# A look at World Happiness from 2015 to 2019

01.08.2020 to 21.08.2020

Roxanna Wijtsma
Data Science
The Bridge
@RoxannaW

## General vision

As part of the Data Science bootcamp, The Bridge, I will be working on an individual project regarding the global world happiness ranking.

## Goals

The goal for this project is to put into practise the concepts we have learned during the first 7 weeks of the Data Science bootcamp: obtaining data, data wrangling, data mining, visualization of the data and analysing the data to draw conclusions on a hypothesis.

For this project I have chosen to focus on all the required steps of option C.

## Specifications

In order to do this project and make the most out of the delivery, I used the following requirements:

### Software

Visual Studio v1.48.0

Git Bash

### Hardware

Computer with i5 processor and at least 4G RAM.

### Requirements

Data required:

- World Happiness Reports from 2015 to 2019: https://www.kaggle.com/mathurinache/world-happiness-report
- Global Peace Index, scraped from: https://en.wikipedia.org/wiki/Global_Peace_Index
- Unemployment rate: https://www.ilo.org/shinyapps/bulkexplorer15/?lang=en&segment=indicator&id=SDG_0852_SEX_AGE_RT_A

Programming language required:

- Python v3.8.3

Libraries required:

- Pandas
- Numpy
- Seaborn
- Matplotlib
- Plotly
- Bubbly

# Steps

## I. Research the context and defining the hypothesis for the project

I have researched what the World Happiness score exactly is and what the reports include. Besides that I have further investigated which other factors might have had an influence on happiness and decided which factors I wanted to include in my datasets.

I have defined the following hypotheses:

- The happiest country in the world does not change over the years.
- The most important factor to influence world happiness is the GDP per capita.

During this project I will analyze the data to see if the hypotheses are correct or not.

## II. Get Data

To obtain the data I have searched different websites. The World Happiness reports were easy to find at Kaggle, but finding data on the Global Peace Index and the unemployment rate were harder as sometimes not all the years were available or the source was not free of charge. I eventually found my data of the Global Peace index on Wikipedia and took the data from there. The information on the unemployment rate I found in Json format on the website of The international Labour Organisation.

## III. Data Wrangling

As I used datasets from different sources, I first had to clean each dataset individually. After that, I made sure the data was merged together in one dataframe per year. Finally, I also made a dataframe with all the years combined. I used the next steps before merging the data together:

- Change column names so each year has the same corresponding column names which are easy to understand.
- Removing any unnecessary columns from the datasets which are not useful in the project.
- Setting the correct index and adding a column with the year for each data frame as reference.

## IV. Data Mining / Cleaning

In order to check if my data frames were clean I performed the next steps:

- Checking the column types
- Checking for any NaN values
- Checking for duplicates in the data

In order to clean the dataset I have changed the column types where necessary, removed a column of the dataset which had too many NaN values which made the column unusable. Lastly, I filled in any remaining NaN value in the best corresponding way.

## V. Data Visualization

To obtain insights and analyse the data I have used data visualization with charts, graphs and maps. I have divided this part into two sections:

1. Analyzing the data frames in general and having a look at the correlation between and the tendency of  the different columns.

2. Specifically visualizing and analysing the data to give an answer to the hypothesis.

## VI. Conclusions

After analyzing the data and visualizing it, I have written the conclusion on each of the hypotheses.

## Answering project questions

As part of the mandatory steps of the project of option C, I will be answering the following questions:

- Was it possible to demonstrate the hypothesis? Why?

Yes, with the visualization and analysing of the data I was able to answer both of the hypotheses with a clear conclusion.

- What can you conclude about your data study?

In general we can say that each year a different country has the highest happiness score and the ranking is different. Only in the last two years, Finland had the highest score in both years. If this was a stroke of luck or not, will need to become clear with future data.

Also we can conclude that there are several factors that have an important influence in the world happiness score: The GDP per capita, Healthy life expectancy and Social support are the three most important factors. The GDP per capita, however,  is the most important factor each year, even though the importance is slowly decreasing over time.

- What would you change if you need to do another EDA project?

In a future project I would choose to not use several datasets from different sources and different years, I would focus more on one dataset (of one year) in particular to be able to dive deeper in the analysis and also have clearer graphs and visualization.

- What did you learn doing this project?

During this project I learned to put into practise the different steps of an EDA project and the different elements we learned during the first weeks of the bootcamp. I learned how to handle problems that occurred during the process and how to effectively resolve this without losing too much time and getting stuck in the process.

Also I have developed new skills in the visualization part of the data analysis and used a library I did not use before: Plotly, to make interactive graphs.

## Future steps

I am aware that with drawing the conclusions to my hypotheses and finalizing the project for now, it will never really be completely finished. I will be continuing  working on, and improving, this project in the future. Some aspects I would like to focus on are:

- Using the feedback I will get on this project to make improvements in any area possible.

- Adding the newest year: 2020 to the dataset and see if there are any interesting observations to be made, taking into account the current global situation (Covid-19).

- If possible, use the future knowledge we will obtain during the next weeks of the bootcamp and apply it to this project. Such as machine learning and making predictions on the future of world happiness.

-