Project 2 (25%)
STQD6014
SEMESTER 1 2024/2025

As a Data Scientist, your primary role is to uncover and communicate the narratives hidden within datasets, delivering insights to stakeholders through impactful visualizations. This assignment challenges you to demonstrate these skills while exploring a topic relevant to the industry you aim to join after completing the MASTER OF SCIENCE (DATA SCIENCE AND ANALYTICS) program. Choose a publicly available open-source dataset that aligns with your chosen industry or area of interest. Your tasks encompass:

1. Data Cleaning:

- Thoroughly clean the **RAW DATA** to ensure accuracy and reliability in subsequent analyses.

2. Visualizations:

- Create a **minimum of five distinct visualization plots**, employing techniques such as pie charts, bar plots, box plots, etc.
- Utilize **Matplotlib**, **Seaborn** or other relevant packages of your choice for creating these visualizations.

3. Insights and Explanations:

- Accompany each plot with a **compelling insight and explanation** that is both relevant and significant. Provide clarity on the story the data tells. You are encouraged to substantiate your observations with other published information online.

4. **Google Colab Notebook**:

- Use the **Google Colab Notebook** as the platform for your assignment.
- Structure the Notebook like a book, encompassing sections such as:
    - **Introduction**: Present the purpose and context of the analysis.
    - **Problem Statement**: Clearly define the problem or question you aim to address.
    - **Results and Discussion**: Showcase your visualizations, insights, and explanations.
    - **Conclusion**: Summarize key findings and any implications.

5. Tools:

- You have the flexibility to choose relevant packages and tools for analysis, with a suggestion to consider **Seaborn** for visualization.

6. Deadline:

- The submission deadline for the Notebook is set for **2025-01-26**.

7. Submission:

- Share your **Google Colab Notebook** by adding me as a collaborator on GitHub using the email: bernardlkb@ukm.edu.my.

Approach this assignment with the goal of producing a comprehensive and well-organized document. Focus on making your analysis clear, accessible and engaging for stakeholders. **Remember to state clearly the link to your selected datasets for reproducibility.**

Criteria	Marks		
Reproducibility	3 The notebook is 100% reproducible	2 The notebook is reproducible with a few missing steps	1 The notebook is not reproducible
Plots	20 All the plots are i. suitable, ii. easy to understand, iii. observations are properly explained	15 Some of the plots are i. suitable, ii. easy to understand, iii. observations are properly explained	10 The plots are i. not suitable, ii. hard to understand, iii. observations are poorly explained
Notebook presentation	2 The overall notebook is i. properly structured, ii.each section neatly organized, iii. easy to follow	1 Part of the notebook is i. properly structured, ii.each section neatly organized, iii. easy to follow	0 The notebook is i. poorly structured, ii. each section is not organized, iii. hard to follow