Saikat Halder
CareerFoundry

Exercise 6.1: Sourcing Open Data

World Happiness Report (2015-2023)

Background

The World Happiness Report surveys happiness levels around the globe. This is an essential factor in shaping government policies. It assesses the current state of happiness worldwide and illustrates how the study of happiness can contribute to personal and national variations in happiness.

Data Source

This data has been taken from Kaggle. The original source is from World Happiness Report. This data in turn was generated from Gallup World Poll. They are based on answers to the main life evaluation question. The 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. They are then asked to rate their own current lives on that 0 to 10 scale. The rankings are from nationally representative samples over three years. For more information regarding how data is collected please visit the following link:

Gallop World Poll Methodology

Reliability

The Gallop World Poll survey is an external source and is reliable and trustworthy. Gallop has very little biasness as the survey data represents 95% of the world's adult population. To mitigate biasness, they have the following methodology:

- The target population is the entire civilian, non-institutionalized, aged 15 and older population.
- There is a standard set of core questions used around the world.
- The questionnaire is translated into the major languages of each country.
- Interviewing supervisors and interviews are trained not only on the questionnaire, but also on the execution of field procedures. This interviewing training usually takes place in a central location.
- Telephone surveys are used in countries where telephone coverage represents at least 80% of the population or is the customary survey methodology. In countries where telephone interviewing is employed, Random-Digit-Dial (RDD) or a nationally

representative list of phone numbers is used. Telephone methodology is typical in the United States, Canada, Western Europe, Japan, Australia, etc.

- In the developing world, including much of Latin America, the former Soviet Union countries, nearly all of Asia, the Middle East, and Africa, an area frame design is used for face-to-face interviewing.
- Quality control procedures are used to validate that correct samples are selected and that the correct person is randomly selected in each household.

Data Collection

The primary data collection method is Surveys and Interviews. There are two ways this data is collected.

- Face to face interviews (usually for 1 hour)
- Telephone surveys (usually for 30 minutes)

Sampling

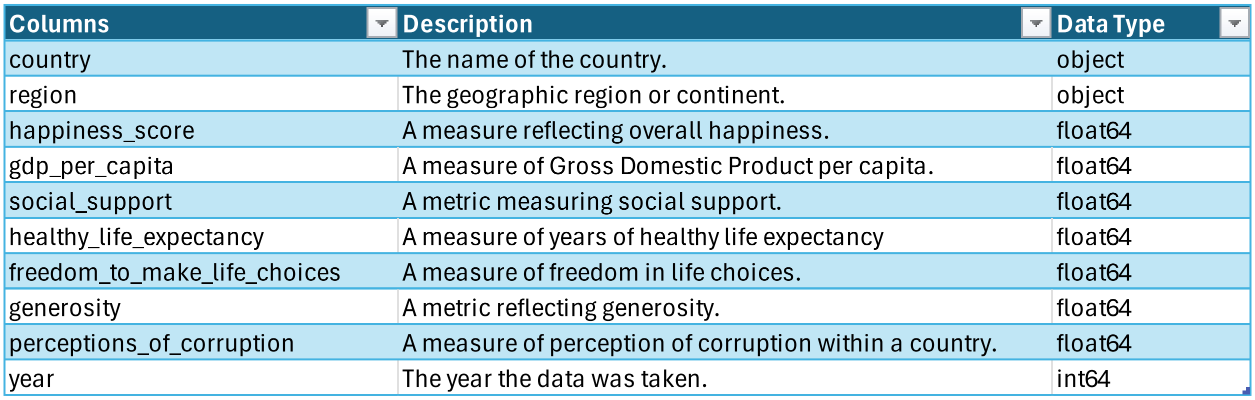
The number of people and countries surveyed varies year to year, but by and large more than 100,000 people in 130 countries participate in the Gallup World Poll each year. The typical World Poll survey includes at least 1,000 surveys of individuals. In some countries, oversamples are collected in major cities or areas of special interest. Additionally, in some large countries, such as China and Russia, sample sizes of at least 2,000 are collected. Although rare, in some instances the sample size is between 500 and 1,000. They are based entirely on the survey scores, using the Gallup weights to make the estimates representative.

Limitations

Even though precautions are taken to reduce biasness, there is always a possibility of human error. Since the data is collected from surveys, there is always a possibility for non-response. While conducting surveys, the questions and answers are translated and hence, this is also subject to biasness. Depending on the country, there is always a issue of coverage (scarcely populated areas where transportation is lacking and the general population is not easily reachable).

Saikat Halder
CareerFoundry

Data Contents



Columns	Description	Data Type
country	The name of the country.	object
region	The geographic region or continent.	object
happiness_score	A measure reflecting overall happiness.	float64
gdp_per_capita	A measure of Gross Domestic Product per capita.	float64
social_support	A metric measuring social support.	float64
healthy_life_expectancy	A measure of years of healthy life expectancy	float64
freedom_to_make_life_choices	A measure of freedom in life choices.	float64
generosity	A metric reflecting generosity.	float64
perceptions_of_corruption	A measure of perception of corruption within a country.	float64
year	The year the data was taken.	int64

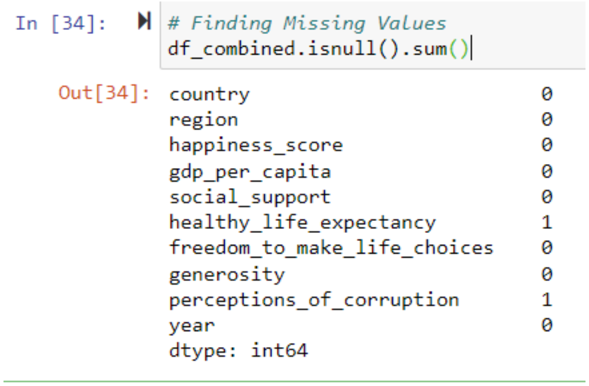
Data Profile

Mixed Type Data

There are no mixed data type columns.

Missing Values

Finding Missing Values



```
In [34]:   # Finding Missing Values
           df_combined.isnull().sum()

Out[34]:   country                         0
           region                          0
           happiness_score                 0
           gdp_per_capita                  0
           social_support                  0
           healthy_life_expectancy         1
           freedom_to_make_life_choices    0
           generosity                      0
           perceptions_of_corruption       1
           year                            0
           dtype: int64
```

We have one count missing for both ‘healthy_life_expectancy’ and ‘perceptions_of_corruption’.

Treatment of missing Values

healthy_life_expectancy

Since there is only one count of State of Palestine, we will take the mean value of the region Middle East and North Africa (2015-2023)

perceptions_of_corruption

Since we have records for 'United Arab Emirates' for all the years except for 2018, we will use mean grouped by Country.

Duplicate Values

There are no duplicate values.

Key Questions

- Does the Mean Happiness Score vary from region to region? Are certain regions consistently happier than others?
- Which factor contributes the most towards the Happiness Score of a Country?
- How Covid 19 Pandemic affected the Happiness Score?

Bibliography

World Happiness Report up to 2023. (2023, October). Retrieved from Kaggle: https://www.kaggle.com/datasets/sazidthe1/global-happiness-scores-and-factors/data