1. Objective

The aim of this assignment is to apply statistical inference techniques to a **data set of your choice**. On this data set, you will have to propose research questions similar to the examples and exercises done in class so you will answer them by applying the methodology in statistical inference developed in this course.

All analysis must be completed using Python, and your write up must be an Jupyter Notebook document (.ipynb extension).

2. Sections and grades

The objective of this project is not to apply as many statistical techniques as possible, but to do so with rigor and judgement. The project will contain (as a minimum) the following sections:

a. **Data** (10 points). Look for a data set on which to launch research questions that will allow you to propose different methods of statistical inference. You have to clearly indicate where you got the information from, and provide links or references to consult that dataset. There are a number of public repositories where you can search. Some recommendations:

   - *Google Dataset Search*
   - *UCI Machine Learning Repository*
   - *OpenIntro Data sets*
   - *Statistics: Unlocking the Power of Data*

b. **Research questions** (10 points). Come up with a research question that you want to answer using these data. You should phrase your research questions in a way that matches up with the scope of inference your dataset allows for. You are welcomed to create new variables based on existing ones. Along with your research question include a brief discussion (1-2 sentences) as to why this question is of interest to you and/or your audience.

c. **Exploratory Data Analysis** (EDA) (20 points). Perform EDA that addresses the research questions you outlined above. Your EDA should contain numerical summaries and visualizations.

d. **Inference** (50 points). Perform inference that addresses the research question you outlined above. Each Python output and plot should be accompanied by a brief interpretation.

e. **Member contribution** (mandatory). You must clearly indicate the contribution of each member of the group.

In addition to these parts, there are also 10 points allocated to format, overall organization, and readability of your project. Total points add up to 100 points

3. Submission

Your submission should be a single Jupyter Notebook document (.ipynb extension) that contains all your code, cell output and plots, and your narrative. Your submission has to be sent to *felipe.alonso@urjc.es* before the 22nd May 2020.

4. Review criteria

You are not asked to produce an extensive assignment, but just the opposite. It is expected a concise, clear, and a well presented project, where the approaches made are justified. The following points list some of the criteria that will be taken into account when evaluating the work:

- Are there well-defined and clearly stated research questions?

- Are the hypotheses stated clearly and do they match the research question?
- Did the authors provide background on the research question as to why they care and why others should also care?
- Are the appropriate method(s) the authors will be using stated?
- Was the correct code used and output provided for all required techniques?
- Are correct interpretations and conclusions for all output provided?
- Is whether or not results from hypothesis test and confidence interval agree stated?
- Is a discussion of what was learned about the research question provided?
- Are ideas for possible future research and/or discussion of additional synthesis or possible shortcomings of study provided?