STAT0032: INTRODUCTION TO STATISTICAL DATA SCIENCE - GROUP PROJECT 2021-22

Outline of the Project

The Problem Climate change is a topic of increasing geopolitical importance for most of the world. In Europe, one of the main challenges is the increased likelihood of extreme weather events such as heatwaves, storms, strong winds or even lightning. These extreme weather events can have devastating consequences for the economy, as well as the well-being of citizens. European governments are therefore keen to understand how climate change may impact their individual countries, and what they can do to mitigate some of this impact.

Your group is a team from a leading data science consultancy firm, and you have been hired to help the government of Portugal prepare for the impact of climate change in the north of the country. There, one of the anticipated risks is an increased likelihood of forest fires. The government is particularly concerned due to the large number of wildfires which took place in Greece and Turkey over the summer 2021. Your company has been hired on a multi-year project and tasked with understanding the current risks associated with large forest fires. In particular, the government is interested in gaining a better understanding of the current scale of those fires throughout a typical year.

Data In order to help answer this question, you have been given access to the "Forest Fires Data Set" from the UCI Machine Learning Repository. See `https://archive.ics.uci.edu/ml/datasets/Forest+Fires` for a description of the dataset as well as details of how to download it. This dataset was made public following the publication of the paper:

P. Cortez and A. Morais. A Data Mining Approach to Predict Forest Fires using Meteorological Data. Proceedings of the 13th EPIA 2007 - Portuguese Conference on Artificial Intelligence, December, Guimarães, Portugal, pp. 512-523, 2007.

which may also contain relevant background information. This dataset contains data on 517 forest fires, and the variable *"area"* indicates the size in hectares of a given fire. Since you are

only interested in large fires, you should discard any fire with area smaller or equal to 0.01 hectares.

Objective Your task is to study the distribution of the area of large forest fires in the north of Portugal, and how this distribution may differ over the summer months. The outcome of this project should be a report describing the analysis performed including any statistical tools used, as well as recommendations for the local government authority. **The report should clearly state all your statistical assumptions, as well as any limitations of the analysis for decision making**. In terms of statistical analysis, the report should include:

- A study of the distribution of the size (as given by the "area" variable) of forest fires for the months of August and September, which tend to have the largest fires. One important question to answer here is whether the distributions of forest fires area follows a log-normal distribution, and you should test this separately for each month. 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 underlying assumptions differ!*).
- A study of how the distribution of the size (as given by the "area" variable) of forest fires differs for the months of August and September. One important question to answer here is whether the distributions for these two months 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!*).

Your report should start with a cover page which contains the title and the student ID of all members of the group. The main body of your report should be maximum four A4 pages long, with a minimum font size of 11. You do not need to share your code, and no appendices should be used (unless they fit within the four page limit). The only exception on the page limit is for references, which can be supplied on a separate page, and the cover page which also does not count towards this limit. Finally, you will also be asked to attach an additional page describing the contribution of each group member; this page does not count towards the four page limit and more details will be given below.

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. You will therefore want to carefully describe your hypotheses, test statistic (and its distribution), and any other detail essential to understanding the tests.

Administrative details

Basic details

- This assessment counts for 20% of your final mark for STAT0032.
- Groups will consist of 4-6 students and will be assigned (at random) at the start of term. The final mark will not be adjusted according to the number of students in a group.
- Groups will be expected to meet at least once a week for an hour over term 1. You are free to meet in person or online. All group members are expected to attend these weekly group meetings, and it is your responsibility to schedule these at a time and in a way which is appropriate for everybody. Please be mindful that some students may prefer to meet online than in person; if that is the case, students should not be pressured into meeting face-to-face.
- The teaching assistants for STAT0032 will be available to answer any questions on the group projects during office hours. 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 to be mindful about contributing fairly to this piece of groupwork. In exceptional circumstances, if a student has not sufficiently participated, I reserve the right to adjust marks accordingly.
- If you feel that one or more of your peers is not contributing fairly, please raise this with them directly and make sure they are fully aware of the problem. Very often this is due to a difference in expectations and can be resolved within the group. If this does not resolve the problem, please contact me by email BEFORE SUBMISSION of the report and as early as possible. Students are expected to participate in their project throughout the entire term (although it sometimes takes a week or two at the start of term for group allocations to be settled).

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.

The additional 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 (and it is therefore your job to very the work of your peers before submission).

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 prior to the submission date (more details to follow on Moodle), which will be during the last week of term 1.

How will the report be marked?

Your report will be marked according to three main criteria, each associated to a different weighting:

- *40% of the marks will go 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. Also, when using mathematical notation, please ensure this notation is defined clearly.
- *60% of the marks will go for the 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.

Dr. François-Xavier Briol