DS 2002: Data Project 1
Members: Shaina Banduri, Diya Gupta, Maya Shah
Reflection Report: ETL Pipeline Implementation

Throughout the implementation of our ETL (Extract, Transform, Load) pipeline, we encountered various challenges and insights that shaped our learning experience.

Challenges Encountered
One of the major challenges we faced was handling API data retrieval from OpenWeatherMap. Since API requests can fail due to rate limits or invalid responses, implementing the proper error handling was crucial. Additionally, merging the weather data with our local CSV dataset proved to be more complex than expected, as the two sources had different structures and required transformation before integration. Another challenge that we faced was ensuring data completeness. Some records lacked necessary latitude and longitude fields, making it difficult to fetch corresponding weather data.

Aspects That Were Easier Than Expected
Certain aspects of the implementation were surprisingly straightforward. Basic data transformation tasks, such as dropping unnecessary columns and filtering records using Pandas, were relatively simple for us. Additionally, converting data formats and storing transformed data into an SQLite database was more intuitive than anticipated, due to built-in Pandas functionalities.

Aspects That Were More Difficult Than Expected
Unfortunately, error handling for API calls turned out to be more complex than expected. Managing failed requests required implementing logic to retry requests or handle missing data appropriately. Furthermore, we faced an additional challenge in ensuring that the final dataset was structured correctly, as aligning data from different sources required extra processing steps.

Future Applications of This ETL Utility
This ETL pipeline can be a valuable tool for future data projects, particularly those that involve integrating multiple data sources. The ability to automate data extraction, transformation, and storage makes it scalable for larger datasets and different APIs. Additionally, the error-handling mechanisms and data-cleaning processes implemented in this project can be adapted for future applications to ensure robust data quality and integrity. This type of utility can be useful in data-driven decision-making, automating workflows, and supporting real-time data analysis.

Overall, this project provided valuable hands-on experience in designing an ETL pipeline, overcoming implementation challenges, and understanding best practices for data integration and transformation.