MATH513 Big Data and Social Network Visualization Coursework

Academic Year 2024-25

1 Coursework Information

Please read the following points before attempting the coursework:

- The deadline for this assignment is **2pm** on **Friday, December 13th, 2024**. You should submit your work through the MATH513 Big Data and Social Network Visualization DLE site. Your submission will be marked anonymously.
- **This is a group coursework. Please work in self-assigned groups of four people.**
- Along with this assignment, you need to submit the **minutes of 4 of the meetings you have held to produce the coursework**, where each time a different member of the group is chair and minute-taker. Note that 5 marks (out of 100) are allocated to the submissions of the minutes.
- You should keep notes of all your meetings. Each member of the group will receive the same mark, unless any member chooses to make use of the Peer Assessment option. If you wish to make use of the Peer Assessment option, you will need to notify the Module Leader Dr Malgorzata Wojtys by 12th December, 2024.
- This assignment counts for 100% of your final mark on this module. Marks will be assigned according to the marking grid on page 6. Please note that your work will be assessed as a whole and you will not be given separate marks for every sub-question.
- Please note that 5 marks (out of 100) are allocated to the submission of the Ground Rules Contract for your group by the **19th November, 2024**, as discussed in class and as explained on the DLE.
- Marked scripts will be returned within **20 working days** of the submission date. In particular, you will get full feedback on your work by January 17th, 2025.

- You are reminded of the **University's Academic Regulations**:

  Academic offences occur when activity is undertaken which could confer an unfair advantage to any candidate(s) in assessment. The University recognises the following (including any attempt to carry out the actions described) as academic offences, regardless of intent:

  a. Plagiarism, which is copying or paraphrasing of other people's work or ideas into a submitted assessment without full acknowledgement. More information on plagiarism is available here:
  https://www.plymouth.ac.uk/student-life/your-studies/essential-information/regulations/plagiarism

  b. Collusion, which is unauthorised collaboration of students (or others) in producing a submitted assessment. The offence of collusion occurs if a student copies any part of another student's work, or allows their own work to be copied. Collusion also occurs if other people contribute significantly to work that a student submits as their own.

  The complete list of regulations can be found here:
  https://www.plymouth.ac.uk/student-life/your-studies/essential-information/regulations

  Use of AI

  AI tools such as ChatGPT must not be used to produce commentary or interpretations to the numerical or graphical results. You may use AI tools for tasks such as proof reading and improving the structure of the report. If you are in any doubt, you should seek clarification from the module lecturers.

  By submitting this coursework, all group members confirm that they have understood the University's policy on plagiarism and collusion.

We now state the relevant MATH513 Big Data and Social Network Visualization Assessed Learning Outcomes (ALOs) for this assignment.

At the end of the module the learner will be expected to be able to:

ALO1 Critically select and use a broad range of techniques to perform Big Data manipulation and visualization;

ALO2 Perform exploratory analyses to extract information, insight and innovation from data;

ALO3 Collaborate with others to produce and document **R** code and to present its professional use for Big Data or Social Network Visualization.

You should keep these ALOs in mind when doing this coursework.

2 Questions

This coursework comprises of two Questions, with equal contributions into the mark. You need to produce a report of your work following the instructions below. Please note that one report is required. Your single report should contain a description of your work for both tasks including your commentary, graphs and conclusions. In your pdf file, you should present code snippets, however **the code for producing similar graphs should not be displayed repeatedly in the pdf report**.

Moreover, your submission should contain an `Rmd` file with well presented and annotated **R** code for all of your analyses.

The **page limit** for the pdf report is **30 pages**. Please note that this is the upper limit and not the target. Please do not submit an additional appendix as it will not be considered. Reports that contain irrelevant or vague comments will be penalized. It is not necessary to repeat figures or code that are very similar.

Question 1: Customer satisfaction survey

A large internet provider wishes to analyse the satisfaction levels of its customers. A survey was performed for 140 business customers and 515 individual customers where customers were asked how satisfied they are with their internet speed and with the company's Customer Service. The data are collected in two files `business_customer_survey.xlsx` and `individual_customer_survey.xlsx` available on the module's DLE website. The files include the following variables:

- `term`: the length of a contract (in months),
- `bill`: the monthly bill (in pounds)
- `cs_satisf`: satisfaction with Customer Service,
- `speed_satisf`: satisfaction with internet speed.

(a) Write an `R` code to create a single dataset based on the two files provided. The newly created dataset should include a new variable indicating the type of the customer:

- `type`: the type of the customer ("business" or "individual").

(b) Briefly summarise and describe the data collected for each variable using numerical and graphical tools that are appropriate for its type. Comment on the findings.

(c) Explore how customer satisfaction depends on the type of customer. Use appropriate numerical summaries and visualisations that you learnt on this module. Interpret the findings.

(d) Explore the relationship between the bill and the length of contract in the two groups of customers. If appropriate, use a linear regression model. What can you learn from the results?

(e) Is the average bill the same for the two types of customers (business and individual)? Perform a t-test and interpret its result.

Question 2: Business news articles

The file `Business_articles.csv` available on the module's DLE website contains news articles from 2015 to 2017 related to business from all over the world.

The dataset includes the following variables:

- `Article`: the text of a news article, which also contains the place where the article was published;
- `Heading`: the heading/title of a news article;
- `Date`: the date when an article was published.

You will compute and present some numerical and graphical summaries of this data set. When analysing free text data, make sure that you perform any necessary text preprocessing and data cleansing (such as tokenization, stop words removal etc.).

(a) Write an `R` code to create a new variable indicating the location of the article. Hint: as a part of your code, you could use the function `regexpr()` that looks for a specific character in a text and returns its position. For example, the command `regexpr(":", "strong>PARIS: Militant")[1]` returns the value 13.

(b) Briefly summarise and describe the obtained locations using suitable numerical and graphical summaries and appropriate comments.

(c) Explore and summarise the length of articles (i.e. the number of words per article) using suitable numerical and graphical summaries and appropriate comments.

(d) Explore the most popular words in the articles and article headings. Consider how they change in time (for example, by year). Provide suitable visualisations and appropriate comments.

(e) Suppose that we would like to track how the usage of the word "trading" in articles changes in time. Produce an appropriate visualisation and provide a commentary.

(f) Perform sentiment analysis of the news articles. Interpret the results.

3 What You Need to Submit

One member of your group needs to submit the following files electronically on the DLE website.

- A Portable Document Format (pdf) file containing your report produced by RMarkdown `Report_First.Second.Third.Fourth.StudentID.pdf`,
  where you substitute in the Student Identification Numbers of all the group members. For example, `Report_11034023.12045043.12830176.13643987.pdf`.
- The RMarkdown file that produces your report
  `Report_First.Second.Third.Fourth.StudentID.Rmd`
  where you substitute in the Student Identification Numbers of all the group members. For example, `Report_11034023.12045043.12830176.13643987.Rmd`.
- The minutes of four of your meetings, named similarly as explained above. For example,
  - `Minutes1_11034023.12045043.12830176.13643987.pdf`
  - `Minutes2_11034023.12045043.12830176.13643987.pdf`
  - `Minutes3_11034023.12045043.12830176.13643987.pdf`
  - `Minutes4_11034023.12045043.12830176.13643987.pdf`

If anything is unclear, you should ask **without delay**.

4 Marking Grid

MATH513 Big Data and Social Network Visualization: Coursework Marking Grid

Mark Band	**R code** (40%)	**Analysis** (40%)	**Report Style** (20%)
Above 80	Exceptionally well written, correct, very clear, tidy and very well commented.	Correct choice of tools. Broad, highly insightful and critically reflective discussion. Very well justified conclusions. Almost no technical errors.	Exceptionally well structured and exceptionally well written report, with outstanding figures. Almost no presentational or grammatical errors in the text. Well formatted references used where needed.
70 to 80	Very well written, correct, clear, tidy and well commented on.	Correct choice of tools. Insightful explanations of concepts and interpretations of results. Well justified conclusions. Very few minor technical errors.	Very well written and very well structured report, with very good figures. Minor grammatical or presentational errors in the text.
60 to 70	Mostly correct, tidy, commented on, a few minor errors permitted.	Correct choice of tools. Some critical and insightful commentary, but perhaps not so deep or lacking detail in some places. Well justified conclusions, with some limitations. Some less minor technical errors permitted.	Well written and well structured report, with good figures. Some less minor grammatical or presentational errors in the text.
50 to 60	Generally correct, may contain some serious errors.	Mostly correct choice of tools. Some discussion, but lacking insight or critical reflection. Limited or poorly justified conclusions. Generally correct critical understanding of the methods. Some more serious technical errors present.	A report with logical structure, mainly correct English and some good figures. Some more serious grammatical or presentational errors in the text. Some spurious or unnecessary R output included in the report.
Below 50	Mostly incorrect, contains many major errors.	Incorrect choice of tools. Poor analyses and muddled discussion. Unclear or very limited conclusions. Many technical errors.	A report with poor structure, poor English or badly produced figures. Many grammatical or presentational errors in the text. A lot of spurious or unnecessary R output included in the report.