**INF4000 Data Visualisation Coursework Brief**

Please note: This coursework involves using the same dataset as your INF6027 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.
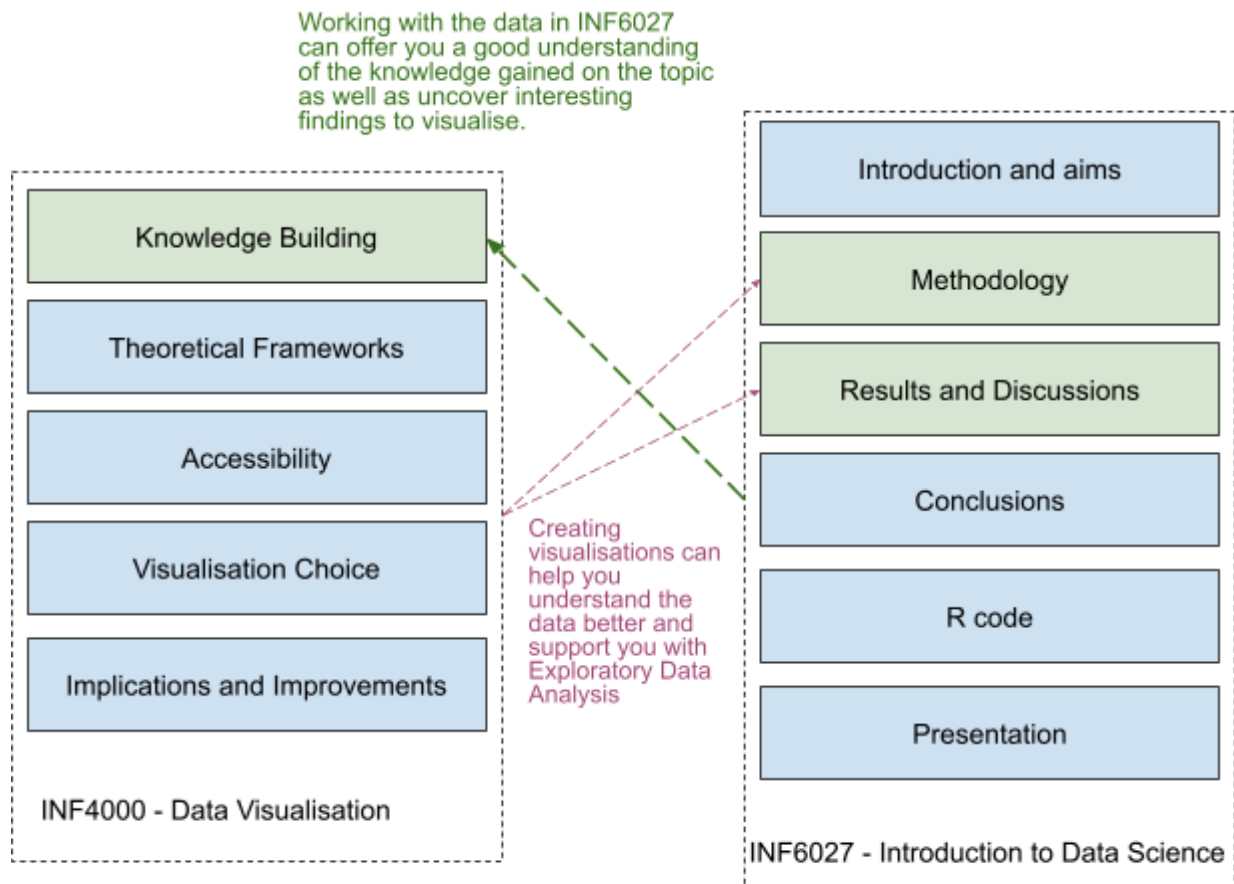
The primary purpose of the INF4000 module is to **evaluate your ability to effectively present interesting findings using visualisations, your understanding of theories for constructing them, your rationale behind their design, your skill in critiquing them with real-world examples, and how they enhance 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.

INF4000 Data Visualisation Coursework Brief

1. General information

The assessment of the module tests not only your ability to create informative visualisation but also the knowledge you have gained throughout the lectures on important visualisation topics and how much you have engaged with the visualisation literature. The assessment will consist of a document (written in Word or similar document writing software) consisting of a composite visualisation and related reflections. The reflections need to be your individual take on the topics discussed in the seminars and lectures. Those reflections need to be put in a wider context using relevant literature sources. **Keep in mind that visualisations that are not your work cannot be used in this assessment.**

The assessment should be **3,000 word** long, divided into five sections of approximate equal length. This assessment will account for 100% of the mark of INF4000 Data visualisation. A mark of 50 is required to pass the module. The deadline for the assessment is <mark>**21st January 2025 at 2:00pm**</mark> via the Turnitin link found in the Blackboard page of module INF4000 Data visualisation.

2. Report format

In the report, you will include the composite visualisation right after the cover page, followed by five sections.

- The cover page should contain module code (i.e., INF4000), word count, your registration number and your topic.

- The composite visualisation should contain at least **4 charts**, which belong to the **same topic**. Each chat contributes **extra information** that you want to communicate.

- The five sections are (1) knowledge building, (2) theoretical frameworks, (3) accessibility, (4) visualisation choice and (5) implications and improvements. The details of each section are given below.

You are recommended to upload your R code, the composite visualisation, along with descriptions and key messages, to GitHub. While this exercise is not graded, it will help you develop skills in presenting yourself and your experience to potential employers or clients. The code should be well-commented, properly indented, and use clear, appropriate variable names, following the standards outlined in the INF6027 coursework brief. The INF4000 project page (either hosted within your profile or linked from it) should include:

- a brief introduction (3-4 lines), including the question(s) you hope to answer using the charts and key messages
- the composite visualisation
- the R code

3. Dataset

You should use the same dataset(s) you choose for INF6027.

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.

INF4000 Data Visualisation Coursework Brief

- 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.

The list of datasets are in **Appendix A**

4. Sections of the report

You should consider all sections equally important and devote around 600 words to each section, for a total of 3,000 words. You need to cover:

- the entire composite visualisation (4 charts) in Section 1.

- Sections 2 – 5 should each mainly focus on **different charts** in the composite visualisation (**one chart per section**). You should choose which chart to discuss in which section, but you have to state it clearly in the sections. Each of the sections should focus on only one visualisation.

The five sections are:

1. **Knowledge building** – Describe how the composite visualisation generates new insights to address a specific question. The question doesn't have to be a research question from INF6027; it could be a component or a side question. The question could also come up as a sub-topic of interest as you explore the dataset for either of the assignments. The key is to ensure that the question is not too broad and needs multiple visualisations to answer it, as well as how you build your knowledge.

   Start by briefly explaining the question and its significance in relation to existing literature on the topic. You should explain the detailed knowledge that each chart contributes to the chosen question. It may be helpful to think about the 'A' in ASSERT to clarify what question the visualisation aims to answer. Additionally, link your discussion of the visualisations to relevant literature, highlighting where the cases might agree or disagree with your analysis.

Examples of what we will be looking for:
   - Clearly state the topic and explain why the topic is important
   - Explain what the visualisation is showing and what new knowledge it provides about the topic
   - Appropriate literature sources
   - Sufficient discussion

2. **Theoretical frameworks** – Describe your visualisation and the way you created it using theoretical frameworks. You will need to refer to both the ASSERT framework and the grammar of graphics. You will need to explain how you followed each stage of the ASSERT framework to create the visualisation and explain the elements of your visualisation using the grammar of graphics (GoG) elements (e.g. geometries, aesthetics, coordinate systems). You will need to explain why you choose these GoG elements with references to literature.

Examples of what we will be looking for:
   - Clearly state the question to answer

- Detailed description for each of the six stages in the ASSERT framework
- Discussion using grammar of graphics (multiple occasions), specifically explaining the different elements of grammar of graphics
- Appropriate literature sources

3. **Accessibility** – Describe what accessibility means in visualisation and whether the visualisation you developed is accessible. You will also need to critique your visualisation with regards to accessibility, discussing whether it is accessible and elaborate on what design choices you made helped or hindered accessibility.

Examples of what we will be looking for:
- Describe what accessibility means in visualisation
- Discussion on whether the visualisation is accessible and what design choices you made helped or hindered accessibility
- Appropriate literature sources
- Sufficient discussion

4. **Visualisation choice** – Describe and justify your choice of visualisation type based on the goal of the visualisation. You will need to describe and justify why you chose the type of charts and discuss possible alternatives. You will need to emphasise the positive and negative aspects of each type of visualisation mentioned and show an awareness of when each visualisation is appropriate based on the data shown and the goal of the visualisation. You can refer to visualisation taxonomies to support your discussion.

Examples of what we will be looking for:
- Justify why you chose the type of visualisation
- Discuss possible alternatives (at least two)
- Appropriate literature sources
- Sufficient discussion related to the goal of the visualisation

5. **Implications and Improvements** – Describe the ethical implications of using the visualisation in the topic you identified. You could discuss how visualisations could be used to (mis)inform the public, or arrive at (in)accurate conclusions. You could also discuss relevant (preferably on your topic or using the same dataset) examples where this has happened in the past, with links to the literature. You should propose changes to justify how the visualisation could be improved. You do not need to provide an improved visualisation, but should provide ideas on datasets, visual design, accessibility.

Examples of what we will be looking for:
- Discussion and reflection on visualisations in the topic you choose
- References to examples or discussions (news, literature)
- Propose improvements on the visualisation you created

Each section should demonstrate that you engaged with relevant literature. That means you are expected to have citations to relevant articles in the literature of visualisation that support or provide context to your reflections in each section. Each section of the assessment will be explained and covered throughout the module in lectures and seminars. Engaging in these will help you produce a good assessment.

5. Information School Coursework Submission Requirements

Information School Student Handbook > PGT Assessment > <u>Submitting your work</u>

Last submission penalties: 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. Coursework submitted after 2pm 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 zero marks. 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.

Word count penalties: Your assignment has a 3,000 word limit. A deduction of 3 marks will be applied for coursework that is 10% or more above or below the word count as specified above or that does not state the word count.

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
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 population and mobility, retail futures, mobility, finance and digital	https://data.cdrc.ac.uk/search/type/dataset

INF4000 Data Visualisation Coursework Brief

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

INF4000 Data Visualisation Coursework Brief

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 profit, R&D expenses, etc.	https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018