CSMODEL Term 3, AY 2020 – 2021

Project 1 Specifications – Statistical Inference

Groupings: 3 members in a group
Deadline: August 10, 2021 (Tuesday) 11:59 PM
Percentage: 20% (part of the 40% for Projects)

Deliverables:

`Zip` file containing:

- Jupyter Notebook file – `ipynb` file
- Other Python 3 files (if necessary) – `py` files
- Dataset files – `csv` files

Submission guidelines: Submit the `zip` file to AnimoSpace
Filename format: `CSMODEL-Project1-<Section>-Group<#>.zip`

SPECIFICATIONS

You are tasked to go through the process of selecting a dataset, formulating research questions, analyzing data, modelling data, 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.

Dataset Representation

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). The dataset should contain both numerical and categorical variables.

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.

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 the 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, you need to 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.

Data Cleaning

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. The exploratory data analysis should guide you in formulating the research questions of the project.

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

- Identify 3 interesting 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 present to answer each question. The whole process should be supported with verbose textual descriptions of your procedures and findings.

Research Question

Come up with research questions to answer using the dataset. The research questions should arise from the exploratory data analysis.

- The first research question should be answerable by performing statistical inference on means.
- The second research question should be answerable by performing statistical inference on categorical data.
- The research questions should be within the scope of the dataset.
- For each research question, you must indicate its importance and significance to the community.

Statistical Inference

Perform the necessary steps in answering each of the research questions.

In this section of the notebook, please take note of the following:

- If needed, you may perform preprocessing techniques to transform the data to the appropriate representation before performing statistical inference to answer the research question. This may include binning, log transformations, conversion to one-hot encoding, normalization, standardization, interpolation, truncation, and feature engineering. You may also need to check and prove if the data is from a normal distribution to perform some statistical inference techniques.
- The techniques used to answer the research question are limited to statistical inference methods discussed in class.
- The technique that you will apply should be appropriate to answer the research question.

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.

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.

HONESTY POLICY AND INTELLECTUAL PROPERTY RIGHTS

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 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 from scratch by yourselves, i.e., in cooperation with your groupmates.

According to the handbook (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."

RUBRIC FOR GRADING

Criteria	Ratings			Points
Description of Data and Method of Collection	**COMPLETE** **5 pts** An overview or description of the data is provided in the Notebook, including the data collection process, and its implications on the types of conclusions that could be made from the data.	**INCOMPLETE** **2 pts** An overview or description is provided but lacks details, or the description does not include the data collection process and its implications to the conclusions.	**NO MARKS** **0 pt** No overview or description of the data is provided.	5 pts
Description of Variables / Observations / Structure of the Data	**COMPLETE** **5 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** **2 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.	5 pts

Data Cleaning	**COMPLETE** **10 pts** The necessary steps for cleaning are performed, including explanations for every step. If no data cleaning is done, there should be a justification on why it is not needed.	**INCOMPLETE** **5 pts** Preprocessing and cleaning steps are performed but lacks explanation, or the cleaning done is insufficient for the dataset.	**NO MARKS** **0 pt** No cleaning is 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 1	**COMPLETE** **10 pts** The first exploratory data analysis question is sufficiently answered, and both 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 is 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 both 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 is 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 both 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 is not presented.	**NO MARKS** **0 pt** There is no analysis done for the third exploratory data analysis question.	10 pts
Research Question 1	**COMPLETE** **5 pts** The first research question is clearly defined. The importance to the researcher and the community is explained convincingly. The research question arose from the EDA.	**INCOMPLETE** **2 pts** The first 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 first research question is not defined.	5 pts
Research Question 2	**COMPLETE** **5 pts** The second research question is clearly defined. The importance to the researcher and the community is explained convincingly. The research question arose from the EDA.	**INCOMPLETE** **2 pts** The second 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 second research question is not defined.	5 pts

Statistical Inference 1	**COMPLETE** **10 pts** The appropriate statistical inference technique is used to answer the first research question.	**INCOMPLETE** **5 pts** The statistical inference technique that is used to answer first research question is applied in an insufficient way. Some preprocessing steps are not performed to prepare the data for statistical inference to answer the first research question.	**NO MARKS** **0 pt** Statistical inference is not performed to answer the first research question.	10 pts
Statistical Inference 2	**COMPLETE** **10 pts** The appropriate statistical inference technique is used to answer the second research question.	**INCOMPLETE** **5 pts** The statistical inference technique that is used to answer second research question is applied in an insufficient way. Some preprocessing steps are not performed to prepare the data for statistical inference to answer the second research question.	**NO MARKS** **0 pt** Statistical inference is not performed to answer the second research question.	10 pts

Insights and Conclusion 1	**COMPLETE** **10 pts** The insights and conclusions to the first research question are stated clearly and backed up with statistical evidence when needed.	**INCOMPLETE** **5 pts** The insights and conclusions to the first research question are stated but not clearly enough, or some statistical evidence is lacking.	**NO MARKS** **0 pt** No insights or conclusions are presented for the first research question. The insights or conclusions are based on an inappropriate statistical inference technique applied to answer the first research question.	10 pts
Insights and Conclusion 2	**COMPLETE** **10 pts** The insights and conclusions to the second research question are stated clearly and backed up with statistical evidence when needed.	**INCOMPLETE** **5 pts** The insights and conclusions to the second research question are stated but not clearly enough, or some statistical evidence is lacking.	**NO MARKS** **0 pt** No insights or conclusions are presented for the second research question. The insights or conclusions are based on an inappropriate statistical inference technique applied to answer the second research question.	10 pts
			Total points:	100