*Technical report: The technical report is simply an essay you can return in PDF (max 5 pages).*

*This is the place to use the technical terminology, and refer to all the methods you used backend and normally wouldn't interest your target audience.*

*Reflect on your initial project proposal: what worked? What didn't (and why)?*

*What changes did you have to make from your initial plan and why?*

*Tell us about your experience, technical implementation of the project and reflect on your learning outcomes.*

*The structure of the report is up to you.*

*If it helps, you can use the elements of the project proposal, e.g. data collection, preprocessing, visualisations, machine learning, communication of results, building the platform, etc.*

-

**Project working methodology**

Given the very open nature of this course project, we begun exploring options and discussing our interest at very early stage. We all seemed to share the opinion, that the best way forward is to do small scale experimentation on possible topics. In this preliminary phase, we explored a number of topics from our joint interest areas. At later stage, we then iterated on a few of the most promising topics to see if they were proper fit for the course.

Collaboration environment was the second thing we begun thinking early. All of us coming from a bit different background it took a moment to find something that is available to all of us - and suitable for project. As for the initial topic exploration and to-do list we first gave OneDrive a go, but then ended up using Google Docs for easy access. When moving on to the further steps of the project, we needed something more robust to host collaborative coding and multiple data sets. Again, tried few, and ended up using GitHub for simplicity.

**Topic selection process**

As mentioned previously, we iterated few topics a bit further to determine proper fit. These topics included sleep schedule suggestions, efficiency of public responses on covid-19 pandemic, customer churn in telecom setting and weather-based suggestions for planting different crops around the world. All of them were interesting topics, but either they were already well-addressed, or the data was too complex or scarce. We then settled with something very concrete: comparison of two note-taking methods from ecological point of view.

After we had decided on a question, we wanted to address in our group work, we begun deeper data exploration to see if we can find evidence to build a proper argument. It posed a real challenge. Production processes for tablets and paper are complex and diverse. And same goes for energy production and recycling. Respectively, data were overwhelming. We skimmed through around ten data sources to get an idea what we are looking at.

After we had a bit of general understanding, it was time to start reiterating the question and decide onto what we really want to focus on. Narrowing the project down was hard. In the end, we deemed that production process was just too broad issue to be handled within the boundary conditions presented by the course schedule. Thus, we decided to look at ecology of use through energy consumption and add paper recycling as a second point of view.

Keeping our story and point of view in mind, when doing initial data analysis and pre-processing, we drifted a bit. Even after narrowing the project down, data was partially a bit overwhelming and thus made it difficult to draw clear connections and conclusions to our original idea. Example of this would be the generality of the *owid energy data*, which was harder than expected to connect into our train of thought.

**Data acquisition and pre-processing**

Data acquisition begun after we had narrowed the project down. We had four primary sources (see <https://github.com/TimoKoski/IDS_project> for more info):

1. Energy data from <https://ourworldindata.org/energy>
2. Refined version of energy data via <https://github.com/owid/energy-data>
3. Paper production from [www.statista.com](http://www.statista.com)
4. Recycling rates in Europe from <https://ec.europa.eu/eurostat>

All source data sites had good options for exporting the data. We used .csv and .xlsx formats. Latter was mainly used to gain insight on *owid energy data*, which proved to be more complex and contain a lot of redundant data for us. Excel was faster for quick filtering and glimpsing through values in multiple columns. This step was important and gave us valuable information on the nature and sufficiency of data. One supplementary source was referred after this phase (Cepi’s Key Statistics 2020).

Some of the source files required a bit of a content clean up before being ready to be used in Jupyter Notebook. Mainly these cleanups were headings or reference information given in .csv file. All the source files were then added to our repository.

We began preprocessing with Python in Jupyter Notebook and Pandas & Matplotlib were the main libraries utilized. For initial data analysis we loaded the data into Data Frames and familiarized ourselves with the content of the four source data sets. Overall analysis included slicing, plotting, sampling and simple calculations to understand how to best prepare the data for further next steps.

Generally, the data we used was of proper quality and sufficiently simple. In addition to cleaning up the files, pre-processing only included null value management and time series restrictions. We had two major findings in this phase. For recycling data set it was less obvious what data set should we use (ended up identifying paper & cardboard packaging data as best option). *Owid data set* continued to be quite complex, and we needed to slice it down for smaller segments to identify what story it was telling us, and in what way we could derive some value from it.

After initial commit of the pre-processing, we developed a more formal folder structure in git. This was to enable better collaboration and general easiness of use. A number of bug fixes (e.g. file extension problems) was also introduced later on.