Flinders University
2023 S1 COMP2031 Data Engineering
Project Description + Marking Criteria

Overview:

In this final project, you will work in a group of 4-5 students with selected data collections from one of the sample datasets available on MongoDB. You will use R code that can run on RStudio, including all required packages. You will be graded on the quality of your data transformation and analysis, the accuracy and effectiveness of your data models and visualizations, and the clarity and persuasiveness of your report and presentation.

Data selection:

You will select one or more than one of the following data collections from only one dataset of MongoDB's sample datasets.

MongoDB's sample datasets	Data collections
sample_airbnb	listingsAndREviews
sample_analytics	accounts customers transactions
sample_geospatial	shipwrecks
sample_guides	planets
sample_mflix	comments movies sessions theaters users
sample_ restaurants	restaurants neighborhoods
sample_supplies	sales
sample_training	companies grades inspections posts routes trips zips
sample_weatherdata	data

When selecting data collections, groups should consider the scope of the data and the types of data included to ensure that they will be able to complete all required tasks. It is recommended that groups choose collections with reasonable sizes and complexity that will allow for meaningful analysis and modelling. The chosen collection should also have enough data to support supervised and unsupervised machine learning tasks, which are required for the data modelling component of the project.

Descriptions of each dataset can be found at the following link:
https://www.mongodb.com/docs/atlas/sample-data/

Project requirements:

The project requirements are to perform the following tasks:

1) (20%) Data Wrangling: Loading and tidying the dataset to ensure it is in a clean and usable format.
   a) Loading the data accurately (5%)
   b) Handling missing or incorrect data appropriately (5%)
   c) Tidying the data effectively (10%)
2) (20%) Data Transformation: Using data transformation techniques to transform the dataset into a format suitable for analysis.
   a) Using appropriate data transformation techniques to prepare the data for analysis (10%)
   b) Creating new variables as needed (10%)
3) (20%) Data Analysis: Analyse the dataset using appropriate statistical methods and create data visualizations to support your analysis.
   a) Using appropriate statistical methods to analyse the data (10%)
   b) Creating accurate and effective data visualizations (10%)
4) (30%) Data Modelling: Creating supervised and unsupervised machine learning models to predict or cluster relevant variables in the dataset.
   a) Creating supervised machine learning models to predict relevant variables in the dataset (10%)
   b) Creating unsupervised machine learning models to cluster relevant variables in the dataset (10%)
   c) Evaluating the performance the models (5%)
   d) Interpretability of the models (5%)
5) (10%) Data Communication: Describing your performances on the data and your findings/results, including relevant visualizations and models.
   a) Writing a clear and concise report summarizing your findings, including relevant visualizations and models. Limit the report to a maximum of 20 pages, not including references or appendices. Limit the word count to 5000 words, excluding figures and tables.
   b) Providing the R source code file used to generate your results.
   c) Creating an engaging and informative presentation of your findings. The presentation should be no longer than 25 minutes, with each team member's presentation no longer than 5 minutes.

Although this is a group project, each group member will be assessed based on their individual contribution to the project, as well as their performance during the presentation. This means that all group members are expected to actively participate in all aspects of the project and each group needs to submit a group contribution statement.

Marking criteria:

Tasks	Report (50 marks)	Presentation (20 marks)	Demo (10 marks)
1a	5%	5%	5%
1b	5%	5%	5%
1c	10%	10%	10%
2a	10%	10%	10%
2b	10%	10%	10%
3a	10%	10%	10%
3b	10%	10%	10%
4a	10%	10%	10%
4b	10%	10%	10%
4c	5%	5%	5%
4d	5%	5%	5%

5	10%	10%	10%

Deliverables:

1) An R source code file containing your group's code and analysis, along with any additional files, packages or data needed to run the code. The source code file must be runnable and include all necessary code, and comments to allow the markers to reproduce the data transformation, analysis, and modelling results.
2) A written report summarizing your group's findings, including relevant visualizations and models. The report should not exceed 20 pages (excluding references or appendices) and 5000 words (excluding figures and tables).
3) A presentation of your group's findings to the audience. The presentation should be no longer than 25 minutes, with each team member's presentation no longer than 5 minutes.
4) A Group Contribution Statement detailing each member's contribution to the project.

Note: Please ensure that you include clear and detailed explanations of your code and analysis in your R code file, report, and presentation.

Group registrations and group presentation modes:

Students are expected to form groups within their registered class. There are three classes in total: COMP2031 online, COMP2031 on-campus, and COMP8031 on-campus. Each group will present their work either online or on-campus depending on the class they are enrolled in. Groups of the COMP2031 online class will present their work online, while groups of the COMP2031 on-campus class and the COMP8031 on-campus class will present their work on-campus. Group registration will be managed through the course website, and students are responsible for forming their own groups.

Plagiarism:

Plagiarism, which includes copying from internet resources or from other group members, is not acceptable and will result in a grade of zero for the entire project. While using external sources for reference and inspiration is allowed, all work submitted must be original and produced by the group members themselves. Any external sources used must be appropriately cited and referenced in the project report. It is important that all group members understand the requirements for academic integrity and work together to ensure that all deliverables must be original and produced by the group members themselves.

Using identical demonstrations or operations on the same data as other groups is also not allowed. Each group is expected to perform their own unique analysis and modelling on their chosen data. While there may be some overlap in the techniques used, the overall approach and conclusions drawn should be distinct for each group. Any evidence of copying or using identical examples from other groups or external sources will result in a grade of zero for the entire project.

Groups are not allowed to use identical examples or demonstrations on the same data from the teaching materials in the workshops and tutorials. While these resources are provided to aid learning and understanding, the purpose of the project is to assess the group's ability to independently apply the concepts and techniques covered in class.