Interactive Workgroup 1 - Instructions

Data Visualization

Introduction

During the interactive workgroups you will have the opportunity to apply the concepts learned in the lectures to explore a dataset of your interest. The aim of this first workgroup is to find a dataset that you will use for this and the upcoming workgroups. Today, you will conduct preliminary explorations on the variables you select. As we progress to the next workgroups, you will delve deeper into the same dataset by examining relationships among its variables.

Dataset Selection

For these workgroups we will use data from the European Statistical Office (Eurostat), because their data base contains a lot of interesting information, many variables. More importantly, these datasets are easily loaded into R. To make the assignments more engaging, you are free to choose the data you will use from Eurostat. Therefore, select topics of interest. However, keep in mind that your chosen data must include:

- plenty of variables so that there is enough to explore (e.g. at least 10 variables);
- both continuous and categorical variables, so you can practice making plots with both types of variables;
- variables that you think may be related to each other, as you will explore any relationships in the next workgroups;
- data from multiple countries, so you can, at a next workgroup, plot your data on maps of Europe;
- data from multiple dates, e.g. variables that are measured every year so you can explore trends at a next workgroup;

To download data from Eurostat, visit the Eurostat website (https://ec.europa.eu/eurostat/databrowser) where you can explore numerous datasets. Each dataset is associated with a unique identifier (`"<codename>"`) which you can use to retrieve the data using the `get_eurostat("<codename>")` function from the `eurostat` package in R. An example is provided at the end of this document.

Instructions and timeline for this workgroup assignment

As mentioned, the main aim of this workgroup is to select data that you will use throughout the interactive workgroups series, and do initial explorations to familiarize yourself with the data. The assignment of this workgroup consists of the following steps.

9.00 - 9.20h:

1. Together in class, we do the example below to get an idea of the aim of today's workgroup and to learn how to work with the Eurostat datasets.

9.20 - 11.20h:

2. For reproduciblity, work in a .Rmd file. Open a new .Rmd file and save that file in a separate folder on your computer. In that file, combine text (to explain what you are doing) with code chunks (in which you run your code). Save datasets, plots, and other results in the same folder for easy retrieval in subsequent workgroups. *Note: regularly knit your .Rmd file to check whether it can run. Make sure that your code can run on the computer of someone else, e.g. do not include any laptop-specific paths to files.*

3. Open the Eurostat data archive and look for topics and datasets that interest you. Choose datasets that satisfy the requirements listed above and load them into R with the `eurostat` package. You may also download multiple datasets if multiple datasets interest you and the variables across these datasets are potentially related (you can then explore that in the next workgroup).

4. Explore your chosen dataset by visualising the variables' distributions and proportions using the `ggplot` functions that you learned in the previous classes. You do not have to explore any relationships between variables yet, we will do this in the next workgroup. In some cases, you will have to make a selection of your data. For example, when a variable is measured over time for several countries, you may need to select either one year or one country to show a distribution that is sensible. In such cases, you can also make multiple plots, e.g. for several countries or for multiple years (see example below). Use the data explorations to check whether your data satisfies the requirements listed above.

11.20 - 12.00h:

5. Make a final selection of the data that you are interested in. Download those datasets to your computer to avoid re-downloading them for subsequent workgroups and to prevent conflicts in case of Eurostat updates.

6. Beautify the 3-5 most relevant data visualizations for these datasets.

12.00 - 12.40h:

7. We will invite some pairs to briefly present their chosen data by showing their data visualisations.

12.40 - 12.45h:

8. Submit your .Rmd file via Brightspace.