FDS Coursework 3, 2022/23

Semester 2 Final Project

Released: Monday 6 March 2023
Submission deadline: Tuesday 4 April 2023 at 12:00 UK time

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 1: Extensions are permitted (7 days) and Extra Time Adjustments (ETA) are permitted and can be combined.

Penalty: If assessed coursework with a numerical mark is submitted late without an approved extension, it will be recorded as late and a penalty of **5% per calendar day will be applied for up to 7 calendar days from the original deadline, after which 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 ITO <ito@inf.ed.ac.uk> 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. Historical trends in movies according to IMDb
2. Chess gameplay analysis
3. Airbnb listings in Edinburgh

More details of each project are given later in this document.

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.

Each individual or group is required present their progress on the project, including ideally at least one visualization, at the workshop sessions in week 9 or 10. The purpose of the presentation is to help you reflect on your progress, and to get feedback from your tutor and peers. Your tutor will not give you a numeric mark, but they will check your attendance. If you have difficulties meeting this requirement, please contact Anna Hadjitofi <a.hadjitofi@ed.ac.uk> to discuss alternative arrangements.

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 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

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 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.

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 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 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/vMJ8dPPCfY
  We will try to find you group members with similar project interests. Please fill in this form by 9am on Thursday 9 March. 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 trends in movies according to IMDb

Leading streaming services, such as Netflix, Amazon Prime and Disney Plus, use recommendation systems to provide suggestions to users of films that they predict they will enjoy. This is an especially difficult task in the case of movies that haven't been seen by a wide audience yet: what can we say about the success of a movie before it is released? Are there certain companies that have found a consistent formula? The Internet Movie Database (IMDb) has made available its data on movies released between 2006 - 2016, including their popularity ranking as recorded by users on their website: https://www.kaggle.com/datasets/gan2gan/1000-imdb-movies-20062016

Everybody (individuals and groups): We would like you to explore what makes a movie popular and/or successful. From the data available, what factors predict the revenue of a film and how well (if at all) can we predict the revenue of a film? Additionally, given that major films costing over millions to produce can still 'flop' in the box office, can we predict which films will be highly rated by users, regardless of if they are a commercial success? We would also like you to visualise the distribution of ranked position and number of votes, and comment on the relationship between them.

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

- Does the year of release or current trends in genre have any influence on a movie's preceding popularity rating or revenue?
- You could also choose to take a 'deep-dive' into the work of one (or a collection of) actor(s) / actress(es) or director(s) and examine trends across their apparent most popular movies.
- Any other questions that arise as you explore the data.

Project option 2: Chess gameplay analysis

Chess is a two-player tactical game played on a 64 squared 8x8 checkerboard, wherein both players start with 16 pieces and each piece has its own role to play. Both players have alternative turns starting from white, moving only one piece at a time, except castling, the goal of the players should be to checkmate the opponent's king. Chess is primarily a game of patterns, and depends on calculations, analysis and decisions. Chess.com has made available their data resulting from 60,000+ chess games played by visitors to their website: https://www.kaggle.com/datasets/adityajha1504/chesscom-user-games-60000-games

Everybody (individuals or groups): We would like you to perform an analysis exploring what allows a player to win. This is an intentionally broad and open question, enabling you to choose what aspects / possibilities interest you the most. For example, the relationship between openings and victory for black and white, or whether a players (ELO) rating can predict whether they are more likely to win.

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

- Is it possible to predict the (ELO) rating of the player from the series of their moves?
- Within the context of this dataset, do individual users show particular play styles?

- You could also choose to take a ‘deep-dive’ into one of the features (e.g. the “time control”) to analyse their apparent influence on the game.
- Any other questions that arise as you explore the data.

Project option 3: Airbnb listings in Edinburgh

Founded in 2008, Airbnb has undoubtedly changed the global travel industry. Functioning as a listing site for private B&Bs, it offers tourists the opportunity to stay in private properties as an alternative to hotels. Airbnbs can now be found in 191 countries across the world and whilst the company’s reach is enormous, it also has profound effects on the housing market, pushing up rents and shrinking local property markets. In a report commissioned by the Scottish Government in 2019, just over half of all active Airbnb listings in Scotland were in the City of Edinburgh or Highland Council areas [1]. We would like to explore the following quarterly data published via Inside Airbnb (http://insideairbnb.com/edinburgh) which contains rich details about recent Airbnb listings and usage in Edinburgh.

Everybody (individuals or groups): We would like you to investigate what makes a good property for Airbnb in Edinburgh. How well can features of a property listing be used to predict its popularity or short-term rental price? Are particular areas or neighbourhoods more sought after or expensive than others?

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

- Have you identified anything unusual about listings which suggest they are not genuine / true rentals?
- Some Airbnb hosts have multiple listings and may be running them as a business. Do their listings seem more popular than more hosts with single listings?
- The housing policies of cities and towns can be restrictive of short-term rentals, to protect housing for residents. Do features, such as the ‘minimum nights’ setting for listings reveal any patterns that may be indicative of such policies in particular neighbourhoods?
- 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).

Since InsideAirbnb updates its available data quarterly, we have saved the snapshot of the data as released on the 16th of December 2022 and made it available here: https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2022-2023/airbnb. Note: we suggest you do not commit the full data file to GitHub. Instead, you can save a copy of the files locally (but not commit to GitHub) or load the files into Python like this:

```
dat =
pd.read_csv('https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2022-2023/airbnb/listings.csv.gz')
```

Criteria for Evaluation

We will consider the following criteria when marking:

- Presentation in week 9 or 10 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

References

[1] Communities and third sector, Housing. “Short-term lets - impact on communities: research” In: *The Scottish Government* (2019). Retrieved on 16 Jan 2023. URL: www.gov.scot/publications/research-impact-short-term-lets-communities-scotland.