Introduction to Data Science

Lesson 1: Introduction to Data Science
Activity 2 – Suggested student responses
Country Data (MEI data set 4)

Answer the following questions:

1. The sheet or sheets in the data set are **rectangular**: they contain columns of data for the **fields** (properties or attributes) and rows for the **records** (each row represents a single instance that the data is being recorded for).

    - How many fields are there in each sheet?

      There is 1 sheet of data. The sheet has 20 fields (including the "no" field).

    - How many records are there in each sheet?

      There are 236 records in the sheet.

2. Which of the fields are numerical (showing measures or quantities) and which are categorical (text or options from a list of items)?

    Numerical: no, population, birth rate per 1000, death rate per 1000, median age, labor force, unemployment, GDP per capita, physician density, Health expenditure, Total area, Life expectancy at birth 1960, Life expectancy at birth 1970, Life expectancy at birth 1980, Life expectancy at birth 1990, Life expectancy at birth 2000, Life expectancy at birth 2010

    Categorical: Country, Region, Land borders

    A useful way to decide is to consider whether numerical or statistical operations, such as multiplying or finding the average of two values, would make sense. If they would then the field can be classified as numerical; if not it can be classified as categorical.

3. Why has this data been stored?

    This data could be used for a number of purposes. The data could be used explore how life expectancy varies in different countries or suggest the impact on life expectancy of various factors.

    There are many other possible reasons for storing this data and ways it could be used.

4. The video contained some questions that you could answer using this data. What other questions could you answer with this data.

    Some possible questions:
    - Which regions have the highest life expectancy?
    - Is there a correlation between life expectancy and GDP?
    - Is there a correlation between land area and population?

- Is there a correlation between birth rate and GDP?
- Do countries that spend more of their GDP on health have higher birth rates/life expectancy?
- Does a higher physician density have a positive impact on countries?
- Does a high unemployment rate have a negative impact on countries?

5. This data is a subset of a much bigger data set. How big is the data set it is taken from? How many fields and records could be involved?

   The CIA world factbook contains over 60 recorded values for each country. These could be explored for each year which would further increase the number of fields. There are many other values that can be recorded for countries, such as those in the DataBank World Bank.

   The number of countries is relatively stable; however, if the data was explored
6. *for individual regions within countries this would increase the number of records.*

   There are many other fields that could be recorded for countries in various categories, for example:
   - *health (infant mortality, obesity rates, …);*
   - *economic (taxes, imports, …);*
   - *energy (consumption, emissions, …).*

   The data recorded would affect the analysis that could be performed.