

MASTER IN  
BUSINESS INTELLIGENCE  
& BIG DATA ANALYTICS

# Homework 1

PELUCCHI – VACCARINO- NOBANI



# Homework

Imagine a scenario for analysis in the Big Data field.

The following datasets are allowed:

- Constructed starting from a scraping phase or from social data (Twitter, etc.)
- Retrieved from Open Data portals (e.g., Regione Lombardia, Comune di Milano, etc.)
- Taken from a collection of public (Kaggle, KDnuggets, etc.) or private datasets (from your company)

Be mindful of cardinality: **IF TOO HIGH, REDUCE THE SIZE!!!**

Set the objective of your analysis and the questions you want to answer.

- Prepare, using the methods and tools covered in the lessons:
- The scraping phase (if necessary)
- Data profiling (what kind of problems are there in the dataset?) and the treatment methods to be adopted
- The ETL phase (**mandatory**) to ensure data quality and create: source tables, staging area, and data warehouse.

# Details

Prepare a presentation (maximum 20 slides) that describes:


- **Scenario, Technologies, and Adopted BI Architecture:**
  - - Detail the context and purpose of your Big Data analysis.
  - - Outline the technologies and tools used (e.g., Hadoop, Spark, BI tools).
  - - Describe the architecture, including data sources, ETL processes, data storage, and BI tools.
- **Objective and Questions:**
  - - Clearly state the objective of your analysis.
  - - List the specific questions you aimed to answer through the analysis.
- **Dataset Characteristics:**
  - - Provide an overview of the dataset, including its structure and source.
  - - Highlight any issues or problems identified in the dataset (e.g., missing values, inconsistencies).
- **Treatment Methods Adopted:**
  - - Describe the data cleaning and preprocessing steps taken to address the identified problems.
  - - Include any transformations, normalizations, or other processing steps performed during the ETL phase.
- **Emerging Critical Issues:**
  - - Discuss any challenges or critical issues encountered during the project.
  - - Explain how these issues were addressed or mitigated, if applicable.

*Ensure that your presentation is clear, concise, and well-organized, with visual aids such as charts, diagrams, and tables where appropriate to enhance understanding.*

# What can't be missed in your projects

- Goal
- The stakeholders
- Usefulness!
- The objectives achieved and the objectives not achieved
- Where you went wrong and how you can improve it

# Submission

- The presentation (PDF format) must be uploaded to the e-learning platform
- It is also possible to upload the work and part of the source code to GitHub 
- During the next lessons following the delivery you will present the most interesting works (1 or 2 presentations)

*NOTE: from this delivery it will be possible to start the second homework on data visualization and spark*