DEPARTMENT OF
SOFTWARE TECHNOLOGY

CSMODEL

Project – Case Study

Major Details

Groupings:	At most 4 members in a group
Deadline:	Phase 1 – October 18, 2024 (Friday) 6:00 PM
	Phase 2 – November 15, 2024 (Friday) 6:00 PM
Demo Schedule:	Phase 1 – October 21 – 25, 2024 (Week 8)
	Phase 2 – November 18 – 22, 2024 (Week 12)
Percentage:	Phase 1 – 20%
	Phase 2 – 20%
Submission guidelines:	Submit the `zip` file to AnimoSpace
Filename format:	`CSMODEL-Project-<Section>-Group<#>.zip`

Deliverables

`Zip` file containing:

- Jupyter Notebook file – `ipynb` file
- Other Python 3 files – `py` files
- Dataset files – `csv` files

Specifications

You are tasked to go through the process of selecting a dataset, formulating a research question, analyzing data, modelling data, hypothesis testing, and extracting insights from the data.

The project is to be submitted as a Jupyter Notebook and, optionally, some Python 3 source files. The notebook should be a self-explanatory document containing a report of the entire process undertaken to come up with the generated insights from the raw dataset. It should contain markup cells explaining the processes undertaken in the project, as well as code cells showing all the code that was performed. Please make sure that the code cells could be successfully run sequentially to replicate the processes done in the project.

Phase 1

The first phase of the case study involves four sections – (1) dataset description, (2) data cleaning, (3) Exploratory Data Analysis, and (4) research question.

Dataset Description

Each group should select one real-world dataset from the list of datasets provided for the project. Each dataset has a description file, which also contains detailed description of each variable.

In this section of the notebook, you must fulfill the following:

- State a brief description of the dataset.
- Provide a description of the collection process executed to build the dataset.
- Discuss the implications of the data collection method on the generated conclusions and insights.
- Note that you may need to look at relevant sources related to the dataset to acquire necessary information for this part of the project.
- Describe the structure of the dataset file.
    - What does each row and column represent?
    - How many observations are there in the dataset?
    - How many variables are there in the dataset?
    - If the dataset is composed of different files that you will combine in the succeeding steps, describe the structure and the contents of each file.
- Discuss the variables in each dataset file. What does each variable represent? All variables, even those which are not used for the study, should be described to the reader. The purpose of each variable in the dataset should be clear to the reader of the notebook without having to go through an external link.

Data Cleaning

For each used variable, check all the following and, if needed, perform data cleaning:

- There are multiple representations of the same categorical value.
- The datatype of the variable is incorrect.
- Some values are set to default values of the variable.
- There are missing data.
- There are duplicate data.
- The formatting of the values is inconsistent.

Note: No need to clean all variables. Clean only the variables utilized in the study.

Exploratory Data Analysis

Perform exploratory data analysis comprehensively to gain a good understanding of your dataset. This step should help in formulating the research question of the project.

In this section of the notebook, you must fulfill the following:

- Identify at least 4 exploratory data analysis questions. Properly state the questions in the notebook. Having more than 4 questions is acceptable, especially if this will help in understanding the data better.
- Answer the EDA questions using both:
    - Numerical Summaries – measures of central tendency, measures of dispersion, and correlation
    - Visualization – Appropriate visualization should be used. Each visualization should be accompanied by a brief explanation.

To emphasize, both numerical summary and visualization should be presented for each question. The whole process should be supported with verbose textual descriptions of your procedures and findings.

Research Question

Come up with one (1) research question to answer using the dataset. Here are some requirements:

- Important: The research question should arise from exploratory data analysis. There should be an explanation regarding the connection of the research question to the answers obtained from performing exploratory data analysis.
- The research question should be within the scope of the dataset.
- The research question should be answerable by performing data mining techniques (i.e., rule mining, clustering, collaborative filtering). Students cannot use other techniques that are not covered in class.
- Make sure to indicate the importance and significance of the research question.

Phase 2

The second phase of the case study involves three sections – (1) data modelling, (2) statistical inference, and (3) insights and conclusions.

Data Modelling

Perform the necessary steps in answering the research question that you have identified. In this section of the notebook, please take note of the following:

- If needed, perform preprocessing techniques to transform the data to the appropriate representation before performing modelling to answer the research question. This may include binning, log transformation, conversion to one-hot encoding, normalization, standardization, interpolation, truncation, and feature engineering.
- Tip: Some algorithms require the values to be scaled. Make sure to consider this before performing data modelling.
- You are encouraged to use the code that you have from your assignments.
- For rule mining: Rules should be generated from the dataset. Description of the rules based on the support and the confidence should be provided.

    - For example: {High population, High GDP} -> {First World Country} is one of the rules that we could extract from the dataset with the highest support and highest confidence.
- For clustering: Clusters should be generated from the dataset. Composition of each cluster should be described. Characterization of each cluster should be provided as well.
    - For example: Cluster 1 is mainly composed of countries in Asia. About 80% of the observations in cluster 1 are countries in Asia.
    - Another example: Cluster 2 has the highest average population count among all the clusters derived from the dataset. The average population count from this cluster is around 432,152,060.
- For collaborative filtering: Sample recommendations should be generated from the dataset. Imputed ratings should be provided as well.
- Use data modelling techniques that are discussed in class. The technique should be appropriate to answer the research question. Students cannot use other techniques that are not covered in class.

Statistical Inference

Perform hypothesis testing to support your answer to the research question. In this section of the notebook, please take note of the following:

- Use statistical inference methods discussed in class.
- Properly state both hypotheses.
- Important: Make sure to show that necessary assumptions and requirements about the statistical test and the data are checked. This will greatly affect the output of the statistical test.
- Show necessary pre-processing steps before computing for the p-value.
- Explicitly mention important values such as the resulting p-value and the significance level.

Tip: Note that there might be a need to check and prove if the data is from a normal distribution to perform some statistical inference techniques. This is especially true for performing statistical inference for means.

In some cases, statistical inference may be performed before data modelling.

Insights and Conclusions

Clearly state your insights and conclusions from the data to answer the research question. Make sure that the conclusion is backed up with statistical evidence using hypothesis testing.

Project Demos

Here are some guidelines regarding the final project presentation:

- Each group is given 40 minutes: 20 minutes to present, and 20 minutes for Q&A.
- Presentations will be done either online or face-to-face.
- Open your Jupyter Notebook before your allotted presentation time slot. Do not wait until the presentation itself to load anything.

- All members should be present and should discuss a part in the final project presentation.
- Kindly read the rubrics to check different requirements and expectations on the project presentation. Your time management skill is also being graded in the rubrics.

Provision for Bonus Points

Groups are encouraged to set-up a consultation with your instructor. During consultation, you need to discuss your partial work. Then, your instructor will give feedback to improve your work. Consultations should be made at least 1 week before the deadline to qualify for bonus points. Your instructor is allowed to give at most 4 points per phase as bonus point for consultation.

Working With Groupmates

For this project, you are encouraged to work in groups of at most 4 members. Make sure that each member of the group has approximately the same amount of contribution for the project. Problems with groupmates must be discussed internally within the group, and if needed, with the instructor.

Deliverables

Submit a `zip` file containing the source code files via AnimoSpace. All exploratory data analysis, data modelling, and core algorithms should be performed using Python 3 code and integrated into the Jupyter Notebook. Other code that you used for the project other than those in the Notebook should also be included in the submission of the project.

Academic Honesty Policy

Honesty policy applies. Please take note that you are NOT allowed to borrow and/or copy-and-paste – in full or in part – any existing related program code or solutions from the internet or other sources (such as printed materials like books, or source codes by other people that are not online). You should develop your own codes and solutions from scratch by yourselves.

The student handbook states that (Sec. 5.2.4.2):

"Faculty members have the right to demand the presentation of a student's ID, to give a grade of 0.0, and to deny admission to class of any student caught cheating under Sec. 5.3.1.1 to Sec. 5.3.1.1.6. The student should immediately be informed of his/her grade and barred from further attending his/her classes."

The student handbook also states that (Sec. 10.3):

A student caught cheating, as defined in Sec. 5.3.1.1., shall be penalized with a grade of 0.0 in the requirement or in the course, at the discretion of the faculty member, without prejudice to an administrative sanction. In cases of alleged cheating, the faculty member should report the incident to the Student Discipline Formation Office (SDFO).

Use of Generative AI

Generative AI tools may be used to assist the group in completing this assessment. However, AI tools should only be used to generate code snippets and should not be used to generate the whole code. More specifically, majority of the code (at least 80%) should still be written by the students. Non-code blocks, i.e., text blocks, should be entirely written by the students without the help of any AI tool.

Authors must disclose the use of generative AI and AI-assisted technologies in the writing process by adding a statement at the end of their manuscript in the core manuscript file, before the References list. The statement should be placed in a new section entitled 'Declaration of Generative AI and AI-assisted technologies in the writing process'.

Statement: During the preparation of this work the author(s) used [NAME TOOL/SERVICE] in order to [REASON]. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

This declaration does not apply to the use of basic tools for checking grammar, spelling, references etc. If there is nothing to disclose, there is no need to add a statement.

RUBRIC FOR GRADING

Phase 1

Criteria	Ratings				Points
Overview of the Dataset	**COMPLETE** **2 pts** Comprehensive overview of the dataset is provided in the notebook.	**INCOMPLETE** **1 pt** Overview of the dataset is provided but lacks details.	**NO MARKS** **0 pt** Overview of the dataset is not provided.		2 pts
Data Collection Method	**COMPLETE** **2 pts** The data collection process is explained in detail in the notebook.	**INCOMPLETE** **1 pt** The data collection process is provided but lacks details.	**NO MARKS** **0 pt** The data collection process is not provided.		2 pts
Implications of Data Collection Method	**COMPLETE** **2 pts** Implications of the data collection method on the conclusion of the study are properly explained in the notebook.	**INCOMPLETE** **1 pt** Implications of the data collection method on the conclusion of the study are provided but lacks details.	**NO MARKS** **0 pt** Implications of the data collection method is not provided.		2 pts
Description of Variables / Observations / Structure of the Data	**COMPLETE** **2 pts** A description of the variables, observations, and/or structure of the data is provided. It should be clear to the reader what each part of the dataset represents without having to go through external resources.	**INCOMPLETE** **1 pt** A description of variables, observations, and/or structure is present but is missing for some aspects of the dataset.	**NO MARKS** **0 pt** No overview or description of the variables / observation is provided.		2 pts
Sufficiency and Correctness of Data Cleaning	**COMPLETE** **6 pts** Preprocessing and cleaning are sufficiently and correctly performed	**INCOMPLETE** **4 pts** Preprocessing and cleaning are insufficiently or incorrectly performed	**INCOMPLETE** **2 pts** Preprocessing and cleaning are insufficiently or incorrectly performed	**NO MARKS** **0 pt** No preprocessing and cleaning are done, and no	6 pts

	for all used variables. If no preprocessing or cleaning is done, there should be a justification on why it is not needed.	for less than half or half of the number of used variables.	for more than half of the number of used variables.	justification is provided as to why it was not done.	
Justification and Description of Data Cleaning Methods	**COMPLETE** **6 pts** Justification and description for preprocessing and cleaning are properly and correctly provided for all used variables. If no preprocessing or cleaning is done, there should be a justification on why it is not needed.	**INCOMPLETE** **4 pts** Justification and description for preprocessing and cleaning are not properly and correctly provided for less than half or half of the number of used variables.	**INCOMPLETE** **2 pts** Justification and description for preprocessing and cleaning are not properly and correctly provided for more than half of the number of used variables.	**NO MARKS** **0 pt** Justification and description for preprocessing and cleaning are not provided at all.	6 pts
Exploratory Data Analysis – Numerical Summary	**COMPLETE** **5 pts** All exploratory data analysis questions are sufficiently answered with appropriate numerical summaries.	**INCOMPLETE** **3 pts** Less than half or half of the exploratory data analysis questions are not sufficiently answered with appropriate numerical summaries.	**INCOMPLETE** **1 pts** More than half of the exploratory data analysis questions are not sufficiently answered with appropriate numerical summaries.	**NO MARKS** **0 pt** Numerical summary is not computed at all.	5 pts
Exploratory Data Analysis – Visualization	**COMPLETE** **5 pts** All exploratory data analysis questions are sufficiently answered with appropriate visualizations.	**INCOMPLETE** **3 pts** Less than half or half of the exploratory data analysis questions are not sufficiently answered with appropriate visualizations.	**INCOMPLETE** **1 pts** More than half of the exploratory data analysis questions are not sufficiently answered with appropriate visualizations.	**NO MARKS** **0 pt** Visualization is not shown at all.	5 pts

Sufficiency and Correctness of Exploratory Data Analysis	**COMPLETE** **5 pts** EDA is sufficiently and correctly performed on the dataset to come up with a research question.		**INCOMPLETE** **2 pts** EDA is not sufficiently nor correctly performed on the dataset to come up with a research question.	**NO MARKS** **0 pt** EDA is not performed at all.	5 pts
Research Question	**COMPLETE** **5 pts** The research question is clearly defined, and the importance of the questions to the researcher and the community is explained convincingly. The research question arose from the EDA.		**INCOMPLETE** **2 pts** The research question is defined but either is not clear or its significance is not explained convincingly. The research question did not arise from the EDA.	**NO MARKS** **0 pt** The research question is not defined.	5 pts
Notebook	**COMPLETE** **5 pts** The notebook properly discusses all steps in the project.		**INCOMPLETE** **2 pts** The report failed to discuss some steps in the project.	**NO MARKS** **0 pt** No steps are discussed in the notebook.	5 pts
Demo Presentation	**COMPLETE** **5 pts** The presenter seldomly looks at notes. The presenter displays a relaxed, self-confident nature about self, with no mistakes.		**INCOMPLETE** **2 pts** The presenter looks at his notes most of the time. The presenter displays mild tension; has trouble recovering from mistakes.	**NO MARKS** **0 pt** The presenter reads the entire report from his notes. The presenter displays tension and nervousness; has trouble recovering from mistakes.	5 pts
Demo Time	**COMPLETE** **2 pts** The group was able to discuss all necessary contents of the notebook within the allotted time.			**NO MARKS** **0 pt** The group failed to discuss all necessary contents of the notebook within the allotted time.	2 pts
Demo Q&A	**COMPLETE** **8 pts** The group convincingly answered all questions	**INCOMPLETE** **5 pts** The group convincingly answered more than half	**INCOMPLETE** **2 pts** The group convincingly answered less than half of	**NO MARKS** **0 pt** The group failed to answer any question	8 pts

	about both the code and the data modelling process.	or half of the number of questions about both the code and the data modelling process.	the number of questions about both the code and the data modelling process.	about the code and the data modelling process.	
				Total points:	60

Phase 2

Criteria	Ratings			Points
Preprocessing for Data Modelling	**COMPLETE** **5 pts** Preprocessing steps are performed sufficiently as needed before data modelling. If no preprocessing is done, there should be a justification on why it is not needed.	**INCOMPLETE** **2 pts** Some preprocessing steps are not performed to prepare the data for the modelling technique to answer the research question.	**NO MARKS** **0 pt** Preprocessing is not performed at all. No justification is provided as to why preprocessing was not done.	5 pts
Appropriateness of the Data Modelling Technique	**COMPLETE** **2 pts** The data modelling technique used is appropriate to answer the research question.		**NO MARKS** **0 pt** The data modelling technique used is inappropriate to answer the research question.	2 pts
Correctness of Data Modelling Technique	**COMPLETE** **8 pts** The data modelling technique is applied in a sufficient and correct way.	**INCOMPLETE** **4 pts** The data modelling technique is applied in an insufficient or incorrect way.	**NO MARKS** **0 pt** No data modelling is done to answer the research question.	8 pts
Interpretation of Results	**COMPLETE** **5 pts** Rules / clusters / recommendations are generated in the notebook. Generated results are interpreted sufficiently and correctly.	**INCOMPLETE** **2 pts** Rules / clusters / recommendations are generated in the notebook. However, interpretation of generated results might be insufficient or incorrect.	**NO MARKS** **0 pt** Rules / clusters / recommendations are not generated in the notebook.	5 pts

Preprocessing for Statistical Inference	**COMPLETE** **3 pts** Preprocessing steps are performed sufficiently before statistical inference. If no preprocessing is done, there should be a justification on why it is not needed.		**NO MARKS** **0 pt** Necessary assumptions and requirements about the statistical test and the data are not checked. No justification is provided as to why preprocessing was not done.	3 pts
Applicability of Statistical Inference Method	**COMPLETE** **2 pts** The hypothesis testing method performed is appropriate and applicable to the data.		**NO MARKS** **0 pt** The hypothesis testing method performed is not appropriate nor applicable to the data.	2 pts
Hypotheses	**COMPLETE** **2 pts** Both hypotheses are stated correctly.	**INCOMPLETE** **1 pts** At least 1 hypothesis is not stated correctly.	**NO MARKS** **0 pt** Both hypotheses are stated incorrectly.	2 pts
Correctness of Statistical Inference Method	**COMPLETE** **5 pts** Hypothesis testing is correctly performed. The p-value is computed correctly.	**INCOMPLETE** **2 pts** Some aspects of the hypothesis testing method are incorrectly performed, e.g., wrong input, wrong arguments to parameters, etc.	**NO MARKS** **0 pt** The p-value is incorrectly computed.	5 pts
Correctness of Statistical Inference Conclusion	**COMPLETE** **3 pts** The conclusion of the hypothesis testing method is correctly and clearly stated.		**NO MARKS** **0 pt** The conclusion of the hypothesis testing method is incorrectly stated.	3 pts
Insights and Conclusion	**COMPLETE** **5 pts** The insights and conclusions to the research question are stated clearly and correctly and are backed up with statistical evidence.	**INCOMPLETE** **2 pts** The insights and conclusions to the research question are stated but not clearly enough, or statistical evidence is lacking.	**NO MARKS** **0 pt** No insights or conclusions are presented for the research question. The insights or conclusions are based on an inappropriate or incorrect data modelling technique or incorrect hypothesis testing method.	5 pts

Notebook	**COMPLETE** **5 pts** The notebook properly discusses all steps in the project.	**INCOMPLETE** **2 pts** The report failed to discuss some steps in the project.	**NO MARKS** **0 pt** No steps are discussed in the notebook.	5 pts
Demo Presentation	**COMPLETE** **5 pts** The presenter seldomly looks at notes. The presenter displays a relaxed, self-confident nature about self, with no mistakes.	**INCOMPLETE** **2 pts** The presenter looks at his notes most of the time. The presenter displays mild tension; has trouble recovering from mistakes.	**NO MARKS** **0 pt** The presenter reads the entire report from his notes. The presenter displays tension and nervousness; has trouble recovering from mistakes.	5 pts
Demo Time	**COMPLETE** **2 pts** The group was able to discuss all necessary contents of the notebook within the allotted time.		**NO MARKS** **0 pt** The group failed to discuss all necessary contents of the notebook within the allotted time.	2 pts
Demo Q&A	**COMPLETE** **8 pts** The group convincingly answered all questions about both the code and the data modelling process.	**INCOMPLETE** **5 pts** The group convincingly answered more than half or half of the number of questions about both the code and the data modelling process.	**INCOMPLETE** **2 pts** The group convincingly answered less than half of the number of questions about both the code and the data modelling process.	**NO MARKS** **0 pt** The group failed to answer any question about the code and the data modelling process.	8 pts
			Total points:	60