# **Project 3 | Web Scraping**

# **Introduction**

The goal of this project is to apply your web scraping skills that you have learned so far to real-world data both to practice and to add to your portfolio.

## **Tasks**

For this project, you will choose either an API/JSON to obtain data from or a web page to scrape. An additional open source dataset to complete the data analysis has to be used.

## **Technical Requirements**

The technical requirements for this project are as follows:

* You must propose a scope in which you’d use such analysis.
* You must find an open dataset to start your analysis on
* You must obtain data from an API/Web Page using Python.
* You must scrape, clean, and analyze the web page/site using Python.
* You must prepare a presentation (5 minutes). This presentation can include GUI and slides.
* Your code should be saved in a Github Repository
* You should include a README.md file that describes the steps you took and your thought process for obtaining data from the API and web page.

## **Deliverables**

* Code
* Presentation

## **Deadline**

**Approval of the project** by 5th of April 2023

**Presentation** Friday 21st of April 2023 around 2pm.

## **Timeline**

By wednesday: have a proposal with **data checked**

By half second week: EDA and GUI on open dataset done

By end of 2nd week: web scraping + DA + descriptive statistics + idea of delivery

By end of 3rd week: presentation and GUI

## **Teams**

* Thierno, Antonio, Joao C
* Gabriel, Joao M, Louis
* Bengu, Rui, Mauriçio

## **Resources**

### For dataset:

Data NOT TO USE (Iris +) <https://analyticsindiamag.com/10-most-popular-datasets-on-kaggle/>

Some sources for data: <https://geekflare.com/open-datasets-for-data-science/>

Some more ideas: <https://www.springboard.com/blog/data-science/15-fun-datasets-to-analyze/>

You can also Googled “open data [CITY\_NAME]”

<https://databank.worldbank.org/home.aspx>

<https://data.world/nrippner/titanic-disaster-dataset>