MSCS 634 — Project Deliverable 1: Data Collection, Cleaning, and Exploration

Author: Samrat Baral
Course: MSCS634 - Advanced Big Data and Data Mining
Deliverable: 1 — Data Collection, Cleaning, and Exploration
Repository: https://github.com/baralsamrat/MSCS634-Project
Date: November 03, 2025

Dataset Chosen

- **Title:** Organ Donation and Transplantation Dataset
- **Source:** Kaggle — https://www.kaggle.com/datasets/iamsouravbanerjee/organ-donation-and-transplantation
- **Granularity:** One row ≈ one transplant case (donor-recipient pair with metadata)
- **Suggested CSV filename:** `organ_donation.csv` (place it in `data/`)
- **Why appropriate:** ≥500 records, ≥10 attributes; mix of numeric/categorical fields enabling preprocessing, feature engineering, classification (`Outcome`), regression (`Wait_Time_Days`), clustering, and association rules (organ-type × blood-type × outcomes).

What This Deliverable Includes

- **Jupyter Notebook (`notebooks/Deliverable1_EDA_OrganDonation.ipynb`)** with:
  - dataset description & justification,
  - robust loading (auto-detect CSV in `../data` or use a synthetic fallback sample),
  - data cleaning (missingness, duplicates, types, categorical normalization),
  - EDA (histograms, boxplots, correlation heatmap, categorical counts),
  - insights for later modeling (features to engineer, targets to try),
  - a **rubric summary** that prints in the notebook **and** saves to `reports/Deliverable1_Rubric_Summary.txt`.
- `requirements.txt` and `.gitignore`
- Folder structure: `data/`, `figures/`, `notebooks/`, `reports/`

How to Run

```
cd MSCS_634_ProjectDeliverable_1
python -m venv .venv
# Windows: .venv\Scripts\activate
# macOS/Linux:
source .venv/bin/activate
pip install -r requirements.txt
jupyter lab  # or jupyter notebook
```

Open `notebooks/Deliverable1_EDA_OrganDonation.ipynb` and run cells top-to-bottom.

Key Steps (for your submission)

- **Notebook**: Keep comments and rationale for every cleaning action. Ensure figures are labeled and saved to `figures/`.
- **README**: Summarize dataset, list cleaning steps and insights, and note any challenges.
- **Repo**: Make it public or grant instructor access.