Faculty of Engineering, Environment and Computing

7153CEM Big Data Analytics and Data Visualisation

Assignment Brief — JANMAY2324

Module Title Big Data Analytics and Data Visualisation	Individual	Cohort: JANMAY2324	Module Code 7153CEM
Coursework Title: **Dataset Analysis and Visualization Using Big Data Program**			Hand out date: 25/04/2024
Lecturer Dr. Anup Pandey			Due date and time: 04/04/2024 **Time: 18:00:00**
Estimated Time (hrs): 25 h **Word Limit: 3000 words**	Coursework type: **Written report**		**100 % of Module Mark**

Very Important

Submission arrangement online via Aula:
Submit before 1800, late work will receive a mark of zero.
File types and method of recording: **WORD** using the "Assignments" link in 7153CEM.
Mark and Feedback date (DD/MM/YY): 3 weeks after submission **26/04/2024)**
Mark and Feedback method (e.g. in lecture, electronic via Aula): electronic via Aula

The reports that are not in Word format will receive a ZERO mark.
The word count needs to include the Table of Contents, Code, and Materials in the Appendix.
Please place all screenshots, results graphs, and any other images in the Appendix section of the report.
Refer to these visuals in the main body of your report.

Module Learning Outcomes Assessed:

3. Demonstrate sound knowledge of different data analytical techniques for different structured and unstructured big data sets to support decision-making.
4. Critically identify and select appropriate analytical technique for big data analysis using examples from case studies
5. Critically evaluate and apply appropriate methods that are suitable for visualising big data

Task:

1. Select a dataset of your choice from Kaggle. The dataset should be suitable for Big Data analytics (Please see description below).

2. Use PySpark (**exclusively**) to analyze the dataset. You should perform at least one of the following data analysis tasks (regression, clustering, classification, etc). You have to explain your choice of the techniques used.

3. Use Tableau (**exclusively**) to explore your dataset and/or to show the results of your analysis. **All** figures must be created using Tableau

4. Critically analyze your findings: the results and the methods used.

Procedure:

- The dataset should be freely accessed (no registration is required). If there are several files in the link, you have to **clearly** specify which one you plan to work on.
- Your final CW submission will include a report (up to 3000 words – strict limit) where you present your work.

Clarifications:

- You can use any operating system that you prefer to install your program.
- Coding the task you are performing yourself is a plus.
- Given the nature of this module and the task, you should document everything you do.
- **Everything you do should be reproducible**: The link to the dataset should be provided (direct link to the dataset itself not the site where it is hosted). The code used, **in its totality** (as text **NOT** as file), should be included in the appendix using the right tool to include code in a WORD file (WORD code syntax highlighter, for example). If you use a code that is not yours, whether totally or partially, this should be **very clearly indicated**.
- You should provide in the appendix clear evidence, using screen captures, that you installed the software and ran every part of the experiments. The screen captures should also clearly show the device (i.e. user ID) on which the experiments were conducted/the software was installed
- Except for the dataset, **NO LINKS** of any kind to your work are allowed. Everything should be included in the report itself (the body or the appendix)
- P**lagiarism and collusion are taken extremely seriously**. Any part, from any source, of any type, in any language, should be **COMPLETELY AND CLEARLY** citated. If you use a figure/table/image that is not yours, this should be indicated in the caption.
- **Copying a large chunk of text, even with proper referencing, will still be classified as plagiarism.**
- **Do not use ChatGPT or any other tool**
- The use of any online platform (AWS, Google Colab, Databricks, etc) is not recommended or needed. However, if you need to do that you have two options ONLY:
    - Either, you install all the software required for the CW on your device, **CLEARY** showing your name as the user, with all the required libraries and modules required to do your CW to show that you can install and configure all the software required for the CW. You may after that, if you choose, install all that again on an online platform and your CW on that platform
    - Or, you use a platform that shows CLEALY, without any ambiguity, that this is your personal account and your name clearly appears on screen captures you include on your CW, showing all that.

Report Structure:

Your report should typically have:

- A title page (should also have word counts)
- Abstract
- Introduction -An introduction in which you briefly describe your project, why this is important, briefly about the dataset you are working on, the data analysis task(s) you are performing and the software you are using.
- A background /related work/data analysis section, focusing on what is related to your project.
- Dataset section- In the dataset section, the details of the dataset, such as the number of rows, columns, and attributes, should be presented in the initial lines. This section should include data processing details.
- Methodology – Describe the systematic approach and techniques used for achieving the objectives. Includes details about the methods and program to/the data analysis task you are performing, the program you are using – with description and figures showing how it is installed, configured and how it works.
- Experimental section – Includes full description of the experimental protocol, how you conducted the experiments, and the results.
- Result Discussion - A discussion of your findings.
- A conclusion and Future works
- Social Impact of this project
- References.
- Appendix (not included in the word count)

Dataset description:

The dataset should have at least 10 attributes (columns) and at least 500 rows. It should at least have three of the following data types (characters, strings, doubles/floats, integers, Booleans, dates, categories). This last point is particularly important.

It is strongly recommended that you use the dataset as it is. If the dataset has too many rows, you can delete some (as long as the remaining dataset will have at least 500 rows). As for the columns, it is particularly recommended that you keep all the original columns, if you need to delete some however, you can do that under two conditions: 1- You should keep at least 80% of the original number of columns 2- The remaining dataset should have the variety of data types described above

The dataset should have previously been on Kaggle (i.e. you cannot create your own dataset and upload it). No merging of two or more datasets is allowed

Mark distribution:

Dataset description	10 Marks	Clearly define the chosen dataset, dataset processing and raw data handling
Proof of Installation of HADOOP and SPARK	10 Marks	Provide proof that HADOOP and SPARK were installed on your Laptop/Machine.

Technical quality	30 Marks	This aspect concerns the depth of the information presented in the report.
Difficulty	10 Marks	This aspect concerns the difficulty of the program used or the analysis applied/the complexity of the dataset/applying several data analysis tasks/programming the method by the student himself/herself.
Visualization (Tableau)	20 Marks	This aspect concerns the quality of visualization produced.
Reproducibility	10 Marks	This aspect concerns using screen shots of all the steps taken, providing the code (using the right tool to include code in a WORD document, e.g. using WORD code syntax highlighter), clear and straightforward explanation of the steps taken to reproduce the results, including the figures
Style and format	10 Marks	

Notes:

1. You are expected to use the Coventry University APA style for referencing. For support and advice on this students can contact Centre for Academic Writing (CAW).
2. Please notify your registry course support team and module leader for disability support.
3. Any student requiring an extension or deferral should follow the university process as outlined here.
4. The University cannot take responsibility for any coursework lost or corrupted on disks, laptops or personal computer. Students should therefore regularly back-up any work and are advised to save it on the University system.
5. If there are technical or performance issues that prevent students submitting coursework through the online coursework submission system on the day of a coursework deadline, an appropriate extension to the coursework submission deadline will be agreed. This extension will

normally be 24 hours or the next working day if the deadline falls on a Friday or over the weekend period. This will be communicated via your Module Leader.

6. You are encouraged to check the originality of your work by using the draft Turnitin links on Aula.
7. Collusion between students (where sections of your work are similar to the work submitted by other students in this or previous module cohorts) is taken extremely seriously and will be reported to the academic conduct panel. This applies to both courseworks and exam answers.
8. A marked difference between your writing style, knowledge and skill level demonstrated in class discussion, any test conditions and that demonstrated in a coursework assignment may result in you having to undertake a Viva Voce in order to prove the coursework assignment is entirely your own work.
9. If you make use of the services of a proof reader in your work you must keep your original version and make it available as a demonstration of your written efforts.
10. You must not submit work for assessment that you have already submitted (partially or in full), either for your current course or for another qualification of this university, with the exception of resits, where for the coursework, you maybe asked to rework and improve a previous attempt. This requirement will be specifically detailed in your assignment brief or specific course or module information. Where earlier work by you is citable, i.e. it has already been published/submitted, you must reference it clearly. Identical pieces of work submitted concurrently may also be considered to be self-plagiarism.

Mark allocation guidelines to students

0-39	40-49	50-59	60-69	70+	80+
Work mainly incomplete and /or weaknesses in most areas	Most elements completed; weaknesses outweigh strengths	Most elements are strong, minor weaknesses	Strengths in all elements	Most work exceeds the standard expected	All work substantially exceeds the standard expected

Marking Rubric

GRADE	ANSWER RELEVANCE		ARGUMENT & COHERENCE		EVIDENCE		SUMMARY	
First **≥70**	Innovative response, answers the question fully, addressing the learning objectives of the assessment task. Evidence of critical analysis, synthesis and evaluation.		A clear, consistent in-depth critical and evaluative argument, displaying the ability to develop original ideas from a range of sources. Engagement with theoretical and conceptual analysis.		Wide range of appropriately supporting evidence provided, going beyond the recommended texts. Correctly referenced.		An outstanding, well-structured and appropriately referenced answer, demonstrating a high degree of understanding and critical analytic skills.	
Upper Second **60-69**	A very good attempt to address the objectives of the assessment task with an emphasis on those elements requiring critical review.		A generally clear line of critical and evaluative argument is presented. Relationships between statements and sections are easy to follow, and there is a sound, coherent structure.		A very good range of relevant sources is used in a largely consistent way as supporting evidence. There is use of some sources beyond recommended texts. Correctly referenced in the main.		The answer demonstrates a very good understanding of theories, concepts and issues, with evidence of reading beyond the recommended minimum. Well organised and clearly written.	
Lower Second **50-59**	Competently addresses objectives, but may contain errors or omissions and critical discussion of issues may be superficial or limited in places.		Some critical discussion, but the argument is not always convincing, and the work is descriptive in places, with over-reliance on the work of others.		A range of relevant sources is used, but the critical evaluation aspect is not fully presented. There is limited use of sources beyond the standard recommended materials. Referencing is not always correctly presented.		The answer demonstrates a good understanding of some relevant theories, concepts and issues, but there are some errors and irrelevant material included. The structure lacks clarity.	
Third **40-49**	Addresses most objectives of the assessment task, with some notable omissions. The structure is unclear in parts, and there is limited analysis.		The work is descriptive with minimal critical discussion and limited theoretical engagement.		A limited range of relevant sources used without appropriate presentation as supporting or conflicting evidence coupled with very limited critical analysis. Referencing has some errors.		Some understanding is demonstrated but is incomplete, and there is evidence of limited research on the topic. Poor structure and presentation, with few and/or poorly presented references.	
Fail **<40**	Some deviation from the objectives of the assessment task. May not consistently address the assignment brief. At the lower end fails to answer the question set or address the learning outcomes. There is minimal evidence of analysis or evaluation.		Descriptive with no evidence of theoretical engagement, critical discussion or theoretical engagement. At the lower end displays a minimal level of understanding.		Very limited use and application of relevant sources as supporting evidence. At the lower end demonstrates a lack of real understanding. Poor presentation of references.		Whilst some relevant material is present, the level of understanding is poor with limited evidence of wider reading. Poor structure and poor presentation, including referencing. At the lower end there is evidence of a lack of comprehension, resulting in an assignment that is well below the required standard.	
Late submission	0		0		0		0	