Module 3: Data Quality Assessment and Data Exploration

Overview

In Module 3 we discuss the dimensions used to assess data quality offering a definition and examples. Then, we focus on and the task of getting to know our data, highlighting exemplary, widely used, exploratory data analysis (EDA) functions. Finally, we perform EDA basic data cleaning and data transformations on the data ingested from the previous module.

Objectives

Upon completion of this module, students will be able to:

1. Describe the concept of data quality assessment
2. Understand the role of data quality in data-intensive problems and the implications of its absence
3. Identify the mechanism of Exploratory Data Analysis (EDA)
4. Discuss examples of EDA methods
5. Synthesize EDA and data transformation methods

Readings (1 hour)

- **(https://artint.info/2e/html/ArtInt2e.Ch7.html) Big Data: Trade-off between Data Quality and Data Security, M. Talha et. al., 9th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS) April 29 – May 2, 2019, Leuven, Belgium (https://canvas.vt.edu/courses/176740/files/28995277?wrap=1)**
- **Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets, N. Gupta, et. al., arXiv: 2108.059335v2 (https://canvas.vt.edu/courses/176740/files/28995298?wrap=1)**
  Towards Implementing Responsible AI, C. Sanderson et. al., ArXiv 2205.04358v4, 2023 (https://canvas.vt.edu/courses/176740/files/28995280?wrap=1)

Watch (30 minutes)

- **Lecture 3 Video (https://canvas.vt.edu/media_objects_iframe/m-55CMQbPZ8vP6SJfGvguRM7ShmqoR4dgD?type=video?type=video)**

- **Lecture 3 Slides (https://canvas.vt.edu/courses/176740/files/28995311?wrap=1)**

Assignment (Lab & Homework) (1 hour)

- **Assignment 2 (https://canvas.vt.edu/courses/176740/assignments/1816311)**

Recitation (1 hour)

The recording of the optional synchronous ZOOM session for this lecture will be linked here.