ETL Extract Lab

Description

This project demonstrates **Full Extraction** and **Incremental Extraction** in the context of ETL (Extract, Transform, Load) using a retail sales dataset. The lab is designed for **DSA 2040A – Data Warehousing and Data Mining**, and helps reinforce the practical aspects of extraction within the ETL pipeline.

Tools Used

- Python
- pandas
- Jupyter Notebook

Files Included

File Name	Description
`etl_extract.ipynb`	Main notebook with full and incremental ETL
`custom_data.csv`	Dataset used (realistic sales data)
`last_extraction.txt`	Stores the last extraction timestamp
`.gitignore`	Git ignore file for unnecessary files
`README.md`	This documentation

Transformations (Lab 5)

This lab extends the ETL pipeline by applying three transformation techniques:

1. **Cleaning**: Removed duplicate rows and filled missing values in `unit_price` and `quantity`.
2. **Enrichment**: Created a new `total_price` column by multiplying `unit_price * quantity`.
3. **Structural**: Converted the `date` column to a proper datetime format.

Transformed datasets are saved as:

- `transformed_full.csv`
- `transformed_incremental.csv` No new incremental data to transform.

How to Run

1. Ensure Python and Jupyter are installed
2. Install required packages: `pip install pandas`
3. Open the notebook: `jupyter notebook etl_extract.ipynb`

4. Run all cells sequentially

🧪 How to Reproduce

1. Clone the repository:

```
git clone https://github.com/aykahsay/ETL-Extract-AmabchowKahsay.git
cd ETL-Extract-AmabchowKahsay
Source: Downloaded from Kaggle
```