ADS2 Group Exercise ICA 2023-24

Understanding substance use

This data analysis exercise uses data about prevalence of disease and death related to the use of two types of substances: alcohol and opioids.

You will be asked to work with a real-life dataset about worldwide prevalence data. This dataset comes from the following resource:

Global Health Data Exchange (2021). Global Burden of Disease Results Tool.

Available from https://vizhub.healthdata.org/gbd-results/

You are asked to work in groups. It's up to you how you organise your group, but it is important that you work together as a group. This exercise has two parts. In part 1, you are asked to answer a few well-defined questions. Do make sure that the methods you use to answer those questions are well explained and documented, including all the work you do to process and clean the data. Use plots to visualise the data and support your argument if you need to.

In part 2, you are asked to formulate your own question and answer it based on the dataset.

What we look for

As graders, we will look for clarity of the explanations, completeness and correctness of the code, and thoughtfulness in the interpretation of results. A simple question, thoroughly answered, can get you more points than a very ambitious project that is poorly executed.

As a group, you are asked to submit an .Rmd file with code and with text explaining the code and interpreting the results. This file should fulfil the following:

- Markers should be able to knit it into a pdf file that shows code, code outputs and explanatory texts, and is well formatted.
- The knitted pdf should be a MAXIMUM of 15 pages, with all figures and text comfortably legible.
- The file should serve as a template for future reports. In particular, if a marker changes the data file to a more updated data file from the same resource, your code should run just the same without needing to be manually updated. (An updated data file will contain the same columns, but will have more rows, for instance because it includes additional years). Of course, you need to draw the conclusions you draw based on the dataset you have been given. But the code itself should run with updated versions of the dataset as well.

If you do need to use additional data files with your submission, then you may submit a zip file instead of an Rmd file. This zip file should contain only your Rmd file and the additional data files, in a format that is ready to your run with your R markdown code.

Note that one **and only one** person per group should submit the assignment. You should agree early on who you choose to submit the assignment on behalf of the group. The assignment file should show the group number, both on the document title and in the filename, but should not show the names of the group members, so that marking can be anonymous.

Data import and description of the dataset

The data you are asked to work with are provided in file substance_use.csv

It contains the following columns:

- **measure**: "Deaths" (deaths) or "Prevalence" (prevalence, i.e. how many people are affected at a given point in time))
- **location**: One of seven world regions, as defined by the World Bank.
- **sex**: In this dataset, this takes the value "Male" or "Female".
- **age**: Age groups in 5-year intervals between 25 and 69 (note that for the purpose of this exercise we are ignoring people aged younger than 25 or older than 69)
- **cause**: "Alcohol use disorders" or "Opioid use disorders"
- **metric**: Percent. (We are only looking at the percentage of the population affected, not at other measures such as absolute numbers)
- **year**: A year between (and including) 1990 and 2019.
- **val**: The best estimate for the value (i.e. the percentage of either Deaths or Prevalence of Alcohol use disorders or Opioid use disorders in a given year, a given world region, and a given sex)
- **upper**: Upper bound to the confidence interval around the value given above
- **lower**: Lower bound to the confidence interval around the value given above

You may have to clean or re-arrange the dataset for your purposes. Do whatever it is you need to do, but make sure you include the code for everything you do in this assignment. Code for importing data should assume that the data file is in the user's working directory. Please do not include file paths that may reveal the identity of one of your group members.

Part 1: Exploring the data

- In 2019, what region of the world has the highest rate of alcohol-related deaths among men aged 40-44?
- Looking at the prevalence of alcohol-related disease in the East Asia and Pacific region, how has this changed over time and in the different age groups? Is there a difference between men and women?
- In the United States, there is talk of an "Opioid epidemic". Part of the problem is that since the late 1990s, doctors have increasingly been prescribing pain killers which can be highly addictive. Looking at the data from the United States, can you confirm an increase in the prevalence of diseases related to opioid use? What age group is the most affected?

Part 2: Ask your own question

- Ask a question that can be answered using this dataset (and only this dataset) and choose a suitable method to answer it. Explain (briefly!) why you are interested in this particular question. Explain your method in words, and show and interpret your results.
- The "method" you use could be, for instance, an inferential hypothesis test, a model simulation, or maybe an especially clever way to visualise data - it is up to you. There is no need to look for methods outside of what you have learnt in your course so far, or during this project - see what you can do with the toolkit you have available!
- There is a lot you can do with this data set without additional information. But if your analysis requires the use of additional information (e.g. population sizes in each of the world regions), please provide this information in your report, along with a reference to where you found it.