Instituto Superior de Engenharia de Lisboa
Mestrado em Engenharia Informática e de Computadores
Mestrado em Engenharia Informática e Multimédia
Mestrado em Matemática Aplicada para a Indústria

Big data mining (MDLE)

Laboratory Class #4 — Data manipulation and Pipelines

2nd semester, 2022/2023 (May, 16)

Code and Report are due by —

1. **Data Resources and Software Tools.**
   For this laboratory class, you will need the following software:
   - Access to RServer (`http://datalys.dyn.fil.e.ipl.pt:8787`) or,
   - R, https://cran.r-project.org/ and
   - R Studio, https://www.rstudio.com/;
   - SPARK ≥ 3.3;
   - Read documentation at https://www.tidyverse.org/.

2. `sparklyr` can be used along with the `dplyr` package to perform data transformations. Considering the following mapping between operations:
   - `select` $\sim$ `SELECT`
   - `filter` $\sim$ `WHERE`
   - `arrange` $\sim$ `ORDER`
   - `summarise` $\sim$ `aggregation operators`
   - `mutate` $\sim$ `calculations`

   Check the `dplyr` documentation to see the possibilities of data manipulation.

3. Check the codes `C1.R` to `C7.R`. Analyse each line of code, and their outcome.

4. Take particular attention to the pipelines and how they can be use in a learning environment.

5. Apply these methodologies to your practical assignment.