Analyzing Pollution Data in Madrid

- The work consists in analyzing a big dataset with hourly pollution data from Madrid, in the period 2011 to 2016.
- In the content area “workgroup_assignment” at the campus there’s 72 csv files, each one containing a piece of the information. Specifically, each csv contains the raw data for a concrete month of a concrete year (6 years, 12 months per year).
  - ↘ Raw data is at an hourly level.
  - ↘ There are many stations that measure air pollutants in Madrid (around 24).
  - ↘ There are several pollutants measured (around 10).
  - ↘ Hourly data is available for every pollutant, for every station, for every hour, for every day of each month-year.
- You will also find two additional datasets:
  - ↘ Weather.xlsx, containing daily time series for min, average, and max temperature, precipitations, humidity and wind in Madrid.
  - ↘ parameters.png (image), containing the key for every pollutant code (eg. “08” ➔ NO_2)
- The scope of the statistical analysis is open, so each group decides the deepness of it. <u>At least</u>, you should:
  - ↘ Read and process all the information in the most automated manner.
  - ↘ Create a dataset with daily information on NO_2, SO_2, O_3, $PM_{2.5}$, and all the weather variables.
  - ↘ Analyze the relations among this variables, creating a descriptive analysis and a revealing visualization, showing also the evolution of each series over time.
  - ↘ Create a multiple linear regression explaining NO_2 with the rest of the variables.
  - ↘ Create a report (doc/pdf/Rmarkdown) with code, results and conclusion for each step of the project.
- “At least” means you are welcome to explore the possibilities of creating an additional analysis at a weekly or monthly level, work with Min, Mean and Max meassures on the pollutants, create a set of possible models, interactive charts etc…

Analyzing Pollution Data in Madrid

- So there are four parts in which this project can be divided in (besides from the report):
  1. Reading every piece of raw data and creating the whole initial raw_data set.
  2. Processing raw_data to create a daily dataset, by averaging each hourly measure, and containing also the weather variables and the names for each pollutant parameter.
  3. Generating a descriptive analysis with correlation matrices, scatterplots, time series charts …
  4. Creating a linear regression model that explains NO_2.

- I may randomly select some of you individually to explain how any of these parts of the code work, and what's the purpose of it.

- The final report may be done in Rmarkdown; this way, you can have code, comments and conclusions on just a single document. If you go with a standard .doc / .ppt / or .pdf, I will ask you to sumbit all .R scripts too.