Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Experimental Methods in Computer Science
(and in Informatics Engineering)

DEI-FCTUC, 2023/2024

Assignment 1 – Exploratory data analysis

The exploratory data analysis (EDA) is the process of systematically applying statistical and/or logical techniques to data sets in order to extract/summarize their main characteristics and provide a condensed view of the information contained in the data. This is often done with visual methods that may include all sorts of graphic types (i.e., statistical graphics), as well as classic charts, tables, and statistical analysis (i.e., more traditional data analysis). The data can include either field data sets collected from real setups or data gathered from designed experiments.

The main goal of exploratory data analysis is to see what the data can tell us about the system ("system" understood in general terms) or the specific aspect of the system that is portrayed by the data. The exploratory data analysis may complement the more formal modelling or hypothesis testing approaches (not addressed in this assignment) or be the main analysis technique.

1. Assignment goals

The goal of the assignment is **to select a dataset (or a few datasets) and plan and perform exploratory data analysis to extract as much information and conclusions as possible from the dataset under analysis**. The students should also discuss the impact of the observations and conclusions extracted from the dataset, considering the context of the systems/software that have originated each the data set.

Concerning the selection of the dataset (or datasets) used in the assignment, the students have two options:

a) Propose a dataset at your choice.

Since there are many datasets freely available on the Internet, students can select and propose a dataset to be used in this assignment. This option must fulfil two requirements:

1. The dataset must contain data about computer systems or software (e.g., performance data, failures observed, bugs found, etc.). The reason why we restrict the choice of the dataset to data related to computer systems or software is to keep the results of the exploratory data analysis within the context of the master's in informatics engineering (MEI).

2. The selected dataset must be approved by the teacher. This approval will be done in the PL classes.

b) Use one (or a few) datasets from the following repository.

Use one (or a few) curated datasets available in the repository maintained by the Experimental Systems Lab from The Hebrew University in Israel. This repository contains data from workload logs collected from real (production) large scale parallel systems installed in various places around the world. The repository is available in the following link: http://www.cs.huji.ac.il/labs/parallel/workload/logs.html, which also contains detailed descriptions of the data, the different formats available, descriptions of the parallel systems, and the logged data itself in several formats.

As the repository has many log files, collected from many parallel systems over a considerable time span (from 1993 to 2018), each student group must select just a few files and perform the exploratory data analysis over the date from such files. Keep in mind that each file represents a different parallel system installation and, although it is possible to compare data collected from different systems, this cross analysis among different systems is in general very limited. In other words, what is expected is a set of nearly independent exercises of exploratory data analysis, each one focusing on a specific log file/parallel system.

The data sets available in the link mentioned above from The Hebrew University include both raw data (or raw date in the Standard Workload Format - SWF) and cleaned data, at least for some of the systems. In fact, a typical first step before analysing data is to perform some cleansing and basic data treatment, if needed. Real data is often imperfect in the sense that it contains noise, some inconsistencies or is incomplete. Manual inspection of the data (or a sample) is a good practice to identify basic cleansing that may be required. It is important to take this into account if you select systems whose data has not been cleansed yet.

It is also worth saying that exploratory data analysis presumes that the analyst knows relatively well the object under analysis. In the case of Assignment 1, this means that you will have to learn the basics about parallel systems, workload scheduling and workload models. This is a fun part of the assignment. ☺

2. Outcome

The outcome of the assignment is a written report (PDF file). This report should be self-contained (to a reasonable extend), which means that it should include the presentation of the data sets, a brief description of the systems/context where the data were collected from, provide the necessary background on terms and concepts needed for the data analysis, present the results, and discuss the key evidence and information that have been extracted from the data sets.

There is no suggested template for the report structure. The objective is to proactively engage students, preventing them from simply filling in the report structure. On the contrary, defining the best structure to present the exploratory data analysis is part of the goals of the assignment. The teacher is fully available to discuss the structure with the

students, as well as any other aspect of the report. For example, the amount of "questions to the data" to be covered, the type of graphics and tables used to present the data, the adequacy of the statistical analysis, etc. Some of the "PL classes" are specifically devoted to providing such support to students.

The heavy use of charts (e.g., X-Y graphs, bar charts, scatter plot, histograms, pie charts etc.) and tables in data analysis imply the use of some basic rules for writing a good report. For example, all charts (figures) and tables must have an adequate caption and should be mentioned (presented) in the text. Remember that figures and tables are a powerful way to convey information but very rarely a figure or a table is fully self-described. Furthermore, the key conclusions must be provided in the body (text) of the report and summarized in the typical "Conclusion" section.

Although not included in the assignment outcome, the environment used to analyse the data (i.e., tools, tables, charts, etc.) should also be kept by the students, as this material will be used in the oral defence of the assignment.

3. Resources

Students should select the EDA tool they want to use in the assignment. There are many tools available on the Internet that can be easily find using key words such as *tool exploratory data analysis*. For example (just a few examples, although some of these tools are proprietary):

- 18 Free Exploratory Data Analysis Tools For People who don't code so well
  https://www.analyticsvidhya.com/blog/2016/09/18-free-exploratory-data-analysis-tools-for-people-who-dont-code-so-well/

- 9 Quick & Simple Exploratory Data Analysis Tools (2023)
  https://www.polymersearch.com/blog/exploratory-data-analysis-tools

- 11 Open Source Data Exploration Tools You Need to Know in 2023
  https://opendatascience.com/11-open-source-data-exploration-tools-you-need-to-know-in-2023/

An example of a very simple tool is Mondrian, which is an interactive general-purpose statistical data-visualization tool (https://en.wikipedia.org/wiki/Mondrian_(software).

Obviously, classic and widespread tools such as R (http://www.r-project.org/), or even Microsoft Excel™, can also be used. In case of R, it is necessary to learn the R language and the R environment for statistical computing and graphics, which requires some effort. But R is a very powerful tool and is, for sure, a worthwhile investment.

4. Calendar and miscellaneous

The assignments must be done by groups of up to 3 students according to the following calendar:

- **Wednesday, October 4** – The students should submit (Inforestudante) a short PDF document containing the following:
  - Identification of the members of the group (names and emails).

    - The dataset (or datasets) that the group is going to use. Include the link to the dataset and a short paragraph describing the data contained in the dataset.
    - The EDA tool (or tools) selected by the students.
- **PL classes until November 3** – Follow up and support for the work developed by each group. Note that Assignment 2 will be launched in the PL classes on October 19 and 20, which means that the work for both assignments will be overlapped for two weeks, from October 20 to November 3.
- **Friday, November 3** – Submission of the assignment report (PDF file) at Inforestudante.
- **November 19 and 20** – Oral discussion/defence. These defences will be held during the PLs classes and outside the classes' times in these days, as the time available in the PLs is not enough. It is mandatory that all the members of each group attend the assignment defence.

Enjoy the work! ☺