# IN3061/INM430 Principles of Data Science (PRD1 A 2019/20)



## Coursework description: A Tiny Data Science Project

You carry out a data science project from start to finish. You will decide on a domain/dataset, come up with a domain-relevant analysis strategy, perform data wrangling, carry out your analysis/modelling as needed, validate your results and communicate your observations along your journey.

You will broadly follow the whole data science process as discussed in the lectures. So don't just predict something. Do **data analysis** and identify **findings** of **relevance to your application domain**

Make sure you **enjoy** this Tiny Data Science Project! The best projects are those where you enjoy the topic and know something about the application domain.

You will need to deliver:

1. **Submission 1: Plan & progress update:** a project progress update (unassessed but mandatory) on **Sunday 3rd November**, on which you will get individual feedback.
2. **Submission 2: The final analysis**: a final submission (computational notebook & report) on **Sunday 15 December**. Your **whole module mark** will be based on this.

---

## Approach

We suggest the following approach:

- **Stage 1: Identify the application domain and datasets.** You are free to choose the application domain and the dataset that you want to investigate. Some ideas are listed below. Make sure that data are sufficiently complex. Combining one or more datasets may help this so that you can investigate the demonstrate observations that are not obvious by performing trivial operations in a system such as Excel. Source your data from wherever you want (as long as its legal and ethical!). Make sure you state the source of your data. Here are some ideas:
  - A challenging option is Yelp Data Challenge provides datasets on business reviews, social network activities, check-ins, etc. This is an ongoing challenge with a deadline on the last day of the year. *(If you're interested in actually entering, please get in touch.)*
  - For a financial focus, here is a good resource on industries and companies.You can enrich your analysis by adding data and new perspective from data repositories of World Bank and IMF.
  - data.gov.uk is a great place to grab datasets. There's a range in quality, where some is too simple (highly aggregated tables) and some is rather unstructured so might take time to prepare. But generally the data sets are relevant and updated.
  - For a more local focus, London Datastore is a great place to start. You might even link some of these datasets to the Yelp ones.
  - Airline Data Project is an interesting initiative by MIT where they gathered several sources of information related to the aviation industry. The datasets can be accessed here. And there is even a glossary to get you introduced to the domain.
  - Visualizing.org has a nice collection of data resources which might help you to find relevant domains and datasets.
  - Kaggle competitions are a great resource to find relevant data sources and problem descriptions. However, if sample solutions for a competition is released, you need to clearly cite any ideas you borrow from the solutions and justify how your solution differs from theirs.

  **Stage 2: Identify questions and analysis tasks.** Ask yourself **"What do I want to achieve/demonstrate in this analysis?"**. **Don't just predict something** - you goal is to do **data analysis and identify findings** of **relevance to your application domain**. You will be expected to answer this by documenting: a brief **overview of the domain, analytical questions** that are being asked, the list of your **objectives**, and expected output of your analysis. Familiarise yourself with the problem domain by reading related academic papers, related articles and maybe even interacting with people active in the domain of interest. As the project is "tiny" we don't expect you to be exhaustive, but it will help inform and justify your analysis strategy and findings.
- **Stage 3: (Initially explore) and develop an analysis strategy (plan)**. Your plan should be designed to address the tasks you identify and designed to fulfil your stated objectives in your selected problem domain. This should be informed by an **initial investigation** of the data sources and the characteristics of the data. Plan which data processing steps you will need to perform, how you will transform the data to make it useable, which data analysis algorithms you could and what sorts of observations these may lead to. Remember that the data analysis process is highly agile, you will find yourself iterating through several different methods and changing your initial plans frequently.

- **Stage 4: Performing the Analysis**. This is the **core** of your analysis which will include getting the data ready for analysis, carrying out your analysis/modelling as needed, validating your results and communicating observations. It will involve many iterative steps. We **expect** to see evidence of:

  - **data processing** (wrangling/merging) to the extent necessary (not all datasets are messy) to prepare useful and robust data to work with
  - **data derivation** to support your analysis. This may include feature engineering. Be creative
  - **building models** that help explain the underlying phenomena. These will help you to explain the data better and/or perform predictions. Models include those using regression, classification or clustering methods
  - **evaluating (and validate) your results** using formal methods and try to minimize uncertainties in your findings.

  **Suggestions**:
  - Feel free to collect and link additional data, if appropriate.
  - You might consider applying dimension-reduction where you have several data variables and need to find out factors that affect the phenomenon that are not directly visible from raw data.
  - You might consider converting data into a graph and performing some graph theoretical analysis. For instance, transport connections can be treated as a graph and some analysis can be carried out on top.
  - Apply a number of different methods to solve the same analytical tasks and compare results to gain trust in the results.
- **Stage 5: Findings and reflections**. Identify the findings that are worthy of reflections and reflect on them, articulating them and relating them to the analysis tasks, questions and objectives you have identified. Focus on the most significant ones. How are your observations useful and actionable? Can your observations provide new insight, inform decision making, and if so how? Once you are at this stage, you may still find yourself going back to the previous stages.

---

## General points

- **Individual work.** This is an **individual piece of work**. You work must be carried out individually, though of course you can discuss general concepts about your work with each other and with the teaching team).
- **Python and Python notebook**. We expect you to present you analysis in a **Python notebook**. However, feel free to utilize supplementary tools, computational methods, or software to enrich your analysis (include details and files where appropriate in your notebook submission ZIP).
- **Analysis not yielding good findings.** Not all analysis leads to the observations and findings you expect or wish for. If this happens, **don't panic!**. As long as the methods and plan you have following is reasonable and you can offer explanations as to why the analysis did not give the results you expected, you will not lose marks. Reasons may include issues to do with data, analysis methods and false expectations. Also, showing **a lack of relationship** is also a finding!
- **Reference your sources!** Clearly reference any resources you make use of in your analysis. This should include URLs of websites that have helped you. This is general good practice. (And the university take this seriously and have a range of sanctions they can apply to people who fluat these.)
- **Communicate your findings well.** It is in your interest to ensure the marker understands your work and its implications easily.
- **Describe the implications** of your findings to your application domain - i.e. its value.
- **Align the stages of your projects.** Ensure that the goals, analysis and findings follow from each other. Link your findings to the results of your analytical steps and your objectives.

---

## Submission 1: Plan and progress update (Sunday, 3rd November, 23:55)

Your plan should demonstrate that you've made a start on your project and provide enough detail for useful formative feedback from us. This submission is mandatory, but will not count towards your module marks. It is an opportunity to get individual formative feedback on your idea, plan and its scope. The brief feedback will ensure that your work is progressing in the right direction. It is also designed to encourage you to start work early and already have some initial observations.

Using the **Jupyter IPython notebook** template (see below), you need to demonstrate that you've chosen your data and application domain, have a plan, have started working with the data and have done some initial investigations. As provided by the template, your report needs to have four sections:

- **Section 1: Data source, application domain and analytical questions (maximum 150 words):** This is a very brief overview of *Stages 1 and 2* (as described above). Briefly describe your data sources and the domain they are related to. Try to write at least **one key question** you want answer with your analysis.
- **Section 2: Analysis Strategy and Plans (maximum 200 words):** This is a very brief overview of *Stage 3* (described above). Briefly comment on your overall strategy is, what transformations you will need to do, what analytical tools you plan to use.
- **Section 3: Initial investigations on the data sources (maximum 150 words):** Report some very brief early observations here. Some examples could be: How many features you have, what are the data types, do you observe several problematic data features, will you need significant data transformations? The answers might come from the code you include in the following section.
- **Section 4: Python code for initial investigations:** Here, we only want you to demonstrate that you've loaded your data into Python for initial investigation and have some ideas of how to approach it. Some code where you load your data and get the first statistical and/or visual summaries to provide you the insight you document in will be sufficient. We just want to see evidence of your progress.

**Submission:** Download the IPython notebook template and fill in the required sections. Export the final notebook as an HTML file, zip the whole folder (containing both the HTML and IPython notebook, but **not** the data) and submit as a *zip* file.

---

# Submission 2: Project Final Submission (Sunday, 15th December, 23:55)

You are expected to deliver your work in two complementary parts:

1. **A report**: introducing the domain, analytical objectives and your analysis plan, findings and reflections. It needs **2 main sections**, but feel free to split them into subsections:

   - **Analysis Domain, Questions and Plan** (maximum 500 words): containing the Stages 1, 2 and 3 as described in the approaches above.
   - **Findings and reflections** (maximum 1000 words): containing Stages 5 as described in the approaches above. You can include relevant figures from your computational notebook.

   **Submission**: Submit as a PDF to the submission area
2. **A computational notebook** The computational notebook should demonstrate all the steps of your data preparation, analysis and computational modelling. You are expected to present a **computational narrative** that mixes: code, resulting figures, and verbal discussions that explains and comments on the observations made. This should be produced using **Jupyter Notebook written in Python**.

   - **Data preparation, analysis and computational modelling** (code, figures and a maximum 1500 words of text): this should include Stage 4. Use sub-titles in your notebook to explain the different phases you go through. Make use of visualisations effectively. Justify your analytical steps (e.g. why transform a variable) and discuss observations using clear and well justified arguments.

   **Submission:** Submit a ZIP of your Python notebook folder containing your .ipynb, an HTML version of it and any support files required. **We don't need the data.**

Submission 2 will constitute **100% of your overall module mark** and will be based on the computational notebook and the report as a whole. We will look for the following:

- **Analysis Domain, Questions, Plan (20%):**
  - The text gives a **clear and informative overview of the domain**
  - The text identifies a **clear and convincing motivation** for the analysis
  - **Analytical problems** are clearly described and listed
  - The report includes a **well-considered plan/strategy** on how the analysis will be carried out
  - Clear objectives for the analysis are presented
- **Analytical Process (50%)**:
  - A **well-considered analysis methodology overall** which matches well with the motivation
  - Evidence of **effective data gathering**, **wrangling** and **transformation** steps
  - A set of data transformations and derivation steps adds richness to the analysis
  - A suitable set of analytical tools and methods have been utilised with convincing justifications
  - Incorporates **effective modelling of the data** using one or more of the methods covered, e.g., regression, clustering, dimension red., etc.
  - A number of complementary investigations that build a **holistic understanding of the data**
- **Findings and Reflections (20%)**:
  - The findings **critically evaluated**, discussing any potential **biases** and **limitations**
  - Objectives and overall motivations are **revisited** and **critically evaluated**
  - Includes reflection on the potential impacts and uses of the findings
  - Evidence of a **holistic understanding** of the data being analysed
- **Clarity and Technical Soundness in Presentation/Argumentation (10%)**:
  - The report indicates **technically well-informed** decisions
  - Arguments and analytical steps are properly **justified**
  - Facts and findings are clearly explained and demonstrates a **good level of understanding**
  - Effective use of **visuals, charts and factual tables**
  - Effective use of references where needed

For postgraduate students, the pass mark is 50%. For undergraduate students, the pass mark is 40%.

# General grading criteria

| | | | | | |
|---|---|---|---|---|---|
| PG: Distinction<br>UG: First class | 85-100 | A | A+ | Outstanding | The analysis is impressive and the report/notebook are written either at a professional/academic level. The student shows great capability in applying the different analytical skills gained through the lectures. The analysis domain is nicely introduced, analytical problems are identified and objectives are clearly set.<br>Data collection and processing are done at a professional level. Data analysis is carried out at a professional level where all details are given and all choices are justified. Good level of creativity is shown in data analysis and the report being robust and correct, also increases interest and suggests new mechanisms for data analysis. Good use of visualisation and communication skills contributes to the quality of the submission. |
| | 80-84 | | | Excellent | |
| | 75-79 | | A | Very Good | The report/notebook are very nicely written and almost at a professional/academic level. The student shows great capability in applying the different analytical skills gained through the lectures. The analysis domain, the analytical problems and objectives are clearly introduced in general although some details might have been left out.<br>Data collection and processing are done at a very good level and appropriate actions are taken. Analysis is carried out at a very good level level where all details are given and all choices are justified. Visualisations have been used nicely as a means of analysis and communication. Although the analysis shows attention to detail and clarity, more imagination and creativity could have improved this submission. |
| | 70-74 | | A- | | |
| PG: Merit<br>UG: Upper second class | 67-69 | B | B+ | Good | The report/notebook is nicely written and is at a good level. The student shows capability in applying several analytical skills gained through the lectures. The analysis domain, the analytical problems and objectives are introduced but some details might have been omitted.<br>Data collection and processing are done at a good level and appropriate actions are taken. Analysis is carried out |
| | 64-66 | | B | | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | at a level where most details are given and most choices are justified although there are some unclear discussions in the report. This report ticks almost all the boxes and fulfils requirements to a very acceptable degree. The submission could have been improved with more imagination and creativity.. |
| | 60-63 | | B- | | |
| PG: Credit<br>UG: Lower second class | 57-59 | C | C+ | Satisfactory | The report/notebook shows some analytical capability and knowledge in the methodologies covered in the lectures. Although certain aspects have been covered, there are important missing details about the problem domain and analysis objectives. The report demonstrates that suitable analysis processes have been utilised, however, not all decisions are justified and/or clearly documented. The observations and findings are presented and discussed to some degree but the submission lacks the depth and comprehensiveness that is expected from such a data science project. |
| | 54-56 | | C | | |
| | 50-53 | | C- | | |
| PG: Fail<br>UG: Third-class | 47-49 | D | D+ | Poor | The report/notebook shows some level of analytical skill in applying the methodologies covered in the lectures. Details of the analysis domain, the analytical problems or the objectives have not been properly discussed. Although some analytical tools have been utilised, the justifications on their use and the interpretation of the outcomes are often flawed. The report does not do a good job in communicating and discussing the observations either. |
| | 44-46 | | D | | |
| | 40-43 | | D- | Very Poor | This is a problematic submission in general. Although some knowledge of basic concepts have been demonstrated, no clear analytical/scientific process has been followed. Fundamental problems in the decisions given during the analysis.<br>Or there is no valid submission.<br>*And If there is any evidence of poor academic practice then this will be mentioned in any feedback.* |
| PG: Fail<br>UG: Fail | 20-40 | E | E | | |
| | 0-20 | | | | |

Last modified: Friday, 11 October 2019, 5:41 PM

◄ Lecture Capture

Jump to...

Submission: Plan & progress update ►