further examine the relationship between happiness and life expectancy, we created a scatter plot of the two variables and encoded the geographic region as the color. Many Western European countries have the highest happiness scores and life expectancies. There are a few outliers that have high life expectancies but lower happiness scores. These outlier countries include Greece, Portugal, Cyprus and Italy.

Another interesting pattern we noticed is that Eastern Asia has some of the highest life expectancies but only slightly above average happiness scores. All of the Eastern Asian countries have lower than expected happiness scores but the three with the lowest are Mongolia, China and Hong Kong.

	Life Expectancy	Happiness
Life Expectancy	1.000000	0.766768
Happiness	0.766768	1.000000



Analysis: Happiness Score by GDP Per Capita

The variable with the second strongest relationship to happiness was GDP per capita. In order to plot this, we needed to plot the GDP on a logarithmic axis. We see many of the same groupings from the previous page when we examine the data by GDP. Again Western Europe has some of the highest GDP per capita and happiness scores. While the United States has a slightly lower life expectancy than other developed countries, the GDP per capita is one of the highest in the world.

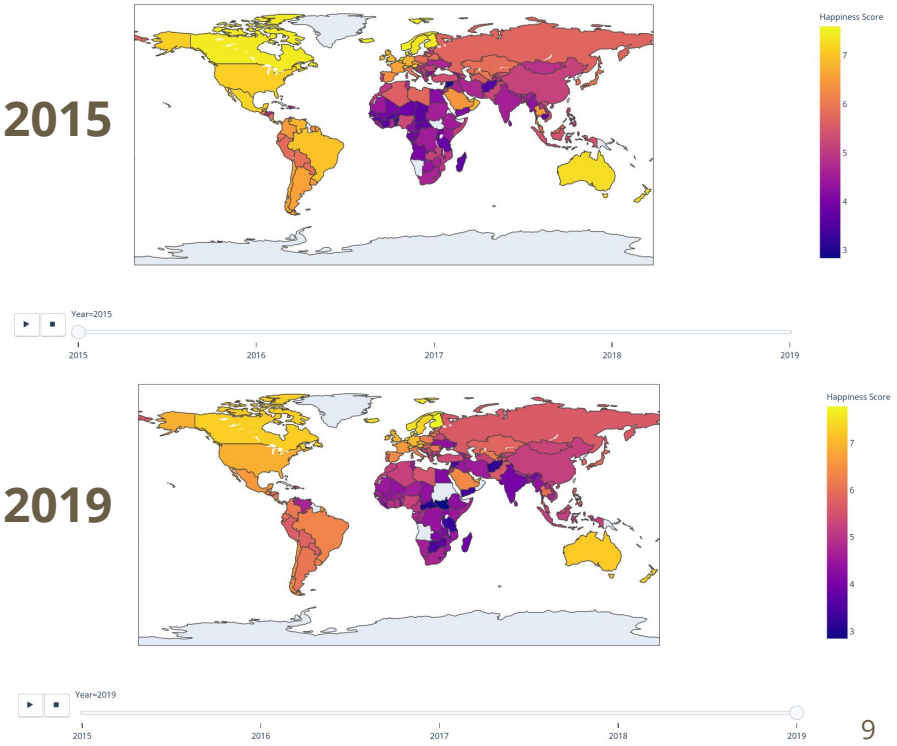
We see the widest range of values for the Middle East and Northern Africa with the GDP per capita ranging from $1,000 USD all the way to nearly $100,000 USD. The countries on the low end were surprising to us - Yemen, Egypt and Morocco. The countries on the highest end of the GDP spectrum were more expected - Qatar, United Arab Emirates and Israel.

Analysis: Choropleth Map of Happiness Over Time

The last interesting finding we discovered relates to the happiness of the world from 2015 - 2019. We were able to animate this plot in Python but were not able to include the animation in our final report, so for the purposes of this document we have a static image from 2015 and 2019 to the right. Visually it appears that the world became less happy between 2015 and 2019.

Specifically this shift appears visually in North America, South America and Africa. Global happiness appears to be almost flat over this time period (5.37 in 2015, 5.40 in 2019), but if we group by region, we do indeed see a decline in happiness scores in all three regions. South America drops by 0.19, North America drops by 0.18, and the Middle East and Northern Africa drops by 0.17.

Central and Eastern Europe and Western Europe are the only two regions with significant increases in happiness over this the same time period: +0.22 and +0.17 respectively. The smaller geographic footprint of Europe makes sense why that would be less apparent on a Mercator projection map.



Statement of Work

Andre set up the Google Colab environment and imported and cleaned the datasets. He also performed much of the exploratory data analysis and contributed to the draft and outline of the final report.

Trevor wrote the bulk of the project proposal, created the visuals, contributed to the analysis, and wrote the final draft of the report.

Both partners contributed equally to the topic selection, discovery of the datasets, and overall strategy and direction of the project.

References

"Global Research." Gallup Inc. https://www.gallup.com/analytics/318875/global-research.aspx. Accessed 17 Sept. 2021.

Horton, M., & El-Ganainy, A. "Fiscal Policy: Taking and Giving Away". International Monetary Fund. https://www.imf.org/external/pubs/ft/fandd/basics/fiscpol.htm. Accessed September 16, 2021.

"World Development Indicators". The World Bank. https://datacatalog.worldbank.org/search/dataset/0037712/World-Development-Indicators. Accessed 17 Sept. 2021.