Foundations of Data Engineering and Analytics

Department of Computer Science
Universität Innsbruck

Eva Zangerle, Monika Steidl, Maximilian Mayerl

Sheet 1: Datasets

Exercise 1 (Jupyter Setup) [2 Points]

This exercise is all about getting your working environment for the course up and running. Set up an installation of Jupyter that you will work with for the rest of the course. Your environment should include proper dependency management. We recommend that you use a tool like `pipenv` or `poetry` to get an environment that can be easy transferred and reproduced on another machine — this also makes working in teams a lot easier.
Optionally, though highly recommended, you can also have a look at additional tools such as nbval[1] for validating your notebooks or nbQA[2] for further quality assurance tools (code formatting, type annotations, static code analyses, etc.).

Exercise 2 (Dataset Wrangling) [5 Points]

In this exercise, you will work with your first dataset. You can consider this exercise a warm-up for future tutorials; it is rather loosely defined to let you freely explore the datasets, coding environment, and libraries. We will learn more structured approaches towards dataset analyses later on in this course. I strongly recommend using this exercise to dive into using pandas[3] for analyzing and manipulating structured data.

a) *1 Point* To collect a variety of datasets during the tutorial, choose any single dataset of your liking from e.g., Kaggle[4], Hugging Face[5], data.gov[6], UC Irvine Machine Learning Repository[7], or FiveThirtyEight[8].

b) *1 Point* Read the data into a suitable data structure and manually inspect the data contained (columns, data formats, etc.).

c) *2 Points* Explore the dataset and its characteristics (distributions, aggregations, plots, etc.). Try to get a sense of the dataset.

d) *1 Point* Reflect on the metadata that you were presented with for the dataset. Were the contents, formats, origin, etc. of the dataset clear or was data missing that you would have required?

[1] https://github.com/computationalmodelling/nbval
[2] https://github.com/nbQA-dev/
[3] https://pandas.pydata.org/
[4] https://www.kaggle.com/
[5] https://huggingface.co/
[6] https://www.data.gov/
[7] https://archive-beta.ics.uci.edu/
[8] https://data.fivethirtyeight.com/

Exercise 3 (Interactive Visualization) [3 Points]

In exercise 2, you already obtained a dataset and create some static visualizations for it in 2c. In this exercise, you are going to extend this and add *interactive* visualizations that allow the user of your notebook to easily explore the dataset. You can choose any library that you prefer for this. Our recommendation is Plotly[9], which is widely used in the data science community.
Use your library of choice to build a dashboard-style visualization of the dataset you already used in exercise 2. Provide the user with interactive widgets to control what the dashboard shows. Like exercise 2, this is also a rather loosely defined exercise. The main goal is to familiarize yourself with the library that you have chosen, so that you can use it effectively for future exercises. However, please keep in mind that your submission has to show enough work to be worth 3 points.

Hint

Follow the checklist (our definition of done) we agreed upon in the first tutorial session to make sure that your notebook is well-structured, well-written, and fulfills our (minimal) criteria regarding quality.

Submission

- `dea01_dataset_analyses.ipynb`
- `README.md` including instructions on how to set up the required environment
- `dependency_specification` pipfile, requirements.txt, or similar

Important: Submit your solution to OLAT and mark your solved exercises with the provided check-boxes. The deadline ends at 23:59 on the day before the discussion.

[9] https://plotly.com/python/