

For our project, we used two datasets: the first one was an Olympics medal dataset in CSV format, which provided information on Olympic medals (including which countries have won the most gold, silver, and bronze medals) and the second one was a weather dataset in JSON format, which we retrieved via an API call.

The primary challenge we encountered was sourcing a suitable dataset in JSON format. While CSV datasets were widely available, JSON datasets that met our requirements were more difficult to find. We decided to use one from an API – WeatherStack. We anticipated this to be difficult due to our limited experience in API calls, but after exploring them a bit it turned out to be far easier than we expected.

The second challenge was processing the two datasets in different formats. While converting the CSV to a different format (like JSON) was easy, transforming JSON data into a table-like structure required more thought, particularly when dealing with nested JSON objects and trying to delete a column. We had to normalize the JSON datasets to delete a specific row inside a column for better data manipulation. Understanding how to normalize and flatten the JSON data so that it fit well into a tabular format was one of the harder parts of the project. We utilized available libraries to help us, such as SQLite3 and pandas.

A utility like this—being able to handle multiple data formats and convert between them—will prove invaluable in future data projects. It allows flexibility in dealing with different data sources, whether it's CSV files, JSON data from APIs, or SQL databases. Many real-world projects involve different data sources, and this experience demonstrated how crucial it is to normalize and transform data into usable formats efficiently. Additionally, API-based data usage offers access to live, up-to-date data, which is very important for time-sensitive analytics projects.