Faculty of Engineering, Environment and Computing

5014CEM Data Science for Developers

Assignment Brief

Module Title: **Data Science for Developers**	Individual /Group: **Individual**	Cohort: **September 2021 (Resit/Deferral in April 2022)**	Module Code: **5014CEM**
Coursework Title: **Report on applying the data science lifecycle (Resit/Deferral in April**			Hand out date: **24/01/2022**
Lecturer: **Mark Johnston**			Due date and time: **04/04/2022 at 6pm**
Estimated Time (hrs): **75 hours** Word Limit*: **3000 words**	Coursework type: **Report (Applied Core)**		Credit value assessed: **15 credits**
Submission arrangement: **Aula** File types: **pdf, docx, odt** Mark and Feedback date (DD/MM/YY): **18/04/2022** Mark and Feedback method (e.g. in lecture, electronic via Aula): **Aula**			

Module Learning Outcomes Assessed:

ILO1. Understand and apply the components of the data mining lifecycle to real-world big data problems.
ILO2. Analyse, design, implement, manage, and critically evaluate a database solution for a specified commercial or scientific objective, using state-of-the-art tools such as R, OpenRefine or Python.
ILO5. Show systematic knowledge of concepts in statistical analysis including experimental design, statistical modelling, probabilities, p-values, categorical data, t-tests, and Pearson correlation; and critically select and justify use of appropriate methods for a given problem space.

Task and Mark distribution:

In this *Individual Coursework* you will work through the phases of the *data science lifecycle* applied to a real-world task. You will obtain, combine and analyse datasets from a range of sources. We are primarily interested in the processes you follow, although you will need to find your own datasets, explain the code used to implement your system, and communicate the results of your analyses. You are encouraged to explore the topic, use your initiative, and show some originality, within the time available. Ensure that you clearly address the module learning outcomes listed above and that you reference any sources you have used.

Please submit one report, e.g., as a single Microsoft Word document. Make sure you include any code snippets, and selected output and plots, directly in the report so that they can be clearly read. The word limit is a maximum rather than a target. Concentrate on producing a clear and concise answer to each subtask. Please also include a full listing of your code as an Appendix. Code developed should be in R or Python or a combination of the two.

SCENARIO

Suppose one of your friends is looking to buy a house somewhere in the neighbouring counties of **Norfolk and Suffolk** in the United Kingdom. They want you to recommend suitable local authority districts within those two counties (not towns or individual houses) in which to focus their house search, based initially on these attributes: house prices, crime in the area, and quality of local schools. They want you to help them compare districts and investigate the tradeoffs between these attributes across these districts.

TASK

The task is to apply each phase of the data science lifecycle (obtain, scrub, explore, model and interpret) to the scenario described above. You are required to write a report that covers the report structure and subtasks given below. Marks for each section are also given below.

1. *Introduction (5 marks)*
   Clearly set out the problem that is being addressed. Include a list of districts in each county (Norfolk has 7 districts and Suffolk has 5 districts). It would be helpful to provide a map showing these districts. Give an outline of the structure of the remainder of the report.

2. *Obtain (10 marks)*
   Give a description of the datasets obtained and specify exactly where each dataset was obtained from (it must be possible to find exactly these datasets again if necessary). Clearly explain why you selected these particular datasets and give a justification of the suitability of each dataset for the task. Note that you may need to also consider population in order to translate crime totals into crime rates. You must only use datasets that are published by the UK government, either centrally or through a public body that would be available to a member of the UK public. You may find the websites https://data.gov.uk/ and https://data.police.uk/data/ useful starting points.

3. *Scrub (20 marks)*
   Give a detailed description and justification of how the data was checked, cleaned and pre-processed to extract any relevant variables from each dataset in a form that will be useful for further analysis. Also, combine the datasets at the level of districts. Explain how you have combined the datasets geographically; you will need to briefly explain the geocoding hierarchy including county, district, MSOA, LSOA, postcode, etc. The ONS has some nice tools to help with things like converting postcodes to MSOAs, electoral wards, etc., at their Open Geography Portal (http://geoportal.statistics.gov.uk/). Aim to do as much of the data pre-processing as possible in R or Python. You can use other software (such as Microsoft Excel) to help with this phase, but you must carefully describe the steps you have taken so that they are repeatable. Clearly show how the final pre-processed dataset is organised at the end of this phase. *You should give enough detail through description and snippets of code that it should be possible to reproduce the process you have used.*

4. *Explore (20 marks)*
   Carry out Exploratory Data Analysis (EDA) on the datasets, i.e., appropriate graphical plots and summary statistics to investigate the distribution of single variable data (including looking for outliers) and investigate the relationships between variables (scatterplots and correlation coefficients). Make sure you describe and discuss what you observe in the plots and summaries you produce. Be clear which variables in a dataset are categorical and which are quantitative.

5. *Model (20 marks)*
   Apply appropriate statistical models and methods to your datasets, e.g., fitting linear models to investigate the statistical relationships between the variables and show how these can be used to make predictions. Assess, interpret and discuss the results obtained including residuals, diagnostic plots, comparing models, and p-values. Also design and implement a simple recommendation system, e.g., it might determine a value in the range 0–10 for district on each variable, and combine these into a score for each district. Include a discussion of the results, an assessment of the degree to which it achieves its goal, and how it could be improved.

6. *Interpret (15 marks)*
   Discuss the legal and ethical issues relating to the data you are using and your recommendation system. Summarise what you have learned from the datasets chosen relevant to the task and draw some conclusions from your investigations. Reflect on how well you were able to apply the data science lifecycle to the problem, and make some recommendations to improve or extend what you have done in the future.

The remaining 10 marks will be assessed based on:

- the quality, presentation and organisation of the report and correct use of referencing (5 marks)
- the quality and functionality of the code produced (5 marks).

Notes:

1. You are expected to use 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 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 coursework 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 proofreader in your work you must keep your original version and make it available as a demonstration of your written efforts. Also, please read the university Proof Reading Policy.
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 may be 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.

Marking Rubric (UG)

Mark band	Outcome	Guidelines
90-100% 1st	Meets learning outcomes	1st - Exceptional work with very high degree of understanding, creativity and critical/analytic skills. Evidence of exceptional research well beyond minimum recommended using a range of methodologies. . Exceptional understanding of knowledge and subject-specific theories. Demonstrates creative flair, a high degree of originality and autonomy. Exceptional ability to apply learning resources. Demonstrates well-developed problem-solving skills. Work completed with very high degree of accuracy and proficiency and autonomy. Exceptional communication and expression, significant evidence of professional skill set. Student evidences deployment of a full range of exceptional technical and/or artistic skills.
80-89% 1st		1st - Outstanding work with high degree of understanding, creativity and critical/analytical skills. Outstanding understanding of knowledge and subject-specific theories. Evidence of outstanding research well beyond minimum recommended using a range of methodologies. Demonstrates creative flair, originality and autonomy. Outstanding ability to apply learning resources. Demonstrates clear problem-solving skills. Assessment completed with high degree of accuracy and proficiency and high-level of autonomy. Outstanding communication and expression, evidence of professional skill set. Student evidences deployment of a full range of technical and/or artistic skills.
70-79% 1st		1st - Excellent work with clear evidence of understanding, creativity and critical/analytical skills. Thorough research well beyond the minimum recommended using methodologies beyond the usual range. Excellent understanding of knowledge and subject-specific theories with evidence of considerable originality and autonomy. Excellent ability to apply learning resources. Demonstrates consistent, coherent substantiated argument and interpretation. Demonstrates considerable creativity and clear problem-solving skills. Assessment completed with accuracy, proficiency, and considerable autonomy. Excellent communication and expression, some evidence of professional skill set. Student evidences deployment of a highly developed range of technical and/or artistic skills.
60-69% 2:1		2:1 - Very good work demonstrating strong understanding of theories, concepts and issues with clear critical analysis. Thorough research, using established methodologies accurately, beyond the recommended minimum with little, if any, irrelevant material present. Very good understanding, evidencing breadth and depth, of knowledge and subject-specific theories with some originality and autonomy. Very good ability to apply learning resources. Demonstrates coherent substantiated argument and interpretation. Demonstrates some originality, creativity and problem-solving skills. Work completed with accuracy, proficiency, and autonomy. Very good communication and expression with evidence of professional skill set. Student has a thorough command of a good range of technical and/or artistic skills.

50-59% 2:2		2:2 - Good understanding of relevant theories, concepts and issues with some critical analysis. Research undertaken accurately using established methodologies, enquiry beyond that recommended may be present. Some errors may be present and some inclusion of irrelevant material. Good understanding, with evidence of breadth and depth, of knowledge and subject-specific theories with indications of originality and autonomy. Good ability to apply learning resources. Demonstrates logical argument and interpretation with supporting evidence. Demonstrates some originality, creativity and problem-solving skills but with inconsistencies. Expression and presentation mostly accurate, proficient, and conducted with some autonomy. Good communication and expression with appropriate professional skill set. Student consistently demonstrates a well-developed range of technical and/or artistic skills.
40-49% 3rd Class		3rd - Meet the learning outcomes with a basic understanding of relevant theories, concepts and issues.. Demonstrates an understanding of knowledge and subject-specific theories sufficient to deal with concepts. Assessment may be incomplete and with some errors. Research scope sufficient to evidence use of some established methodologies. Some irrelevant material likely to be present. Basic ability to apply learning resources. Demonstrates ability to devise and sustain an argument. Demonstrates some originality, creativity and problem-solving skills but with inconsistencies. Expression and presentation sufficient for accuracy and proficiency. Sufficient communication and expression with basic professional skill set. Student demonstrates technical and/or artistic skills.
30-39% Fail	Fails to achieve learning outcomes	Fail – Very limited understanding of relevant theories, concepts and. Little evidence of research and use of established methodologies. Some relevant material will be present. Deficiencies evident in analysis. Fundamental errors and some misunderstanding likely to be present. Limited ability to apply learning resources. Student's arguments are weak and poorly constructed. Very limited originality, creativity, and struggles with problem-solving skills. Expression and presentation insufficient for accuracy and proficiency. Insufficient communication and expression and with deficiencies in professional skill set. Student demonstrates some deficiencies in technical and/or artistic skills.
20-29% Fail		Fail - Clear failure demonstrating little understanding of relevant theories, concepts and issues. Minimal evidence of research and use of established methodologies and incomplete knowledge of the area. Serious and fundamental errors and aspects missing Little evidence of ability to apply learning resources. Students arguments are very weak and with no evidence of alternative views. Little evidence of originality, creativity, and problem-solving skills. Expression and presentation deficient for accuracy and proficiency. Insufficient communication and expression and with deficiencies in professional skill set. Student demonstrates a lack of technical and/or artistic skills.

0-19% Fail		Fail - Inadequate understanding of relevant theories, concepts and issues. Complete failure, virtually no understanding of requirements of the assignment. Material may be entirely irrelevant. Assessment may be fundamentally wrong, or with major elements missing. Not a serious attempt. No evidence of research. Inadequate evidence of ability to apply learning resources. Very weak or no evidence of originality, creativity, and problem-solving skills. Students presents no evidence of logical argument and no evidence of alternative views. Expression and presentation extremely weak for accuracy and proficiency. Communication and expression very weak and with significant deficiencies in professional skill set. Student evidences few or no technical and/or artistic skills