Visualization
Prof. Bernhard Schmitzer, Uni Göttingen, summer term 2023

Problem sheet 1

- *Submission by **2023-05-10** 18:00 via StudIP **as a single PDF/ZIP**. Please combine all results into one PDF or archive. If you work in another format (markdown, jupyter notebooks), add a PDF converted version to your submission.*
- *Use Python 3 for the programming tasks as shown in the lecture. If you cannot install Python on your system, the GWDG jupyter server at https://jupyter-cloud.gwdg.de/ might help. Your submission should contain the final images as well as the code that was used to generate them.*
- *Work in groups of two to three. Clearly indicate names and matrikelnr of all group members at the beginning of the submission.*

Exercise 1.1: applying Tufte's design principles.
The files

- `energy.png` (https://www.umweltbundesamt.de/themen/klima-energie/erneuerbare-energien/erneuerbare-energien-in-zahlen#uberblick)
- `baseball.png` (https://benfry.com/salaryper/)
- `machine-learning.png` (Bertolini et al.: Machine Learning for industrial applications: A comprehensive literature review, Expert Systems with Applications 175, 2021, https://doi.org/10.1016/j.eswa.2021.114820)

contain three examples of statistical charts found 'in the wild'. Apply Tufte's design principles on minimalism, data ink, and chart junk to *two out of three* of them. This means: list/describe/mark which parts of the charts are not data ink, and describe/sketch how an improved version of the chart could look like.
Remark: It is not necessary to re-create the charts with plotting software. A simple 'dissection' and a sketch, as shown in the lecture, using e.g. Paint or Gimp is fully sufficient.

Exercise 1.2: a first visualization task.
The file `mpg-data.csv` contains the `mpg` example dataset from the `ggplot2` library (https://ggplot2.tidyverse.org/reference/mpg.html). The dataset contains information about the fuel efficiency of various car models. It can be imported as `pandas` data frame via the `read_csv` function. For each car/model, the column `class` gives the category/class (e.g. 'suv' or 'compact'), the column `displ` gives the engine displacement in liters, and the column `hwy` gives the fuel efficiency in miles per gallon on highways.

1. Split the dataset into different car classes. For each class, perform a linear regression on the dependency of `hwy` on `displ`.
2. Give a scatter plot of `hwy` against `displ`. Make sure that the class of each car can be determined from the plot. Add straight lines showing the regression lines for each class. Make sure that your plot has appropriate axes labels and legends.

Hint: Code from the examples `2023-04-17_Example-Anscombe` and `2023-04-17_WorldDemographics-Test-01` can be helpful for this exercise.