DSA 2040 A US 2025 Mid Semester Exam

DATA WAREHOUSING AND MINING
VENUE: LAB 9
INSTRUCTOR: AUSTIN ODERA
Answer all the questions

ETL Pipeline (Extract → Transform → Load)
Total Marks: 40
Submission: GitHub Public Repository (link to be submitted on LMS)

Objective

To assess your mastery of the **ETL process** through a hands-on mini-project that extracts data, applies useful transformations, and loads it into a structured format — all while maintaining professional documentation and GitHub hygiene.

Folder & File Naming Convention

Use this structure **exactly** for consistency and automated assessment.

```
ETL_Midterm_<FirstName>_<IDLast3Digits>/
├── data/
│   ├── raw_data.csv
│   └── incremental_data.csv
├── transformed/
│   ├── transformed_full.csv
│   └── transformed_incremental.csv
├── loaded/
│   ├── full_data.db or full_data.parquet
│   └── incremental_data.db or incremental_data.parquet
├── etl_extract.ipynb
├── etl_transform.ipynb
├── etl_load.ipynb
├── README.md
└── .gitignore
```

Example

```
ETL_Midterm_Austin_840/
```

Use **only your first name** and the **last 3 digits** of your student ID for privacy.

Project Data

A clean `raw_data.csv` and a smaller `incremental_data.csv` will be provided on BlackBoard or by your instructor.
You may also **simulate your own** small dataset using tools like:

- Mockaroo
- Random Excel data (5–10 columns, 50–200 rows)

ETL Task Instructions
EXTRACT – `etl_extract.ipynb`

- Load and preview `raw_data.csv` and `incremental_data.csv`.
- Display a `.head()` and `.info()` of each.
- Add observations (e.g., missing values, suspicious columns, duplicates).
- Save raw copies to `data/` directory.

Hints:

- Comment your code clearly!

2. TRANSFORM – `etl_transform.ipynb`

Apply **at least 4 meaningful transformations** to both datasets.

Category	Examples
Cleaning	Handle missing values, remove duplicates
Enrichment	Add `total_price = quantity * unit_price`
Structural	Convert dates, change data types
Filtering	Drop irrelevant columns or rows
Categorization	Create age bins, customer tiers

- Save transformed files to `transformed/` folder as:
  - `transformed_full.csv`
  - `transformed_incremental.csv`

Notes:

- Show before and after for each transformation.
- Explain what and why you are transforming.

3. LOAD – `etl_load.ipynb`

Load both transformed files into either:

- **SQLite** using `sqlite3` or `SQLAlchemy`, OR
- **Parquet** using `pandas.to_parquet()`

Preview the stored results using:

- SQL query (`SELECT * FROM full_data LIMIT 5`)
- Or `pd.read_parquet()` then `.head()`

Save the outputs in the `loaded/` folder.

README.md Instructions

Include the following sections in your `README.md`:

1. **Project Overview** – What the ETL lab does
2. **ETL Phases** – Description of each notebook and tasks done
3. **Tools Used** – Python, Pandas, SQLite, Parquet, etc.
4. **How to Run the Project** – Step-by-step instructions
5. **Screenshot** of data or chart

Data Privacy Rules

Use first name only in folders and repo names
Use only last 3 digits of Student ID
Don't share personal or real customer info
Make the GitHub repository **public**
Do not hardcode local file paths like `C:/Users/...`

GitHub Submission Instructions

1. Create a public GitHub repo named: `DSA2040A_ETL_Midterm_<FirstName>_<ID3>`
2. Push your entire folder (with all `.ipynb`, `.csv`, `.db`, etc.)
3. Commit logically (extract, transform, load)
4. Submit the repo **link** on BlackBoard

Marking Rubric (Total: 40 Marks)

Section	Criteria	Marks
Extract (5)	Data loaded, inspected, observations made	5
Transform (15)	≥4 transformations, before-after, explained	15
Load (10)	Data correctly loaded, verified, reproducible	10
GitHub (4)	Organized repo, good commits, clean structure	4
README.md (4)	Clear, useful, well-written	4
Bonus (2)	Visualization or unique transformation	+2

Final Checklist for Students

- Folders structured correctly?
- At least 3 notebooks?
- 4+ transformations explained?
- Final files in `transformed/` and `loaded/`?
- Public GitHub repo created and pushed?
- README present and helpful?