FDS Coursework 3, 2021/22

Semester 2 Final Project

Released: Monday 7 March 2022
Submission deadline: Tuesday 5 April at 16:00 BST

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

Late submission rules

This coursework uses the Informatics Late Submission of Coursework Rule 3 with **a maximum 6 day extension**:

- Extensions, Extra Time Adjustments (ETAs) for Extra Time and for Extra Time for Proof Reader/Interpreter are permitted, but cannot be combined. The maximum extension is up to 6 days, or fewer if specified.
- Penalty: If assessed coursework is submitted late without an approved ETA extension, it will be recorded as late and a penalty of 5% per calendar day will be applied for up to the specified number of calendar days (≤6), after which a mark of zero will be given.
- For electronic submissions, the last version that has been submitted by the deadline will be the one that is marked (late submission will only be accepted if no submission in time has been made).
- If a student with an extension or either type of ETA submits late beyond the specified extended deadline a mark of zero will be given.

If you are granted an extension for assessed group work please inform the other members of your group and the Course Organiser so they are aware of the extension.

Good scholarly conduct

It's not a nice 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 a zero mark. We expect that you will have read the page on academic misconduct before starting work on this coursework: https://web.inf.ed.ac.uk/infweb/admin/policies/academic-misconduct

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

However you obtain the assignment, 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. Performance of Scottish A&E services
2. Worldwide trends in music streaming according to Spotify
3. Student learning on the EEdi educational platform

For more details of each project, see later in this document.

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

Initial progress for the project, including at least one visualization, will be presented in a dedicated workshop in week 9.

Submission

We will ask you to submit:

1. A short report of your project written in LaTeX, using the supplied template (**available in Overleaf** – https://www.overleaf.com/read/brpnfsptvxnp) and word limits. The report will be assessed according to the criteria below. The report will be submitted via Turnitin.
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 via github-classroom.

3. If you are doing a project in pairs or threes, you will each need to write a short individual statement how you divided the work, and what were the individual contributions of each member of the group. This can be a brief statement of contributions, e.g. "X & Y designed 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.

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

Report Structure

Format

You must use the LaTeX template we supply, and not change margins or font-sizes. You can either "Copy Project" from the Overleaf menu to start editing your own version or download the source as a zip if you wish to edit it locally using another LaTeX editor. This link contains useful documentation on LaTeX, including how to do citations and references.

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

- 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 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 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 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 Q&A session recording from after CW1 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 come up with your own teams.

- If you know who you want to work with, please set up the team on github-classroom at this link: https://classroom.github.com/a/PAGpFI-i
- This page has instructions on how to do this
- If you want to work alone, use the link above and set up a team called “Individual <your username>”.
- If you haven’t found anyone to work with but would like to find prospective team members, please use this form:
  https://forms.office.com/r/0mDjL2dBEm
  We will try to find you teammates with similar project interests. Please fill in this form by 9am on Thursday 10 March. We will form the teams on Thursday afternoon.

- We recognise that individual schedules, preferences, and other constraints might limit your ability to work in a team. 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 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: Performance of Scottish A&E services

Waiting times are important to patients and regular reporting of such measurements enables monitoring of how the NHS is responding to demand for services. Since 2007, the national standard for Accident and Emergency (A&E) is that new and unplanned return attendances at an A&E service should be seen and then admitted, transferred, or discharged within four hours. Data on attendances at A&E services across Scotland is available at: https://publichealthscotland.scot/publications/ae-activity-and-waiting-times/ae-activity-and-waiting-times-month-ending-31-december-2021 (see "A&E activity waiting times statistics data CSV | 1.6MB" in the data files section of the webpage).

Everybody (individuals and groups): We would like you to explore the data on monthly activity and the compliance with the 4-hour standard of these services across Scotland, using tables, summary statistics and/or visualisations. How well have services complied with the 4-hour waiting time standard? Are there any interesting patterns in compliance that you can identify over time? For example, you might want to consider comparing pandemic and pre-pandemic periods and think about how confident we can be about the size of any effects observed. Does the activity at some treatment locations change more than others over time?

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

- Is there a relationship between the discharge destination and whether an attendance meets the 4-hour waiting time aim?
- Is there an observable effect of the time of year on the number of attendances or waiting time?
- Any other questions that arise as you explore the data.

If you are testing hypotheses, you should of course report on all the hypotheses you have tested.

Project option 2: Worldwide trends in music streaming according to Spotify

Music streaming is now ubiquitous and streaming services provide easy access to massive databases of music of many genres. Currently, Spotify plays an important part in music streaming. Spotify's 'Worldwide Daily Song Ranking' dataset contains the daily ranking of the 200 most listened songs from several countries around the world from 2017 and 2018 by Spotify users (https://www.kaggle.com/edumucelli/spotifys-worldwide-daily-song-ranking).

Everybody (individuals or groups): We would like you to explore how artists and tracks' popularity varies in time across the global stage. For example, do regions share the same top-ranking artists or songs? Does the popularity of a track in one region predict its (upcoming?) popularity in other regions?

You could also visualise the distribution of ranked position and number of streams, and comment on the relationship between them.

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

- Can a track's name alone predict its popularity? You might want to consider computing features of a track's name and think about what other variables in the data should be controlled for.
- Are there any patterns in what day(s) of the week and/or months experience the most streams? Do these patterns, if any, differ by region?
- Does reaching the top *N* (e.g., 5, 10, 20) positions have any relationship with the subsequent lifetime of a track in the data?
- You could also choose to take a 'deep-dive' into the work of one (or a collection of) artist(s) and examine trends across their apparent most popular tracks.
- Any other questions that arise as you explore the data.

You may wish to find additional data for these tasks (but are not obliged to).

Note: we suggest you do not commit the data file from Kaggle to Github. Instead you can load the file into Python like this:

```
dat =
pd.read_csv('https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2021-2022/spotify/data.csv.zip')
```

Project option 3: Student learning on the EEdi educational platform

Online education systems offer students the opportunity to learn using technology. Many platforms use multiple-choice questions where incorrect answers are chosen to highlight common misconceptions. By construction, this reveals something about the nature of the students' knowledge that platforms could hope to utilise to then seek out relevant material to help resolve the misunderstanding. We would like you to study a dataset provided by the online educational platform EEdi, offering data from students' answering mathematics questions. The data was originally published as part of the NeurIPS 2020 Education Challenge (https://eedi.com/projects/neurips-education-challenge). We have extracted a subset the data for you to work on: https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2021-2022/eedi.

Everybody (individuals or groups): Investigate how students' self-reported confidence level when answering a question relates to their learning behaviour. For example, how does the level of confidence correlate to the true response? When are students' over-confident and under-confident, and can this be used to identify any wider observations about specific questions, such as their difficulty? Is there a relationship between the available metadata (e.g., date of birth) and self-reported confidence when answering questions?

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

- Can you compare the effects of different aspects (e.g., time of day, demographics, income support) on student performance?

- Are there any subjects within the overarching topic of mathematics which students especially struggle or perform well with?
- Any other questions that arise as you explore the data.

Note: we suggest you do not commit the full EEDI data file to Github. Instead you can save a copy of the files locally (but not commit to Github) or load the file into Python like this:

```
dat =
pd.read_csv('https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2021-2022/eedi/task_3_4.csv.gz')
```

Criteria for Evaluation

We will consider the following criteria when marking:

- Presentation in week 9 workshop is an essential requirement, but we will not mark the quality of the presentation
- Content:
    - Clear and complete overview
    - Clear description of context and objectives in the introduction
    - Clear description of where the data has come from and how you have processed it
    - Overall quality of exploration using visualisations, tables and descriptive statistics – how well the story of the data is told
    - Techniques from descriptive and inferential statistics and machine learning have been applied appropriately
    - Interpretation of the results is accurate
    - The work has been critically evaluated, I.e. limitations have been considered or has been discussed in the light of at least one other finding relating to the question
- Presentation of report:
    - The report is written in LaTeX
    - Figures meet guidelines for font sizes
    - Figures have meaningful labels and captions
    - Writing is clear, including being spell checked
- Code has been supplied
- Originality and good scholarly practice
    - Previous work cited clearly and correctly

Resources

- University of Edinburgh digital skills guide: LaTeX for Beginners using Overleaf