Chicken Breeder ETL data pipeline Project

This project demonstartes an ETL (Extract, Transform, Load) pipeline that extracts a 100 messy chicken dataset.

How it's Made

- Programming Language : Python
- Database: PostresSQL
- Data Source:
    - Created by AI. - As it was my first time using Pandas i wanted a simple set of Data
- Python Libraries:
    - pandas : For data manipulation and cleaning
    - psycopg2 : For Database interaction
    - dotenv: To handle enviromental variables

Prerequistes

- Python 3.13
- Docker
- Required Python packages (see installation)

Installation

1. Clone the repo

```
bash
git clone
cd chicken-breeder-ETL
```

2. Install Python dependencies
   ```
   bash
   pip install pandas psycopg2-binary python-dotenv
   ```

3. Set up enviromental variables - Create a .env file in the Db directory with:

```
env
postgres_host=localhost
postgres_user=postgres
postgres_pass=password123
```

```
postgres_db=chicken
postgres_port=5432
```

Running the Project

Step 1: Start PostgreSQL Database
Run PostgreSQL in Docker:

```
bash
docker run --name postgres-chicken -e POSTGRES_PASSWORD=password123 -p 5432:5432 -d postgres
```

Create the chicken database:

```
bashdocker exec -it postgres-chicken psql -U postgres -c "CREATE DATABASE chicken;"
```

Step 2: Clean and Load Data
Navigate to the Db directory and run the ETL pipeline:

```
bashcd Db
python db_chicken_alt_solution.py
```

This will:

Clean the messy chicken data (removes invalid records, standardizes formats)
Create the database tables
Load approximately 79 cleaned chicken records into PostgreSQL

Step 3: View Data (Optional)
Option A: Using Adminer (Web UI)

```
bashdocker run --name adminer -p 8080:8080 --link postgres-chicken:db -d adminer
```

Then visit: http://localhost:8080
Connection details:

System: PostgreSQL
Server: postgres-chicken
Username: postgres
Password: password123
Database: chicken

Lessons Learned :

- Pandas
  This was my first time learning and using pandas. In my previous project, I used python alone to clean the data. For example, using things like .strip() etc. After researching how data engineers/ how data is transfomed in the ETL process, I came across Pandas. I learnt the basics and the syntax through w3schools pandas teaching segment, and i made use of various methods such as isnull(), df.info(), .notna(). Before learning about Pandas i was curious as to how data was cleaned. I used to think, how would you know what parts of the data are missing, incorrect, misrepresented, however, pandas opened my eyes to how data can be analysed without having to look through every single

column. It did this by having the data in a dataframe(df). And when you used the method df.info(), it would show details about the entire csv i extracted, specifically, the amount of Nulls in each column/s.

Im excited to develop this newly aquired skill and use it with real life datasets.