Foundations of Data Science
Semester 2 Final Project

The FDS final project is now available. This is a **marked** assignment which will count towards **40%** of your final grade for **Inf2-FDS**. The deadline for submission is **Monday 5 April at 16:00 GMT.**

The usual penalties in the Informatics Late Coursework & Extensions policy apply:

- 5 percentage points will be deducted for every calendar day or part thereof it is late, up to a maximum of 7 calendar days.
- If you have not submitted coursework within these 7 days, a mark of zero will be recorded.

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 5-week 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 to undertake the project in self-selected groups of two or three. However, as we have previously indicated this would be an individual project, 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. First project. UK Higher education data
2. Second project: Scottish Munros
3. Third project: Just Eat Cycle Data

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 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 10.

Submission

We will ask you to submit:

1. A short report of your project written in LaTeX, using the supplied template and word limits. The report will be assessed according to the criteria below.
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 pass compilation 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.

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.

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: 6 pages
- Pairs: 8 pages
- Threes: 10 pages

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 this Q&A session recording 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.

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/g/vM6MAk9U
  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, please use this form: https://forms.office.com/Pages/ResponsePage.aspx?id=sAafLmkWiUWHiRCgaTTcYTypls-7hrNNpEglAlZTPD1UNFVRUzY1VjdEUU5LRDNPUzJETzVOMkswQi4u
  to find prospective team members. We will try to find you two teammates with similar timezones and project interests. If you are looking for team-mates, please fill in this form by 9am on Thursday 4 March. We will form the teams on Thursday afternoon.
- We recognise that individual schedules, different time zones, 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: Variability of UK Higher Education grades

There is a wealth of data on Higher Education in the UK at: https://www.hesa.ac.uk/data-and-analysis/ Over the past few years, the number of students obtaining a first class degree has increased from 28% in 2018/19 to 35% in 2019/20 (https://www.hesa.ac.uk/data-and-analysis/students/outcomes). However, these overall statistics do not show how much variability there is in degree classifications between different "Higher Education Providers" (I.e., Universities and Colleges), or why this might be the case.

Everybody (individuals and groups): We would like you to visualise the distribution of degree classifications achieved in different providers using tables, summary statistics and/or visualisations. We would like you to identify interesting features, for example, which providers have the highest and lowest fractions of first class and other degrees? Are any differences you see statistically significant?

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

- Are there any interesting patterns in grades that you identify in the data over time?
- Have some institutions changed more the fraction of firsts over time than others?
- Can we explain differences between institutions, for example the sort of subjects studied (science versus arts), the size of the institution, the country in the UK?
- 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: Popularity of Munros

Munros are hills in Scotland taller than 3000 feet (914.4m) that are "sufficiently separated" from their nearest peaks. (https://www.smc.org.uk/hills/). The Scottish Mountaineering club maintains a list of

Munros at https://www.smc.org.uk/hills/ The walkhighlands website contains a user-generated list of the number of times that walkhighlands users have climbed various Munros (https://www.walkhighlands.co.uk/munros/most-climbed).

Everybody (individuals or groups): What makes a Munro popular? To address this question you might start by visualising the distribution of Munro heights and frequency of them being climbed. Are there any obvious outliers in the data, and can you think why this might be the case? Does it look as though there is there a statistically significant relationship between height and frequency of climb? How much can we trust the conclusions given the nature of the data?

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

- Can you identify any other factors that could lead to the popularity of Munros?
- Can you cluster Munros according to their features?
- Any other questions that arise as you explore the data.

You may need to find additional data for these tasks.

Project option 3: Edinburgh Cycle Hire in the pandemic

The Edinburgh "Just Eat" cycle hire scheme provides data about cycle hire trip from September 2018 until now: https://edinburghcyclehire.com/open-data/historical . During the pandemic, there has been a huge shift in the way people move around, with many more people cycling for leisure. On the other hand, people may be more nervous about using shared bikes for hygiene reasons, and maybe there are fewer situations when people want to use a shared bike.

Everybody (individuals or groups): What has the effect of lockdown on bicycle usage in Edinburgh? You might answer this question in terms of the number of trips, popular stations or popular hire times. You might want to consider how to make your comparisons between pandemic and pre-pandemic fair and to think about how confident we can be about the size of any effects observed.

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

- What's the difference between winter and summer, patterns of usage within the week?
- Visualize the popular cycling routes as they change over time.
- Any other questions that arise as you explore the data.

Criteria for Evaluation

We will consider the following criteria when marking:

- Presentation in week 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