Exercises

We will conduct the exercises using Gapminder Tools [the online version is available at https://www.gapminder.org/tools/; the offline version at https://www.gapminder.org/tools-offline/].

Module 1 – Data Insight Fundamentals

Take some time to explore the tool. In the online version, the default starting point is a bubble chart of 2020 life expectancy vs. income, per country (with bubble size associated with total population). In the offline version, select the "Bubbles" option.

1. Can you identify the available variable categories and some of the variables? [You may need to dig around a bit.]
2. Why do you think that Gapminder has selected *Life Expectancy* and *Income* as the default plotting variables?
3. Replace *Life Expectancy* by *Babies per woman*. Observe and discuss the changes from the default plot.
4. Formulate a few questions that could be answered with the default data.
5. Formulate a few questions that could be answered using some of the other variables.
6. At what point in the data science workflow do you think that visualizations of this nature could be useful?
7. Do these visualizations provide a sound understanding of the system under investigation (the geopolitical Earth)?

Module 2 – Data Collection and Data Management

1. What do you think the data sources are for the underlying dataset? [You may need to dig around the internet to answer this question].
2. Are all variables and measurements equally trustworthy? How could you figure this out?
3. Is the underlying dataset structured or unstructured?
4. Provide a potential data model for the dataset.

Module 3 – Data Processing and Data Cleaning

1. Explore the dataset with the Gapminder Tools in its default configuration. Do you think that there could be problems with the reported values? For instance, select Sweden and the United States from the checkbox menu on the right and follow their path from 1799 to 2018/2020. From what point onwards are the values sensible? What do you think is happening at the start of the time series?
2. Follow Eritrea for the same duration. Look up the country's independence date from Ethiopia. What do you think the measurements prior to that date represent?
3. Follow Austria for the same duration. Look up the historical timeline of the country's boundaries (Austria-Hungary, Anschluss, modern borders, etc.). What does that imply for the measurements?
4. Follow Finland for the same duration. What happens in 1809? Does that tell you anything about the way data is coded in the dataset?
5. De-select all countries and let the simulation run from 1799 to 2018/2020. Can you identify instances where a large subset of observations behaves in unexpected manners? If so, do you think that this is due to data cleaning/data processing issues?
6. Continue exploring the dataset. You may change which variables are displayed or work with some of the other visualization methods. Overall, do you think that the dataset is sound? Would you use it to run analyses? What are some of its strengths and weaknesses?

Module 4 – Basic Data Analysis

1. What are the types of the 4 default variables (*Life Expectancy*, *Income*, *Population*, *World Regions*)?
2. Play around with the charts for a bit. Can you find pairs of variables that are positively correlated? Negatively correlated? Uncorrelated?
3. Among those variables that are correlated, do any seem to you to exhibit a dependent-independent relationship? How could you identify such pairs?

4. Can you provide an eyeball estimate of the mean, the median, and the range of various numerical variables?
5. Can you provide an eyeball estimate of the mode of the categorical variables?
6. Can you identify epochal moments (special temporal points) in the data, where a shift occurs, say?

Module 5 – Statistical Learning

1. In the default configuration, we can identify some potential association rules. Using visual and ballpark estimates, evaluate the performance of the following rules:
    i. *Income* > 8000 → *Life Expectancy* > 70
    ii. *Income* < 8000 AND *Life Expectancy* < 70 → *World Region* = Africa
2. Play around with various charts and variables and identify/evaluate 5+ additional association rules.
3. Identify groups of similar countries, in 2018 [be sure to validate your groups using various charts].
4. In the default configuration, follow the trajectories of Finland, Sweden, Iceland, Norway, and Denmark between 1900 and 2018. Do the countries appear to follow similar trajectories? Are there outliers or anomalous trajectories?
5. Repeat step 4 for Brazil, Paraguay, Uruguay, Venezuela, Colombia, Peru, and Ecuador.
6. Based on your results in steps 4 and 5, would you expect the trajectory for Argentina to be more like those of the Nordic countries or those of the South American countries? Or perhaps neither? Is your answer the same over all time horizons?

Module 6 – Data Visualization Basics

1. Now that you've had some more time to play with the dataset, let us revisit a question you answered in Module 1: at what point in the data science workflow do you think that visualizations of this nature could be useful?
2. What are the ways in which observations could be anomalous? Have you found any such anomalies? Do you have explanations? [Think back to the South African example in class]
3. Pick 2+ "definitive" visualizations (methods, variables, etc.) other than the default configuration. What are some important insights?
4. How would you describe the insights of step 3 without resorting to visual vocabulary?
5. Can you think of ways in which the data of interest to you in your day-to-day activities could benefit from the same treatment? What situations could you explore in such a scenario? How would that help your team better understand the system under consideration?