Lab I - Statistics and Data Analysis II

November 4, 2023

To be handed in (Lisam) no later than November 11. The submission should include code, relevant output, as well as answers to questions. Preferably, all combined in a ".pdf" file. We recommend the use of RMarkdown to create the report.

---

1. In this lab, you will be working with two datasets: `hsbc_basic.csv` and `hsbc_health.txt` — both originating from a survey dataset, HSBC, containing information about the health of sample of Swedish students in 2014. Your first task is to import them to R. Note that columns are separated by *white space* in `hsbc_health.txt` and by *commas* in `hsbc_basic.csv` (Hint 1: you may use `read.table` and `read.csv` to import .txt and .csv files respectively) (Hint 2: ensure that `header=TRUE` to keep column names). Assign the imported data to R objects with the same names as the file names, i.e. `hsbc_basic` and `hsbc_health`.

2. After importing, your next task is to present the following *basic information* about the two datasets (Hint: `str()` is useful here):

   a. The number of rows and columns
   b. The number of `numeric`, `integer`, `character`, and `Factor` variables.

3. As you might already have noticed, both datasets contain the column `id4`. This is a variable that uniquely identifies the respondee of the survey. Your task is now to perform an *inner-join* to merge the two datasets, using `id4` as they key (Hint: use the `merge()` function). Call the resulting data.frame `hsbc`. Report the number of rows and columns of `hsbc`. Explain why it has the number of rows it has (Hint: think about the key characteristic of an inner-join).

4. Next up is data cleaning! Specifically, you are to investigate whether there are any rows of `hsbc` that contain *missing values*. If you find such instances, state which *column(s)* that are affected, and then filter out the *rows* with missing data.

5. Once you have ensured that `hsbc` does not contain any missing values, your next task is to produce a set of *variable-level summaries*. Specficially, report:

   a. The average life satisfaction (`lifesat`). (Hint: use the `mean()` function)
   b. The total number of observations in each age-category (`AGECAT`). (Hint: use the `table()` function). Which age-category have the most observations?

6. Building on *5b*, examine which age-category (`AGECAT`) that have the highest recorded *number* of bullied kids (`bully_dummy==1`). (Hint: you may again use the `table()` function).

7. Next, you are to perform a *counting exercise* that involves both *continuous* and *categorical* variables simultaneously. Use conditional subsetting to report the following (Hint 1: you can use `nrow()` to count the number of rows of a `data.frame` | Hint 2: use `&` to combine boolean tests):

   a. How many bullied kids (`bully_dummy==1`) there are with a `lifesat` score lower than 7.

b. How many girls (`sex==Girl`) there are in age-category 13 (`AGECAT==13`) that have a `lifesat` score greater than 8.

8. Create a new column in `hsbc` that is set to 1 if `health_index` is greater than or equal to 7, and set to 0 otherwise. Call the new column `health_index_binary`. (Hint: use `ifelse()`)

9. Compute the *conditional mean* of `lifesat` given the two different statuses of `health_index_binary` (0/1). For which out of the two do you find the highest average life satisfaction? (Hint: you may use `aggregate()` to compute the conditional means)

10. Next up is plotting! As preliminaries, first, load the `ggplot2` package. Second, format the variable `health_index_binary` as a `Factor` (Hint: using the `factor()` function). The latter step is performed to make `ggplot2` aware that `health_index_binary` is a discrete variable, and not a continuous one.

11. Construct a *density plot* of `lifesat` (Hint: use `geom_density()`). How would you characterize its distribution?

12. Extend the plot in *11* by colouring the distribution based on the membership to either of the `health_index_binary` categories (0/1). (Hint: both `color` and `fill` are arguments in `aes()` that can be used to color plots conditional on other variables). Interpret.

13. As a final task, export `hsbc` to your hard-drive (where exactly, you decide). You may export it either as `.txt` or `.csv`.