# CSMODEL Case Study Specifications

## Case Study Overview

For your case study, you will be going through the process of selecting a dataset, formulating research questions, analyzing data, modelling data, and extracting insights from the data. The case study is an expression of what you have learned in CSMODEL, not only in terms of the application of specific techniques and algorithms, but more importantly in terms of your understanding of the intuitions behind data modelling.

## Groupings

This project is to be accomplished by group, with each group having a maximum of three members. The group assignments are to be determined on a per section basis; please consult with the instructor.

## Dataset

Each group should select their own real-world dataset to work with. When selecting a dataset, please ensure that the dataset was collected properly. Please make sure that the dataset contains enough variables for you to play around and explore in your analysis. As a rule of thumb, a good number would be at least 10 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:

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

You may explore other sources aside from the ones listed above.

## Project Requirements

The project is to be submitted as a Jupyter Notebook and, optionally, some Python 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. The Notebook 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.

The final output for your project should include the following information clearly.

1. **Data Description**. Describe how the dataset was collected and the implications of the data collection method on the generated conclusions and insights. Note that you may need to look at the relevant sources related to the dataset you are

working on to be able to provide the necessary information for this part of the project.

2. **Exploratory Data Analysis.** Perform data cleaning (if needed) and perform exploratory data analysis. This part of the project should include numerical summaries and visualizations whenever appropriate. Each visualization should be accompanied by a brief explanation. The exploratory data analysis should guide you in formulating the research questions in the next step.

   You should perform exploratory data analysis comprehensively to gain a good understanding of your dataset. However, for the Notebook, please choose only 3 of the most interesting exploratory questions that you did for your dataset. For these exploratory questions, please include numerical summaries and visualizations that address these questions along with textual descriptions of your processes and findings.

3. **Research Question**. Come up with at least 2 research questions that you want to answer using the dataset. The 2 research questions should be answerable using different methods (i.e., do not choose 2 questions that could be answered using the exact same approach but just changing some of the parameters). You should select research questions that are within the scope of the dataset you are working on. For each research question, you must indicate why this question is of interest to you and the community.

4. **Data Modelling**. Perform the necessary steps in answering each of the research questions you have identified. This includes cleaning and validating the data, performing exploratory data analysis, and applying the appropriate data modelling technique for the dataset and each of the research questions that you aim to address.

   For this step, please take note of the following:
   - The data modelling approaches and techniques that you can use include but **are not limited to** making inference, data mining techniques, text analysis, time series analysis, graph analysis, and image analysis.
   - Feel free to explore techniques that are not directly discussed within CSMODEL.

5. **Insights and Conclusions.** Clearly state your insights and conclusions from the data to answer each research question you have defined. Make sure that all conclusions are backed up with statistical evidence when necessary.

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.

**Submissions**

The following are the due dates for this project.

***By September 1, 2020 (Saturday), 11:59 PM***, you are to submit a brief description of the project you intend to make via AnimoSpace. This includes the following information:

- The dataset that you intend to use
- The potential research questions that you are going to explore

***By September 22, 2020 (Thursday), 11:59 PM***, all projects should be submitted via AnimoSpace. When submitting the project, please submit the following files:

- the .ipynb file containing the main content for the project
- the dataset file/s used in the Notebook (if the datasets are too large to upload, please upload them to Google drive and include a link to download them in the Notebook instead)
- all other Python source files used in the Notebook

**Rubric for Grading**

The following rubric will be used for grading the case study. Please use this as reference as well to make sure that your project is able to comply with all the requirements.

| **Criteria** | **Full Marks** | **Partial Marks** | **No Marks** |
|---|---|---|---|
| **Description of Data and Method of Collection** | **5**<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. | **3**<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 conclusions. | **0**<br>No overview or description of the data is provided. |
| **Description of Variables / Observations / Structure of the Data** | **5**<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. | **3**<br>A description of variables, observations, and/or structure is present but is missing for some aspects of the dataset. | **0**<br>No description of variables, observations, and/or structure is provided. |
| **Preprocessing and Cleaning** | **10**<br>The necessary steps for preprocessing | **7 or 4**<br>Preprocessing and cleaning steps were | **0**<br>No preprocessing and cleaning were |

| | | | |
|---|---|---|---|
| | and cleaning are performed, including explanations for every step. If no preprocessing or cleaning is done, there should be a justification on why it is not needed. | performed but lacks explanation. Or, preprocessing and cleaning done was insufficient for the dataset. | done, and no justification was provided as to why it was not done, or the justification is weak or incorrect. |
| **Exploratory Data Analysis 1** | **10**<br>The first exploratory data analysis question was sufficiently answered, and the appropriate numerical summaries and visualizations were presented. | **7 or 4**<br>The first exploratory data analysis question was not sufficiently answered, or the appropriate numerical summaries or visualizations were not presented. | **0**<br>There was no analysis done for the first exploratory data analysis question. |
| **Exploratory Data Analysis 2** | **10**<br>The second exploratory data analysis question was sufficiently answered, and the appropriate numerical summaries and visualizations were presented. | **7 or 4**<br>The second exploratory data analysis question was not sufficiently answered, or the appropriate numerical summaries or visualizations were not presented. | **0**<br>There was no analysis done for the second exploratory data analysis question. |
| **Exploratory Data Analysis 3** | **10**<br>The third exploratory data analysis question was sufficiently answered, and the appropriate numerical summaries and visualizations were presented. | **7 or 4**<br>The third exploratory data analysis question was not sufficiently answered, or the appropriate numerical summaries or visualizations were not presented. | **0**<br>There was no analysis done for the third exploratory data analysis question. |
| **Research Question 1** | **5**<br>The first research question was clearly defined, and the importance of the questions to the researcher and the | **3**<br>The first research question was defined but either was not clear or its significance was not | **0**<br>The first research question was not defined. |

| | | | |
|---|---|---|---|
| | community is explained convincingly. | explained convincingly. | |
| **Research Question 2** | **5**<br>The second research question was clearly defined, and the importance of the questions to the researcher and the community is explained convincingly. | **3**<br>The second research question was defined but either was not clear or its significance was not explained convincingly. | **0**<br>The second research question was not defined. |
| **Data Modelling and Algorithms 1** | **10**<br>The appropriate data modelling and techniques were used to answer the first research question. | **7 or 4**<br>The data modelling and techniques that were used to first research question has some idea to it but were applied in an insufficient way. | **0**<br>No data modelling was done to answer the first research question. |
| **Data Modelling and Algorithms 2** | **10**<br>The appropriate data modelling and techniques were used to answer the second research question. | **7 or 4**<br>The data modelling and techniques that were used to second research question has some idea to it but were applied in an insufficient way. | **0**<br>No data modelling was done to answer the second research question. |
| **Insights and Conclusions 1** | **10**<br>The insights and conclusions to the first research question were stated clearly and backed up with statistical evidence when needed. | **7 or 4**<br>The insights and conclusions to the first research question were stated but not clearly enough, or some statistical evidence is lacking. | **0**<br>No insights or conclusions were presented for the first research question. |
| **Insights and Conclusions 2** | **10**<br>The insights and conclusions to the second research question were stated clearly and backed up with statistical evidence when needed. | **7 or 4**<br>The insights and conclusions to the second research question were stated but not clearly enough, or some statistical evidence is lacking. | **0**<br>No insights or conclusions were presented for the second research question. |

Prepared by: Thomas James Tiam-Lee and Arren Antioquia