Nepanji Davis

Azmir Suljic

Alex Mull-Dreyer

December 9, 2021

A Comparative Analysis of World Happiness in relation to Global Cause of Death Metrics Other Than Diseases: Is there a correlation?

**The World Happiness Report**

**History**

The first World Happiness Report was released in April 2012. The UN General Assembly originally adopted the Happiness: Towards a Holistic Definition of Development resolution inviting countries to participate in a measure of the happiness of their people. The data collected was used to help guide public policy. An annual issue of the report began in 2016 and it coincides with the UN’s International Day of Happiness.

**Methodology**

The World Happiness Report asks the respondents to think of a ladder when they rate their own current lives on a scale of 0 to 10, with the best possible life for them represents 10, and the worst possible life being a 0. This is called the Cantril ladder method.

**Global Cause of Death Other Than Diseases Report**

**History**

The Global Cause of Death other than Diseases report was a part of a Global Disease Burden project completed in 2017. The data was compiled to help investigate which global cause of death is impacting which country.

**Methodology**

The Global Cause of Death Report data was captured from Institute for Health Metrics and Evaluation; Global Disease Burden.

**The Proposal**

The source data will include the World Happiness Report results paired with the Global Cause of the Deaths Other Than Diseases Report.

**Data Source #1**

The first data source file selected was the World Happiness Report CSV file retrieved from [kaggle.com](https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report.csv). This CSV file includes the annual reports from each participating country from 2008 - 2017. The data used to rank countries in each report is drawn from the Gallup World Poll. Metrics reported are as follows: (1) Country name, (2) Year, (3) Life Ladder, (4) Log GDP per capita, (5) Social support, (6) Healthy life expectancy at birth, (7) Freedom to make life choices, (8) Generosity, (9) Perceptions of corruption, (10) Positive affect, (11) Negative affect.

**Data Source #2**

The second data source file selected was the Global Causes of Death other than Diseases CSV file retrieved from [kaggle.com](https://www.kaggle.com/tahminashoaib86/global-cause-of-the-deaths-other-than-diseases). This CSV file includes the number of deaths within each country for each year, in addition to, the cause of death such as conflict and terrorism, famine, pandemic, natural disaster, and other injuries from 1980 - 2020. Metrics reported are as follows: (1) Country name, (2) Year, (3) ISO\_CODE, (4) Total # of Deaths, (5) Male POP, (6) Female POP, (7) Total POP, (8) GDP, and (9) GDP per capita (PCAP).

**Project Goals**

For this project, our goal is to see if there is a correlation between a country’s happiness compared to the causes of death in said countries outside of natural diseases. Using the Kaggle datasets, [world happiness dataset](https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021?select=world-happiness-report.csv) (2008-2017) and the [cause of death dataset](https://www.kaggle.com/tahminashoaib86/global-cause-of-the-deaths-other-than-diseases) (2008-2020), we will use Pandas to transform CSV files and SQL to load the joined file. These datasets were joined on the year and country. In order to use the data to investigate whether there is a correlation between a country’s happiness compared to the global cause of deaths, we must first analyze and bring the datasets together using the following steps:

* **Extract:** Indicates original data sources and how data was formatted
* **Transform:** Explains what data cleaning or transformation was required
* **Load:** Explains the final database, tables/collections, and why topic was chosen

**Data Cleanup**

**Extract**

* We started with two original data sources located on Kaggle.com as CSV files.
* Both files were extracted from Kaggle.com and downloaded to the computer as CSV files and loaded into Jupyter Notebook using Pandas.
  + **Original File Format Details**
    - **World Happiness Report**: from 2008-2020 with 12 columns and 1950 rows of data
    - **Global Cause of Death**: from 1980-2017 with 10 columns and 36,860 rows of data

**Transform**

* After each CSV file was downloaded, they were imported into Pandas. Each file included columns with lengthy titles and decimal places.
* The data cleanup process included: (1) shortening of column titles in both files, (2) formatting of column cells to reduce the number of decimal places, (3) renaming the key record column to match in both tables
  + **Cleaned File Format Details**
    - **World Happiness Report**: from 2008-2020 with 12 columns and 1950 rows of data
    - **Global Cause of Death**: from 1980-2017 with 10 columns and 36,860 rows of data

**Load**

* + Upon completion of the data cleanup step, both files were prepped and ready for load into SQL.
  + Each file possessed the common attribute of country and year. This common format structure allowed for us to use a key record to create a relational database.
  + An inner join was executed in order to return all rows from both participating files where the key record of one table is equal to the key record of the other table.
    - **Cleaned and Loaded File Format Details**
      * **World Happiness Report**: from 2008-2017 with 7 columns and 1,570 rows of data
      * **Global Cause of Death:** from 2008-2017 with 6 columns and 9,700 rows of data