Exercises 3

1. Descriptive statistics and estimation in literature

Identify the visualisations and numerical summaries used in the given parts of the following two articles. Are the reported numbers used for descriptive analysis or estimation? (NOTE: You may have to browse through the whole article in order to fully understand the study.)

- Article 3 (Figures 1 & 2 and related discussion): P. Andre *et al.*, Globally representative evidence on the actual and perceived support for climate action, DOI: https://doi.org/10.1038/s41558-024-01925-3
- Article 4 (Table 1): M.S. Venäläinen *et al.*, Easy-to-use tool for evaluating the elevated acute kidney injury risk against reduced cardiovascular disease risk during intensive blood pressure control, DOI: https://doi.org/10.1097/HJH.0000000000002282

2. More cyclists

Load the data available in the file `more-cyclists.csv`, which contains daily numbers of cyclists spotted on selected streets.

- For each weekday (Monday - Sunday), calculate the mean daily number of cyclists for each street. Plot the means as a bar plot.
- For each month (January - December), calculate the mean daily number of cyclists for each street. Plot the means as a bar plot.
- For each pair of streets, calculate the Spearman's correlation coefficient and create a scatter plot.
- What information do these plots reveal?

3. Running

The file `running.csv` contains simulated lap times of 50 runners measured before and after training. You can assume that the sample is representative of the population.

- Estimate the mean and variance of lap times in the population before and after training.
- Estimate the mean change in lap times due to training.
- What conclusions can you make about the population and/or the training based on these estimates?
- (BONUS) Use confidence intervals to estimate the mean change. What conclusions can you make?
  - See the `statsmodels.stats.weightstats.DescrStatsW` class and its `tconfint_mean` method available in the `statsmodels` library (https://www.statsmodels.org/stable/index.html).

4. Electric bikes (continues)

Continue to analyse the electric bike data from the earlier exercises.

- Create both visualisations and numerical summaries to explore the data.
- Make comments about the data. Is there anything odd or interesting in the data?
- Process irrelevant records and invalid values like previously.

5. Electric bikes (continues)

Continue to analyse the electric bike data from the earlier exercises.

- For each ticket type, calculate the number of trips made, the total distance travelled, the total time travelled and the total amount of fees paid. Examine your results and make comments about the customers.
- For each ticket type, visualise the monthly rental activity in terms of the total distance travelled. Examine your results and make comments about the development of the rental activity over time.
- For each trip in the data set, calculate its net energy gain, which is defined as the difference between the energy collected and consumed during a trip. Visualise the distribution of this new variable. Make comments about how much battery levels tend to change during trips.