CBB634 Final Project Report

Panos Ketonis

Source code: `github.com/pketonis/cbb634-final-project`

December 18, 2024

1 Introduction

1.1 Background

The World Happiness Report is a publication that contains rankings of national happiness, based on respondent ratings of their own lives. Since 2024, the report has been published by the Well-being Research Center at the University of Oxford, in partnership with Gallup, the UN Sustainable Development Solutions Network, and an independent editorial board. The report primarily uses data from the Gallup World Poll. Interestingly, as of March 2024, Finland has been ranked the happiest country in the world seven times in a row.

The History of the publication dates to an adopted resolution by the UN General Assembly from 2011: 65/309 *Happiness: Towards a Holistic Definition of Development*. The first World Happiness Report was released on April 1st 2012, as a foundational text for a UN meeting and drew international attention. Since 2016, it has been issued on an annual basis on March 20th, to coincide with the UN's International Day of Happiness.

1.2 Motivation

The World Happiness report reflects a worldwide demand for more attention to happiness and well-being as a 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 [1]. My motivation to investigate this topic is twofold. First, I believe that happiness is an important part of health, especially mental health, that is often overlooked in today's world. A favorite quote of mine by John Steinbeck highlights this: *A sad soul can kill you quicker, far quicker than a germ.* Second, the intersection of Happiness Data and health outcomes rests at the intersection of many CBB themes. Analyzing happiness through an informatics lens can shed light on important insights in global health.

2 Data

2.1 Data Source and Methods

The data was sourced from an online repository on Kaggle [2]. The rankings of national happiness are based on a happiness measurement survey undertaken world-wide by Gallup, Inc. Nationally representative samples of respondents are asked to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their current lives on that scale. The life factor variables used in the reports are reflective of determinants that explain national-level differences in life evaluations across research literature. However, certain variables, such as unemployment or inequality, are not considered because comparable data is not yet available across all countries. The variables used illustrate important correlations rather than causal estimates [1]. These variables currently include: real GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, and perceptions of corruption. The six metrics are used to explain the estimated extent to which of each of these factors contribute to increasing life satisfaction, but they themselves do not have an effect on the total score reported for each country [2]

2.2 Data FAIRness

The dataset on Kaggle is clearly organized and easy to access and use. The website includes some interactive functionalities that allow sorting and inclusion/exclusion criteria. The data is not standardized, however, this is expected for this type of data (mostly meant to be interpreted cardinally). The license for the data is CCO (no copyright), ensuring that the data is as publicly available and accessible as possible. This meets all four aspects of FAIR as the data is easily findable, easily accessible, interoperable, and reusable. A future directive for Kaggle would be to include more recent reports of the publication, as the most recent report lised was 2019. This was the publication used for this report.

2.3 Data Preprocessing

The score data was standardized by converting the scores into percentiles. This was done to aid in percentileRank and for color-coding various graphics on the web interface. All countries that were included in the dataset were confirmed to have a rank, an overall score, and a score for each category. Thus, there was no missing data in the dataset. However, not every country is present in the original dataset, leading to some countries being left out of the analysis and therefore not available on the web interfaces. Validation was performed through various basic sanity checks and independent calculation of variables (ex. calculating distances separately for KNN analysis).

3 Analysis

3.1 Summary Statistics

The first summary statistics created were scatter-plots used to display correlations between the socioeconomic factors and happiness rank. There were no notable issues in creating the summary statistics, either with the scatterplots, or the homepage country statistics analysis. The creation of the scatterplots was inspired by the goal of revealing insights into which socioeconomic factors correlate most strongly with happiness. This can be a guiding factor in directing government funding and/or social work.

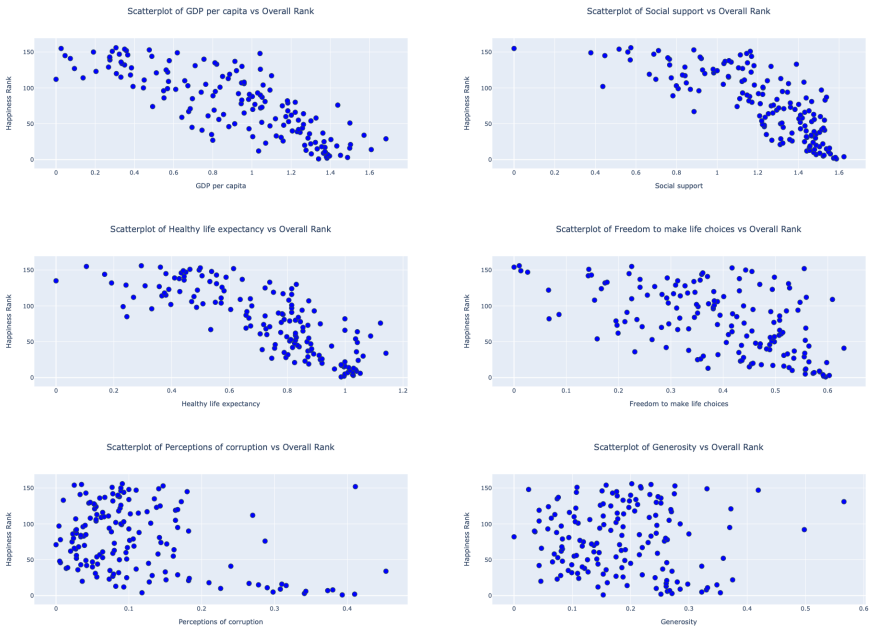

Figure 1: Socioeconomic factor correlation with Happiness Ranking.

Happiness ranking was found to correlate most strongly with GDP per capita (r=-.80), Healthy Life Expectancy (r=-.79), and Social Support (r=-.77), shown in Figure 1. It is interesting to note, that even factors with a high correlation, such as GDP, do not match exactly with the lowest happiness rank (highest score) when approaching the right side of the scatter plot. Some notable outliers visible in the graph include Qatar, Luxembourg, and Singapore, countries with very high GDP per capita, but necessarily high happiness score. This is probably due to the countries small population relative to their unusually high GDP, however, could also be indicative of insufficient funding of healthcare and well-being and/or social initiatives within those countries.

3.2 K-nearest Neighbors

The K-nearest neighbors analysis that was implemented takes a country as an input and predicts its happiness score based on the happiness scores of its 'k' nearest neighbors. For example, one can input 'Mexico' and 'k=5', to predict the happiness score of Mexico based on its 5 nearest neighbors. This is shown below in Figure 2.

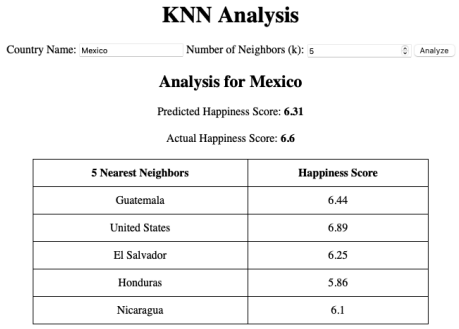

Figure 2: K-nearest Neighbors Analysis with input 'Mexico' and 'k=5'

The K-nearest neighbors analysis correctly finds the 5 nearest countries to Mexico and lists their happiness scores. It then averages these scores and yields a predicted happiness score of 6.31. This score is remarkably close to the actual happiness score of 6.6. The accuracy of this predictor varies with the region of the world the country resides in and works best on countries situated centrally and inland, with multiple close neighbors. Happiness scores change most dramatically between continents when compared to intercontinental variance, thus the analysis is worse at countries near the edges of certain continents. The analysis is inspired by the real-world similarities that can often be found by neighboring countries and likewise by the trends shown in the interactive map. The analysis can help find anomalous regions of either extremely high or low happiness compared to a country's neighbors to investigate.

3.3 K-means Clustering

The K-means clustering analysis clusters countries based on their geographic locations and scores. After the clusters are created, they are ordered by their average score, with Cluster 1 being the highest scoring. The clusters are displayed on a world map accompanied by a color-coded legend. This analysis hopes to investigate if countries can be grouped into larger regions to gain further insights. For example, global organizations might be able to seek solutions for certain regions of countries that face similar issues affecting their happiness rankings, and thus their peoples quality of life.

The result of a sample query of this function is displayed below in Figure 3.

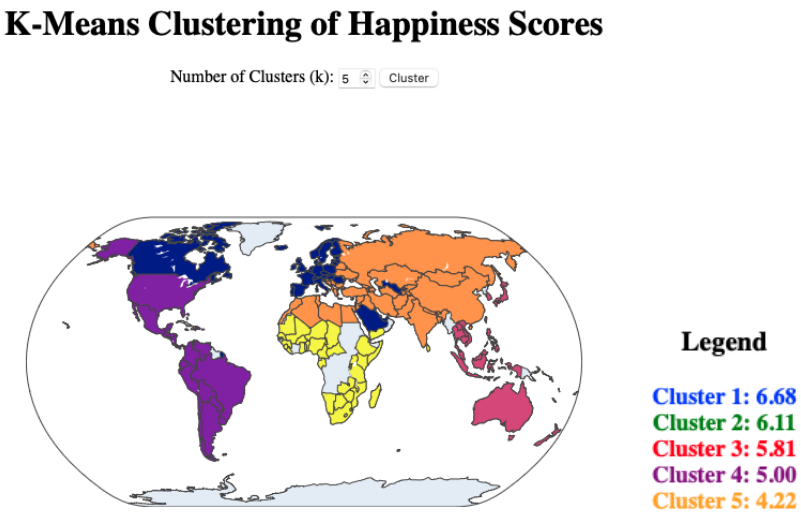

Figure 3: K-means Clustering Analysis with input 'k=5'

4 Online Implementation

4.1 Front End

The front end of the website includes an interactive landing page, as well as various linked pages that are available for analysis. At the top of the page, there are buttons to redirect to these linked pages as well as a text box used to enter the name of a country to display its rank, score, and socioeconomic factor scores. The webpage will also display the country's flag and create a header notifying the user which country is currently being displayed. Percentile scores for each category are also visible and color-coded for easier visualization and interpretation. This is shown below in Figure 4.

Enter a country name to get its statistics:

Finland Get Statistics

Statistics for Finland

Category	Value	Percentile
Overall Rank	1	-
Score	7.769	100%
GDP per capita	1.34	85.9%
Social support	1.587	99.36%
Healthy life expectancy	0.986	82.37%
Freedom to make life choices	0.596	97.44%
Generosity	0.153	40.71%
Perceptions of corruption	0.393	98.08%

Figure 4: Requested Statistics for Finland

The homepage hosts the main interactive map that allows visualization of the happiness scores for every country present in the dataset. The map is shown below in Figure 5.

The interactive map includes a hover functionality that displays the country's name, rank, and overall happiness score. The map also includes a click functionality that links to the main statistical display that provides the statistics shown in figure 4.

The homepage also includes the data source, license, and reference for the country flags used.

Overall Trends is the first analysis page that is available to access from the homepage through the 'Overall Trends' button. This page allows users to create scatter plots for the various socioeconomic factors to see how they correlate with a country's happiness ranking. The page also contains a table that lists the correlation coefficient for each scatter-plot.

KNN Analysis is the second analysis page that is also accessible through a button on the homepage named 'KNN Analysis'. The page contains two text entry boxes, allowing users to enter a country name as well as set the parameter value for 'k'. The website then returns a table as seen in Figure 2.

K-means Clustering is the third analysis page that can be accessed through the use of the 'K-means Clustering' button on the homepage. It takes users to a page in which they can select a value of 'k' to create 'k' clusters. These are

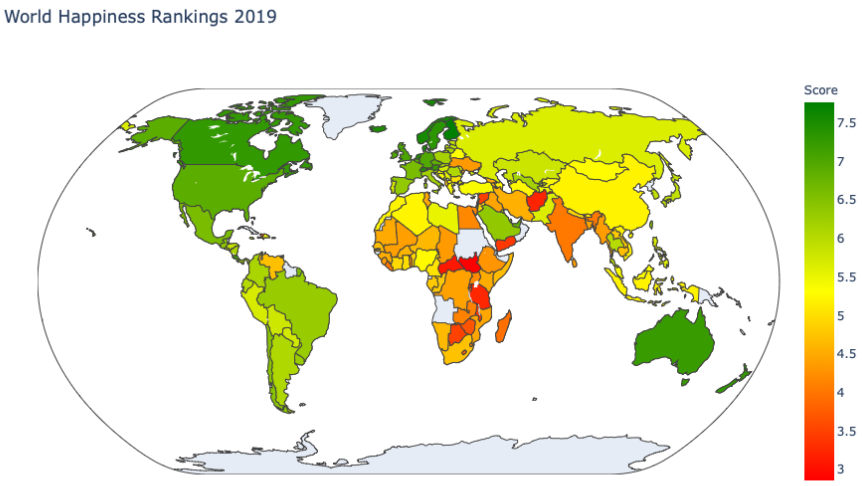

Figure 5: Interactive Map

Correlation of Categories with Overall Rank

Category	Correlation Coefficient
GDP per capita	-0.80
Healthy life expectancy	-0.79
Social support	-0.77
Freedom to make life choices	-0.55
Perceptions of corruption	-0.35
Generosity	-0.05

Choose a category to plot: GDP per capita View Scatterplot

Back to Main Page

Figure 6: Overall Trends Page

visualized on a map similar to the interactive map found on the homepage. The clusters are ordered by average score and color-coded according to the legend visible under the map.

4.2 Back End

APIs allow for users to access data and the statistical analyses tools mentioned through their computers. A country data API allows the retrieval of summary statistics about a country similar to the analysis available on the homepage (returns rank, total score, and score of factors). A KNN analysis API returns the result of a K-nearest neighbors request and takes a country

and a value for 'k' as an input. A K-means clustering API takes 'k' as an input and returns the countries sorted into 'k' clusters with the average score for each cluster. An example usage of the API is shown below in Figure 7.


```
{
  "Country": "Finland",
  "Freedom to make life choices": 0.596,
  "GDP per capita": 1.34,
  "Generosity": 0.153,
  "Healthy life expectancy": 0.986,
  "Perceptions of corruption": 0.393,
  "Rank": 1,
  "Score": 7.769,
  "Social support": 1.587
}
```


Figure 7: API result from retrieving country data. Input is 'Finland'. http://127.0.0.1:5000/api/country-data?country=Finland

References

[1] Wellbeing Research Centre at the University of Oxford, UK. (n.d.). The World Happiness Report: About. The World Happiness Report. https://worldhappiness.report/about/
[2] Network, S. D. S. (2019, November 27). World happiness report. Kaggle. https://www.kaggle.com/datasets/unsdsn/world-happiness/data?select=2019.csv