Foundations of Data Engineering and Analytics

Department of Computer Science
Universität Innsbruck

Eva Zangerle, Monika Steidl, Maximilian Mayerl

Sheet 2: Data Preparation and Data Quality

Exercise 1 (Data Wrangling) [5 Points]

Download the data set (file `dessert.json`) from Olat. Transform the data set that it will be suitable for data analysis by formating, transforming and reshaping the data. Argument why your chosen approach is necessary and how it improves the dataset.

a) *1 Point* Handle missing data appropriately. Which method did you choose and why?

b) *1 Point* Which dessert types exist and how are they distributed?

c) *1 Point* Display the median of vitamins by dessert type.

d) *2 Points* Are there outliers in the amount of available vitamins?

- Please use several outlier detection methods to justify your answer.
- How do the outlier-detection algorithms and their results differ?

Exercise 2 (Social Data) [5 Points]

Read the following paper regarding biases in social data, especially Section 3-8, to learn which biases social datasets may be exposed to. Come up with your own hypothetical **social** study, research question, data collection strategy etc. Identify which issues or challenges your study and dataset may be exposed to and identify how to mitigate these issues or challenges.
`Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological`
`pitfalls, and ethical boundaries. Frontiers in Big Data, 2, 13.`

a) *1 Point* How will you create your dataset. Which **general biases and issues** does your dataset have? How can you mitigate these?

b) *1 Point* Which issues at the **data source/origin level** does the dataset have? How can you mitigate these?

c) *1 Point* Does your **data collection** and **processing** introduce additional biases? If so - which and how can you counteract these?

d) *1 Point* Which type of **analysis** would you choose for your study? Can a combination of different types mitigate the risk of a biased insight?

e) *1 Point* How do you plan to **evaluate and interprete** the findings? Which issues may occur and how can you mitigate these?

Submission ↑

- `dea02_ex1_dataset_analyses.ipynb` with thorough explanations of your decisions and undertaken steps for data wrangling
- `dea02_ex2_social_bias.pdf` describe your chosen study and throughly answer the respective questions.

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.

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