# STAT0032: INTRODUCTION TO STATISTICAL DATA SCIENCE - GROUP PROJECT 2020-21

## Outline of the Project

**The Problem** Your group is a team from a leading data science consultancy company, and you have been hired to help a small business owner improve their business. The business in question is a small wine shop specialising in red wines from all over the world. The business owner is a wine expert with several decades of experience in sourcing wines from a variety of vineyards. She also understands very well what her clients tend to like, but most of this knowledge is anecdotal and has been acquired through discussions with clients. However, the market conditions are challenging, and the shop is experiencing intense competition from supermarket chains, which are often able to buy at larger scales and hence source wines at lower prices.

The business owner has hired your company as she would like to use data science in order to better understand the wines that are likely to sell well. This would then allow her to better compete with the large supermarket chains. During your initial meetings with the business owner, one point that came up regularly is the acidity of the wines. The business owner has received conflicting feedback on what clients like in terms of acidity level. She would therefore like to understand how acidity impacts the client's appreciation of a wine in more detail.

**Data** In order to help answer this question, you have been given access to the "Wine Quality Data Set" from the UCI Machine Learning Repository. See `https://archive.ics.uci.edu/ml/datasets/wine+quality` for a description of the dataset as well as details of how to download it. This dataset was made public following the publication of this paper: Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553, which may also contain relevant background information. The variable *quality* indicates how good a wine is. The wine shop owner has indicated that she would consider any wines with quality value less or equal to 4 to be of "low quality", whereas any wine with value greater or equal to 7 would be considered of "high quality".

**Objective** Your task is to study the distribution of acidity levels and how these may impact sales for the wine shop owner. The outcome of this project should be a report (up to four A4 pages) describing the analysis performed including any statistical tool used, as well as recommendations for the wine shop owner. The report should clearly state any limitations of the analysis. In terms of statistical analysis, the report should include at least the following:

- A study of the distribution of pH values of wines, including for the whole dataset and for the "low quality" and "high quality" wines. One important question to answer here is whether the distributions of pH values follows a normal distribution. To answer this question, you should look into hypothesis tests that fall in the category of "goodness-of-fit tests". You should use at least two such tests, which should be described in detail and compared (including a discussion of how any underlying assumptions differ).
- A study of how the distribution of pH values of wines differs for the "low quality" and "high quality" wines. One important question to answer here is whether the distributions of pH values are the same or whether they differ. To answer this question, you should look into hypothesis tests that fall in the category of "two-sample tests". You should use at least two such tests, which should be described in detail and compared (including a discussion of how any underlying assumptions differ).
- Any additional statistical analysis on this dataset which may help the shop owner to get further insights into wine preferences, and as a result become more profitable. This may be based on any statistical tool (whether discussed in the module or not) as long as these tools are clearly explained, and clearly justified for the task at hand. Preference should be given to the relevance of the tool for the problem at hand, rather than using an advanced method whose use is not properly justified.

The report should be written at a level appropriate for anyone with a basic understanding of statistical data science (for example, the level of a STAT0032 student by the end of term 1), but not any specific knowledge of the methods that you decide to use. For example, the report can assume basic knowledge of the general framework of hypothesis testing, but not of the specific tests being used.

# Administrative details

## Basic details

- This assessment counts for 20% of your final mark for STAT0032.
- Groups will consist of 5-6 students and will be assigned at the start of term. Groups will be expected to meet at least once a week for an hour over term 1.

- The teaching assistants for STAT0032 will be available to answer any questions on the group projects during these meetings. They will however not comment on any draft reports. Note that it may not be appropriate to answer all your questions, but they will do their best to be as helpful as possible in a manner which is fair to all groups.

## Additional Page

In addition to the report, all groups must submit an additional page where each group member briefly describes their contribution to the project.

- You will need to agree this in your groups *before* submitting the report.
- Note that I do not plan to mark this page, nor allocate different marks to different group members based on this. The purpose is to encourage you all to be mindful about contributing to this piece of groupwork. In very exceptional circumstances, if a student has not sufficiently participated, the marks of everyone in the group will be adjusted accordingly.
- If you feel that one or more of your peers is not contributing fairly, please contact me by email in the first instance BEFORE SUBMISSION of the report and as early as possible.

You should insert student ID numbers of all students in your group on the report, but **do not write your names**. Your report will be marked anonymously.

This page should also include a sentence stating that you are fully aware of the content of the "Plagiarism and Collusion" section in the Taught Postgraduate Student Handbook for the Department of Statistical Science (You may find the handbook online here: `https://www.ucl.ac.uk/drupal/site_statistics/sites/statistics/files/migrated-files/pghb.pdf`). In particular, the responsibility for any academic misconduct will be shared by the entire group.

## Submitting your work

The report and additional page should be submitted as a single pdf document. Details of how to submit will be announced a few days prior to the submission date (more details to follow on Moodle).

## How will the report be marked?

Your report will be marked out of 50, with allocation as follows:

- *15 marks for the presentation of the report.* This includes the structure of the report, how easy it is to read and understand, good use of plots/tables, adequately sized graphics with suitably informative captions and labelling, and so on. Please do not make the font or margin too small or you will be penalised.
- *25 marks for the mandatory statistical analysis, including the goodness-of-fit and two-sample tests.* This includes a detailed and relevant description of the research problem and dataset, a clear description of the methods used, whether you have selected appropriate information and supporting evidence to present, and whether your results are accurate.
- *10 marks for any additional statistical analysis.* This includes a detailed and relevant description of the methods used, whether you have selected appropriate information and supporting evidence to present, and whether your results are accurate.

Dr. François-Xavier Briol