This project organizes data into three levels: **Bronze**, **Silver**, and **Gold**. We will use different libraries like **Pandas**, **PySpark**, and **Polars** to work with three datasets from Kaggle. The goal is to clean, transform, and analyze these datasets step by step.

**Datasets**

1. **Brewery Operations and Market Analysis**  
   [Dataset link](https://www.kaggle.com/datasets/ankurnapa/brewery-operations-and-market-analysis-dataset)  
   Processed with **PySpark**.
2. **Steam Games Dataset**  
   [Dataset link](https://www.kaggle.com/datasets/fronkongames/steam-games-dataset)  
   Processed with **Pandas**.
3. **Houston Housing Market 2024**  
   [Dataset link](https://www.kaggle.com/datasets/datadetective08/houston-housing-market-2024)  
   Processed with **Polars**.

**Project Layout**

This project has three main layers:

1. **Bronze Layer**: This is the raw data we download directly from Kaggle.
2. **Silver Layer**: This is the cleaned and transformed data after we process it.
3. **Gold Layer**: This is the final data we use for analysis and insights.

Here's how the project is organized:

bash

Copy code

/project-root

│

├── Bronze/ # Raw data files

│ ├── brewery.csv

│ ├── steam\_games.csv

│ └── houston\_housing.csv

│

├── Silver/ # Cleaned data

│ ├── brewery\_cleaned.parquet

│ ├── steam\_games\_cleaned.csv

│ └── houston\_housing\_cleaned.parquet

│

└── Gold/ # Final data for analysis

├── brewery\_insights.parquet

├── steam\_games\_analysis.csv

├── houston\_housing\_trends.parquet

└── houston\_housing\_summary.csv

**What Are the Layers?**

**Bronze Layer**

This is the **raw data** we downloaded. It’s uncleaned and has all the columns and rows just like in the original files. The files are placed in the Bronze/ folder.

**Silver Layer**

In this layer, we **clean and transform the data**:

* **Remove irrelevant columns** we don’t need.
* **Handle missing values**: we either remove them or fill them with default values.
* **Normalize** the data if needed (for example, making sure all dates follow the same format).

This cleaned data is saved in the Silver/ folder, ready for further analysis.

**Gold Layer**

The **Gold Layer** is where we get the data ready for analysis. This includes:

* **Aggregating** or summarizing the data (grouping, getting averages, etc.).
* **Creating new columns** for insights.

The results are saved in the Gold/ folder.

**How to Use This Project**

**Steps to Follow:**

1. **Install the required libraries**  
   Make sure you have the required libraries installed by running:

bash

Copy code

pip install -r requirements.txt

1. **Bronze to Silver (Cleaning the Data)**  
   Each dataset will be cleaned using a different tool:
   * **Brewery Dataset**: We use **PySpark** to clean brewery.csv and save the cleaned version as brewery\_cleaned.parquet in the Silver/ folder.
   * **Steam Games Dataset**: We use **Pandas** to clean steam\_games.csv and save it as steam\_games\_cleaned.csv in the Silver/ folder.
   * **Houston Housing Dataset**: We use **Polars** to clean houston\_housing.csv and save it as houston\_housing\_cleaned.parquet in the Silver/ folder.
2. **Silver to Gold (Final Analysis)**  
   After cleaning, we will process the Silver data to generate insights:
   * **Brewery Dataset**: Analyze trends in brewery operations using **PySpark**, and save the results as brewery\_insights.parquet in the Gold/ folder.
   * **Steam Games Dataset**: Analyze game trends and save the insights as steam\_games\_analysis.csv in the Gold/ folder.
   * **Houston Housing Dataset**: Analyze housing trends and save the results as houston\_housing\_trends.parquet and houston\_housing\_summary.csv in the Gold/ folder.

**Important Files:**

* **Brewery Data Processing**: brewery\_processing.ipynb
* **Steam Games Data Processing**: steam\_games\_processing.ipynb
* **Houston Housing Data Processing**: houston\_housing\_processing.ipynb

**Key Decisions**

1. **Handling Missing Data**: We either removed rows with missing data or filled in the gaps with default values.
2. **Dropping Unnecessary Columns**: Some columns didn’t add value to our analysis, so we dropped them to keep the data clean.
3. **Normalization**: We standardized some data to ensure consistency (for example, converting date formats to be the same).
4. **Merging Data**: In some cases, we merged multiple tables or datasets to gain better insights.