Softwarica College of IT & E-Commerce

ST5014CEM Data Science for Developers

Assignment Brief [March 2024]

in collaboration with

Softwarica College of IT & E-commerce | Coventry University

Module Title: Data Science for Developers	**Ind/Group:** Individual	**Cohort:** March 2024 - Regular	**Module Code:** ST5014CEM
Coursework Title: Town Recommendation System and Report on an Individual Data Science Project			**Handout Date:** 07/01/2024
Lecturer: Siddhartha Neupane			**Due Date:** 08/19/2024
Estimated Time (hrs.) [NA] **Word Limit: 3000**	**Coursework Type:** Assignment		**% Of Module Mark** 70%

Submission arrangement online via schoolworkspro:

The link to your GitHub classroom repository containing app source code must be submitted on schoolworkspro.

File types: Please submit your report in PDF format only.
Mark and Feedback date: 3 weeks after submission
Mark and Feedback method: written feedback using schoolworkspro

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.

Submit one report as a single pdf 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. Make sure you 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
Imagine your international friends are planning to make a significant investment by purchasing a property in either **Bristol or Cornwall**, two beautiful regions in the United Kingdom known for their distinct charm and appeal. As they value your expertise in data analysis and your knowledge of the area, they've turned to you for recommendations on suitable locations based on various factors.

First and foremost, they are interested in housing prices within these counties. They understand that property prices can vary significantly depending on the location within each region and want to ensure they make an informed decision that aligns with their budget and investment goals. They are open to exploring both urban and rural settings, depending on the affordability and potential return on investment.

Additionally, your friends are also concerned about the availability and reliability of internet connectivity in the areas they are considering. As individuals who rely heavily on a stable internet connection for work and leisure activities, they want to ensure that they will have access to high-speed internet services without any disruptions.

Moreover, safety and security are paramount for your friends, especially since they are considering relocating to a new country. They are keen on understanding the local crime rates in the neighborhoods they are considering to gauge the overall safety of the area and make an informed decision about the security of their future home.

While housing prices, internet speed, and local crime rates are significant factors in their decision-making process, they are also open to considering other factors that might impact their choice of location. Factors such as proximity to amenities (e.g., schools, hospitals, shopping centers), public transportation options, recreational facilities, and the overall quality of life in the area are also essential considerations for them.

Given these criteria, your friends are eager to receive your recommendations and insights based on your analysis of the data available for Bristol and Cornwall. They trust that your thorough examination of these factors will guide them towards making the right decision for their future home and investment.

Tasks
You must only use datasets that have been released by the UK government, either centrally or through a public entity that is accessible to the public in the UK. The website https://data.gov.uk/ might be a good place to start. You should clean the data, create a uniform data model, use exploratory data analysis to examine the dataset, investigate statistical links between the attributes, and create a basic recommendation system utilizing this knowledge. The processes you use must be documented and explained.

You must ensure that:

- The system is developed in R, Python or a combination of the two.
- All data used is normalised to at least 3NF and stored in an appropriate database. You must use the three key characteristics described above (house prices, broadband speed, and crime in the area) and at least one additional characteristic (but you must justify your reason for including it).
- You must include a simple recommendation system that determines a value in the range 0–10 for each of the characteristics, combines these into a score for each town, and displays the top three towns in order.
- The towns used are in the counties given. You may restrict the number of towns you look at to main towns, but you must justify your selection in your report.

The main element must be done in R or Python, but the data cleaning may be done in any language as long as you explicitly explain the processes you took so they can be repeated. You'll need to search up the districts in the two counties, and the big towns may be found from the districts. The

ONS' Open Geography Portal (http://geoportal.statistics.gov.uk/) includes several useful tools for things like converting postcodes to MSOAs, electoral wards, and so on.

Submission details
You should submit a single ZIP file containing the following:

- An electronic copy of your code (.R or .py files).
- An electronic copy of your report as detailed above (.pdf format).
- An electronic copy of your datasets.

Report Template

- An introduction that clearly sets out the problem, a description of the datasets obtained, a justification of the suitability of the data for the task, and exactly where the datasets were obtained from.
- A clear, detailed description and justification of how the data was checked, cleaned and preprocessed, and a description of the data model, including why it is organised in this way (this can be thought of as a description and rationale for your database tables).
- Exploratory data analysis (EDA) undertaken on the datasets, i.e., 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).
- Interpretation and discussion of results obtained by applying appropriate statistical models and methods to your datasets, e.g., fitting linear models and discussing the output, diagnostic plots, comparing models, and statistical tests (interpreting the p-values).
- Design of your simple recommendation system, a discussion of its results, and an assessment of the degree to which it achieves its goal.
- An overview of the design of your code (with justification) and details of testing.
- A discussion of the legal and ethical issues relating to the data you are using and your recommendation system.
- Conclusion giving an analysis and reflection on how well you were able to apply the data mining lifecycle to the problem. Summarise the conclusions that you have made about your data and make some recommendations to improve or extend what you have done in the future.
- References.

Marks are distributed as follows:

- Identification, justification and gathering of data (10 marks)
- Data cleaning, pre-processing, and data model (15 marks)
- Exploratory data analysis (15 marks)
- Application and interpretation of statistical models, linear regression, and other methods (15 marks)
- Design and effectiveness of the recommendation system (10 marks)
- Code readability (10 marks)
- Discussion of legal and ethical issues (5 marks)
- Conclusions, analysis, and reflection (10 marks)
- Report quality, presentation, organization, and referencing (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 academic services team and module leader for disability support.

3. The college 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 cloud-based services.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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 college, 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.

Mark allocation guidelines to students (to be edited by staff per assessment)

0-39	40-49	50-59	60-70	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	Report	Code And Results	Data Mining Lifecycle
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 report. Engagement with theoretical and conceptual analysis. Correctly referenced.	Code is well written and follows a logical structure. Analysis of the data is clear with a range of statistical methods applied.	All stages of the data mining lifecycle have been correctly applied to all datasets. A range of appropriate datasets, including the key datasets, have been chosen, cleaned and applied.
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. Correctly referenced in the main.	Code is readable and functions as expected. An appropriate range of statistical methods are applied, but analysis is not as good as it could be.	All states of the data mining lifecycle have been correctly applied to all datasets. The three key datasets have been identified, cleaned, and applied.
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 shows only a partial understanding of the key concepts. Referencing is not always correctly presented.	Code functions as expected. A statistical method is correctly applied with some analysis.	All states of the data mining lifecycle have been correctly applied to all datasets. The three key datasets have been identified and used.
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.	Limited understanding of the theoretical concepts. Limited justification of method and results. Referencing has some errors.	Code has most of the functionality implemented. A statistical method is applied with limited analysis.	Most of the stages of the data mining lifecycle have been applied to most datasets. Two of the three key datasets have been identified and used.
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. At the lower end displays a minimal level of understanding. Poor presentation of references.	Code has some functionality implemented. A statistical method is applied with little or no analysis.	The majority of the stages of the data mining lifecycle have been applied to some datasets. One of the three key datasets has been identified and used.
Late Submission	0	0	0	0