Tutorial 2:

Introduction to Data Science

Why use Data?

Ernst C. Wit

September 23, 2024

Exercises

1. **What is an Astronomical Statistical Unit?** Youden (1994) collected data from various studies trying to determine the distance from the earth to the sun. This is called an *astronomical unit* (AU) and is fundamental in astronomy to describe distances. The data is available as `AU.csv` on the course website.

	study	year	million miles
1	Newcombe	1895	93.28
2	Hinks	1901	92.83
3	Noteboom	1921	92.91
4	Spencer 1	1928	92.87
5	Spencer 2	1931	93.00
6	Witt	1933	92.91
7	Adams	1941	92.84
8	Brouwer	1950	92.98
9	Rabe	1950	92.91
10	Millstone Hill	1958	92.87
11	Jodrell Bank 1	1959	92.88
12	STL	1960	92.93
13	Jodrell Bank 2	1961	92.96
14	Caltech	1961	92.96
15	Soviets	1961	92.81

   a. Identify the statistical units and their features in this data.

   b. If you had lived in 1961, how would these data help you to determine an AU unit?

c. Load the data and try to determine a value for the AU. Modern calculations put it at

$$AU = 92,955,807.2730\text{miles}$$

Determine how much closer your choice is than any of the individual measurements.

2. **Standard soldier.** Armies around the world employ various *standards* to select their soldiers. The following table describes the minimum length required by the French army over the years. The dataset `soldier.csv` is available on the class website.

	year	height
1	1780	178
2	1789	165
3	1818	157
4	1852	156
5	1862	155

(a) The structure of the data is very similar to the data from the previous exercise. Moreover, they both are dealing with standards. What is the main difference between the two datasets?

(b) What do you conclude when you look at the data? Can you explain why this may be the case?

3. **Consider your own data records.** In the last tutorial, you created a dataset with at least two features for each statistical unit associated with the content of the following painting:

(a) Do your data on each of the two features point at possible *standards* (e.g. standards of 17th century painting)?

(b) Analyse your data within R/Python and conclude what standards you might have found.

(c) What pitfalls might your analysis have encountered? Discuss whether any of the following fallacies may have been relevant:

- False cause fallacy (post hoc ergo propter hoc)
- Slippery slope fallacy
- Unrepresentative sample fallacy
- Cherry picking fallacy
- Overgeneralization
- False analogy
- Confirmation bias
- Misleading vividness
- Common cause fallacy
- Bandwagon fallacy