CDS6214 Data Science Fundamentals

Project (40%)

The aim of this project is to propose a problem that can be solved by undergoing the data science process. You are to be involved in the entire process from figuring out the problem and question you intend to ask, understand, collect data, perform cleaning/pre-processing, explore the data, build data-driven models and visualize the data meaningfully.

General Instructions

1. This is a group project to be completed in a group of a max of 4 students within the SAME tutorial section. Please refer to your tutor on instructions to register as a group for Project.

2. Based on your tutorial section, your group can only work on dataset within the domain that has been assigned. The original dataset prior to data pre-processing should be of at least 2000 rows.

Tutorial Section	Domain
TT1L	Agriculture
TT2L	Accommodation and Tourism
TT3L	Financial / Banking
TT4L	Retail
TT5L	Nutrition and Diets
TT6L	Sports
TT7L	Climate and Environment

3. With the assigned domain, propose a *problem* to solve and to uncover insights from your data. Under that main problem, design four questions that you would like to answer (note: go for a variety of different types of questions).

4. Submission and evaluated components will be in the form of:
   a. *Written– softcopy*
   Note: Your report should not be more than 15 pages. The Cover Page, Task Distribution Page, Table of Content and References are not counted as part of the 15 pages. Please use the template cover page. Include the Task Distribution Page immediately after the Cover Page and indicate as specific as possible, the tasks done by each member of the group. For References, please use APA style. Excluding figures and tables, the content of your report should be written within Normal margin with Times New Roman, font size 12.

b. *Supporting materials (e.g. dataset, additional datasets, codes etc.) – softcopy.*
   Note: You are REQUIRED to use **Python** for code development and data visualization. Make sure that all the codes are clearly documented and consolidated in a Python Notebook

c. *Presentation*
   Note: Record your group's presentation, upload it to YouTube and kept as unlisted. The maximum duration is 10 minutes per group. All members in the group must do presentation and introduce themselves during the presentation. The YouTube link should be included on the Cover Page of the report.

Softcopy submissions in a zipped file are to be done via **eBwise** by Wednesday of Week 14. Only one submission is required per group i.e.: elect one of the group members to submit the zipped file. Late submission is acceptable with penalty up to Monday of Week 15. After Monday of Week 15, zero marks will be awarded for this Project.

Be aware that plagiarism is a serious offence. Cite all your references! This includes, but not limited to:

- Materials taken from websites, articles,
- Research papers, books,
- Images, videos (YouTube etc.) and other media.

Penalties

- 3 marks will be deducted for each day late after the deadline.
- 0 mark will be awarded for this Project if the content of this Project is plagiarised from any sources
- 0 mark will be awarded for this Project if the group submit the Project after Monday of Week 15.
- 3 marks will be deducted for the video that exceeds 10 minutes
- 3 marks will be deducted for exceeding 15 pages of the report excluding cover page and references
- 3 marks will be deducted for submitting a slide deck that exceeds 10 slides.
- 3 marks deducted for not having a Cover Page.
- 20 marks will be deducted for working with data in different domain other than being assigned per tutorial section

Evaluation Mark Breakdown

Deliverable		**Marks (/40)**	**Totals**
Report	- Motivation, Problem - Impact on the Proposed Problem to be solved to communities / society / nation - Description and implementation of Pipeline solutions for each DS process component. o (Questions, Data collection, Data pre-processing, EDA, Data mining/data modelling – analysis / assessments of solutions, Data visualization) o Snapshots of codes and output can be included to support description. - Challenges encountered and limitation of current work - Organization, Clarity and Language - References	22	
Code	- Documentation within the Python Notebook - Code implementation of Pipeline solutions for each DS process component - Deployment o How to connect to your web API? o Hosted streamlit? o Real-time prediction and visualization?	8	
Presentation	- Oral Presentation - Visual content & impact	8	
Others	Supporting materials (dataset, additional datasets, etc)	2	
	TOTAL		40

Note

Presentation component is assessed individually. For Report, Code and Others, these components are assessed as a group.