NDB048 Data Science – Structure of Report

1. Introduction

    - Reasons for carrying out analysis
    - What do we expect / what story are we telling?

2. Data

    - Data size, format, fields, data types, …
    - Data source, reliability?
    - Basic statistics + visualizations
    - Sanity checks
    - Missing values, outliers, …

3. Methodology

    - Description of data selection, cleaning, transformation
    - Discussion of the choice of analyses for the selected task
    - Description and assessment of the results
    - Description of usage of the approaches for Big Data processing (MapReduce / Spark / multi-model DB) for a selected part of the analysis
        - Note: Add the used source codes/scripts as an attachment
    - Possibly description of the iterations made

4. Summary

    - Summarization of the findings
    - Possible exploitation, data/methodology extensions

Typical problems:

1. Useless graphs/tables (Does it carry any useful information? Why is it there? Is it worth it?)
2. Too many similar graphs, inappropriate type of graphs
3. No comments on the graphs/tables (What insights a graph provides should be summarised in text, too.)
4. Graphs without appropriate description (axis labels, titles, population constraints, ...)
5. Graph inconsistencies across a whole document (color, scales, graph types, …)
6. No summary/conclusion
7. Useless information (copying information from slides, list of used technologies, …)
8. Reproducibility problems, code behind the report issues