Cairo University
Faculty of Engineering
Computer Engineering Department

CMP4007
Spring 2023

Data Science Project Document

Description:

For this project, you are expected to pick a real-life challenging problem that you really aim to well understand, analyse, and utilise your perspective in data analysis techniques to propose some innovative solutions that you are able to interpret, know how to implement, and communicate to the owners.

Purpose:

The main purpose of this project is to walk through the activities of the data analysis cycle you studied in the course lectures, and perform the typical epicycle of data analysis through each of these activities.

You also have to highlight the knowledge you extracted and the insights you gained, then clarify how you would build your business solutions upon your findings.

Project details:

Idea selection: picking a real problem is a MUST, working for a real client will be regarded as a **bonus.**

Dataset selection: you can pick the datasets that support your problem as you wish, you can merge 2 or more ones, or you can do data scraping on your own. Quality of the final dataset is highly graded.

Data preparation: you will have to do enough pre-processing, data cleaning, organisation and visualisation. Working directly on raw data is not accepted.

Stating questions: you need to state and refine your questions as per the data analysis cycle, you have to ask various types of questions.

Data description/exploration: you will need to use some tool or specialised framework for that purpose, obtain really useful insights, and make meaningful dashboards and analytical reports.

Information collection: No restrictions on data processing, you can extract your findings through any model (machine learning, statistical, time series forecasting...etc.) whatever fits your questions and provides suitable answers.

Testing: You will need to apply testing (hypothesis, statistical, …etc) (Covered in the course lectures in detail)

Results interpreting: you have to propose real findings, practical solutions to the problem and effective answers to your questions.

Results communication: you need to show how you communicate your results to the client, or demonstrate them through your presentation.

Notes and restrictions:

- Every team is requested to state 10 - 12 questions, based on the workload of each. For those teams who are doing data scraping rather than using ready datasets, 6 - 8 questions are enough.

- Refining the questions is accepted as per the epicycle, but it needs appropriate justifications. Neither changing the whole question nor omitting one of them is accepted.

- Every team member must have a CLEAR part

- Every team consists of 3 – 4 members.

Phases and deliverables:

- **Phase 1:** Deliver a proposal report by end of week 5 containing:
    - **i.** Team members
    - **ii.** Idea/problem description
    - **iii.** All questions stated
    - **V.** Proposed work plan for each question
- **Phase 2:** Deliver by end of week 12 the following:
    - **i.** Final report including:
        1. Clear contribution of each member.
        2. your effort in applying the data analysis cycle and epicycle in detail.
        3. Knowledge and insights extracted throughout the project ordered chronologically and aided visually with graphs, charts, diagrams … etc.
        4. Final findings and results.
        5. Future work and enhancements
    - **ii.** All the codes and tool-specific files collected in an organised way.
    - **iii.** A video presentation where you show your idea, explain your work, and conclude your results. No special PPTX file is essentially required, you can present your report if done accordingly.

Suggested data sources:

A- https://data.gov

B- https://www.kaggle.com/datasets

C- https://data.world/datasets/data-science

D- https://datacatalog.worldbank.org/home

E- https://datasetsearch.research.google.com