FDS Coursework 2, 2024/25

Data Science Project

Released	Wednesday 12 February 2025 (Week 4)
Submission deadline	Friday 28 March 2025 (Week 10) 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.
Formative feedback	Monday 3 March (Week 7) at 12:00 UK time: deadline for submission of project plan/update Week 8: feedback on plan/update Week 8 and 10: Project presentations
Expected effort	30 hours for a student up-to-date with course activities. There will be variance, but if you're spending longer than 30 hours, you should consider how many marks you are gaining for each extra hour of work.

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

Good scholarly conduct

As with all work for credit, you are expected to undertake this coursework in line with the University's policy on 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.

Generative AI: The School of Informatics default policy is that such Generative AI must not be used, unless explicitly allowed in writing by the course organiser, and where allowed, must be cited and explained as described in the University guidance. In this coursework, we allow you to use Generative AI for the following purposes:

- Helping to improve your grammar, spelling, and writing
- Overcoming writer's block through dialogue with a GenAI tool
- Help with writing, de-bugging code and logical reasoning

If you are going to use Generative AI, please read the University's guidance on Generative AI in full, including how to cite use of AI: https://information-services.ed.ac.uk/computing/comms-and-collab/elm/guidance-for-working-with-generative-ai**.**

If you use Generative AI, we strongly suggest that you try writing your report without the use of Generative AI first – part of the benefit of writing is that it helps you develop your thinking, and this coursework provides a valuable opportunity to learn how to do this.

Whatever tools you use, you are responsible for all aspects of your work.

You may publish your report three weeks after the submission deadline. If you choose to publish your report, we would appreciate it if you could let us know and give permission to use your work as an example of previous assessment.

Assessment criteria

This coursework is assessing the following course learning outcomes:

1. Describe and apply good practices for storing, manipulating, summarising, and visualising data

2. Use standard packages and tools for data analysis and describing this analysis, such as Python and LaTeX.

3. Apply basic techniques from descriptive and inferential statistics and machine learning; interpret and describe the output from such analyses.

5. Complete a data science project and write a report describing the question, methods, and results.

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

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.

Project options

We are offering a choice of three project options:

1. Heat and electric data from Appleton Tower since 2016
2. Video game data from the Steam catalogue
3. University of Edinburgh course data

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

Feedback on your progress

We offer the opportunity to share any progress on your project either via a mini one-page progress update (due Week 7) or by presenting at a workshop in weeks 8 or 10. Neither the progress update nor the presentation are for credit; their purpose is to help you reflect on your progress, and to get feedback from your tutor and peers (or a FDS staff member for those who submit a progress update). 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.
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.

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. We reserve the right to penalise submissions that do not use the template.

1. To get the template, firstly find the template in Overleaf: https://www.overleaf.com/read/yzbyfvyvtyjg#0f70cd
2. "Copy Project" from the Overleaf menu to start editing your own version
3. 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 question(s) 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?
    - **Data description**, e.g. variables in each table, number of records.
    - Description of how you have **processed** the data, e.g., cleaning, removing missing values, joining tables
- **Exploration and analysis** (suggested 500 words): A data science analysis of the paper that addresses the question(s) in the Introduction, including:
    - Visualisations and tables
    - Description of how you have analysed the data to address the question(s) posed in the Introduction. This analysis may use the statistical and ML methods learned in FDS, but it is not required to use inferential statistics or ML if they do not help address the question.
    - Interpretation of the findings
- **Discussion and Conclusions** (Suggested 400 words):
    - **Summary of findings –** a short summary of what you can conclude about the objectives and questions in the Introduction
    - **Evaluation of own work: Strengths and limitations** – the extent to which the conclusions are supported by the evidence, i.e. the data and your analysis of it. For example, there could be limitations in the data, e.g. missing data or problems with the data collection, that could affect your conclusions.
    - **Comparison with any other related work –** for example if there are people who have done similar work, are your conclusions similar or different? If there is no similar work, perhaps you can set your work in the wider context.
    - **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 using BibTeX. Please ask if you need more help with citing.

Overall the text, visualisations and tables in the report should tell a story: i.e. a coherent narrative that sets up a question(s), allows the reader to understand what the data are, and addresses the question.

Page limits

We will limit the report length to **6 pages**. The references do not count towards the page limit. To be clear this means that 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.

Figure & Table format

- For figures, follow the FDS visualisation principles that were used in Coursework 1
- Ensure that the font size in the figures is at least 8pt in the actual PDF file you submit (not just specified as 8pt 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.

Project option details

Project option 1: Heat and electric data from Appleton Tower

For every student in the University of Edinburgh, 1.51 tonnes of CO2 equivalent (CO2e) was produced in 2022-23. Electricity and Gas account for 82% of the University's carbon emissions (26,893 tonnes CO_2e from Electricity and 34887 tonnes CO_2e from gas; 74,986 tonnes CO_2e in total).

The University generates electricity by burning gas in a Combined Heat and Power plant close to George Square. Waste heat from this process is delivered to buildings via a heat network of pipes; in colder weather a gas-fire boiler is used to supplement the heat.

Energy consumption in many buildings in the University is monitored every half hour. The University Estates Department have given us electricity and heat consumption from Appleton Tower. We would like you to explore this data. You could ask one or more of the following questions, or pose your own question related to this data:

- How has the electricity and/or heat consumption changed over time?
- Are there any changes that you cannot explain?
- Are there any anomalies in the data?
- You may want to understand how the consumption varies with factors such as temperature and semester time, perhaps using a linear model.

A challenge with this data will be data cleaning and understanding its meaning – we know from Estates that there is some missing data

The data and a data description are in this Sharepoint folder.

Project option 2: Video game data from the Steam catalogue

Steam is the largest digital platform for the distribution of PC games. It currently has over 100,000 games, with over 130 million monthly active users in 2021.

An independent Steam user used Steam's API (https://steamcommunity.com/dev, under the license in https://steamcommunity.com/dev/apiterms) to scrape publicly available data from Steam during October 2024. They then cleaned the data, organised it into an easier-to-use database, and made it publicly available. The same user also did some analysis of the data, which can be seen in this YouTube video: https://youtu.be/qiNv3qv-YbU.

We would like you to provide further insights about the data. Don't just repeat what the above video has done but go beyond it (you may use it as inspiration). You could ask one or more of the following questions, or pose your own question related to this data:

- What is the relationship between review scores and game tags?
- Does release date affect anything significantly?
- Are there specific games that caused a big impact to their genre?

Note that the dataset is larger than the others and is presented in the form of several different .csv files that you would need to combine cleverly to answer most of the interesting questions.

A GitHub repository containing the data, along with a description of the dataset, can be found here: https://github.com/NewbieIndieGameDev/steam-insights. Note that the dataset only contains information about the games themselves at the time it was scraped, and not about playtime or about temporal changes to the data.

Project option 3: University of Edinburgh Course data

Having moved to a new Timetabling system in 2024/25, the University of Edinburgh is running a project to improve Timetabling and Course Selection processes. The aim is to "deliver a stable, detailed and accurate timetable for students before [students] start the academic year. It is also working to ensure students choose from a suite of courses that are available to them within a reasonable set of parameters and that this is explained and transparent to all."

There are a number of challenges in timetabling and course selection, including: (1) Courses offered change from year to year; and (2) the course catalogue and Degree Programme Tables should be stable from the publication date each year (around 11 April), but there are changes to courses offered, the delivery semester, and availability to visiting students after this point.

We have provided course catalogue data dumped at various times through the planning period in 2022/23, 2023/24 and 2024/25. In addition, you could try scraping data from DRPS from the current and previous years – this information is Copyright University of Edinburgh. You could ask one or more of the following, or pose your own question related to this data.

- How stable is our offering of courses over the years, e.g. how many courses courses open and close each year and what's the fraction of 10, 20 and 40 credit courses?
- How many changes in the DRPS were there in the past 3 years after the publication date, and when did these changes happen?
- Potentially more detailed questions about the nature of courses

The data provided should be fairly clean, but a challenge is thinking what metrics you could use to show changes. You could try some web scraping to get more information about courses from their DRPS entries – if so, please be mindful to follow the principles of ethical webscraping in the lecture notes.

The data and a data description are in this Sharepoint folder.

Feedback via written update or presentations (not for credit)

At the beginning of week 7 (Monday 3 March, 12 noon), you will be required to let us know whether you will either:

- be attending a week 8 or 10 workshop to present an update on your project (e.g. at least one visualisation)
- or submitting a mini one-page 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 8 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/ktmrsbwgmwjn#9f6061 . We will aim to provide feedback on written project updates by the beginning of week 8.

Please contact Bobby Ikomonov <Borislav.Ikonomov@ed.ac.uk> if you wish to discuss alternative arrangements for feedback.

Resources

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