Visualising Data – Dataset Description

Kalyan Kola Cahill – C16357813

For this project, a number of different datasets were gathered from various sources before being cleaned and merged together for analysis. Discussed below are the three gathered datasets, their respective source, along with an in-depth description of both the organisation which produced the dataset, as well as what each feature in the set represents.

**Freedom House – Freedom in the World**

Freedom House is a US Government-funded non-profit organization which conducts research and publishes reports and statistics on a number of issues related to international democracy, political freedom, human rights, and civil liberties. Their flagship publication *Freedom in the World* acts as a standard comparative assessment of each countries political and civil rights. Additionally, they also produce a number of smaller annual reports on topics such as *Freedom on the Net* and *Freedom of the Press.*

*For this project, the FIW* data for the years 1972-2017 was used, which contains 18 features both numerical and categorical ratings for 195 countries and 15 territories on matters such as Political Rating and Civil Liberties. In total it contains 9385 rows, with the most notable features being:

* Country Name/Country Code - The name and unique code for each of the 195 analysed countries.
* Year - Temporal data representing the specific year each of the observations were made.
* PR - The continuous, numerical, political rights rating, scored between 1 and 7, where 1 represents a greater degree of political rights, and 7 represents a severe lack of political rights.
* CL - The continuous, numerical, civil liberties rating, scored in the same way as the previously mentioned PR rating.
* Status - A categorical value denoting the overall freedom level of the specific country for that year, either **F** [free], **PF** [partially free], or **NF** [not free].

**Source:** Freedom House. 2018. *Freedom in the World, 1973-2018.* Edgell, Amanda B. 2018. “Stata Friendly Freedom House Data, 1973-2018” available at [www.acrowinghen.com/data](http://www.acrowinghen.com/data)

**SDSN – The Happiness Index**

The World Happiness Report is an annual publication produced by the Sustainable Development Solutions Network which contains scores for 156 countries based on how happy their citizens perceive themselves to be. The data used for this report contains a range of numerical scores for each of these countries from 2006 – 2017.

The dataset comprises 1421 rows across 30 features, the most notable of which are discussed below:

* Country - Text based data denoting the country for which the scores relate to.
* Year - Temporal data representing the specific year the surveys were conducted.
* Social Support - This score denotes the average response to the question: “if you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”, where yes is recorded as a 1, and no as a 0.
* Healthy Life Expectancy - A continuous value representing the average number of healthy years a child is estimated to live, calculated by the WHO based on 100 different health factors.
* Freedom to Make Life Choices - Similar to the previously mentioned Social Support, this is the average response to the question: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?”
* Generosity - This value denotes the average response to the question “Have you donated money to a charity in the past month?”
* Perceptions of Corruption - This value is an average of the answers to the following two questions: 1) “Is corruption widespread through the government or not?” 2) ”Is corruption widespread within businesses or not?”.

**Source:** New York Sustainable Development Solutions Network. Helliwell J., Layard R., Sachs J. 2017. “World Happiness Report 2017.” available at data.world/laurel/world-happiness-report-data

**World Bank – GDP (current US$)**

This dataset contains the GDP in US Dollars for each country in the world from around 1960 – 2018. It is sourced from The World Bank, under their National Accounts Data and OECD National Accounts data files. The dataset is comprised of 11,508 rows across four features as follows:

* Country Name - Text based data representing the country or region the GDP value refers to.
* Country Code - A three-character string unique to each country or region identical to that in the Freedom House dataset previously discussed.
* Year - Temporal data denoting the year the particular GDP value refers to.
* Value - A continuous numeric value which shows the Gross Domestic Product for that country/region for that particular year.

**Source:** The World Bank. 2018. “The World Bank national accounts data, and OECD National Accounts data files” available at data.worldbank.org/indicator/NY.GDP.MKTP.CD

Together, these three datasets provide a very large selection of features, including continuous numeric, categorical, temporal, and location data. They contain potentially common features for merging, though it is clear that a significant deal of cleaning and wrangling will be required.