DEPARTMENT OF
SOFTWARE TECHNOLOGY

CSMODEL

Project – Case Study

Major Details

Groupings:	At most 3 members in a group
Deadline:	Phase 1 – November 21, 2022 (Monday) 8:00 AM
	Phase 2 – December 12, 2022 (Monday) 8:00 AM
Demo Schedule:	Phase 1 – November 21 to 26, 2022 (Week 11)
	Phase 2 – December 12 to 17, 2022 (Week 14)
Percentage:	40%
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 research questions, 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 codes could be successfully run sequentially to replicate the processes done in the project.

Phase 1

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

Dataset Description

Each group should select their own real-world dataset to analyze. When selecting a dataset, please ensure that the dataset is collected properly. The dataset should contain enough variables to explore. As a rule of thumb, a good number would be at least 20 variables (could be actual features from the original dataset or generated features).

There are several online sources for public online datasets. Some of them are as follows:

- Kaggle (https://www.kaggle.com/datasets)
- Google Public Datasets (https://cloud.google.com/bigquery/public-data/)
- Our World in Data (https://ourworldindata.org)

Datasets from other sources aside from the ones listed above may also be used. Note that each group in a section should work on a different dataset. A sign-up sheet will be provided by your instructor to track all datasets reserved by all groups per section.

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 be able to provide the necessary information for this part of the project.
- Describe the structure of the dataset file. In 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? In this section, 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.

Dataset Description

For each used variable, check for 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 three (3) exploratory data analysis questions. Properly state the questions in the notebook.
- 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:

- The research question should arise from the 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 either performing data mining techniques or any domain-specific data modelling technique (i.e., techniques in modelling text, time-series, graph, or image data) taught in class.
- Make sure to indicate the importance and significance of the research question.

Phase 2

The second phase of the case study involves the 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 transformations, conversion to one-hot encoding, normalization, standardization, interpolation, truncation, and feature engineering.
- Use data modelling techniques that are discussed in class. The technique should be appropriate to answer the research question.

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 the hypotheses.
- 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.

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.

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.

Working With Groupmates

For this project, you are encouraged to work in groups of at most 3 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 lecturer.

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).

RUBRIC FOR GRADING

Phase 1

Criteria	Ratings			Points
Description of Data and Method of Collection	**COMPLETE** **2 pts** An overview or description of the data is provided in the Notebook, including how it was collected, and its implications on the types of conclusions that could be made from the data.	**INCOMPLETE** **1 pts** An overview or description is provided but lacks details, or the description does not include how the data was collected and its implications to the conclusion.	**NO MARKS** **0 pt** No overview or description of the data is provided.	2 pts
Description of Variables / Observations / Structure of the Data	**COMPLETE** **3 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 pts** 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 data is provided.	3 pts
Data Cleaning	**COMPLETE** **10 pts** The necessary steps for preprocessing and cleaning are performed, including explanations for every step. If no preprocessing or cleaning is done, there should be a	**INCOMPLETE** **5 pts** Preprocessing and cleaning steps are performed but lacks explanation. Or, preprocessing and cleaning done are insufficient for the dataset.	**NO MARKS** **0 pt** No preprocessing and cleaning are done, and no justification is provided as to why it was not done, or the justification is weak or incorrect.	10 pts

	justification on why it is not needed.			
Exploratory Data Analysis 1	**COMPLETE** **10 pts** The first exploratory data analysis question is sufficiently answered, and the appropriate numerical summaries and visualizations are presented.	**INCOMPLETE** **5 pts** The first exploratory data analysis question is not sufficiently answered, or the appropriate numerical summaries or visualizations are not presented.	**NO MARKS** **0 pt** There is no analysis done for the first exploratory data analysis question.	10 pts
Exploratory Data Analysis 2	**COMPLETE** **10 pts** The second exploratory data analysis question is sufficiently answered, and the appropriate numerical summaries and visualizations are presented.	**INCOMPLETE** **5 pts** The second exploratory data analysis question is not sufficiently answered, or the appropriate numerical summaries or visualizations are not presented.	**NO MARKS** **0 pt** There is no analysis done for the second exploratory data analysis question.	10 pts
Exploratory Data Analysis 3	**COMPLETE** **10 pts** The third exploratory data analysis question is sufficiently answered, and the appropriate numerical summaries and visualizations are presented.	**INCOMPLETE** **5 pts** The third exploratory data analysis question is not sufficiently answered, or the appropriate numerical summaries or visualizations are not presented.	**NO MARKS** **0 pt** There is no analysis done for the third exploratory data analysis question.	10 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
			Total points:	50

Phase 2

Criteria	Ratings			Points
Data Modelling	**COMPLETE** **20 pts** The appropriate data modelling technique is used to answer the research question.	**INCOMPLETE** **10 pts** The data modelling technique that is used to answer the research question is applied in an insufficient way. Some preprocessing steps are not performed to prepare the data for the modelling technique to answer the research question.	**NO MARKS** **0 pt** No data modelling is done to answer the research question.	20 pts
Statistical Inference	**COMPLETE** **20 pts** Appropriate and applicable hypothesis testing is performed correctly to support the answer to the research question.	**INCOMPLETE** **10 pts** Hypothesis testing is either applied incorrectly or insufficiently.	**NO MARKS** **0 pt** No hypothesis testing is done to support the answer to the research question.	20 pts
Insights and Conclusion	**COMPLETE** **10 pts** The insights and conclusions to the research question are stated clearly and backed up with statistical evidence.	**INCOMPLETE** **5 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 data modelling technique applied to answer the research question.	10 pts
			Total points:	50