Data Science Individual Project

Objective

Identify and address a real-world problem or question that genuinely interests you using data-driven methods. Your goal is to design and execute a full data science workflow — from problem formulation and data acquisition to model development, evaluation, and interpretation — applying tools and techniques discussed in this course (or closely related methods).

Datasets

You must use **dataset(s) different from your group project**. The dataset can come from an open data repository, or an API. It should be rich enough to support meaningful exploration, modeling, and analysis.

Tasks

Your project should include the following key components:

- Problem Definition
    - Identify a topic or issue that genuinely interests you.
    - Formulate **at least four analytical questions** that guide your investigation and connect to your chosen problem.
- Data Acquisition and Preparation
    - Select appropriate credible datasets.
    - Conduct exploratory data analysis (EDA) to understand the structure and quality of your data.
    - Perform necessary data cleaning, transformation, and feature engineering.
- Method Selection and Model Development
    - Choose suitable analytical or machine learning methods to address your questions.
    - **At least one question should involve training or fitting a model**, such as: Classification, Regression, Clustering, Dimensionality Reduction, Natural Language Processing, etc.

  - Implement models using appropriate tools such as Scikit-learn, TensorFlow/Keras, CatBoost, OpenCV, or other relevant Python libraries.
- Analysis and Evaluation
  - Evaluate all results and interpret your findings.
  - Discuss insights gained from your analysis in relation to the original questions.
- Conclusion and References
  - Summarize your key findings and their implications.
  - Cite all data sources, literature and external materials used.

Deliverables

A **Jupyter Notebook** that clearly presents your workflow, analysis, and findings in a structured and reproducible manner. Your notebook should include narrative explanations, code, visualizations, and interpretations.

Deliverables for your project:

- Proposal of topic, **due 11/2 (Sunday), at 11:59 pm**
  - Create a proposal by the provided template
  - Requirement: topic, questions
- Link of the Github repo, Due at 11:59 midngith 11/23 (Sunday):
  - Required: Submit the link of the Github repo by BrightSpace assignment.
  - The word template of the report can be download in BrightSpace -> Content -> Individual_Project
  - In the GitHub repos, it should contain
    1. Folder report: Both word and pdf versions of the draft report (format see the `template.docx` word file) with a draft of the introduction, datasets and methods.
    2. Three folders: Data, picture, Codes (could be empty folder if results are not ready)
- Final report, Due at 11:59 midngith 12/7 (Sunday):
  - No submission. Instructor will use the previous Github link for grading
  - Requirement: Github repo should including the following
    1. Folder report: **both word and pdf versions** of final report, Format: Must use `template.docx` word file format. **Length: No more than 8 pages**

    2. Folder codes: Jupyter notebook with all codes files
    3. Folder data: including all data files
    4. Folder graph: inlcuding all pictures.

- Method
    - Published code, pictures and report to a repository with readme reference
    - if use private Github repo, must add 'pangwit' by the following steps in link

Rubric

Category	Explanation
Introduction	Why was the project undertaken? What was the research question, the tested hypothesis or the
Selection of Data	What is the source of the dataset? Characteristics of data? Any munging, imputation, or feature
Methods	What materials/tools were used in answering the research question?
Results	What answer was found to the research question; what did the study find? Any visualizations?
Discussion	What might the answer imply and why does it matter? How does it fit in with what other research
Coding & Reference	Clear citation at end of the report. ipynb file with clear comments and datafile.

Rubric based on the IMRAD:https://en.wikipedia.org/wiki/IMRAD

Sample

https://github.com/pangwit/DS_Individual_Project_Example/tree/main