6.1: Sourcing Open Data

Data Source

1. Dataset

a. **“World Happiness Report”**
b. Happiness scored according to economic production, social support, etc.
c. **Dataset sourced from:**
   - ✓ World Happiness Report (kaggle.com)
   - ✓ World Happiness Report up to 2022 (kaggle.com)
   - ✓ World Happiness Report 2023 (kaggle.com)
d. **Type of Data :** External data
e. **Owner :** Sustainable Development Solutions Network (Owner)

2. Data Collection Method

a. **Type of data:** Interviews and Surveys (Public Survey data)
b. **Collection Method:** The happiness scores and rankings use data from the Gallup World Poll. 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.
c. **Time Line:** The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. This dataset is reporting the world happiness index from 2015 till 2023.

3. Overview of Data Contents

a. Variables Included:

1. Country : Name of country
2. Region : Region the country belongs to.
3. Happiness Rank : Rank of the country based on the Happiness Score
4. Happiness Score : A metric measured in 2015 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."
5. Standard Error : The standard error of the happiness score.
6. Economy (GDP per Capita) : The extent to which GDP contributes to the calculation of the Happiness Score.
7. Family : The extent to which Family contributes to the calculation of the Happiness Score.
8. Health (Life Expectancy) : The extent to which Life expectancy contributed to the calculation of the Happiness Score.
9. Freedom : The extent to which Freedom contributed to the calculation of the Happiness Score.
10. Trust (Government Corruption) : The extent to which Perception of Corruption contributes to Happiness Score.

Country	String
Region	String
Happiness Rank	Integer
Happiness Score	Integer
Standard Error	Integer
Economy (GDP per Capita)	Integer
Family	Integer
Health (Life Expectancy)	Integer
Freedom	Integer
Trust (Government Corruption)	Integer

Why this dataset was chosen

This dataset, sourced from the Kaggle , meets all project requirements, allowing for advanced analytical techniques such as regression, clustering, and geospatial analysis. It is very simple me to interpret the data accurately and draw meaningful insights, contributing to public heppiness research efforts. Additionally, the dataset is ethically sound, with no personally identifiable information, ensuring responsible use. Analysing this data may yield significant public happiness insights, aligning with my commitment to improving outcomes through data-driven research.

Data Profile

1. Data cleaning process of Dataset from 2015 to 2022

a. Initial Data Exploration:

Inspected the dataset to check the structure and basic information.Checked the numerical statistics and visualized distributions.

b. Handling Missing Values:

Checked for missing values. If There are missing values. Correcting missing data.

c. Filtering and Dropping Data

Dropped the Group column which was no longer relevant to the analysis

1. **2015** : ‘Standard Error’ & ‘Dystopia Residual’
2. **2016** : ‘Lower Confidence Interval' & 'Upper Confidence Interval', ‘Dystopia Residual’
3. **2017** : 'Whisker.high' , 'Whisker.low' & ‘Dystopia Residual’
4. **2020** : 'Standard error of ladder score', 'upperwhisker','lowerwhisker', 'Logged GDP per capita', 'Social support','Healthy life expectancy','Freedom to make life choices', 'Generosity', 'Perceptions of corruption','Ladder score in Dystopia' & ‘Dystopia + residual'
5. **2021** : 'Standard error of ladder score', 'upperwhisker','lowerwhisker', 'Logged GDP per capita', 'Social support','Healthy life expectancy','Freedom to make life choices', 'Generosity', 'Perceptions of corruption','Ladder score in Dystopia' & ‘Dystopia + residual'

6. **2022 :** 'Whisker-high','Whisker-low', 'Dystopia (1.83) + residual'
7. **2023** : 'Standard error of ladder score', 'upperwhisker','lowerwhisker', 'Logged GDP per capita', 'Social support','Healthy life expectancy','Freedom to make life choices', 'Generosity', 'Perceptions of corruption','Ladder score in Dystopia' & 'Dystopia + residual'

d. Add new column

1. Add a new column 'Year' in each data frame. There are dataset of each year ,but no year column.
2. For year 2017 , 2018 & 2019 dataset , adding column 'Region', by deriving Region according to country names **.**

e. Change the order of the columns

Take the 'Year' , 'region' and 'country' column ahead , it will help to get clear view while doing analysis.

f. Check the country names and correcting country name in 2018

g. Check the null value in data

h. Data Type Conversion

In data set of year 2022 , There are different data types of every column. All were Float (int). So , Data types has been converted to expected data types.

Column	Data Type
Rank	int64
Year	int64
Country	Object
Region	Object
Happiness Score	float64
Economy (GDP per Capita)	float64
Family (Social Support)	float64
Health (Life Expectancy)	float64
Freedom (Life Choices)	float64
Trust (Government Corruption)	float64
Generosity (Donations to Charity)	float64

i. **Addressing Duplicates and Mixed-type Data**

Checked for and confirmed no duplicate rows.
Ensured no mixed-type data within columns. Drop the other missing values.

j. **Merging the data sets and Exporting Data**

Conducted final checks on the cleaned data frame for structure, statistics of numerical columns, and unique values in categorical columns. Exported the cleaned data frame to a CSV file for further analysis Merge all data set in on data set and give it a final check .Exported the cleaned merged data frame to a CSV file for further analysis.

2. Data Profile of final cleaned dataset

a. General Information

1. Total Entries: 1359
2. Columns: 11
3. Categorical Variables:2
4. Numerical Variables: 8
5. Date Variables: 1

b. Column Descriptions

Column	Datatype
Rank	int64
Year	int64
Country	object
Region	object
Happiness Score	float64
Economy (GDP per Capita)	float64
Family (Social Support)	float64
Health (Life Expectancy)	float64
Freedom (Life Choices)	float64
Trust (Government Corruption)	float64
Generosity (Donations to Charity)	float64

Questions to explore with the analysis:

a. Geographic Analysis:

Which country / region has highest rank all over these years ?
Which country / region has lowest rank all over these years ?

Wha are the most common conditions contributing in increase or decrease of happiness index in each region or country?

Are there regional patterns in the types of conditions associated with happiness score across the world ?

b. Temporal Analysis

If any country is around same score (constant score) over all these time period , then What are the conditions contributing constant happiness score of particular country ?

If there is spike or down fall in happiness score of any country over all these time period , then What are the conditions contributing the rise or fall in happiness score of particular country ?