Assignment 1 (25%)

STQD6324 Data Management

SEMESTER 2 2024/2025

As you work towards becoming a proficient Data Scientist, it is essential to develop strong **DATA MANAGEMENT** skills, which is an integral component of your training. For this assignment, you are required to **select a topic** related to the **industry you aspire to join** upon completing your **MASTER OF SCIENCE (DATA SCIENCE AND ANALYTICS)** program. You may refer to resources such as "Top Big Data Analytics Companies in Malaysia" or other related media to help identify a suitable industry.

Once you have chosen your target industry, proceed to identify and access **open-source online databases** containing **RAW DATASETS** relevant to that domain. A list of suggested data sources is provided in the **Appendix** below.

You are required to use tools such as **Apache Hive** or **Apache Pig**, as well as **R (via Rmarkdown)** or **Python (via Google Colab)**, or any other tools that you have learnt to complete this assignment. To prepare for future job interviews and to effectively showcase your skills set, you must publish your project on **GitHub**. Your GitHub repository should include the following sections:

- **Data Cleaning**
- **Data Visualizations**
- **Insights and Explanations**
- **Recommendations**
- **Conclusion**
- **Any additional elements that support your analysis**

You are encouraged to organize and present your GitHub project **creatively and professionally**. The submission deadline is **2025-05-17**. Kindly share your completed notebook by adding me as a GitHub collaborator at bernardlkb@ukm.edu.my.

Open-source Online Databases:

1. UNData [https://data.un.org/]
2. Amazon AWS Dataset [https://registry.opendata.aws/]
3. Google Dataset [https://datasetsearch.research.google.com/]
4. Awesome Public Data Sources [https://github.com/awesomedata/awesome-public-datasets]
5. Country Codes List [https://www.nationsonline.org/oneworld/country_code_list.htm#S]
6. Spotify [https://research.atspotify.com/datasets/]
7. Tableau Public Data Sets [https://public.tableau.com/app/learn/sample-data]
8. UC Irvine Machine Learning Repository [https://archive.ics.uci.edu/]
9. United Nations Children's Fund (UNICEF) [https://iatiregistry.org/publisher/unicef]
10. US Census Bureau [https://data.census.gov/]
11. USA Open Data [https://data.gov/]
12. Wikipedia Data Set [https://www.dbpedia.org/]
13. Worldbank dataset [https://data.worldbank.org/]
14. World Health Organization [https://www.who.int/data/gho/]
15. Yelp Dataset [https://business.yelp.com/data/resources/open-dataset/]

Reproducibility	3 The notebook is 100% reproducible	2 The notebook is reproducible with a few missing steps	1 The notebook is not reproducible
Plots	**10** All the plots are i. suitable, ii. easy to understand iii. observations are properly explained	5 Some of the plots are i. suitable, ii. easy to understand iii. observations are properly explained	3 The plots are i. not suitable, ii. hard to understand iii. observations are poorly explained
Style & Clarity	**5** The article is written in an engaging style and tone; free of grammatical and spelling errors	3 The article is written in an engaging style and tone; some grammatical and spelling errors	1 The article is not written in an engaging style and tone; lots of grammatical and spelling errors
Overall GitHub presentation	2 The overall GitHub is i. properly structured, ii.each section neatly organized, iii. easy to follow	1 Part of the GitHub is i. properly structured, ii.each section neatly organized, iii. easy to follow	0 The GitHub is i. poorly structured, ii. each section is not organized, iii. hard to follow