DS372 Project Specifications

Overview

The course project will be a multi-part group project which will ask that you complete an investigation of several related datasets based on a **performance benchmark** (dataset, EDAs, preprocessing, visualisation, models, and measurements in term of accuracy and time) of your design.

Group work requirement: You **must work in a team**.

Learning Goals

This project is designed to give you a "real-world" example of working with an unknown dataset, and doing so in a team. We *do not expect you to have all the answers*, and there are a number of tasks which are not directly covered in class.

You should expect to get practice with:

- Software Engineering:
    - Setting up new tools on your own computer and debugging the error messages that (sometimes) happen
    - Reading documentation for new tools
    - Working with git and GitHub
- Data Engineering:
    - Interpreting an unknown dataset
    - Designing a schema / modeling a database
    - Analyzing professional data systems
- Professional Skills:
    - Working in groups effectively
    - Writing and communicating technical concepts

Get Started

Template Repository

To get started, **one** team member needs to create a **private** GitHub repository. Start from the **DS372 Project Template** (https://github.com/azalhowaide/DS372-Project-Template.git)

Your project repo should be **private.** After creating your repository, add your teammates and me (azalhowaide@gmail.com) as collaborators.

The repository I've provided is a light scaffold to get you started. You are free to adapt it for your team's needs as necessary. However, be careful not to commit very large files to git.

Project Deliverables

You will submit a GitHub repository and written report, where your analysis will include the following steps:

Milestone 1 — Project Proposal & Dataset Selection

Goal: Define the problem, objectives, and data sources.
Deliverables:

1. Problem statement and motivation (why this project matters)
2. Research questions or hypotheses
3. Identification of datasets (must be from multiple sources)
4. Data collection plan (APIs, web scraping, public repositories, etc.)

- **Datasets selection**. Select several datasets to integrate or to compare results on them.
    - Total uncompressed size ≥ 1GB of data
    - Minimum of 1 million records of integrated dataset (may change based on project and dataset type)
    - Minimum of 20 different features' type
    - **How might you decide if an alternate dataset makes sense?** it can be helpful to pick a domain in which you have expertise or deep interest.
    - You should consider the clarity and organization of the raw data. Make sure there is a need to be spend cleaning and interpreting data—though some of that work is always necessary. Look for data which are well documented, have clearly named attributes, and come in easy-to-parse formats like JSON, and CSV.
    - We have **a list of suggested datasets** repositories to use for your project, but you are also welcome to select any source for datasets, while it is publicly available:
        - **General-Purpose Dataset Repositories**
          These host datasets across many domains (finance, health, text, images, etc.):
            - **Kaggle Datasets**
              The largest public dataset hub. Contains thousands of user-uploaded datasets with ready-to-use CSVs, notebooks, and metadata.
            - **UCI Machine Learning Repository**
              Classic ML datasets used in research and education (e.g., Iris, Wine, Heart Disease, etc.).

        - **Google Dataset Search**
          A search engine for datasets across the web (government, research institutions, and public data portals).
        - **AWS Open Data Registry**
          Datasets hosted on Amazon S3 — includes satellite imagery, genomics, economics, and more.
        - **Microsoft Research Open Data**
          Datasets from Microsoft's research projects (AI, NLP, vision, and education).
        - **Zenodo**
          A repository by CERN and OpenAIRE — supports open research datasets with DOI citations.
        - **Figshare** —
          General-purpose academic repository — datasets from universities, journals, and researchers.
        - **Hugging Face Datasets** —
          NLP, computer vision, and multimodal datasets ready for AI/ML training.
        - **Google Cloud Public Datasets** —
          Ready-to-query datasets in BigQuery — e.g., COVID-19, weather, NYC taxi, etc.
        - **Data World** —
          A community data platform with datasets across domains and visualization tools.
    - **Government & Open Data Portals**
        - **data.gov (USA)**
        - **EU Open Data Portal**
        - **UK Data Service**
        - **World Bank Open Data**
        - **UNdata**
        - **OECD Data**
        - **data.gov.uk**
        - **data.gov.sa (Saudi Arabia)**
    - **Financial & economic**
        - **Yahoo Finance API / Datasets** — stock market data.
        - **IMF Data** — global economic indicators.
    - **Environment & Geospatial**
        - **NASA Earth Data**
        - **USGS Earth Explorer**
        - **OpenStreetMap Data Extracts**
        - **NOAA Climate Data Online**

- **Health & Biology**
  - **Kaggle Health Datasets**
  - **PhysioNet** — clinical and physiological data.
  - **CDC Data Portal**
  - **WHO Global Health Observatory**
- **NLP & Text**
  - **Hugging Face Datasets**
  - **Google's Natural Questions**
  - **Common Crawl** — petabytes of web text.
  - **Project Gutenberg** — classic literature text data.
- **Image, Audio & Video**
  - **ImageNet** — image classification benchmark.
  - **COCO Dataset** — object detection and segmentation.
  - **Open Images Dataset**
  - **LibriSpeech** — speech recognition data.
  - **AudioSet** — audio event data.

Milestone 2 — Data Integration, Exploration & Cleaning Pipeline

Goal: Build a repeatable ETL (Extract, Transform, Load) pipeline.
Deliverables:

- Preliminary exploratory data analysis (basic stats, data size, missing values)
- Exploratory data analysis (visualizations, correlations, trends)
- Tools and technologies to be used (ETL tools, cloud, Python libraries, etc.)
- Code or workflow that extracts and integrates data from multiple sources
- Data schema and entity-relationship documentation
- Data cleaning procedures (handle missing, duplicates, outliers)

Milestone 3 — Data Transformation

Goal: Organize, model, and optimize data for analytics.
Deliverables:

- Feature consistency checks and basic data validation
- Store integrated data into a structured format.
- Data transformations and feature engineering steps
- Documentation of data lineage and metadata

Milestone 4 — Data Analytics & Machine Learning

Goal: Derive insights and predictive models.
Deliverables:

- Exploratory data analysis 2(visualizations, correlations, trends), if needed
- Predictive or clustering model (regression, classification, etc.)
- Model evaluation metrics and interpretation
- Comparison between analytical techniques

Milestone 5 — Dashboard, Reporting, and Final Paper

Goal: Communicate results and document the entire pipeline.
Deliverables:

- Interactive dashboard (e.g., Power BI, Tableau, Plotly Dash, or Streamlit)
- Research-style final report (problem, methods, results, discussion, limitations)
- Presentation or recorded demo
- Reflection on challenges, scalability, and potential future work