
DEPARTMENT OF
**SOFTWARE TECHNOLOGY**

# CSMODEL

**Project – Case Study**

## Major Details

| | |
|---|---|
| **Groupings:** | At most 4 members in a group |
| **Deadline:** | Phase 1 – October 20, 2023 (Friday) 6:00 PM |
| | Phase 2 – November 17, 2023 (Friday) 6:00 PM |
| **Demo Schedule:** | Phase 1 – October 23 – 27, 2023 (Week 8) |
| | Phase 2 – November 20 – 24, 2023 (Week 12) |
| **Percentage:** | Phase 1 – 25% |
| | Phase 2 – 25% |
| **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 their own real-world dataset to analyze. When selecting a dataset, please ensure that it is collected properly. The dataset should contain enough variables to explore. Datasets with around 10 to 20 variables are recommended. Datasets with less than or more than this recommended count can still be used.

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. You may check a list of recommended datasets at the last part of this document. 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 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 either performing data mining techniques (i.e., rule mining, clustering, association rule mining) or any domain-specific data modelling technique (i.e., techniques in modelling text, time-series, graph, or image data) taught in class. 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.
- 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.

## 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 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).*

**Sample List of Datasets**

- Complete Pokemon Dataset (Updated 16.04.21)
- The Nutritional Content of Food
- Spotify - All Time Top 2000s Mega Dataset
- Video Game Sales and Ratings
- Filipino Family Income and Expenditure
- OECD PISA 2018
- Anime Recommendation Database 2020
- The Movies Dataset
- Book Recommendation Dataset
- Board Game Database from BoardGameGeek
- Sales Transaction
- Retail Store Sales Transactions (Scanner Data)
- Diamonds
- Diabetes Dataset
- Heart Disease Dataset
- Breast Cancer Dataset
- Red Wine Quality Dataset

**RUBRIC FOR GRADING**

## Phase 1

| Criteria | Ratings | | | | Points |
|---|---|---|---|---|---|
| **Description of Data and Method of Collection** | **COMPLETE**<br>**5 pts**<br><br>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**<br>**2 pts**<br><br>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**<br>**0 pt**<br><br>No overview or description of the data is provided. | | 5 pts |
| **Description of Variables / Observations / Structure of the Data** | **COMPLETE**<br>**5 pts**<br><br>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**<br>**2 pts**<br><br>A description of variables, observations, and/or structure is present but is missing for some aspects of the dataset. | **NO MARKS**<br>**0 pt**<br><br>No overview or description of the data is provided. | | 5 pts |
| **Data Cleaning** | **COMPLETE**<br>**10 pts**<br><br>The necessary steps for preprocessing and cleaning are performed, including explanations for every step for each used variable. If no preprocessing or cleaning is done, there should be a justification on why it is not needed. | **INCOMPLETE**<br>**7 pts**<br><br>Preprocessing and cleaning steps are performed but lacks explanation. Or, preprocessing and cleaning done are insufficient for less than half or half of the number of used variables. | **INCOMPLETE**<br>**3 pts**<br><br>Preprocessing and cleaning steps are performed but lacks explanation. Or, preprocessing and cleaning done are insufficient for more than half of the number of used variables. | **NO MARKS**<br>**0 pt**<br><br>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 |

| | | | | | |
|---|---|---|---|---|---|
| **Exploratory Data Analysis** | **COMPLETE**<br>**15 pts**<br><br>All exploratory data analysis questions are sufficiently answered, and the appropriate numerical summaries and visualizations are presented.<br>EDA is sufficiently and correctly performed on the dataset to come up with a research question. | **INCOMPLETE**<br>**10 pts**<br><br>Less than half or half of the exploratory data analysis questions are not sufficiently answered, or the appropriate numerical summaries or visualizations are not presented.<br>EDA is not sufficiently performed on the dataset to come up with a research question. | **INCOMPLETE**<br>**5 pts**<br><br>More than half of the exploratory data analysis questions are not sufficiently answered, or the appropriate numerical summaries or visualizations are not presented.<br>EDA is not sufficiently performed on the dataset to come up with a research question. | **NO MARKS**<br>**0 pt**<br><br>EDA is not performed at all. | 15 pts |
| **Research Question** | **COMPLETE**<br>**5 pts**<br><br>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**<br>**2 pts**<br><br>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**<br>**0 pt**<br><br>The research question is not defined. | | 5 pts |
| **Demo Q&A** | **COMPLETE**<br>**10 pts**<br><br>The group convincingly answered all questions about both the code and the data modelling process. | **INCOMPLETE**<br>**7 pts**<br><br>The group convincingly answered more than half or half of the number of questions about both the code and the data modelling process. | **INCOMPLETE**<br>**3 pts**<br><br>The group convincingly answered less than half of the number of questions about both the code and the data modelling process. | **NO MARKS**<br>**0 pt**<br><br>The group failed to answer any question about the code and the data modelling process. | 10 pts |
| | | | | **Total points:** | 50 |

**Phase 2**

| Criteria | Ratings | | | | Points |
|---|---|---|---|---|---|
| **Data Modelling** | **COMPLETE**<br>**18 pts**<br>The appropriate data modelling technique is used to answer the research question. Preprocessing steps are performed sufficiently. | **INCOMPLETE**<br>**12 pts**<br>Some preprocessing steps are not performed to prepare the data for the modelling technique to answer the research question. | **INCOMPLETE**<br>**6 pts**<br>The data modelling technique that is used to answer the research question is applied in an insufficient way. Or, the data modelling technique is not appropriate for the data. | **NO MARKS**<br>**0 pt**<br>No data modelling is done to answer the research question. | 18 pts |
| **Statistical Inference** | **COMPLETE**<br>**15 pts**<br>Appropriate and applicable hypothesis testing is performed correctly to support the answer to the research question. Preprocessing steps are performed sufficiently. | **INCOMPLETE**<br>**10 pts**<br>Necessary assumptions and requirements about the statistical test and the data are not checked. | **INCOMPLETE**<br>**5 pts**<br>Hypothesis testing is either applied incorrectly or insufficiently. Or, hypothesis testing is not appropriate for the data. | **NO MARKS**<br>**0 pt**<br>No hypothesis testing is done to support the answer to the research question. | 15 pts |
| **Insights and Conclusion** | **COMPLETE**<br>**5 pts**<br>The insights and conclusions to the research question are stated clearly and backed up with statistical evidence. | **INCOMPLETE**<br>**2 pts**<br>The insights and conclusions to the research question are stated but not clearly enough, or statistical evidence is lacking. | **NO MARKS**<br>**0 pt**<br>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. | | 5 pts |
| **Demo Q&A** | **COMPLETE**<br>**12 pts**<br>The group convincingly answered all questions about both the code and the data modelling process. | **INCOMPLETE**<br>**7 pts**<br>The group convincingly answered more than half or half of the number of questions about both the code and the data modelling process. | **INCOMPLETE**<br>**3 pts**<br>The group convincingly answered less than half of the number of questions about both the code and the data modelling process. | **NO MARKS**<br>**0 pt**<br>The group failed to answer any question about the code and the data modelling process. | 12 pts |
| | | | | **Total points:** | 50 |