title: Term project

Format

The term project will delivered in teams, consisting 3 students each.

Goal

- You will create a data product suitable for analysis and submit it together with documentation outlining the technical choices you made. The purpose of this assignment is for you to apply all the concepts learned during the course.
- Your grade depends on the number of different tools and concepts you apply and how appropriately you do so. The topic of the data product has no bearing on your grade, but you may want to think about what data you are going to work with in your Capstone Project and work with that.

High level requirements

- Combine at least two distinct datasets
- Use an example dataset or bring your own
- Persist at least one dataset in a database (SQL or NoSQL)
- Build an ETL data pipeline with Knime
- Do some data cleaning
- Do some basic analytics or visualization on the end of the pipeline

Delivery

The project artifacts should be stored and handed over in a folder "Term DE2" in a GitHub repo.

The main artifact submitted is a 4-6 page report (including figures and data citation), with the following requirements: * documents the solution provided * documents the the technical choices made * documents the data model (ER diagram for RDBMS) * documents the analytics and/or visualization * indicates which team member did what in the project * can be a README in your git folder

Artifacts to be submitted: * Term project report * Power point presentation - material for a few minutes presentation * Knime workflow file * Source files * Script (or instructions) of data persistence (sql file in case of a RDBMS)

Reproducibility: the project should be reproducible in a straightforward manner. In other words, we should be able to run your code and obtain the same outcome as you.

Grading criteria

- Fitness of the input dataset to the purpose **5 points**
- Complexity of the input data set **5 points**
- Usage of concepts used in the class **10 points**
- Knime pipeline **20 points**
- Using database **10 points**
- Delivery: Naming, structure **10 points**
- Delivery: Report **15 points**
- Delivery: Presentation **10 points**
- Reproducibility **15 points**

Extra points: - Using NoSQL in the project - Using API in the project - Anything special not covered during the course but, makes sense in the project context

Submission and deadlines

11th of December 2020 EOD - Every material should be committed to GitHub. Link to GitHub should be provided into the Google Sheet already shared (Term project teams)

12th of December 2020 17:00 - 19:30 - Teams will present their results into a Zoom session. (Use the Zoom link from Moodle) * The order of presentation will be by team number * Every team will have 7 minutes for presentations, followed by 7 minutes Q&A