PROJECT GUIDELINES

Statistics for Data Analysis

Master on Big Data Analytics

GENERAL COMMENTS:

- The project should be based on a real data base collected by the students based on a real survey.
- The data base is supposed to be a random sample from an infinite population (do not consider finite populations and avoid temporal or spatial data).
- The project should be motivated by a news item from a newspaper or social networks and a link to it shoud be included. One or two hypothesis should be formulated at the beginning of the project.
- It would be recomended to consider one (or two) main continuous variable(s) and various categorical variables that may be related with the main variables.
- Projects should be done in groups of at most four students.
- For the final evaluation, you should upload a pdf/html document with your final project, together with the data file, code and a four minute video presentation. Project length is limited to 15 pages. Although it will be possible to include additionally some supplementary material.

STRUCTURE OF THE PROJECT:

1. Introduction and motivation (one page):

   Briefly explain the motivation of the project including a link to the news item related to it.

   Describe the one or two main hypothesis of the project and survey implementation details. Define clearly each one of the selected variables to be analysed.

2. Descriptive analysis (three pages):

   Using graphs, tables and summary statistics, describe the main features of the main variables. Consider if it would make sense to transform them.

3. Model fitting (one or two pages):

   Use the method of moments and maximum likelihood estimation to fit a univariate distribution model for each main (possibly transformed) variable. Select the best model using the AIC criteria.

4. Statistical inference of one variable (one or two pages):

   Obtain and interpret confidence intervals for the population mean of the variables of interest. Perfom hypothesis testing for your main motivating hypothesis. Illustrate your results with descriptive plots.

5. Statistical inference of two variables (two or three pages):

   Perform interesting hypothesis tests based on two variables. These can be based on the main motivating hypothesis or new ones depending on your data. For example, you may want to test if two categorical variables are related to one another. Or you may want to compare if there are statistical differences in the mean of two (or more) populations. In case you have two paired continuous variables, you may want to test if there exist significant linear correlation between them. Illustrate your results with some descriptive plot.

6. Conclusions (half page):

   Summarize your main results in a conclusion section. Comment on future extensions.

7. References.

8. Appendix.

   Include here some supplementary material, if desired.

EXAMPLES:

- Motivation: `Decrease in lightweight plastic carrier bags in 2020`

  `https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20221116-1`

  Hypothesis 1: Europeans use on average MORE THAN 87 lightweight plastic carrier bags per year

  Hypothesis 2: Spanish people use on average MORE lightweight plastic carrier bags that other Europeans

  - Age
  - Gender
  - Level of studies
  - Country
  - How many lightweight plastic carrier bags do you use per week

- Motivation: When do young Europeans leave their parental home?

  `https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20230904-1`

  Hypothesis 1: Young people across the EU leave their parental home on average at a LARGER age than 26.4 years.

  Hypothesis 2: Spanish people leave their parental home on average at LARGER age than other young Europeans.

  - Age
  - Gender
  - Level of studies
  - Country
  - Have you left your parental home?
  - If yes, at what age did you left your parental home?
  - If not, at what age are you planning to leave your parental home?

- Motivation: Average salary in Spain is 500 euros lower than in the EU as a whole

  `https://www.euronews.com/next/2023/07/22/average-working-hours-in-europe-which-countries-work-t`

  Hypothesis 1: Average hours worked per week in Europe is LESS THAN 36.2 hours

  Hypothesis 2: Spanish people work on average MORE hours than the other european countries

  - Age
  - Gender
  - Level of studies
  - Country
  - Average hours worked per week

- Motivation: Europe is home to the world's heaviest drinkers. Which country drinks the most alcohol?
  https://www.euronews.com/next/2023/06/30/so-long-dry-january-which-country-drinks-the-most-alco

  Hypothesis 1: Europeans drink LESS THAN 190 litres of beer per year

  Hypothesis 2: Europeans drink LESS THAN 80 litres of wine per year

  Hypothesis 3: Europeans drink LESS THAN 24 litres of spirits per year

  - Age
  - Gender
  - Level of studies
  - Country
  - How much beer do you drink weekly (liters)?
  - How much wine do you drink weekly (liters)?
  - How much liquor do you drink weekly (liters)?