STAT0032: INTRODUCTION TO STATISTICAL DATA SCIENCE - GROUP PROJECT 2022-23

Outline of the Project

The Problem The energy and cost of living crisis in the UK is widely predicted to last several years. The issue, coupled with growing concerns over climate change, is leading to a significant push away from polutting, fossil-fuel based, modes of transports. In large cities, this is leading to a renewed interest in alternative modes of transport such as cycling. This move towards cycling is supported in part by the use of city-wide bicycle sharing schemes.

Your group is a team from a leading data science consultancy firm, and you have been hired by Transport for London to provide them with insights which can help them manage their bike sharing scheme. In particular, one of your goals is to understand how much demand to expect at different times of the day, and during different times of year.

Data In order to tackle the task at hand, you have been given access to the "Bike Sharing Data Set" from the UCI Machine Learning Repository, which consist on data about cycling hires through the Capital Bikeshare scheme in Washington D.C. See `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset` for a description of the dataset as well as details of how to download it. This dataset contains data on bike hires and the variable *"cnt"* indicates the total number of bikes which are rented at a given time.

Objective Your task is to study the distribution of cycle hire counts to gain a better understanding of how various external factors might impact the popularity of cycling in London. The outcome of this project should be a report describing the analysis performed, including any statistical tools used as well as (non-technical) recommendations for Transport for London. Transport for London is particularly keen to gain a greater understanding of how much demand to expect during peak commuting hours in the evening. **The report should clearly state how you preprocess your data, describe all your statistical assumptions, and discuss any limitations of the analysis for decision making**.

In terms of statistical analysis, the report should include:

- A study of the distribution of bikes hires during peak commuting times in the evening in spring and summer. One important question to answer here is whether the distributions of bike hires follows a normal distribution, and you should test this separately for each season (i.e. do a first test for spring, and a second test for summer). 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 bike usage differs during spring and summer. One important question to answer here is whether the distributions for these two seasons 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 be maximum six A4 pages long, with a minimum font size of 11. The first page should be a cover page which contains the title and the student ID of all members of the group. This should be followed by a maximum of four A4 pages with the main content of the report. Finally, the last (i.e. sixth) page should include references, a statement describing the contribution of each group member, and a plagiarism statement; more details will be given below. You do not need to share your code, and no appendices should be used.

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.

Final Page

All groups must submit a page where each group member briefly describes their contribution to the project (the sixth page of the report).

- 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 sixth 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 should be submitted as a single pdf document. Details of how to submit will be announced a few weeks 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