Data Analytics Immersion: Task 6

1. **Data source**

World Happiness Report | Kaggle (https://www.kaggle.com/datasets/unsdsn/world-happiness)

Our dataset is a survey of the state of global happiness. It uses data from the Gallup World Poll and ranks 155 countries by their happiness levels from 2015 to 2019. The dataset is released by Sustainable Development Solutions Network (SDSN) which is a non-profit launched by the United Nations in 2012 to promote the implementation of the UN Sustainable Development Goals (SDGs) at national and international levels.

2. **Data collection**

The data was collected by surveying. The scores are based on answers to the main life evaluation question asked in the poll. This question, known as the Cantril ladder, asks respondents to think of a ladder with the best possible life for them being a 10 and the worst possible life being a 0 and to rate their own current lives on that scale. The happiness level is based on six factors: economic production, social support, life expectancy, freedom, absence of corruption, and generosity.

3. **Data contents**

The dataset contains the happiness scores and ranking of 155 countries. The columns following the happiness score estimate the extent to which each of six factors – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – contribute to making life evaluations higher in each country than they are in Dystopia, a hypothetical country that has values equal to the world's lowest national averages for each of the six factors.

4. **Data limitations and ethical consideration**

The data describe an evaluation of the quality of life in 155 countries. The evaluation is based on a survey made within nationally representative samples of the population. This manual collection makes the dataset susceptible to contain human errors.

The answers from the respondents are not based on specific criteria like the 6 factors listed above but are based on a comparison with a dystopia, which is an imaginary country that has the world's least-happy people imaginary lands. Therefore, the happiness score given is more a perceived value making this highly subjective.

The 6 factors (GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption) describe the extent to which these factors contribute in evaluating the happiness in each country.

The dataset is fully anonymised and raised no further ethical considerations.

5. **Data relevance**

This dataset was produced by a non-profit organisation. We can then assume that it's the most trustworthy and complete version of the data available. This dataset is relatively new and presents an innovative approach to evaluate population happiness within countries. For these reasons, this data set is critical to addressing your project objective.

6. Data profiling

Variables and data type

Country: qualitative, time-invariant, nominal

```
count             782
unique            170
top       Switzerland
freq                5
Name: Country, dtype: object
```

Happiness rank: quantitative, time-variant, continuous

```
count     779
unique    156
top         1
freq        5
Name: Happiness rank, dtype: object
```

Happiness score: quantitative, time-variant, continuous

```
count    782.000000
mean       5.379018
std        1.127456
min        2.693000
25%        4.509750
50%        5.322000
75%        6.189500
max        7.769000
Name: Happiness score, dtype: float64
```

GDP per capita: quantitative, time-variant, continuous

```
count    782.000000
mean       0.916047
std        0.407340
min        0.000000
25%        0.606500
50%        0.982205
75%        1.236187
max        2.096000
Name: GDP per capita, dtype: float64
```

Social support: quantitative, time-variant, continuous

```
count    782.000000
mean       1.078392
std        0.329548
min        0.000000
25%        0.869363
50%        1.124735
75%        1.327250
max        1.644000
Name: Social support, dtype: float64
```

Healthy life expectancy: quantitative, time-variant, continuous

```
count    782.000000
mean       0.612416
std        0.248309
min        0.000000
25%        0.440183
50%        0.647310
75%        0.808000
max        1.141000
Name: Healthy life expectancy, dtype: float64
```

Freedom: quantitative, time-variant, continuous

```
count    782.000000
mean       0.411091
std        0.152880
min        0.000000
25%        0.309768
50%        0.431000
75%        0.531000
max        0.724000
Name: Freedom, dtype: float64
```

Perception of corruption: quantitative, time-variant, continuous

```
count    781.000000
mean       0.125436
std        0.105816
min        0.000000
25%        0.054000
50%        0.091000
75%        0.156030
max        0.551910
Name: Perception of corruption, dtype: float64
```

Generosity: quantitative, time-variant, continuous

```
count    782.000000
mean       0.218576
std        0.122321
min        0.000000
25%        0.130000
50%        0.201982
75%        0.278832
max        0.838075
Name: Generosity, dtype: float64
```

7. Data integrity issues

The dataset is having non-uniform variable name:

- 'Health (Life Expectancy)', 'Health..Life.Expectancy', 'Healthy life expectancy' were replaced by 'Healthy life expectancy'.
- Trust (Government Corruption)', 'Perceptions of corruption', 'Trust..Government.Corruption.' were replaced by ‘Perception of corruption’.

- 'Happiness Rank' was replaced 'Happiness rank'.
- 'Happiness Score', 'Happiness.Score' were replaced 'Happiness score'.
- 'Family' was replaced by 'Social support'.
- 'Economy (GDP per Capita)' was replaced by GDP per capita.

The dataset contained irrelevant variables:

- 'Lower Confidence Interval', 'Upper Confidence Interval', 'Dystopia Residual', 'Region', 'Whisker.high', 'Whisker.low', 'Dystopia.Residual' have been removed.

8. Data consistency

The dataset contains 4 missing values which are negligeable and will be given the value 0.

The data type in the column ‘Year’ has been changed into string data type.

No data duplicates were identified.

9. Project overview (Questions to explore)

Motivation

Most governments of world have as main objective the improvement of the quality of life of their population. As parameter of this life quality usually include societal factors like economic production, healthy life expectancy, social support, freedom… our dataset is proposing to directly ask the population about how happy they feel as the ultimate parameter for life quality.

Objective

The main objective of this analysis is to search for any correlation between the perceived happiness from the population and the quantitative parameters of life quality, then establish the pertinence of this method of monitoring life quality.

We will achieve this by examining following questions:

What countries or regions rank the highest in overall happiness and each of the six factors contributing to happiness? How did country rank or score change between the 2015 and 2016 as well as the 2016 and 2017 reports? Did any country experience a significant increase or decrease in happiness? How changes in quantitative parameters affect the ranking of a country?

Scope

The analysis will cover 170 countries in the world.