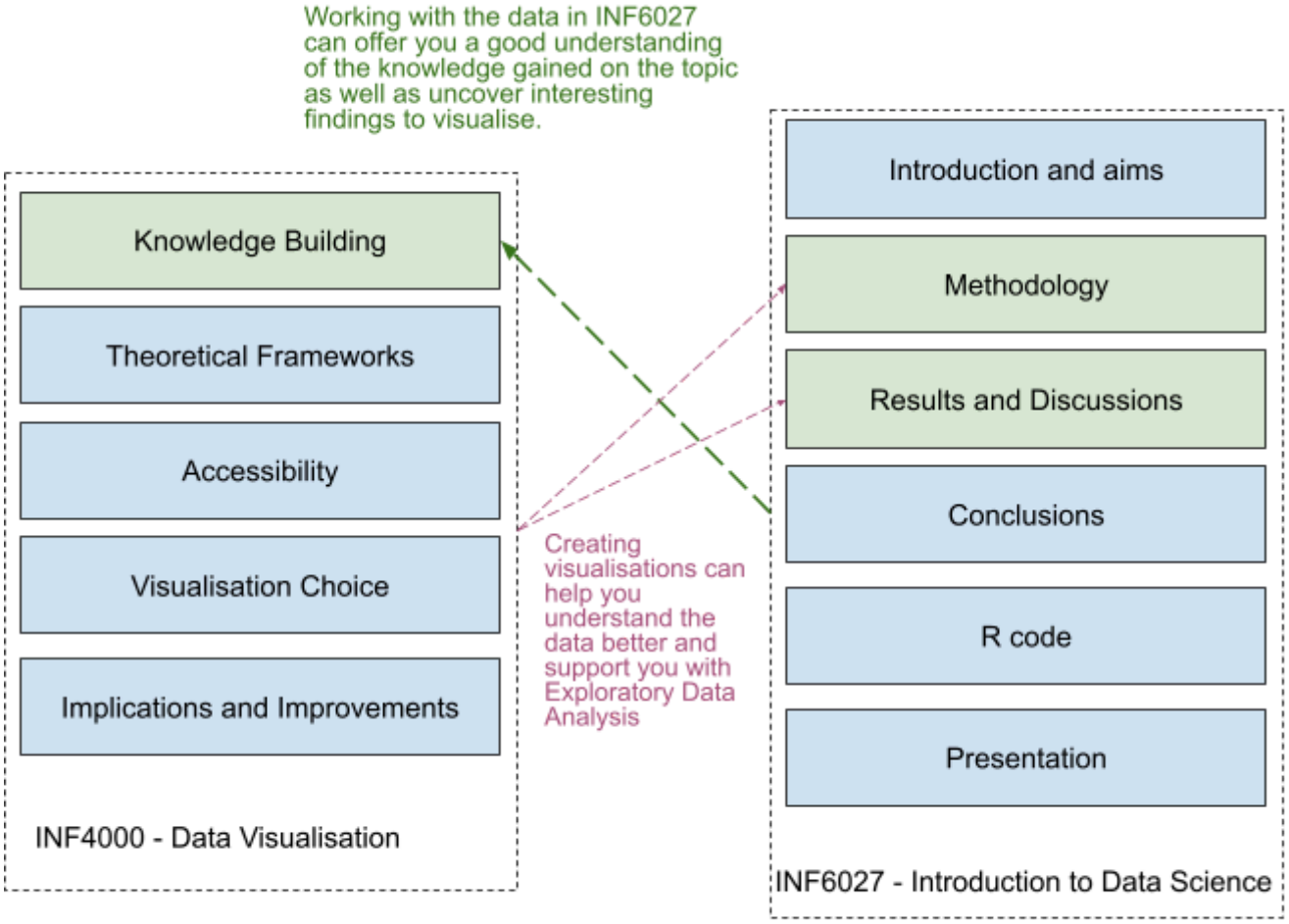
INF6027 – Introduction to Data Science
Coursework Brief (100% of the module credits)

Please note: This coursework involves using the same dataset as your INF4000 assignment.

You must use the same dataset so that you can spend more effort in thoroughly understanding your data better. There are, therefore, two primary areas in the INF4000 and INF6027 assignment where you can benefit from using the same dataset.



The primary purpose of the INF4000 module is to **evaluate your ability to effectively present interesting findings using visualisations, your understanding of theories for constructing them, your rationale behind their design, your skill in critiquing them with real-world examples, and how they enhance topic comprehension**.

The primary purpose of the INF6027 module is to assess how you identify a problem based on a given dataset, and then conceptualise, design, and implement a data science project. We expect visualisations to be created in this module, particularly to help you in exploring the selected dataset(s) and presenting results of your analysis/models.

You can therefore use the same or similar visualisations for the two modules, but they need to be differently contextualised, positioned and discussed.

Introduction

This assessment for INF6027 Introduction to Data Science comprises a piece of individual coursework to assess your ability to analyse data using R/RStudio and to then communicate your findings. Given a specific topic and dataset (see Section 2), you should identify a specific problem or topic you would like to investigate. You will then need to pre-process and analyse the dataset to identify patterns and relationships that address your selected problem/topic. This should involve using techniques learned throughout the practical sessions that will help you to demonstrate your R skills in conducting data science.

This coursework aims to follow the stages involved in a 'typical' data science process:

1) define the question(s) to address (note, sometimes this does not come at the start of the process, but after initial exploration of the data);
2) gather data;
3) transform, clean and structure the data;
4) explore and analyse the data; and
5) communicate the findings of the data analysis.

This often occurs in an iterative manner and centred on one or multiple questions you are seeking to address. For example, the data discovery process in Figure 1 presents an example of the stages involved in data discovery as an iterative process and you can find more details in Section 3. This is also similar to the data science process from the "Doing Data Science" book (O'Neil & Schutt, 2013).

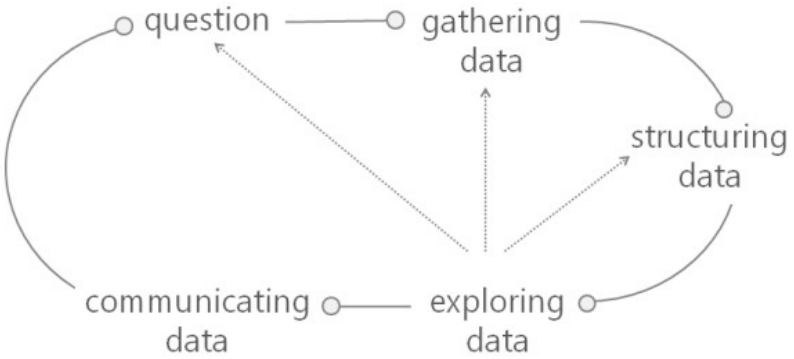


Fig.1 Example data discovery process (Johnes, 2014:p2)

You should write a **2,800 word** structured report (see Section 4) that describes the approach you have taken to explore and analyse the data for the selected problem/topic. Your report should clearly communicate the results of your data analysis and be written in a way that helps the reader interpret your findings. Note: charts, tables, and appendices are not included in the word count.

This assessment is worth 100% of the overall module mark for INF6027. A pass mark of 50 is required to pass the module as a whole. **Submission deadline: 2pm Monday 16th January 2025** via Turnitin. See Section 5 for more general information about Coursework Submission Requirements within the Information School.

Dataset options

There are a number of datasets you can choose from for this coursework. You must:

- Choose **one primary dataset** to analyse in your coursework, although some datasets may contain multiple files that you need to link and integrate.
- You can, if you choose to, combine multiple datasets (strictly within this list) and perform some data analysis. However, your focus of the study should be the primary dataset.
- There are multiple datasets on each topic, spanning over different time periods and each dataset has different characteristics that could be studied. You would likely need to join multiple files from one of the dataset or join multiple datasets.

You must not use datasets outside the ones provided by us. This is because the use of a dataset will need ethics approval, which requires more time than we have for this assessment.
The list of datasets are in **Appendix A**

What you need to do

The following sections describe what you need to do in order to carry out the coursework. This roughly follows the steps shown in Fig. 1, but you don't have to be constrained by this or follow them in this particular order; it is just a suggestion. Also, all the R we have done in the practical sessions should be enough to conduct the coursework, although you may need to investigate certain areas further that relate specifically to the problem you tackle in your investigation.

a) Review the literature and identify research question(s)

As mentioned previously, you should select a specific problem/topic related to the data (the 'question' stage in Fig. 1). To decide what area to focus on you could start by undertaking a brief review of the relevant literature around the broad domain. As examples:

Football: clustering of similar players, analysis of player and team statistics, predictive modelling between player statistics and match outcome, etc.

Maneiro, R. et aol. (2019). Offensive Transitions in High-Performance Football: Differences Between UEFA Euro 2008 and UEFA Euro 2016. Front. Psychol. 10:1230

Sarmento, H. et al. (2014). Match analysis in football: a systematic review, Journal of Sports Sciences, 32:20, 1831-1843

Deprivation: comparison of different areas, correlation or predictive modelling between different indicators (e.g., are there any associations between certain indicators?), clustering of local areas based on different deprivation index, relation between deprivation and other population or socio-economic phenomenons (you may have to search for and join other datasets).

Aungkulanon, S. et al. (2017). Area-level socioeconomic deprivation and mortality differentials in Thailand: results from principal component analysis and cluster analysis. Int J Equity Health 16, 117

Salvatore, M. et al. (2021). Area deprivation, perceived neighbourhood cohesion and mental health at older ages: A cross lagged analysis of UK longitudinal data, Health & Place, Volume 67, 102470

News dataset: analysis of news articles, sentiment analysis of news over time, studying news topics and correlations between topics, comparing manually generated categories with topic modelling.

Rameshbhai, C. J., & Paulose, J. (2019). Opinion mining on newspaper headlines using SVM and NLP. International journal of electrical and computer engineering (IJECE), 9(3), 2152-2163.

Liang, H., Ganeshbabu, U., & Thorne, T. (2020). A dynamic Bayesian network approach for analysing topic-sentiment evolution. IEEE Access, 8, 54164-54174.

Stock/Share: clustering of similar stocks, sector analysis, temporal analysis of individual stocks or sectors, correlation between indicators (e.g., a particular kind of expense and income/profits).

Liu, H., Huang, S., Wang, P., Li. Z. (2021). A review of data mining methods in financial markets. Data Science in Finance and Economics, 1(4): 362-392.

Ng, K. et al. (2017). StockProF: a stock profiling framework using data mining approaches. Inf Syst E-Bus Manage 15, 139–158.

Reviewing past literature will help you understand what kinds of analyses are undertaken in your chosen domain and provide a possible source of ideas for what you could do with the datasets mentioned in Appendix A

You are highly recommended to discuss your ideas with the tutors in-class, as they may give you feedback on the feasibility of the idea and/or difficulties in finding related literature. Do not leave this too late as your tutors will receive an increased amount of queries towards the coursework deadline and this is better discussed through a chat than emails. All submissions must have research questions.

b) Download, pre-process and explore the data

As well as reviewing relevant academic literature you should also download some data as clarified above and perform an exploratory analysis (i.e. 'play' with the data), to better understand the dataset and also help you to identify a particular problem or topic you might want to focus on. This part of your investigation will include steps to pre-process and transform the data, such as cleaning up the data, dealing with missing values, standardising numeric values, etc. This may also include combining or joining the data with another dataset from the list of options (should you choose to do so). This reflects the 'gather' and 'structure' stages in Fig. 1. (Note: this part of the analysis could take a lot of time so don't underestimate how much time you will need to spend on this part of the coursework.)

c) Analyse and explore the data

As you identify a topic of interest for your analysis then you should identify the most appropriate techniques (using R and associated packages) for carrying out your analysis and exploring the data. E.g. for football, you might want to predict match performance for a player based on their statistics. This might also be an iterative process whereby you perform some analysis and then gather (or remove) more data. Where possible relate your analysis to the relevant literature. This relates to the 'exploring data' stage in Fig. 3.

Note that this is often an iterative process: as you explore the data you may end up re-designing your research questions, having to gather more data or having to perform further cleaning as more data quality issues arise. Again, this is all a part of the data discovery process.

d) Write up your findings

Once you have performed analysis on the data and have some results then you need to write up your investigation into a report (this is the 'communicate' stage of Fig. 1). The report should be structured as outlined in Section 4. Writing up will need to be done for two sets of readers: (i) the report will need to present your findings as would be expected from a research paper; (ii) a set of GitHub pages where you present yourself and your project to a prospective employer/client.

A research audience: You will be evaluated on your ability to plan and undertake data analysis and exploration of the problem based on your chosen dataset, your ability to engage with the relevant

literature, your use of R (and appropriate packages) and RStudio to process and analyse the data, and the way in which you communicate your findings within the report for your given problem/topic.

A prospective client/employer: You must also provide your R code, together with a summary of your key findings on GitHub. The code must be commented, appropriately indented, using variable names that are appropriate. You should provide sufficient information on the code so that someone else can follow what you have done. The code should also be consistent (i.e. same standards across all code files).

The GitHub pages should be organised as follows:

- Your own profile page, with your interests and professional skills
- The INF6027 project page (either within your profile page or linked from your profile), where you present:
    - a brief introduction (3-4 lines),
    - your research questions,
    - key findings
    - The R code
    - Instructions for downloading and running the code
- The INF4000 project page (this will be detailed in the INF4000 coursework brief)

Please note: The GitHub pages, including the code must not be changed after the submission deadline. Changes past the deadline will be checked on GitHub history, and lateness penalties will be applied as usual to the mark if changes past the deadline have been made. The penalties are detailed in the section 'Information School Coursework Submission Requirements' below.

The minimum requirement to pass is to perform at least one type of data analysis (e.g., clustering, prediction, time-series analysis, etc.) and include at least two visualisations (e.g., charts, maps, etc.) that communicate the findings of your data science activity in the report. To obtain a higher mark and more effectively communicate your findings, you may decide to use more than one dataset or present more than one type of data analysis and/or use multiple visualisations and/or use multiple strategies for analysis. Again, you should also engage as much as possible with the appropriate literature.

Report structure

You are required to produce a structured report that includes the sections detailed in Table 1. You must state the word count on the first page of the report. As there is a word count limit (2,800 words) you should aim to make your writing as concise and informative as possible. Also note that your work will be assessed taking into account the word limit; therefore, we are not expecting detailed multiple analyses in the report; rather the emphasis should be on the clarity, accuracy and quality in communicating your findings.

Note: words within tables and appendices are not included in the word count. Marking will be qualitative. This means that 'box ticking' (i.e., including content that meets the listed criteria) does not necessarily lead to the full mark being awarded for a section.

Your report must have the following sections

Section	Purpose	Key aspects looked at	Max Marks
Introduction and aims	This section should describe your selected problem or topic addressed in the report and that forms the focus for your data	- Clear statement regarding the overall goal of your investigation. - Brief literature review of the	20

	analysis. This should include a (brief) summary of the literature around your selected topic. You should also state why you chose this problem/ topic and why you think it is an important topic to consider in this dataset (ideally supported by the relevant literature). You must also list your research questions	chosen topic. - Sufficient engagement with the relevant literature. - Appropriate research questions	
Methodology	This section should describe the process you have used to: (i) gather the data, (ii) pre-process and clean the data, (iii) conduct your analyses and visualise the data (note, you could follow the stages in Fig. 1). This will include ways in which you gathered, pre- processed, transformed, and sampled/ filtered the data. You should try to justify your choices and include references to relevant literature where appropriate. This should also include details of the experimental setup, e.g. which R packages you have used etc. Think of it like this, if someone else had to replicate your methodology have you provided enough details (and clearly enough) for them to reproduce your results.	- Clear description of methodology used in your analyses. - Clear list of the datasets used (and links to sources) and variables in the dataset(s). - Clear discussion of methods for pre- processing data (and appropriate use of R packages). - Examples of the data. - Range of techniques used, appropriateness, links to supporting literature etc. (e.g.,methods for trend prediction, spatial data analysis etc.). - Techniques can include types of visualisation and references to which R libraries have been used. - Methods to deal with data quality issues, such as missing values. - Clear link with how the methodology helps answer the research questions	25
Results and discussion	In this section you should present the results of your data analysis and exploration (e.g., statistics, maps, trends, predictions). You should use the results to address the selected problem by presenting and discussing tables and charts as appropriate, and link with the literature. You should present your findings in a way that helps the reader interpret the results. You should focus on effectively communicating the results of the	- Correct use of statistics and visualisations. - Packaging results etc. into tables rather than simply using R output or command line code. - Clear narrative and structure (e.g., adding sections and sub-sections and guiding the reader through the analysis). - Clearly explaining the results and graphics used (e.g., use of legends etc.). - Using graphics that convey information (e.g., combine results) and help identify insights	25

	analysis to the reader by highlighting the trends or patterns you have observed during your data analysis. You should answer your research questions explicitly	(e.g., use of log scales to dampen effects of high values etc.) - Bringing out insights rather than leaving the reader to interpret the findings. - Re-labelling the variable names in graphs and tables where appropriate to improve readability. - Narrative is not just reading out graphs, but explaining what is the meaning of the findings	
R code, GitHub pages	In this section, you should document how you have presented your GitHub resources. You must have a link to your GitHub pages, that present: (i) Your profile; (ii) Your INF6027 project page with findings (iii) Code (iv) Instructions for running the code	- Well-commented, indented code with appropriate variable names. - Clarity of presentation, accessibility of research context and findings. - Consistent style across R files - Clear instructions for running the code	10
Conclusion	In this section you should summarise the project you have done. In bullet points, you should present the main findings of your analysis and lessons learned In bullet points, you should highlight any weaknesses, limitations and assumptions of your analysis You should provide details of any future work	- Summary of the main findings of the analysis with respect to the original aim(s) of the investigation. - Appropriate set of limitations/ weaknesses of your methodology and analysis. - Clarity in the key findings of your research	15
References	List of references used for the research	Appropriate references in APA citation format	5

Information School Coursework Submission Requirements

It is the student's responsibility to ensure no aspect of their work is plagiarised or the result of other unfair means. The University's and Information School's Advice on unfair means can be found in your Student Handbook, available via http://www.sheffield.ac.uk/is/current

Your assignment has a word count limit. A deduction of 3 marks will be applied for coursework that is 10% or more **above or below** the total word count as specified above or that does not state the word count.

It is your responsibility to ensure your coursework is correctly submitted before the deadline. It is highly recommended that you submit **well before the deadline**. Upon successful submission, you will receive an automatically generated email as receipt. You must keep this receipt as evidence of your submission and for future reference. Coursework submitted after 2pm (even if by a few seconds) on the stated submission date will result in a deduction of 5% of the mark awarded for each working day after the submission date/time up to a maximum of 5 working days, where 'working day' includes Monday to Friday (excluding public holidays) and runs from 2pm to 2pm. Coursework submitted after the maximum period will receive 0 mark.

Work submitted electronically, including through Turnitin, should be reviewed to ensure it appears as you intended.

Before the submission deadline, you can submit coursework to Turnitin numerous times. Each submission will overwrite the previous submission. Only your most recent submission will be assessed. However, after the submission deadline, the coursework can only be submitted once.

If you encounter any problems during the electronic submission of your coursework, you should immediately contact the module coordinator and the Information School student support team inf-student-support@sheffield.ac.uk). This does not negate your responsibilities to submit your coursework on time and correctly.

Appendix A - List of Datasets

Topics:

1. Football

Dataset	Brief description	URL
International Football Results	An up-to-date dataset of over 47,000 international football results	https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017?select=shootouts.csv
Transfer Market Data	Clean, structured and automatically updated football (soccer) data from Transfermarkt	https://data.world/dcereijo/player-scores (Might respond with a 404) OR https://www.kaggle.com/datasets/davidcariboo/player-scores/data?select=clubs.csv
UEFA Euro 2020 Dataset	This is a collection of the Euro 2020 tournament data with multiple files including events, match information, player statistics, line-ups etc.	https://data.world/cervus/uefa-euro-2020
UEFA Euro dataset	The dataset contains all players & coaches, all matches & results, and main match events in Football/Soccer UEFA European Championship/EURO (1960-2024), and Nations League (2019-2023)	https://www.kaggle.com/datasets/piterfm/football-soccer-uefa-euro-1960-2024/data

2. Open Data published by UK Government:

Dataset	Brief description	URL
gov.uk	Data available on a range of topics such as transport, health, education, environment, crime	https://www.data.gov.uk/
Office for National Statistics	Census data (2011, 2021 etc.), providing different levels of detail on demographics, labour market, migration, housing etc.	https://www.ons.gov.uk/
Consumer Data Research Centre	A large collection of datasets on a range of topics such as	https://data.cdrc.ac.uk/search/type/dataset

	population and mobility, retail futures, mobility, finance and digital	

3. News

Dataset	Brief description	URL
News Category Dataset	This dataset contains around 210k news headlines from 2012 to 2022 from HuffPost. This is one of the biggest news datasets and can serve as a benchmark for a variety of computational linguistic tasks.	https://www.kaggle.com/datasets/rmisra/news-category-dataset?resource=download
The Multilabled News Dataset	This dataset contains 10,917 news articles with hierarchical news categories collected between January 1st 2019, and December 31st 2019 classified by using NewsCodes Media Topic taxonomy.	https://zenodo.org/record/7394851

4. Music and Films

Dataset	Brief description	URL
Spotify Tracks Dataset	This dataset contains around 210k news headlines from 2012 to 2022 from HuffPost. This is one of the biggest news datasets and can serve as a benchmark for a variety of computational linguistic tasks.	https://hf-proxy-cf.effarig.site/datasets/maharshipandya/spotify-tracks-dataset
MusicOSet An Enhanced Music Dataset for Music Data Mining	An open and enhanced dataset of musical elements (music, albums, and artists) suitable for music data mining. The attractive features of MusicOSet include the enrichment of existing metadata to which it is linked and the popularity classification of the musical elements present in the dataset.	https://marianaossilva.github.io/DSW2019/index.html
IMDB Dataset	The Internet Movie Database data is a large and comprehensive collection of information related to movies, TV shows, video games and audience ratings.	https://developer.imdb.com/non-commercial-datasets/

5. Stock Market

Dataset	Brief description	URL
Stock Market Data (NASDAQ, NYSE, S&P500)	Date, Volume, High, Low, and Closing Price (for all NASDAQ, S&P500, and NYSE listed companies). Updated weekly	https://www.kaggle.com/datasets/paultimothymooney/stock-market-data/data
200+ Financial Indicators of US stocks (2014-2018)	Collection of financial indicators of US stocks/shares over the period of five years: 2014-2018. There are five CSV files, each containing data of a particular year. For each stock, more than 200 financial indicators are recorded, such as sector, revenue, revenue growth, gross profie, R&D expenses, etc.	https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018