Postgraduate coursework, Information School
INF6027 Introduction to Data Science (2020-21)

Analysis of the UK Police Dataset

1. Introduction

This part of the 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 (e.g., where or when particular types of crime occur, or co-occur). 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, such as summarising datasets, statistical modelling or data visualisation, to highlight and illustrate particular aspects of the data you want to communicate (e.g., particular patterns or trends).

This coursework aims to follow the stages involved in a 'typical' data science process: (i) 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); (ii) gather data; (iii) transform, clean and structure the data; (iv) explore and analyse the data; and (v) 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[1] and you can find more details in Section 3. This is also similar to the data science process we have been using in class from the "Doing Data Science" book (O'Neil & Schutt, 2013).

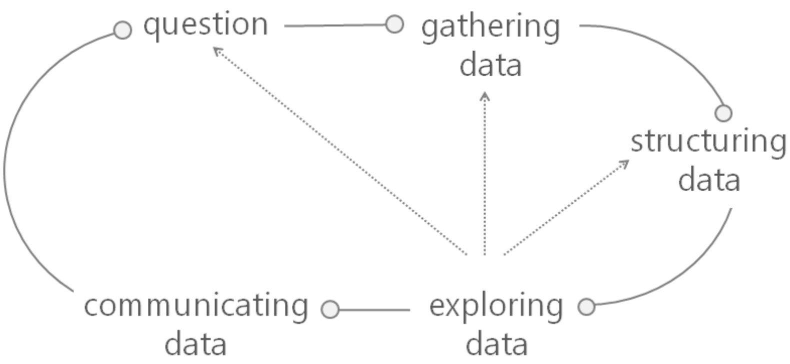

Fig. 1 Example data discovery process (Jones, 2014: p.2)

You should **write a 3,000 word structured report** (see Section 4) that describes the approach you have taken to explore and analyse the data for the selected problem/topic. You 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: 10am Monday 18th January 2021 via Turnitin.** See Section 5 for more general information about Coursework Submission Requirements within the Information School.

2. The UK Police Dataset

The dataset to be used in this assessment is the UK Police Dataset, which has made public crime data since 2011 (this is an example of Open Data). There has been a lot of recent interest in analysing publicly available datasets to identify patterns of crime and gain insights into criminal activity, see for example the crime activity browser by IBM[2]. If interested in the topic you can also find further crime-related datasets produced by the UK Data Service (https://www.ukdataservice.ac.uk/get-data/themes/crime). There is also an increasing use of crime Open Data used in the media to highlight aspects of policing and criminal activities (see, e.g., https://www.bbc.co.uk/news/uk-44044537).

[1] You can find out more about this process in (Jones, 2014: p.2): https://tanthiamhuat.files.wordpress.com/2015/07/communicating-data-with-tableau.pdf

[2] Open Crime Data, Free for All: https://developer.ibm.com/clouddataservices/2016/11/03/open-crime-data/

A description of the data is available here: http://data.police.uk/about/ also including an explanation on how to download the data[3]. The data are provided as CSV files (note that there is also an API available if you prefer) and provide street-level crime, outcome and stop and search information broken down by police force[4] (in the UK there are 45 territorial police forces and 3 Special Forces) and 2011 Lower Layer Super Output Areas (LSOA).

The dataset describes crimes reported to UK police during each month in different areas of the UK. Information in the dataset includes the following: geographical location (longitude and latitude), date (month, year), LSOA code (i.e., the census area), and type of crime (e.g., vehicle crime, burglary, robbery, etc.). You can select any data from the UK Police Dataset. (This may require multiple downloads.) You can also aggregate the dataset with other data sources if you want (e.g., census data), which would demonstrate your ability to join datasets (although you don't have to do this to pass the coursework as the emphasis of the coursework is on how you carry out your analysis in R/RStudio and communicate your findings on the UK Police Dataset).

3. 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 (and the final 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.

3.1. 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 areas, such as analysis of crime data, geographical analysis of crime, predictive policing, crime sensing, analysis of crime statistics, etc. For example, these articles may be a useful starting point:

> Vandeviver, C., and Bernasco, W. (2017) The geography of crime and crime control, *Applied Geography*, Volume 86, pp. 220-225. (Available online: http://www.sciencedirect.com/science/article/pii/S014362281730838X)
>
> Field, S. (1992) The Effect of Temperature on Crime, *The British Journal of Criminology*, Volume 32, Issue 3, pp. 340–351. (Available online: https://doi.org/10.1093/oxfordjournals.bjc.a048222)

Reviewing past literature will help you understand what kinds of analyses are typically undertaken using crime data and provide a possible source of ideas for what you could do with the UK Police Dataset. Examples of possible topics include, but are not restricted to, the following:

- Evolution of crimes in an area over time;
- Trends and predictions of crimes and crime rates;
- Analysis of certain types of crime (e.g., vehicle crimes);
- Comparisons of crime types in a region;
- Clustering and classification of data, e.g., by type of crime;
- Normalisation and integration with other datasets (e.g., LSOA census statistics);
- Focus on a certain census dimension (e.g., age of residents in the area);
- Visualisation of the data (e.g., on maps).

3.2. Download, pre-process and explore the data

As well as reviewing relevant academic literature you should also download some data from http://data.police.uk/ 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. **You must include most recent data in your analysis.**

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 further datasets, e.g. census or deprivation data. 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.)

[3] You can also find an article describing the accuracy of the data here: https://www.tandfonline.com/doi/full/10.1080/15230406.2014.972456

[4] https://en.wikipedia.org/wiki/List_of_police_forces_of_the_United_Kingdom

3.3. 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. you might want to predict crime rates using regression or compare levels of crime types using statistical tests. This might also be an iterative process whereby you perform some analysis and then gather (or remove) more data. Where possible relate you 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.

3.4. 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. You will be evaluated on your ability to plan and undertake data analysis and exploration of crime based on the UK Police 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.

You should also provide your R code as an appendix and marks will be awarded for your clarity, consistency and way in which you comment your R code (see, e.g. http://stat405.had.co.nz/r-style.html). The specific style you use is not as important as how well you comment your code so that someone else can follow what you have done and being consistent in whichever style you adopt.

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.) 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. Again, you should also engage as much as possible with the appropriate literature.

4. 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 (3,000 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 that words within tables and appendices are not included in the word count.

Table 1: Required content of the structured report.

Section	**Description**	**Examples of what we will be looking for and mark allocation**	**Maximum allocated marks**
Structured abstract	This should provide a summary of your report in a structured manner, e.g. objective, methods, results, conclusions. *This is not included in the word count.*	• Brief but informative abstract that is clearly structured.	Required, but 0 marks
Table of contents	This should include section titles and page numbers. *This is not included in the word count.*	• Clearly structured Table of Contents with use of numbering for sections.	Required, but 0 marks
Introduction and aim(s)	This section should describe your selected problem or topic addressed in the report and that forms the focus for your data analysis. This should include a (brief) summary of the literature around analysis of crime data relevant to your selected topic that helps to provide the background to your chosen 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 support by the relevant literature)	• Clear statement regarding the overall goal of your investigation. • Brief literature review of data and crime analysis. • More marks for engagement with the relevant literature.	10 marks
Methodology	This section should describe the process you have used to gather the data, pre-process and clean the data, 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.	• Expect to see a 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-	20 marks

	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. As well as describing the methodology used to generate your results, you should list all the UK Police datasets used (e.g., data covering different regions or time periods). You should also list any additional external datasets used (e.g., shape files or census statistics for LSOA areas). Describe all datasets used, any pre-processing and how they were joined together (e.g., over LSOA area identifiers).	processing data (and appropriate use of R packages). • More marks for examples of the data. • More marks for multiple data sources used. • More marks for the 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 • More marks for the detail of the description provided, e.g., could include use of group_by(), aggregate() etc. • More marks for use of methods to deal with data quality issues, such as missing values. • More marks for discussing use of appropriate techniques for different types of data, e.g. categorical data.	
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. 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 analysis to the reader by highlighting the trends or patterns you have observed during your data analysis.	• More marks for correct use of statistics and visualisations. • More marks for packaging results etc. into tables rather than simply using R output or command line code. • More marks for a clear narrative and structure (e.g., adding sections and sub-sections and guiding the reader through the analysis). • More marks for clearly explaining the results and graphics used (e.g., use of legends etc.). • More marks for using graphics that convey information (e.g., combine results) and help identify insights (e.g., use of log scales to dampen effects of high values etc.). • More marks for bringing out insights rather than leaving the reader to interpret the findings. • More marks for not over-interpreting the results and recognising biases. • More marks for re-labelling the variable names in graphs and tables (rather than using default names). • More marks for how well the data is summarised and made accessible for comparison.	50 marks
Conclusion	In this section you should summarise the main findings of your analysis and lessons learned. You should state the main message the reader should come away with from your analysis. You should also highlight any weaknesses of your analysis and state what you would do to improve your analysis if you had more time.	• Summary of the main findings of the analysis with respect to the original aim(s) of the investigation. • More marks for highlighting limitations/ weaknesses of your methodology and analysis. • More marks for a clear set of take-away messages.	10 marks
R code	You should include the full R code as an appendix.	• More marks for well-commented code. • More marks for clarity of presentation. • More marks for consistent style.	5 marks
Presentation	The overall presentation of the report will be given a separate mark, including how well you have presented your results, clarity of writing and use of literature.	• More marks for use of appropriate references. • More marks for clarity of writing	5 marks

		- More marks for use of appropriate charts and tables and their presentation quality.	

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 5% or more above or below the 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. Coursework submitted after 10am 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 10am to 10am. 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.

During your first Semester at the School, when submitting a piece of work through Turnitin, you will only be able to view a 'similarity report' when submitting your Test Essay. You can then edit and resubmit your Test Essay. For other coursework you will not be able to view a Turnitin 'similarity report'. Details about the submission of work via Turnitin can be found at: http://youtu.be/C_wO9vHHheo

If you encounter any problems during the electronic submission of your coursework, you should immediately contact the module coordinator and one of the Information School Exams Secretaries (Julie Wharton, J.Wharton@sheffield.ac.uk, 0114 2222839 or Corrie Houton, c.houton@sheffield.ac.uk, 0114 2222640). This does not negate your responsibilities to submit your coursework on time and correctly.