University of Helsinki

Introduction to Data Science

Lecturer: Teemu Roos

Autumn 2021, period 1

Technical Report for the Mini-Project

Group members:

Kristina Svetova

Timo Koski

Meike Knull

This is the technical report for the project work “tablet vs. paper: how to take notes sustainably” by Kristina Svetova, Timo Koski and Meike Knull. Following the structure of our process this report first focuses on the topic finding process, then on the collection of the data as well as the data preprocessing. After that, our data analysis and the learning from the data is described. The report concludes with details on how the visualization and the deliverable have been designed.

Project working methodology

Given the very open nature of this course project, we begun exploring options and discussing our interest at a 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 several topics from our joint interest areas. At a later stage, we then iterated on a few of the most promising topics to see if they were a proper fit for this course’s project.

The collaboration environment was the second thing we begun to think about early on. Since we all have different backgrounds, it took a moment to find something that is available to all of us and suitable for the project. As for the initial topic exploration and to-do list we first tried to work with OneDrive, but then ended up using Google Docs for easy access. Moving on with the project, we needed something more robust to host collaborative coding and multiple data sets. Again, we tried a few, and ended up using GitHub because of its simplicity.

Topic selection process

As mentioned before, we worked on a few topics a bit further to determine whether it is a 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: the comparison of two note-taking methods being using a tablet or paper from an ecological point of view.

After we had decided on the question we wanted to address in our group work, we began 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 of what we are looking at.

After we had a bit of a 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 the production process was just too broad to be handled within the boundary conditions presented by the course’s schedule. Thus, we decided to look at the ecology of use through energy consumption and in comparison, paper recycling. Targeting mostly students and generally interested people, we thought that this could, even if it is a minor impact, help reduce the individual's ecological footprint and raise awareness for making improvements for our climate in every part of one’s life.

Keeping our question of interest 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. An 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 began after we had narrowed the project down. We had four primary sources (see https://github.com/TimoKoski/IDS_project for more info and precise links):

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
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 the 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 the .csv file. All the source files were then added to our repository.

We began preprocessing with Python in Jupyter Notebook and Pandas and 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 the recycling data set it was less obvious what data set we should use (and ended up identifying paper and cardboard packaging data as the best option). The *Owid data set* continued to be quite complex, and we needed to slice it down for smaller segments to identify what way we could derive some value from it.
After the 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.

Learning from data

After working with the data, the task was to analyze them. To do this, we selected data from our sources based on EU membership (to simplify the task, a dictionary was created with the name of the country and possible codes that were used to describe the country in tables, for example, iso code, country code).

Attention was drawn to the distribution of indicators such as, for example, energy consumption during the year between countries, for better understanding of the data, visualizations of these data were built to show how countries are changing in these indicators. We also examined the energy consumption in each of the specific countries, namely how the share of energy consumption per person changed over time. We also considered data on renewable sources and compared the distribution of the shares of renewable sources in each country separately. We studied how much the share of renewable energy increased or decreased in countries every year. Similar work was done with the indicators of paper recycling in different countries: such indicators as the volume of paper recycling, the country's rating by recycling were considered, visualizations were built describing the distribution of the volume of recycling over time.

Visualization

To visualize the results most important for our target group, students and the interested public, we wanted to avoid tables, since they take too much effort to read the gain out of it. We had a look at different kinds of graphs and graphics to come to the conclusion that we want to have three main aspects in the deliverable of which one would be a map, one a graph and one was to be a little bit of explaining text.

The map was supposed to show the share of renewable energy of the world using the data from the data set on energy shares from 1900 to 2020. To give as much information as possible without overwhelming the user we chose to reduce the kinds of energy shares displayed to the most relevant one (renewable energy) and ask the user for an input. This input is the year of which the shares are to be displayed. If the project was to be extended, a similar map could be created for the share of recyclable paper, or we could try to visualize both characteristics (share of renewable energy and of recyclable paper) in one map using patterns to show the value for one characteristic and colors for the other. If this was still a good visualization would have to be evaluated when it is implemented because due to the two different scales of the two characteristics it may be harder to gain the added value from the map.

To implement the map, we used the libraries Geopandas, Pandas, Json and Folium. At first, we downloaded a data set including some information on countries including their iso code and their coordinates in the format of a shapefile. After reducing it to the columns needed, we ask the user to put in the year that is of interest. With that information, we reduce the data frame including the data from the data set about shares of different kinds of energy in different countries for 1900 to 2020 to the rows with the year of interest.

Then, we merged the data frame with the geographical information about the countries with the data frame containing the information on energy shares and reduced the merged data frame again to the relevant columns. This way, we were able to connect the geographic of a country very easily to the correct value for the share of renewable energy without taking care of the two data frames having the same order of rows.

Next, we created a json file including the information of the merged data frame because we needed this format to create the map with folium. To visualize the data in the map, we wanted to color the countries in according colors, thus, we chose to create a choropleth map. Trying different color schemas, the one chosen was the one least likely to imply unwanted biases (like for instance red-green schemas would). We then saved the map as an html file just in case we may need that later when we want to put it online and to have the map available without the notebook.

Creating a choropleth map was new to us and thus, included more googling than other parts of the project. Especially, using a json format for that was rather unexpected and working on changing the details from what we found online to what we wanted to see and trying to vary different parameters was very interesting. We also tried to create the map with bokeh which would have been very nice to use since there are tools to, for instance, include a panel to choose the year displayed, which would be more interactive and convenient to use than the way we implemented it. But unfortunately, we did not have enough time to fully understand the relevant modules of bokeh and the available code examples for such interactive maps to get the map with the panel running properly.

The second visualization that we chose to put into the deliverable was a graph. Just like for the map, we wanted to allow the user to put in an input (this time a country) and get a graphic out that shows two graphs: the share of recycling paper and the share of renewable energy over time (1900 to 2020) to see the development. The function to create said graphic has been written during the analysis of and learning from the data that we used. Thus, we now just wrote the code to ask for a country name. To ensure that the country requested is in the data frame we use, the while loop tests whether the name is in the list of the values of the column of country names.

At this point one of the problems we faced appeared. To give an impression of the process of solving problems we came across, this was our procedure to solve this problem: We first tried to check if the user input is in the column of the data frame (‘ while not c_o_i in c_e2020[“country_x”]’) but that loop never ended because c_e2020[“country_x”] is still a data frame of which a string like the input will not be identified as an element of that data frame. To find that out, we wrote a quick test to manually check which part is not working as expected. In this test, we looked at the structure of c_e2020[“ country_x”] which is still a data frame and we made sure that we just save the string in c_o_i, which we did. Thus, the problem had to lie within the actual logical statement ‘c_o_i in c_e2020[“country_x”]’. Looking at which countries are in ‘c_e2020[“country_x”]’ and checking the statement for those and it returning false, supported our thought that the problem lies within this statement. Since c_o_i was indeed a regular string we concluded that the problem probably lay within ‘c_e2020[“country_x”]’ being a data frame. Since we know ‘… in …’ definitely works with lists we made ‘c_e2020[“country_x”]’ into a list. After that the statement was true when c_o_i was a country that was part of our data frame and false otherwise.

Deliverable

As described in the visualization part, we wanted to do an interactive map, a graph created with the data based on an input from the user and some explaining text as our deliverable. Unfortunately, the group member, who wanted to work out how we can put the deliverable online so that it was interactive, got sick and thus, could not do that. Hence, we decided to reduce the deliverable to a static pdf reachable via our public git repository. This pdf includes the same three components as before but static: the map is shown for the most recent year we had data on (2020) and the graph shows the development for Finland. If we had more time or manpower, we would have liked to make it interactive to make it more interesting for our target group and to allow them to have a look at maps from different years and graphs for different countries to get a better feeling for the information delivered and add more value for people all over the world (by not just showing them the development for Finland).

Also, we would have liked to be able to give a specific answer to our question of interest. But even though we already tried to make the question specific, it is still hard to find an exact value to compare the sustainability of paper and tablet usage. Therefore, we designed the deliverable in an informative way. We hope that with the value the target group gains from this deliverable, they will be able to make a more educated decision for their note taking tool and be even more aware of being able to have at least a little bit of an impact with things like note taking on preserving the world.