World Happiness Report Project: Exploratory data analysis (EDA) on the World Happiness Reports 2018/2019

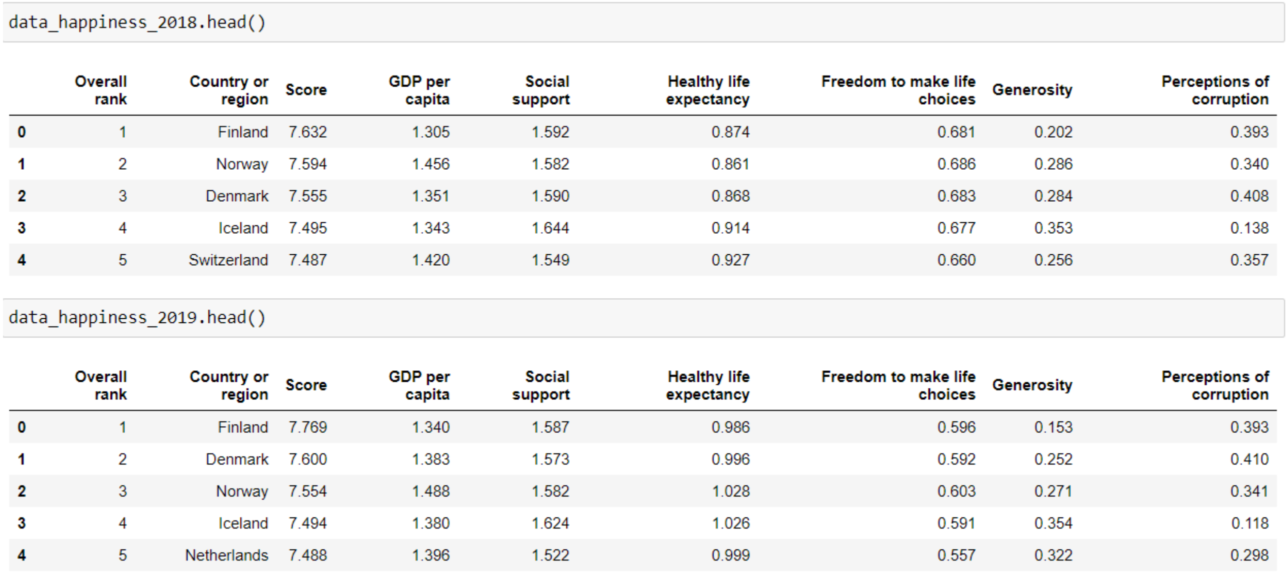
Description of the data set and a summary of its attributes

The World Happiness Report is a survey of the global happiness of the state. To carry out our EDA, we will be using 2 datasets in this project, the first one contains the results of 2018's survey and the second one the results of 2019's survey.

The main columns in these datasets are:

> Overall Rank: Rank of the country based on the Happiness Score / Country or Region / Score = A metric measured by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest" / GDP per Capita / Social support / Healthy Life Expectancy / Freedom to make life choices / Generosity / Perceptions of corruption

Data exploration



```
data_happiness_2018.head()
```

	Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.632	1.305	1.592	0.874	0.681	0.202	0.393
1	2	Norway	7.594	1.456	1.582	0.861	0.686	0.286	0.340
2	3	Denmark	7.555	1.351	1.590	0.868	0.683	0.284	0.408
3	4	Iceland	7.495	1.343	1.644	0.914	0.677	0.353	0.138
4	5	Switzerland	7.487	1.420	1.549	0.927	0.660	0.256	0.357

```
data_happiness_2019.head()
```

	Overall rank	Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	1	Finland	7.769	1.340	1.587	0.986	0.596	0.153	0.393
1	2	Denmark	7.600	1.383	1.573	0.996	0.592	0.252	0.410
2	3	Norway	7.554	1.488	1.582	1.028	0.603	0.271	0.341
3	4	Iceland	7.494	1.380	1.624	1.026	0.591	0.354	0.118
4	5	Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298

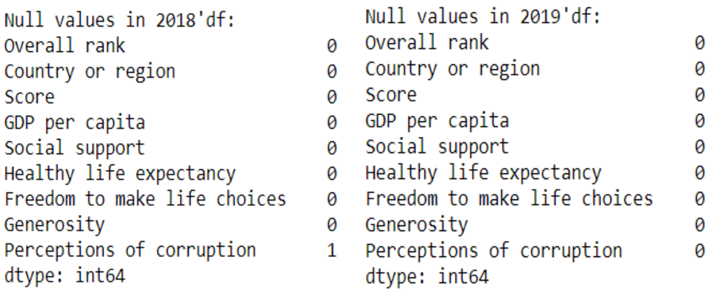
Fig1 - Data exploration 2018/2019

The shape of the 2 datasets is (156, 9): there are 156 rows (countries) and 9 columns (features).

Data cleaning

I start by checking the duplicated data in the data sets: The sum of duplicated data of 2018's df: 0 and The sum of duplicated data of 2019's df: 0. There are no duplicated values.

Then I check the null values in the data sets: There is one country with NaN value in the column 'Perceptions of corruption':



```
Null values in 2018'df:
Overall rank                    0
Country or region               0
Score                           0
GDP per capita                  0
Social support                  0
Healthy life expectancy         0
Freedom to make life choices    0
Generosity                      0
Perceptions of corruption       1
dtype: int64
```

```
Null values in 2019'df:
Overall rank                    0
Country or region               0
Score                           0
GDP per capita                  0
Social support                  0
Healthy life expectancy         0
Freedom to make life choices    0
Generosity                      0
Perceptions of corruption       0
dtype: int64
```

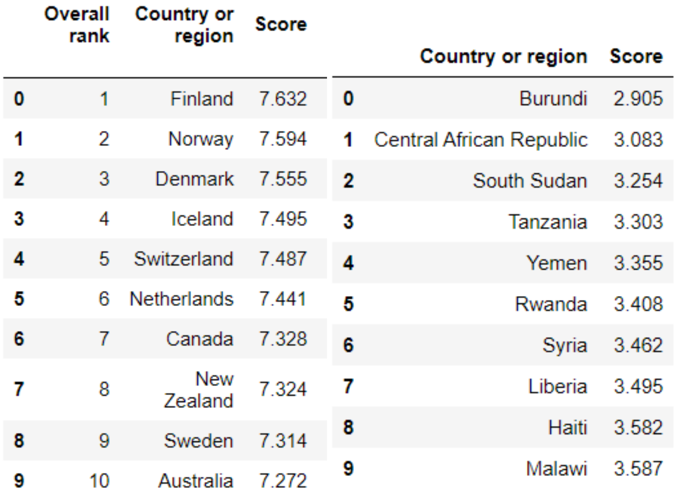
Fig2 – NaN values in 2018/2019

I did replace the NaN value with the mean of the column 'Perceptions of corruption' and there are no more null values. Our data is ready for the analysis.

Data analysis

I. Top 10 Happiest Countries – 2018 and Top 10 Saddest Countries – 2018

In order to determine these two lists, I used the column 'Score' and sort the values (the countries with the highest scores are the happiest): In the first list we find the top 10 happiest countries and in the second one the top 10 saddest countries with the lowest scores



	Overall rank	Country or region	Score
0	1	Finland	7.632
1	2	Norway	7.594
2	3	Denmark	7.555
3	4	Iceland	7.495
4	5	Switzerland	7.487
5	6	Netherlands	7.441
6	7	Canada	7.328
7	8	New Zealand	7.324
8	9	Sweden	7.314
9	10	Australia	7.272

	Country or region	Score
0	Burundi	2.905
1	Central African Republic	3.083
2	South Sudan	3.254
3	Tanzania	3.303
4	Yemen	3.355
5	Rwanda	3.408
6	Syria	3.462
7	Liberia	3.495
8	Haiti	3.582
9	Malawi	3.587

Fig3 – Top 10 happiest countries in 2018 / Top 10 saddest countries in 2018

II. Top 10 Happiest Countries – 2019 and Top 10 Saddest Countries – 2019

I did the same work with data set of 2019.

III. Comparison between 2018 and 2019

1. Top 10 happiest countries

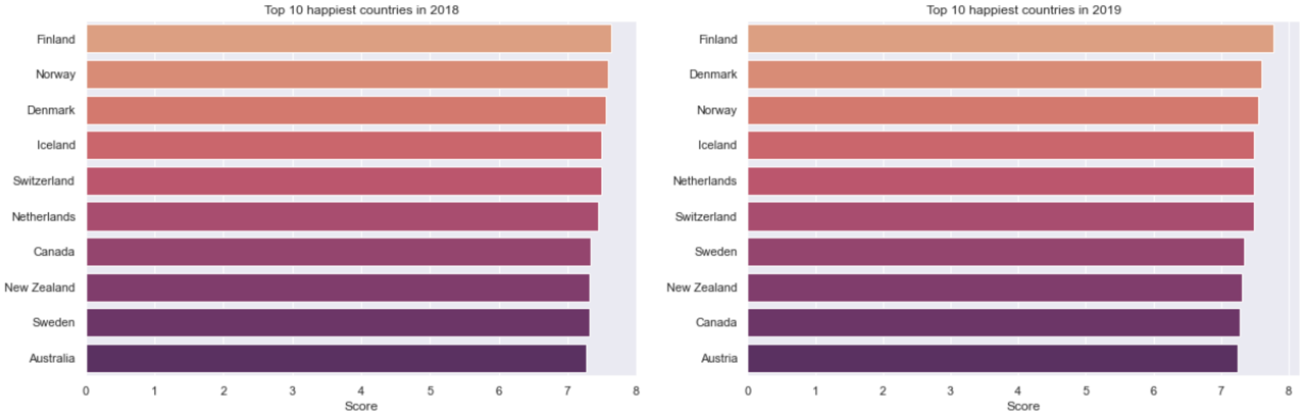


Fig4 – Top 10 happiest countries in 2018 / 2019

- Finland is always the happiest country in 2018 and 2019.
- Norway was the second happiest country in 2018 but in 2019 Denmark took its place.
- Overall, for both years, we still have the same list of countries with a small change in order.

2. Top 10 saddest countries

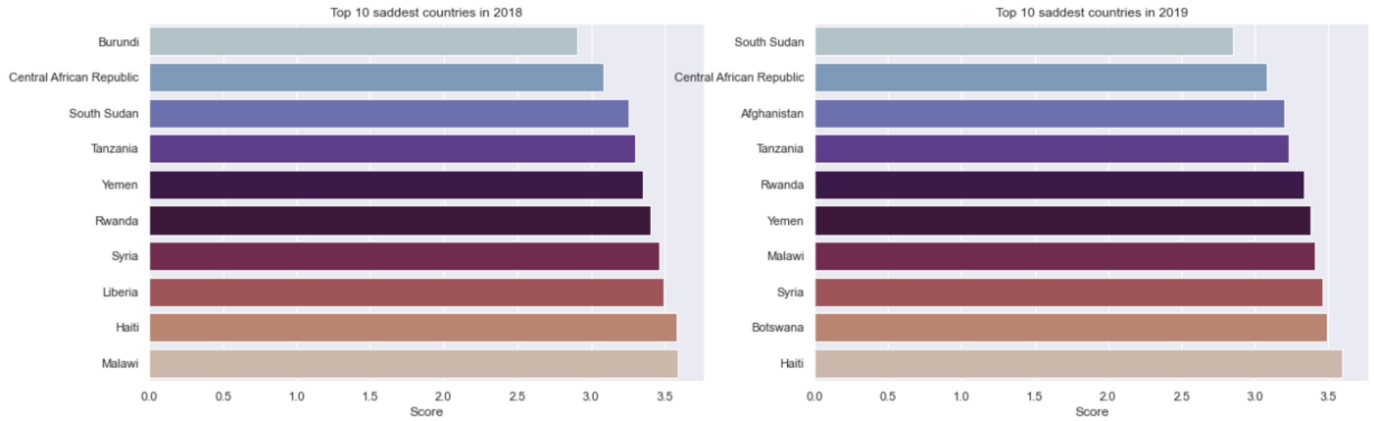
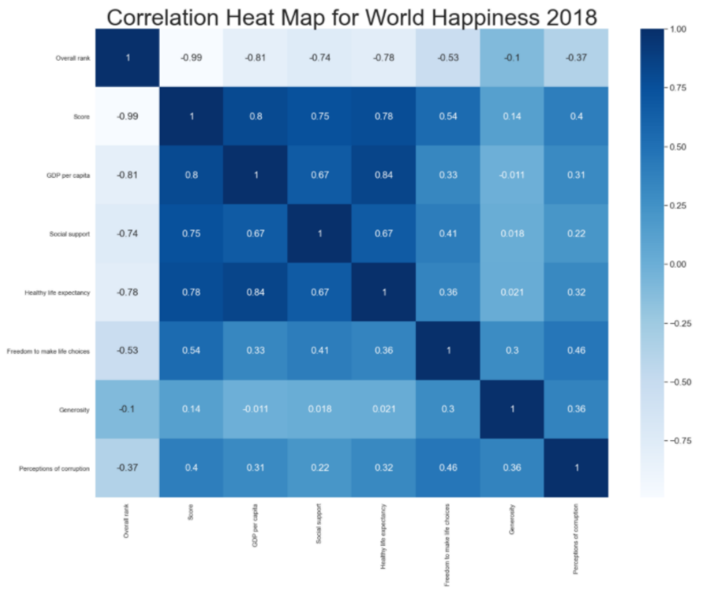


Fig5 – Top 10 saddest countries in 2018 / 2019

- In 2018, Burundi was the saddest country but in 2019 it is no longer in the list of the top 10 saddest countries.
- South Sudan became the saddest country in 2019.
- Overall, for both years, we still have the same list of countries with a small change in order.

IV. Correlations in 2018

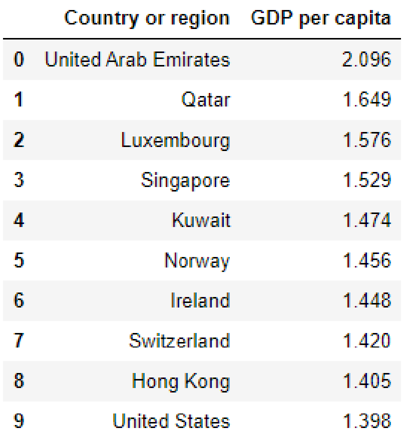


	Overall rank	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
Overall rank	1	-0.99	-0.81	-0.74	-0.78	-0.53	-0.1	-0.37
Score	-0.99	1	0.8	0.75	0.78	0.54	0.14	0.4
GDP per capita	-0.81	0.8	1	0.67	0.84	0.33	-0.011	0.31
Social support	-0.74	0.75	0.67	1	0.67	0.41	0.018	0.22
Healthy life expectancy	-0.78	0.78	0.84	0.67	1	0.36	0.021	0.32
Freedom to make life choices	-0.53	0.54	0.33	0.41	0.36	1	0.3	0.46
Generosity	-0.1	0.14	-0.011	0.018	0.021	0.3	1	0.36
Perceptions of corruption	-0.37	0.4	0.31	0.22	0.32	0.46	0.36	1

Fig6 – Correlations Heat Map for 2018

- We can see that there are strong positive correlations between Score and GDP per Capita, Score and Social support, and Score and Healthy life expectancy.

V. Top 10 countries with highest GDP per Capita in 2018

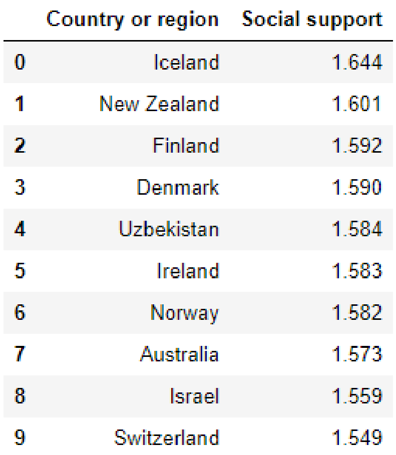


	Country or region	GDP per capita
0	United Arab Emirates	2.096
1	Qatar	1.649
2	Luxembourg	1.576
3	Singapore	1.529
4	Kuwait	1.474
5	Norway	1.456
6	Ireland	1.448
7	Switzerland	1.420
8	Hong Kong	1.405
9	United States	1.398

Fig7 – Top 10 countries with highest GDP per Capita in 2018

- The country with the highest GDP per capita is United Arab Emirates
- From the list of the top 10 happiest countries in 2018 we find: Norway in the 5th rank and Switzerland in the 7th rank, this may prove that a strong economy is not the answer to happiness.

VI. Top 10 countries with highest Social support in 2018



	Country or region	Social support
0	Iceland	1.644
1	New Zealand	1.601
2	Finland	1.592
3	Denmark	1.590
4	Uzbekistan	1.584
5	Ireland	1.583
6	Norway	1.582
7	Australia	1.573
8	Israel	1.559
9	Switzerland	1.549

Fig8 – Top 10 countries with highest Social support in 2018

- The country with the highest Social support is Iceland.
- The top 3 happiest countries in 2018 (Iceland, Finland, and Denmark) are in the top 4 countries with highest Social support, this supports the idea that Social support is key to creating a happier country.

VII. Score and Healthy life expectancy

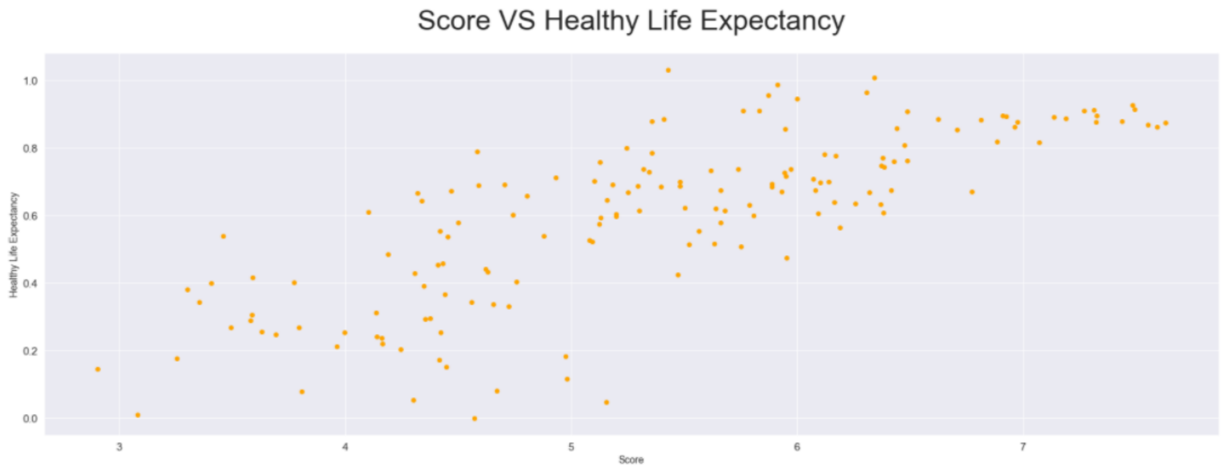


Fig9 – Score vs Healthy life expectancy in 2018

- The graph representing the correlation between happiness score and healthy life expectancy displays a stronger correlation with the higher scores. However, it does seem that the curve flattens as you reach a healthy life expectancy score of 0.8, showing that beyond this point of life, happiness stays the same.

Hypothesis Testing

To apply the hypothesis testing approach seen in the course I will use some new techniques that I have learned from another course. In this project I have chosen to work with hypothesis testing based on mean's Statistic Tests. To do that we can use 2 distributions:

- Student's t-distribution
  When **the sample size N is small** (let's say less than 30 is small), we must draw our test statistic from a t-distribution. The shape of this distribution narrows and becomes more normal like when the sample size

increases or degrees of freedom increases, where degrees of freedom is N-1. We have three methods, provided by the library 'scipy', when we use this distribution:

1. If we have one sample that we want to compare to some specified value, we must do a **one-sample t-test**.
2. If we have two independent samples that we want to compare to each other, we must do **an independent samples t-test.**
3. If we have two dependent samples taken from the same individuals or objects, we must do **a paired samples t-test.**

- Normal z-distribution
  When the sample is large, we use a normal distribution.

To understand the steps of these techniques we will do 3 hypothesis tests:

1. **1st hypothesis testing:**
   A sample of the top 20 happiest countries in 2018 yields a 'GDP per capita' sample mean of sample_mean = 1.4482 . **Can we consider the chosen sample as a representative sample of the population of 2018? (**population_mean = data_2018['GDP per capita'].mean() = 0.89). To test this question, we will use what is refered to as a one-sample t-test.
   First, we state the null hypothesis and alternative hypothesis like this;

- **H0:** The chosen sample is a representative sample of the population of 2018, sample_mean = population_mean.
- **HA:** The chosen sample is not a representative sample of the population of 2018, sample_mean != (not equal) population mean
  Then, we specify **a significance (alpha) level**. Usually, statistical significance is associated with an alpha level of **α = 0.05** or smaller. Next, we use a T-distribution table (or Z-distribution table if we are working with a large sample size) (we can find these tables on the internet) to look up the **critical t-value** that corresponds to this α level (also to the confidence interval). Here we are doing **a two-tailed test** because we do not care if the sample mean is greater than or less than the population mean. We just are testing to see if the two are equal or not.
  Next, we calculate the t-statistic (or z-statistic) or we can use the 'scipy' library to have the result.
  If this t-statistic (or z-statistic) is lower than t-critical (z-critical) then we accept the null hypothesis, otherwise we reject the null and accept the alternative hypothesis.
  In this 1st hypothesis testing and with alpha=0.05 we have t-critical = 2.093 and I found t-statistic = 14.198 (I used the function provided by 'scipy'):
  Conclusion:
  T_statistic is greater than T_critical in magnitude, so there is a statistically significant difference at the alpha = 0.05 level between the sample mean and the population mean. So, we reject the null hypothesis in favor of the alternative.

2. **2nd hypothesis testing:**
   Does the generosity increase from 2018 to 2019?

- H0: It does not increase, sample_mean (generosity_mean_2019) = population_mean (generosity_mean_2018).
- HA: It increases, sample_mean != (not equal) population mean
  We will use α = 0.05 and since the sample size is large (more than 30) we will do a z-statistic test this time.
  The z-critical associated to the chosen alpha is z-critical = 1.65 (from the z-distribution table).
  z-statistic = (sample_mean - population_mean)/(s/sqrt(N)) where s is the sample standard deviation.
  I found z-statistic = 0.5
  Conclusion:
  The z-statistic is less than z-critical then we accept the null hypothesis. The generosity does not increase from in 2019.

3. **3rd hypothesis testing:**
   Previously we have said that the list of the top 10 happiest countries in 2019 is almost the same as the list of 2018. Now **we want to know if those 10 countries were happier in 2019 than in 2018**? To answer this

question, we will use the mean of the 'Social support' as we have already found that the 'Social support' is key to creating a happier country. So;

- sample1 = the top 10 happiest countries in 2018
- sample2 = the top 10 happiest countries in 2019
- H0: The Social support mean is the same for both years (x̄1 = x̄2)
- H1: The social support mean is greater in 2019 (x̄2 > x̄1)
  alpha = 0.05 We are going to compare two independent samples (with small sizes), therefore we will do an independent samples t-test.
  From the t-distribution table: T-critical for specified alpha level: t-critical = 1.833 (one-tailed, right-side)
  And I found t-statistic = 0.99
  Conclusion:
  T-statistic is lower than T-critical in magnitude which means we have to accept the null hypothesis.
  Statistically, we say the sample1 mean is no different than the sample2 mean.

Next steps in analyzing this data

We can look for outliers and make some variable transformations if needed.

The quality of the data set

Personally, I think that these datasets are extremely useful as they help to review the state of happiness in the world today and show how the new science of happiness explains personal and national variations in happiness.

The World Happiness Reports can be used in different fields and as an example the 2021 report focuses on the effects of Covid-19 and how people all over the world have fared.