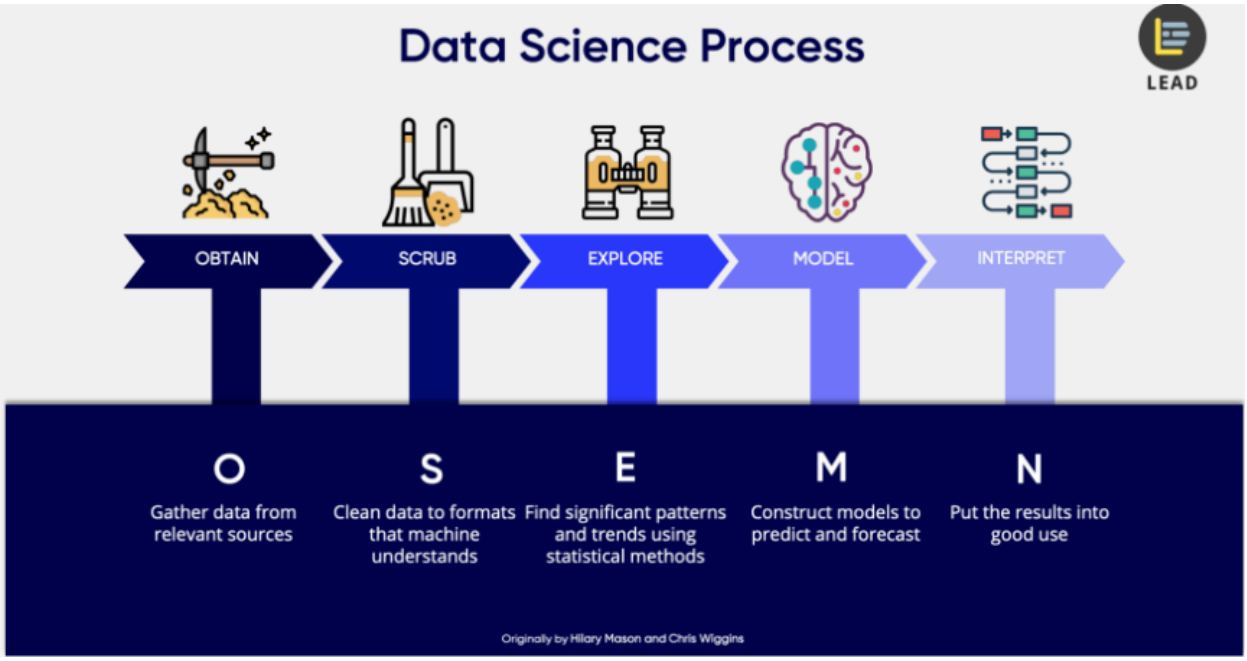
WIA1007/WIE2003 : INTRODUCTION to DATA SCIENCE

Group Assignment (GA- 40 %)

This assignment is divided into 2 parts (GA1 and GA2)

Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. The first task of data scientist is to understand the objectives and requirements of the project. It is important to determine business objectives and define business success criteria. In other words, what should the project try to achieve? The following diagram shows a data science process flow, using OSEMN framework.



As a data scientist , you may do the following tasks : i) Discover patterns and trends in datasets to get insights; ii) Create forecasting algorithms and data models.; iii) Improve the quality of data or product offerings by utilising suitable DS tools and machine learning techniques and finally become the top in field of data science innovations. Domain areas that you can work on are as follows: -

- ❑ Education (Higher Education)
- ❑ Government
- ❑ Health Care
- ❑ Tourism
- ❑ Environmental
- ❑ E-Commerce
- ❑ Human Resources

- ❑ Transportation
- ❑ Financial
- ❑ Others (please specify) (SDG)

This proposal will lead to the development of a model and a data product. You may choose a case study (scenario) from your reading, or actual scenario or event based on your working experience before you come out with your proposal. A case study will also help you to write your problem statement clearly, as case study discusses about problems faced by a business. In addition, a case study will help you to identify why a business needs to analyze the dataset, and gain insight from it.

The tasks for GA1 should include the first three stages of OSEMN Framework:

Obtain – Data Collection

First step is that you obtain the data you need from available data sources. It is very important to collect complete and reliable data

- Kaggle/open gov data

Scrub – Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

- how to choose what parameters → literature review, see others usually use which parameter

Explore – Data Analysis

Exploratory data analysis. This phase allows us to understand the data so that we can figure out the course of actions and areas that we can explore in the modelling phase

GA1 Report Suggestions:

1. Project background - Description of the Data Science Project (suitable for which organization, target users, potential benefits, etc)
2. Problem Statement
3. Project Objectives
4. Project Scope
5. Literature Study / Information Gathering Analysis
6. Description of Methodology
    a. Obtain – Types of Data collected, Sources, Reliability
    b. Scrub – Processes done to clean the dataset, types of imputation used etc

c. Explore – Exploratory Data Analysis to investigate the data in terms of anomalies, and to check assumptions using statistics and graphical representations.

7. Impact of the Project to the society
8. Ethical considerations
9. References

Method & Submission for GA1:

- Submit the report on week 8 (before lecture hour) in a softcopy form. Prepare a power Point slides for 7 minutes presentation.
- Content should include basic explanation about each task and any related points that are suitable.
- You may include related diagrams, charts and any supporting material in your report.
- Report Format :
  - 1.5 spacing, Arial 11, maximum of 15 pages excluding cover page and attachment
- Your report should include cover page, with your group details (Name, matrix no, and topic).

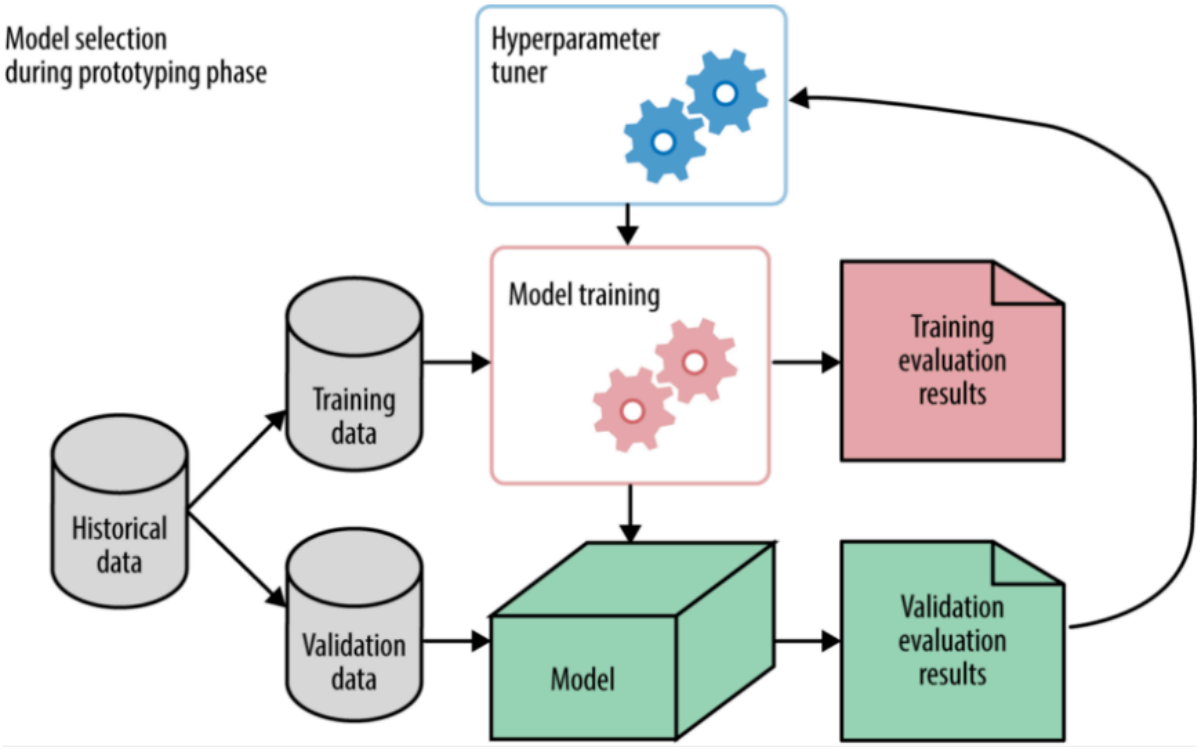
Rubric GA1 (20%)

Bil	**Items**	**Marks %**
1	Project background and Objectives	20
2	Literature Analysis	20
3	3 stages of OSEMN	30
4	Presentation and Q&A	20
5	Group Commitment	10
	Total	100

Group Assignment 2 (20%)

In relation to Assignment 1, the second part of the assignment will be on Modelling and Interpreting the Data. In the fourth step, we use analytic techniques to help in making sense of the data and acquire important insights for data-driven decision-making. This phase as many people would call it, “where the magic happens”.

For instance, regression and predictions are used to forecast future values, and classification identifies and groups the values obtained from the dataset



Interpreting data refers to the presentation of the data to a non-technical layman. The result is delivered to answer the business questions asked when the project first started, together with the actionable insights that are found through the data science process. At this stage, deployment of a data product is crucial. A data product, in general terms, is **any tool or application that processes data and generates results**. Businesses can use the results of such data analysis to obtain useful information like stock forecasting and customer segmentation and use these results to make smarter decisions.

GA2 Report Suggestions:

1. Project background - Description of the Data Science Project (suitable for which organization, target users, potential benefits, etc)
2. Project Objectives
3. Data Modelling
4. Data Interpretation
5. Deployment of Data Product
6. Insights and Conclusion
7. References

Method & Submission for GA2:

- Submit the report on week 14 (during lecture hour) in a softcopy form. Prepare a power Point slides for 7 minutes presentation.
- Content should include basic explanation about each task and any related points that are suitable.
- You may include related diagrams, charts and any supporting material in your report.
- Format :
    - 1.5 spacing, Arial 11, maximum of 15 pages excluding cover page and attachment
- Your report should include cover page, with your group details (Name, matrix no, and topic).

Rubric GA2 (20%)

Bil	Items	Marks %

1	Data Modelling	20
2	Data Interpretation	20
3	Data Product	20
4	Insights and Conclusion	20
5	Presentation and Q&A	10
6	Group Commitment	10
	Total	100