****

**Faculty of Engineering, Environment and Computing**

##### 5014CEM Data Science for Developers

**Assignment Brief**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Module Title:  **Data Science for Developers** | Individual  /Group:  **Individual** | | Cohort:  **September 2024** | Module Code:  **5014CEM** |
| Coursework Title:  **Report on an Individual Data Science Project** | | | | Hand out date:  **Nov 2024** |
| Lecturer  **Mark Johnston, Fung Kai Tat, Johnny Kwong** | | | | Due date and time:  **29 Dec 2024 at 6pm Sun (HKT)** |
| Estimated Time (hrs):  Word Limit: **3000 words** | | Coursework type:  **Report (Applied Core)** | | Credit value assessed:  **15 credits** |
| Submission arrangement: **online via moodle**  File types: **pdf, docx, odt**  Mark and Feedback date (DD/MM/YY): **29 Dec 2024 (Sun) HKT**  Mark and Feedback method (e.g. in lecture, electronic via Aula): **moodle** | | | | |

|  |
| --- |
| 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 develop, analyse and combine 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.  **Scenario**  Suppose one of your friends is looking to buy a house somewhere in the Hong Kong, New Territories district of **New Territories East (NTE)** and **New Territories West (NTW)**. They want you to recommend suitable *districts* within those two locations, based initially on house prices, crime in the area, and quality of local schools.  **Task**  You may find the house price website a useful starting point. The community data such as traffic, crime, community facilities or school distribution data can be found at Hong Kong government website. Using this information, you should clean the data, produce a unified data model, explore the dataset using exploratory data analysis, investigate statistical relationships between the characteristics, and build a simple recommendation system. You must document and explain the processes that you undertake.  Other websites for data reference for the HK house pricing index.  <https://www.property.hk>  Housing Data  <https://data.gov.hk/en-data/dataset/hk-rvd-tsinfo_rvd-property-market-statistics>  School Data  <https://www.edb.gov.hk/en/student-parents/sch-info/sch-search/schlist-by-district/index.html>  Crime Data  <https://www.police.gov.hk/ppp_en/09_statistics/csd.html>  Traffic Data  <https://www.td.gov.hk/en/transport_in_hong_kong/transport_figures/monthly_traffic_and_transport_digest/index.html>  List of All Statistical Products  <https://www.censtatd.gov.hk/en/page_1273.html>  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, crime in the area, and quality of local schools). *You will need to also consider population in order to translate crime totals into crime rates.* * 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. *You must implement this recommendation system in code.* * The locations used are in the city given, i.e., all locations in New Territories East together with all locations in New Territories West. *It would be helpful to provide a map showing these locations.*   You have to do the core part in R or Python, but the data cleaning can be done in whatever you want as long as you clearly describe the steps you have taken so that they are repeatable. 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/>).  **Submission details**  You should submit a single document consisting of your report as detailed below. Include snippets of your code in relevant places in your report, and include a full listing of your code as an Appendix.  **Report template**  The report should contain the follow sections:   1. An introduction that clearly sets out the problem (including a clear list of districts in each county), 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. *It must be possible to find exactly these datasets again if necessary.* 2. A clear, detailed description and justification of how the data was checked, cleaned and pre-processed, 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). Include a brief explanation of how you have linked the data geographically across the datasets (district, town, LSOA, MSOA, etc). 3. 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). 4. 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). 5. Design of your simple recommendation system, a discussion of its results, and an assessment of the degree to which it achieves its goal. 6. An overview of the design of your code (with justification) and details of testing. 7. A discussion of the legal and ethical issues relating to the data you are using and your recommendation system. 8. 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. 9. References. 10. Appendix listing all of your code.   **Marks distribution**  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 and methods (15 marks) * Design and effectiveness of the recommendation system (10 marks) * Code (10 marks) * Discussion of legal and ethical issues (5 marks) * Conclusions, analysis and reflection (10 marks) * Report quality, presentation, organisation, and referencing (10 marks) |
| Notes:   1. You are expected to use the [Coventry University APA](https://libguides.coventry.ac.uk/apa) style for referencing. For support and advice on this students can contact [Centre for Academic Writing (CAW)](http://www.coventry.ac.uk/study-at-coventry/student-support/academic-support/centre-for-academic-writing/?theme=main). 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](https://share.coventry.ac.uk/students/Registry/Pages/Deferrals-and-Extension.aspx). 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 (to be edited by staff per assessment)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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** |  | **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 |  |