Faculty of Computing, Engineering and Science

Assessment Brief

Module Title: Big Data and Analytics

Module Code: IS4S765

Module Leader/Tutor: Paul Jarvis/Rob Berry

Assessment Type: Asynchronous Assessment

Assessment Title: Querying Big Data Sources

Weighting: 50%

Word count/duration/equivalent: 2000 words (20% either side)

Submission Date: 23.59 on Friday 5th May 2024

Return Date: Friday 7th June 2024

## Assessment Description

**Your Task**

For this assignment, you will conduct analysis of one or more data sources in order to answer two research questions using two different analytics techniques that you have chosen. You must write a report that describes the entire process that you followed, detailing the analysis conducted and the conclusions that you have drawn.

The report should be 2000 words in length (within 20% either side). You should work individually on this assessment; it is not team-based.

**Generating research questions**

You should create two specific research questions and make use of R to perform two different analytics techniques that were used in class to answer the questions. You should take your word count limit into account, ensuring that you will be able to explain your analysis and findings clearly.

**Choosing data sources**

You may use a single source for each question, or different sources. The source(s) you use must be different from those used for Coursework 1. Ensure that whichever source(s) you use will be appropriate to the analysis that you wish to conduct. Some potentially suitable datasets can be found at the following repository, but there are many more available across the web if you search appropriately:

* A collection of over 1,300 datasets originally distributed in R packages, curated by Vincent Arel-Bundock from the University of Montreal: <https://vincentarelbundock.github.io/Rdatasets/datasets.html>
* Kaggle: <https://www.kaggle.com/datasets>.
* Google Datasets search: <https://datasetsearch.research.google.com/>

The maximum permitted file size of your data is **100MB**. Please be aware that code will be scrutinised closely to check for evidence of collaboration. You are required to submit both the data and the R code file (.R) – see Submission Details for more information.

**Analytics techniques**

You should demonstrate your ability to use two analytics techniques that have been covered in class, choosing appropriate methods to query relevant datasets. The marking scheme awards credit for the complexity of the methods that you employ. You should ensure that you justify your choices, appraising and contrasting the strategies used, using high quality sources where possible to support you.

You must choose two of the following four techniques:

* Classification Trees
* Linear Regression
* Association Rules
* Cluster Analysis

**Results and Conclusions**

You should present your results clearly and explain how you have drawn conclusions from them.

**The Report**

The report should be structured as follows:

*Introduction*: A short section that sets out the content of the report. You must specify which techniques have been chosen and the nature of the research questions to be tackled.

*Technique 1:*

*Data Source*: A description of the data source analysed using your first chosen technique and a concise, clear statement of its associated research question(s).

*Methods*: A section describing the application of the first technique, with justification or explanation where necessary.

*Results*: A section that presents results from the analysis conducted using the first technique.

*Conclusion*: A conclusion that answers your research question(s) using your first technique.

*Technique 2:*

*Data Source*: A description of the data source analysed using your second chosen technique and a concise, clear statement of its associated research question(s).

*Methods*: A section describing the application of the second technique, with justification or explanation where necessary.

*Results*: A section that presents results from the analysis conducted using the second technique.

*Conclusion*: A conclusion that answers your research question(s) using your second technique.

**Hints for achieving better grades:**

1.Use the lecture notes as a guide on what analytics techniques to use and how to apply them.

2.Remember not to be too ambitious- while you need to demonstrate a variety of skills it is important that you give yourself enough space to discuss what you have done and to present the results.

3.In general, try to keep your writing as concise as possible. Try to avoid unnecessary jargon. Take care of the word count and do not miss out on marks for writing too much or too little.

4.Justify and support your decisions, making use of references.

5.As far as possible, don’t copy/paste material verbatim from existing sources; paraphrase into your own words. Whether you are quoting verbatim or paraphrasing, cite and reference your sources using the Harvard System. A description is available at: <https://library.southwales.ac.uk/collections-subject-guides/referencing/>

6.Look carefully at the assessment criteria to ensure you are covering aspects that will be needed to earn marks

7. Make sure you structure your report in line with the requirements set out above.

A video can be found on the Assessment tab on Blackboard, providing additional guidance.

## Guidance on Format of Assessment

Note: Students are reminded **not** to include this assignment brief with the assignment submission.

Referencing must be completed in line with the USW Harvard style, as outlined here: <https://library.southwales.ac.uk/collections-subject-guides/referencing/>

## Learning Outcomes Assessed

LO1: To appraise and contrast strategies for dealing with Big Data

LO2: To demonstrate an ability to apply Big Data concepts in non-trivial contexts

## Marking Criteria/Rubric

Note: All grades are provisional until they are ratified by the exam board

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Fail | Narrow Fail | 3rd Class / Pass | Lower 2nd Class / Pass | Upper 2nd Class / Merit | 1st Class / Distinction |
| A brief introduction 5% | * Missing or very superficial introduction that does not clearly state which techniques have been chosen and does not identify the nature of the research questions to be tackled | * Gives a superficial insight into the content of the report, does not clearly state which techniques have been chosen or does not identify the nature of the research questions to be tackled | * Gives a basic insight into the content of the report, but may be ambiguous about which techniques have been chosen and the nature of the research questions to be tackled. May contain irrelevant information | * Gives a reasonable insight into the content of the report, stating which techniques have been chosen and the nature of the research questions to be tackled. May contain superfluous information | * Gives a clear insight into the content of the report, explaining which techniques have been chosen and the nature of the research questions to be tackled although there may be ambiguities or the explanation may be too verbose | Clear and concise insight into the content of the report, explaining which techniques have been chosen and the nature of the research questions to be tackled |
| Technique 1 40% | * Unsuitable dataset chosen that does not allow analysis using chosen technique * Justifications and explanations for application methods are absent or contain too many errors to show understanding of technique * Analysis absent, or not producing interpretable results * Interpretations absent or entirely erroneous * Absent or entirely erroneous explanations of conclusions that do not answer research questions | * Unsuitable dataset chosen that does not allow effective analysis using chosen technique * Justifications and explanations for application methods are superficial or contain many errors that show a lack of understanding of technique * Analysis performed to produce superficial, inappropriate or inaccurate results * Interpretations of results include many errors such that they are of minimal value * Superficial or inappropriate explanations of conclusions that do not answer research questions, or contain many errors | * Crude dataset chosen that allows only functional analysis * Justifications and explanations for application methods are vague or contain several errors * Analysis performed to produce basic results that include several errors or are incomplete * Interpretations of results include several errors but are of value * Basic explanations of conclusions that answer research questions to a limited extent, but may also contain several errors | * Suitable dataset chosen that allows appropriate analysis * Justifications and explanations for application methods contain some errors or includes limited detail * Analysis performed to produce mainly accurate results that may include some errors or lack some detail * Interpretations of results may include some errors or lack some detail * Reasonable explanations of conclusions that answer research questions, but may include some errors or lack some detail | * Good dataset chosen that allows detailed analysis * Justifications and explanations for application methods contain only minor errors or ambiguities * Analysis performed to produce accurate results that contain only minor errors or ambiguities * Interpretations of results contain only minor errors or ambiguities * Detailed explanations of conclusions that answer research questions, but may include minor errors or ambiguities | * Advanced choice of dataset that facilitates complex analysis * Justifications and explanations for application methods are clear and concise, with no errors or ambiguities. * Complex analysis performed to an exceptional standard, producing detailed and meaningful results, demonstrating advanced skill level. * Interpretations of results are clear and concise, with no errors or ambiguities. * Clear and concise explanations of conclusions that fully answer research questions, with no errors or ambiguities. |
| Technique 2 40% | * Unsuitable dataset chosen that does not allow analysis using chosen technique * Justifications and explanations for application methods are absent or contain too many errors to show understanding of technique * Analysis absent, or not producing interpretable results * Interpretations absent or entirely erroneous * Absent or entirely erroneous explanations of conclusions that do not answer research questions | * Unsuitable dataset chosen that does not allow effective analysis using chosen technique * Justifications and explanations for application methods are superficial or contain many errors that show a lack of understanding of technique * Analysis performed to produce superficial, inappropriate or inaccurate results * Interpretations of results include many errors such that they are of minimal value * Superficial or inappropriate explanations of conclusions that do not answer research questions, or contain many errors | * Crude dataset chosen that allows only functional analysis * Justifications and explanations for application methods are vague or contain several errors * Analysis performed to produce basic results that include several errors or are incomplete * Interpretations of results include several errors but are of value * Basic explanations of conclusions that answer research questions to a limited extent, but may also contain several errors | * Suitable dataset chosen that allows appropriate analysis * Justifications and explanations for application methods contain some errors or includes limited detail * Analysis performed to produce mainly accurate results that may include some errors or lack some detail * Interpretations of results may include some errors or lack some detail * Reasonable explanations of conclusions that answer research questions, but may include some errors or lack some detail | * Good dataset chosen that allows detailed analysis * Justifications and explanations for application methods contain only minor errors or ambiguities * Analysis performed to produce accurate results that contain only minor errors or ambiguities * Interpretations of results contain only minor errors or ambiguities * Detailed explanations of conclusions that answer research questions, but may include minor errors or ambiguities | * Advanced choice of dataset that facilitates complex analysis * Justifications and explanations for application methods are clear and concise, with no errors or ambiguities. * Complex analysis performed to an exceptional standard, producing detailed and meaningful results, demonstrating advanced skill level. * Interpretations of results are clear and concise, with no errors or ambiguities. * Clear and concise explanations of conclusions that fully answer research questions, with no errors or ambiguities. |
| Presentation, spelling and grammar 5% | * Report structure does not reflect assessment requirements * Many spelling mistakes in each sentence * Many grammatical errors * Writing style unreadable | * Report structure superficially reflects assessment requirements * Many spelling mistakes in each paragraph * Many grammatical errors * Writing sometimes understandable although meaning is often unclear | * Report structure loosely reflects assessment requirements * Frequent spelling mistakes * Several grammatical errors * Writing understandable although meaning may be unclear in places or some text may be verbose | * Report structure somewhat reflects assessment requirements * Some spelling mistakes * Some grammatical errors * Writing understandable throughout | * Report structure largely reflects assessment requirements * Few spelling mistakes * Clear sentence construction, using appropriate punctuation * Brief and clear writing style | * Report structure accurately reflects assessment requirements * Very few/no spelling mistakes in the report * Very clear sentence construction that makes good use of punctuation * Clear and informative writing style of publishable standard |
| Use of references 10% | * No relevant background reading with many statements unsubstantiated * References and citations missing or contain many errors | * Little relevant background reading with many unsubstantiated statements * Few sources cited and referenced in line with USW Harvard guidelines with several errors | * Evidence of reading only web sites and class notes with some unsubstantiated statements * Main sources cited and referenced in line with USW Harvard guidelines with several errors | * Evidence of reading relevant sources (eg. books and quality web sites), with a few unsubstantiated statements * All sources cited and referenced in line with USW Harvard guidelines with a few errors | * Evidence of reading several quality and relevant academic sources (eg. journals, books and quality web sites), used to support analysis * All sources cited and referenced in line with USW Harvard guidelines with only minor errors | * Evidence of reading many high quality and relevant academic sources (eg. journals, books and quality web sites), used effectively to support analysis * All sources cited and referenced correctly in line with USW Harvard guidelines |
|  | | | | | | |

## Submission Details

Submit your work through the appropriate submission slots on Blackboard before the submission deadline.

Your submission should be uploaded to two different submission slots that will be marked clearly:

1. Written Report Submission slot.

Here you should submit your written report in the form of an MS Word document.

1. Data and Code Submission slot

Here you should submit a zipped file containing the following files:

* An R code script file (.R).
* Your data file (preferably CSV, but you are free to use any data format).

## What happens next?

Your marked assessment should be available 20 working days after submission. However, please be advised that this may be subject to change in the event of Bank Holidays, University Closure or staff sickness. If there is something about the feedback you have been given that you are unclear about, please see your module tutor.

## Feedback Method

Feedback will be emailed to you.

## Late Submission

The assessment submission slot on Blackboard will remain open after the deadline has passed. If it necessary for you to submit your work late then you should submit your work through the same method and inform your lecturer when you have done so. Your work may then be assessed, if appropriate.

## Retrieval in the Event of Failure

Resit assessments opportunities will be available in the summer if appropriate.

## Extenuating Circumstances

[https://advice.southwales.ac.uk/a2z/extenuating-circumstances](https://advice.southwales.ac.uk/a2z/extenuating-circumstances/)

## Referencing, Plagiarism and Good Academic Practice

[https://advice.southwales.ac.uk/a2z/referencing-plagiarism-and-good-academic-practice](https://advice.southwales.ac.uk/a2z/referencing-plagiarism-and-good-academic-practice/)

## Learning Support Resources

[https://studyskills.southwales.ac.uk](https://studyskills.southwales.ac.uk/)

## Your Assessment Queries

Assessment related queries should be raised in class ideally. It may be possible to offer additional individual support on request. Contact your lecturer to arrange this if necessary.

## Student Checklist

Ensure that you have submitted:

* 1. A MS Word document consisting of your written report.
  2. A zip file containing:
     1. R code (.R) file.
     2. Data file (data used in the analysis).