Project Worksheet Day 1

August 23, 2021

During the course week you will work on a group project, and you will start to work on it TODAY! You will select a data set to work on, formulate the research questions you want to answer with it, and think about possible (statistical) methods to use. Note that no programming skills are required at this stage!

Project Workspace

For a start, you can simply share documents and other files in your project team on MS Teams, they will appear in the “Files” tab there.

Data Selection

In your project you will do an exploratory data analysis (EDA) on a data set provided by the Centraal Bureau voor de Statistiek (CBS, the national statistics organization of the Netherlands). Exploratory data analysis means to analyze data sets by summarizing their main characteristics using descriptive statistics and visual methods. Formal modeling and hypothesis testing can be done, too, but does not have to be part of an EDA. The CBS provides large amounts of data sets on a variety of topics, and typically in good quality. As you will see, they are however often complex and hard to read manually, so letting a Python program turn them into something readable and informative will make a large difference.

With your project group, go to the CBS Open data StatLine portal (online at https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS) and browse through the topics and tables to find a data set that you find interesting. It can be related to your study subject, but it does not have to be. Make sure that the data set that you choose is not too small. There are some tables which only contain a handful of data rows, that does not give a lot of possibilities for interesting analyses. The “Preview table” under “More information” is very useful for a quick impression of data sets, but you will need the complete set in program-accessible format later. Thus, for the data set of your choice, also download the “Metadata” and raw data (“Original dataset”) files in CSV format. CSV files can be opened by spreadsheet programs like MS Excel or OpenOffice Calc. The metadata file contains information on what data is contained in the data set, which codes are used and what they mean, etc., and is needed to make sense of the data in the actual data file. Write a bit of text (max. 1 page in A4 format) where you briefly state why you find this data interesting and describe the data set in your own words. For example:

For example, we might have chosen the data set “ICT kennis en vaardigheden” (ICT knowledge and skills, available at https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=83428NED&_theme=472). We downloaded the files 83428NED_metadata.csv and 83428NED_UntypedDataSet_30012020_111246.csv. We find this data set interesting because it relates to many discussions that we have among computer science lecturers, like which ICT skills we can expect from incoming students, which skills they should develop through their studies, etc. The data set contains information on the self-reported overall ICT skill levels (no, little, basic, advanced) and the presence of specific ICT-related skills (such as installing

software, working with spreadsheet software, or writing computer programs) for different groups of people (such as men/women, age groups, educational levels). The data have been collected annually since 2015.

Research Questions and Analysis Methods

Now brainstorm about questions that you would like to get answered with the data set that you have selected. Choose at least as many questions as you have members in the group, and formulate them as research questions as shown below. Don't worry about how to possibly implement all that with Python at this moment, we will get to that in the next days.

Some possible research questions for the "ICT kennis en vaardigheden" data set:

> *RQ1: How do the reported ICT skill levels differ between the groups?*
> *RQ2: How have the reported skill levels changed over time?*
> *RQ3: Which skills do most people have, and which lack most often?*

For each of the research questions you came up with, make a list of (statistical) methods that could be used to answer them. As we will focus on programming concepts in this course and not teach data analysis methods as such, please use methods that you are already familiar with. If you want to read up on statistical methods, David Lane et al.'s free online book "Introduction to Statistics" (http://onlinestatbook.com/2/index.html) is a great reference. In particular, Chapter 2 (Graphing Distributions, http://onlinestatbook.com/2/graphing_distributions/graphing_distributions.html) and Chapter 3 (Summarizing Distributions, http://onlinestatbook.com/2/summarizing_distributions/summarizing_distributions.html), might be helpful for you to find suitable descriptive statistics methods (and the correct English terms for them). Pay attention that the chosen methods are applicable to the kind of data you have, but as mentioned before, don't worry about the implementation in Python for now.

For example, the following methods could be used to answer the questions on the "ICT kennis en vaardigheden" data set:

RQ1: How do the reported ICT skill levels differ between the groups?
Possible methods: (Clustered) bar charts per skill with clusters of four bars (for no/low/medium/high skill levels) for each of the considered groups on the x-axis, and the percentage of the reported skill levels on the y-axis.

RQ2: How have the reported skill levels changed over time?
Possible methods: Line graphs per skill and group with the years on the x-axis, the percentage of the reported skill levels on the y-axis and different colors for the lines to represent the different skill levels.

RQ3: Which skills do most people have, and which lack most often?
Possible methods: For the most and least common skills (in a specific year), sort the reported skills by their percentage. Take the top 5 and plot them in a bar chart with the skill names on the x-axis and the percentage on the y-axis. Do the same with the bottom 5 skills to find represent the skills that are lacking most often. The data and development over several years could be represented by a clustered bar chart, where each cluster has the bars for e.g. three different years.