

## IDC 5204, Project: Data/Article description

In alignment with the research questions, expected data and articles from the previous assignment, proceed to:

1. Actually find 1-2 academic papers that could be of relevance, quickly summarizing their main idea (abstract pretty much).
2. Description of the data you've selected for your project, including
  - i. Why it would help answering your research question. If you don't have any data preferences off the top of your head, some examples of data sources are *Kaggle.com*, UCI Machine Learning repository, *sports-reference.com*, among many others.
  - ii. The size of the data (# of observations and # of variables).
  - iii. Description of all variables (similarly to descriptions you encounter for *R* data sets, e.g. <https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Boston.html>)

It's **preferred**, albeit not required, that your data set contains:

- A total of at least 10 variables and 50 observations.
- Several numerical variables, one of which may act as a response variable in linear regression (your main variable of interest).
- At least one categorical variable.

**Note:** It shouldn't be among data sets that we've already used in class, or that are contained in any of the *R* packages. It should be obtained from external source (most likely a *.csv* file, or via web scraping). **ALSO, PLEASE DON'T USE THE PORTUGUESE HIGH SCHOOL DATA or FRAMINGHAM HEART STUDY DATA** - these have been overused for quite a while.

The report should be

- **no more (!) than 3 pages long.**