Homework 1

Foundations of Data Science FS2023 - TA team

Release date: February 28, 2023
Due date: March 9, 2023 at 2pm

1 Introduction

In this homework, you will go back in time and study data from Nobel prizes delivered between 1901 and 2020. After this homework, you will be able to:

1. to use the shape(), unique(), groupby(), min(), max(), isna(), dropna(), value_counts() functions
2. to use seaborn to visualise the data and save outputs;
3. to identify missing data, formulate hypotheses explaining their presence and retrieve the data missing when possible;
4. to identify outliers and aberrant values and propose ways of handling them.

Please submit your homework as a .zip file on Moodle. Your .zip file should include the data, code and output folders and a PDF document with the written answers. Make sure that your code runs in the terminal when using the command:

```
python file_name.py
```

2 General question

- How many columns and rows are there in the data?
- How many unique recipients of a Nobel prize are listed? Does that match the number of rows? Explain.
- How many categories of prizes are part of the data? Does that match your expectations? Which categories are listed? Explain.
- Give the name of the person who was the youngest when they received their Nobel prize. Same question for the oldest recipient.
- Look at the column 'share'. It states with how many other recipients the prize was shared. Do you notice something about those values? How are such values called? What would be you strategy to handle them?

3 Visulisation

- Create a barplot with the number of prizes per category. Save your figure as "counts_category.png" in the output folder. Give the name of the two categories with the least number of entries. Give a probable explanation for each of the two categories.
- Plot the age distribution. Don't forget to add units, axis titles and a plot title. Save your figure as "age_distribution.png" in the output folder.
- Plot the age distribution by sex. Save your figure as "age_distribution_sex.png" in the output folder. Comment on the difference in counts between male and female.

4 Missing data

In this section, you will focus on the columns 'died' and 'diedCountry' and 'diedCountryCode'.

- How many missing values are present in the column 'died'? What could be the reason?
- Create a new dataframe with only the rows with a value in column 'died'. In this new dataframe, how many values are missing in the 'diedCountryCode' column? Can you identify a pattern? Propose a way of dealing with those missing values.