6.1 Sourcing Open Data

World Happiness Rankings

I chose this dataset because I have always been fascinated by understanding what makes societies and cultures thrive and flourish. Having studied Ancient Cultures for my master's degree, I developed a deep appreciation for the diversity of human experiences and the factors that contribute to societal well-being. This dataset provides a modern perspective on how different countries perceive their happiness and well-being, which I find incredibly interesting because it reflects the ongoing evolution of human societies. Analysing this data allows me to explore cultural differences in values, social support, and quality of life across the globe. I see it as an opportunity to deepen my understanding of how cultural, economic, and political factors influence happiness today, bridging my interest in historical cultural studies with contemporary societal analysis.

Data Source:
The dataset is derived from the World Happiness Report 2015-2019, which surveys and ranks the happiness levels of countries worldwide. It contains data from the Gallup World Poll for the years 2013-2016, capturing each country's happiness score along with key contributing factors. The dataset includes variables such as GDP per Capita, Social Support, Life Expectancy, Freedom, Generosity, Trust in Government, and Corruption levels, as well as an estimate called the Dystopia Residual, which explains unexplained variations. The purpose of this dataset is to enable analysis of the factors influencing national happiness, compare countries over time, and understand the interplay between various socio-economic variables and overall well-being.

Data Collection:

The data was collected from the Gallup World Poll, a widely recognized survey that gathers responses from nationally representative samples across countries. Respondents answer questions related to their life evaluations, which are then scaled on a 0-10 ladder (the Cantril ladder). The other variables, such as GDP, social support, life expectancy, and others, are sourced from reliable international indicators and statistical databases. The calculation of the happiness scores and contributing factors allows researchers to analyze national well-being comprehensively while accounting for residuals that reflect unexplained variations in happiness.

Goals of the Project:

- To identify which countries or regions rank highest in overall happiness and in each of the contributing factors.

- To examine how country rankings and scores change over consecutive years (2015-2016, 2016-2017).
- To analyse how the residuals relate to the happiness scores, identifying unexplained factors affecting well-being.
- To explore the influence of economic, social, and political variables on national happiness and identify patterns or outliers.
- To assess whether improvements in factors like GDP, social support, or freedom correlate with changes in happiness scores over time.

Ethical Considerations:

Cultural Sensitivity: Happiness is subjective and influenced by cultural norms. Different societies may interpret or respond to survey questions differently, which can affect cross-country comparisons. It's important to recognise these differences when analysing and presenting results, to avoid unfair stereotypes or assumptions.

Privacy and Data Use: The data is aggregated and anonymised, but any future analysis should continue to respect individual privacy. Care must be taken not to re-identify individuals or use the data in ways that could harm communities.

Responsible Communication: When sharing findings, approach sensitive topics with care, especially for countries with lower scores. Highlight that data are indicative and influenced by multiple factors, and avoid stigmatising nations based on their scores.

Limitations of the Data:

Subjectivity of Happiness Scores: The happiness scores are based on personal self-assessments, which may be influenced by cultural attitudes, personal perceptions, and social desirability. This subjectivity can impact the accuracy of cross-country comparisons.

Unmeasured Influences: The dataset captures key variables like GDP, social support, and health, but many other important factors influence happiness—such as safety, environmental conditions, or political stability—that are not included, and residuals may reflect these unmeasured influences.

Temporal Scope: The data covers selected years (mostly 2013-2016), so it may not fully reflect longer-term trends or recent changes. External events or regional issues occurring outside this period may also influence the results.

Data Quality and Coverage: Variations in data collection quality and sampling methods across countries may introduce bias or inconsistencies, especially in regions with less reliable reporting or access.
Correlation vs. Causation: While the analysis may show relationships (e.g., higher GDP correlates with higher happiness), it does not prove causation. Other underlying factors could be influencing both variables.

Data Profile:
Handling Missing Values:

There was only one missing value across all datasets, which I have chosen to not remove, as it is unlikely to impact the overall analysis. A review of the .describe() output shows that the minimum values for columns like 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', 'Trust (Government Corruption)', 'Generosity', and 'Dystopia Residual' are all 0.000. This is expected and normal, and does not indicate any concern regarding missing data.

Duplicates:
I have checked for duplicate entries in the datasets and found none, ensuring data integrity.

Optimizing Data Types:

Columns in each DataFrame have been reviewed and set with appropriate data types to optimise memory usage and facilitate analysis.

Dropping Irrelevant Columns:

Currently, I have identified some columns, such as 'Region', which might be irrelevant for certain analyses. However, I believe that including region-based insights could be valuable, especially to identify trends within culturally similar areas like Western Europe versus Africa or Asia. Some datasets already include a region column, and for those that do not, I plan to add this information when possible to support regional trend analysis.

Uniformity:
All column names have been renamed to follow a consistent format for clarity (e.g., replacing spaces or special characters). I have also verified that country names are standardised—full names are used rather than abbreviations—to maintain uniformity across datasets.

Integrity Checks:

Preliminary checks indicate the data values are consistent and logical, with no obvious anomalies or issues detected at this stage.

Questions to Explore

- **Which countries or regions rank the highest in overall happiness, and how do these rankings vary across the years 2015-2019?**
  To identify patterns and regional differences in happiness levels over time.
- **How do the six contributing factors (GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust in Government) vary across countries and regions?**
  To understand what factors are most influential in different parts of the world.
- **What is the relationship between each of the six factors and the overall happiness scores? Are certain factors more strongly correlated with happiness than others?**
  To determine which variables have the most significant impact on happiness.
- **Are there any countries that experienced significant increases or decreases in happiness scores between 2015-2019?**
  To identify outliers and analyze potential causes for these changes.
- **How do residuals (unexplained components) vary among countries, and what might larger residuals indicate about unmeasured influences on happiness?**
  To explore the limitations of the model and other factors affecting well-being.
- **Is there a pattern between a country's economic status (GDP per Capita) and its happiness score? Do wealthier countries generally report higher happiness?**
  To analyse the impact of economic wealth on well-being.
- **How does the perception of social support and freedom relate to happiness across different regions?**
  To assess the social and political aspects influencing happiness.
- **Are there differences in happiness scores between countries with high levels of generosity or trust in government?**
  To explore the impact of social cohesion and governance on national happiness.
- **Can we predict changes in happiness scores based on shifts in the six contributing factors?**
  To evaluate the potential for modeling and forecasting happiness trends.