FDS Coursework 3, 2023/24

Data Science Project

Released:	Monday 26 February 2024
Submission deadline:	Tuesday 2 April 2024 at 12:00 UK time
Late submission rules:	Rule 1: extensions (3 days) and ETA (7 days). Late penalties apply. See Late coursework and extension requests for full details of rules and late penalties.
Group or individual work:	If working as a group, please see the details on Learn of how Extensions and ETAs apply to this project.

This is a **marked** assignment which will count towards **40%** of your final grade for **Inf2-FDS**.

Good scholarly conduct

It's not a pleasant topic, but to avoid confusion and issues for us all later it's important that you're aware of the University's policy on good scholarly conduct. As with all work for credit, you are expected to undertake assignment in line with good scholarly conduct. In essence, this means that:

- "You should complete coursework yourself, using your own words, code, figures, etc.
- Acknowledge your sources for text, code, figures etc. that are not your own.
- Take reasonable precautions to ensure that others do not copy your work and present it as their own." (https://web.inf.ed.ac.uk/infweb/admin/policies/academic-misconduct)

If work is not in line with good scholarly conduct, it will be penalised. In serious cases there may be zero mark. We expect that you will have read the academic misconduct policy before starting work on this coursework: https://web.inf.ed.ac.uk/infweb/admin/policies/academic-misconduct

As the policy above states, general discussions (but not specific solutions) are acceptable. Please ask us either privately or on Piazza if anything is unclear.

We forbid the use of Generative AI such as ChatGPT in any part of this assignment. The University's position that "**such systems must not be used, unless explicitly allowed in writing by the course organiser".**

Publishing your solution is not permitted, in line with the policy on Academic Misconduct.

Project description

For your final project in FDS you will work on a data science project. The goal of the project is to go through the complete data science process to answer a question. You will:

- acquire the data, explore and visualise it
- apply one or more basic techniques from descriptive and inferential statistics and machine learning
- interpret and describe the output from your analysis
- communicate the results so that there is a clear story.

To reduce workload, and make the project more enjoyable and potentially interesting, we are encouraging you strongly to undertake the project in self-selected groups of two or three. However, we are offering the option of undertaking the project individually. There will be slight differences between the individual and group projects, as described below.

Project options

We are offering a choice of three project options:

1. Historical and world-wide trends in ultramarathon running
2. The cancellations of planned operations in the Scottish NHS.
3. National university student satisfaction survey data

Later in the document there are more details of each option, including questions to address.

1. If you are working individually, you should address the main question we have supplied.
2. If you are working in a pair, you should address the main question we have supplied, and propose and address an extra question.
3. If you are working in a group of three, you should address the main question we have supplied and propose and address two extra questions.

Feedback on your progress

We offer the opportunity to share any progress on your project either via a mini one-page progress page (due Week 8) or by presenting at a workshop in week 9 or 10. This is not a marked component of the assignment; the purpose of this is to help you reflect on your progress, and to get feedback from your tutor and peers (or another FDS staff member for those who submit a written page). Details of this are outlined in the section below, "Feedback via written update or presentations (not for credit)".

Submission

We will ask you to submit:

1. A short report of your project written in LaTeX, using the supplied template and word limits (see "Report Structure", below). The report will be assessed according to the criteria below. The report will be submitted using Gradescope. For group submissions, only one member of the group submits and should use the Gradescope interface to tag their other group members at the time of submission.
2. Jupyter notebooks and/or python files containing the code. We will not mark the code, but we may wish to run it. The code must run with no errors. The code will be submitted to Learn.
3. If you are doing a project in pairs or threes, you will each need to write a short individual statement about how you divided the work, and what the individual contributions of each member of the group were. This can be a brief statement of contributions, e.g. "X & Y planned the analysis, Y implemented the analysis, X did the visualisations, X & Y wrote the report". This is common practice in scientific reports. This statement will be submitted via a Microsoft Form that will be distributed near the submission date.

Submission details for the report and individual statements will be released closer to the deadline.

Report Structure

Getting and using the LaTeX template

You must use the LaTeX template we supply, and not change margins or font sizes.

To get the template, firstly find the template in Overleaf: https://www.overleaf.com/read/brpnfsptvxnp and then either

1. "Copy Project" from the Overleaf menu to start editing your own version
2. or download the source as a zip file if you wish to edit it locally using another LaTeX editor.

The training resource *LaTeX for Beginners using Overleaf* by the University of Edinburgh Digital Skills & Training Team contains a step-by-step guide to using LaTeX with Overleaf, including how to do equations, tables, citations and references. Tutorials on LaTeX are also available from InfPALS.

Format

The report format is as follows:

- Overview, giving description of problem, work carried out, and results (Maximum 250 words)
- Introduction (suggested 400 words): Background to the question to be read by someone with no prior knowledge of the question. It should give:
    - Context and motivation - what is the area of this data science study, and why is it interesting to investigate?
    - Brief description of any previous work in this area (e.g., in the media, scientific literature or blogs)
    - Objectives of the project – what questions are you setting out to answer?
- Data (Suggested 300 words): A description of the dataset(s), and how you processed it or them:
    - Data provenance: Who created the dataset(s)? How you have obtained it (e.g., file or web scraping), and do the T&Cs allow you to use obtain the data for the project?
    - Description of the variables in each table, e.g. variables in each table, number of records.
    - Description of how you have processed the dataset, e.g., removing missing values, joining tables
- Exploration and analysis (suggested 500 words for individual report; proportionately longer for group projects). A data science analysis of the paper, including:
    - Visualisations and tables
    - Interpretation of the results
    - Description of how you have applied one or more of the statistical and ML methods learned in the FDS to the data
    - Interpretation of the findings
- Discussion & Conclusions (Suggested 400 words)
    - Summary of findings
    - Evaluation of own work: Strengths and limitations
    - Comparison with any other related work
    - Improvements and extensions – note that this is just *discussing* what improvements and extensions you would make if you had more time, not actually implementing them.

- References: A list of work cited – the template has examples of how to cite various types of work. Please ask if you need more help with citing.

Page limits

We will limit the report length depending on whether the project is individual, in pairs, or in threes:

- Individual project: 6 pages
- 2-person project: 8 pages
- 3-person project: 10 pages

The references do not count towards the page limit. To be clear this means that:

- For an individual project you can have 6 pages of the main text, including tables and visualisations, with the references section starting at the top of page 7. However, you can have the references within the 6 pages if you want.
- For a 2-person project you can have 8 pages of the main text, including tables and visualisations, with the references section starting at the top of page 9. However, you can have the references within the 8 pages if you want.
- For a 3-person project you can have 10 pages of the main text, including tables and visualisations, with the references section starting at the top of page 11. However, you can have the references within the 10 pages if you want.

Figure & Table format

- Ensure that the font size in the figures is at least 9pt in the actual PDF file you submit (not just specified as 9pt in matplotlib – see the second visualisation lecture for how to get font sizes correct).
- Do not change the font size in tables.
- All figures and tables should have a meaningful caption and should be referred to in the text.
- Note that the plots do not necessarily need to have a title above them – the figure caption (I.e. everything inside the `\caption{}` in LaTeX) can fulfil that role. However, titles above multiple axes in a figure can make them easier to read.

Forming groups

You can choose your own groups.

- If you haven't found anyone to work with but would like to find prospective group members, please use this form:
  https://forms.office.com/e/r3u5Y1gQW6
  We will try to find you group members with similar project interests. Please fill in this form by 5pm on Thursday 29 February. We will form the groups on Friday morning.
- We recommend setting up a **private** repository on GitHub to keep track of your code within your groups.
- We recognise that individual schedules, preferences, and other constraints might limit your ability to work in a group. The default expectation is that grades for each group member will be same, but if your statements of how you worked as a group indicate that one member did

significantly less than the others, we reserve the right to reduce the mark of that group member.

Please divide up tasks between yourselves, e.g. after an initial discussion, one of you might focus on data cleaning, and another on coding, and another on presentation.

Project options

Project option 1: Historical and world-wide trends in ultramarathon running

Ultramarathon running pushes the limits of human endurance, making it a fascinating subject for data analysis. An ultramarathon is described as any footrace longer than the traditional marathon length (42.2 kilometres) and consequently demands even greater resilience and perseverance of athletes. The following dataset spans over two centuries of ultra-marathon race records and offers a unique opportunity to explore global trends, factors influencing performance, as well as the evolution of the sport, https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running/data

Everybody (individuals and groups): We would like you to explore the evolution of participation and performance in ultramarathons across the world over time. How has finishing time performance changed for different distances and athlete demographics? Have there been any significant shifts in regional participation patterns? Additionally, we would like you to identify whether certain factors (e.g., event characteristics, athlete attributes) predict better performance. For example, can you develop a model to predict finishing times for future ultramarathon events based on available data?

Groups: The extra questions should extend the basic findings. Examples of questions are:

- Can you detect any influence of external factors (e.g., climate, terrain, technology, historical events) on finishing times or participation rates?
- You could also choose to take a 'deep-dive' into a shorter time period or the performance of one (or a collection of) athletes with remarkable achievements.
- Any other questions that arise as you explore the data.

Project option 2: Analysing the cancellations of planned operations in the Scottish NHS.

Healthcare systems worldwide constantly strive to balance patient needs with resource limitations. Understanding the factors contributing to cancelled planned operations is crucial for improving efficiency, reducing patient anxiety, and ensuring equitable access to essential medical services. The following dataset contains information on the number of cancelled planned operations in Scotland by NHS hospitals since 2015, https://www.opendata.nhs.scot/dataset/cancelled-planned-operations. We can gain valuable insights into this issue by analysing the number of cancelled operations, their reasons, and the responsible NHS boards/hospitals.

Everybody (individuals or groups): We would like you to explore the data to report statistics on the number of planned operations and reasons for their cancellation (e.g., clinical, capacity, patient-related) across Scotland, using tables, summary statistics and/or visualisations. Are there any interesting patterns in cancellations that you can identify over time? Additionally, are there significant differences in cancellation rates between different NHS Boards and hospitals? Can you identify regions requiring specific attention?

Groups: The extra questions should extend the basic findings. Examples of questions are:

- Do specific reasons for cancellation dominate in certain regions or hospitals, suggesting variations in practice or resource management?
- Do the number of cancellations and their reasons vary significantly by month, suggesting potential resource strain during specific seasons?
- You could also find additional data to compare the cancellations with a related issue (e.g. drug and alcohol treatment waiting times, https://www.opendata.nhs.scot/dataset/drug-and-alcohol-treatment-waiting-times).
- You could also identify and explore any outliers of particular interest within the data (e.g. relating to COVID-19 measures).
- Any other questions that arise as you explore the data.

Note, that explanatory data dictionaries relevant to the data can be found under the "Explore > Preview" option at the bottom of the given URL.

Project option 3: National university student satisfaction survey data

The National Student Survey (NSS) is an annual survey of final-year undergraduate students in UK higher education institutions. It is a valuable source of data on student satisfaction with their courses and universities. The 2023 NSS data release includes responses from over 339,000 students, making it a rich resource for data analysis: https://www.officeforstudents.org.uk/data-and-analysis/national-student-survey-data/download-the-nss-data/

Everybody (individuals or groups): We would like you to explore what factors contribute to higher student satisfaction within and across different subject areas and universities in the UK. How does student satisfaction vary by subject of study? For example, are students in STEM subjects more or less satisfied than students in humanities subjects? Are there any universities that consistently outperform or underperform in terms of student satisfaction? Where appropriate, make sure to relate your answers to the factors underlying student satisfaction as indicated by the survey's questions.

The website provides various versions of the data; we recommend that you mainly look at the "2023 NSS results by registering provider (full-time) (XLSX, 95.4 MB)" under 'Provider-level data'.

Groups: The extra questions should extend the basic findings to explore advanced relationships in the data. Examples of questions are:

- How has student satisfaction changed over time? Are there any trends that can be identified?
- Are there any trends in satisfaction between England, Scotland, Wales and Northern Ireland?
- Do students in smaller, niche subjects express different satisfaction levels compared to those in larger, 'mainstream' subjects?
- You may also wish to explore the 'student characteristic' data which gives student satisfaction broken down by characteristics such as age, disability, ethnicity (and more).
- Any other questions that arise as you explore the data.

Feedback via written update or presentations (not for credit)

At the beginning of week 8 (mid-day 11th March), you will be required to let us know whether you will either:

- be attending a week 9 or 10 workshop to present an update on your project (e.g. at least one visualisation)
- or submitting a mini one-page document of your update to receive some written feedback on.

This part of the project is optional (and not marked) and is meant to be helpful and informal.

- Please fill in this excel sign-up sheet to attend a week 9 or 10 workshop.
- If submitting a written project update, we ask that you use the following latex template to do so: https://www.overleaf.com/read/rcjrrftvqmkj#ab096. We will aim to provide feedback on written project updates by the beginning of week 9.

Please contact Anna Hadjitofi <a.hadjitofi@ed.ac.uk> if you wish to discuss alternative arrangements for feedback.

Criteria for Evaluation

The rubric for the coursework is available in the project instructions box on Learn.

Resources

- University of Edinburgh digital skills guide: LaTeX for Beginners using Overleaf
- InfPALS provide useful resources and tutorials on LaTeX and git