Progress Report

Student name: Yilong Xu

Project scope update:

The project scope has remained consistent with my original proposal. The main goal is still to investigate how climate change influences agricultural productivity in the United States, focusing specifically on corn and wheat yields. The project aims to understand how long-term temperature trends correlate with fluctuations in crop yields across different U.S. states.

The only adjustment so far is in the data retrieval strategy for NOAA climate data. Initially, I planned to collect statewide temperature data using FIPS location codes. However, I discovered that the NOAA GSOM dataset does not always support statewide aggregation for certain variables like *TAVG*. Because of this, I temporarily shifted to retrieving climate data from a single, reliable weather station (LAX Airport, station ID: GHCND:USW00023174) in order to verify that my API requests, environment configuration, and code structure were working. In the final project, I plan to expand this approach to include multiple temperature stations for each state or use climate division data depending on API feasibility.

Other than this adjustment in the technical approach, the core analytical goals remain the same.

---

Data sources

Data obtained so far

Using my Python scripts, I have successfully retrieved:

1. USDA Crop Yield Data (CORN & WHEAT)
    - Includes annual yield (bushels per acre)
    - Covers years 2010–2024
    - Covers all U.S. states that report crop yields
    - Data is cleaned into a DataFrame with columns:
        - year
        - state_name

        - CORN_yield / WHEAT_yield

2. NOAA Climate Data (Monthly Temperature)
    - Using NOAA GSOM dataset
    - Variable: TAVG (average monthly temperature)
    - Station tested: GHCND:USW00023174
    - Time range: 2010–2010 (for testing)
    - Data includes:
        - date
        - datatype
        - station
        - value (temperature)

All data is successfully saved into the data/ folder and is ready for merging and analysis.

---

APIs used

1. USDA NASS QuickStats API

URL: https://quickstats.nass.usda.gov/api
Purpose:

- Retrieves historical corn and wheat yield data by state
- Very stable and provides complete records

Key parameters used:

- source_desc=SURVEY
- sector_desc=CROPS
- group_desc=FIELD CROPS
- commodity_desc=CORN/WHEAT
- statisticcat_desc=YIELD

- agg_level_desc=STATE

2. NOAA Climate Data Online (CDO) API

URL: https://www.ncei.noaa.gov/cdo-web/api/v2
Purpose:

- Retrieves monthly temperature data for climate analysis

Key parameters used:

- datasetid=GSOM
- datatypeid=TAVG
- stationid=GHCND:USW00023174 (initial testing station)
- startdate and enddate
- API authentication via token header

Both APIs require keys, which I stored securely in a .env file.

---

Issues / difficulties

During development, I encountered several technical challenges:

1. NOAA API limitations

The NOAA GSOM dataset does not always support high-level location IDs such as:

- FIPS:<state code>
- Certain climate division codes
- Certain datatypes combined with certain datasets

This caused repeated 400 and 500 errors until I switched to using station-level queries. This limitation means I will need to rethink how to aggregate statewide temperature data.

2. Some stations lack temperature data

Several stations returned errors because they only report precipitation, not temperature. I had to filter out stations beginning with prefixes like US1, which correspond to CoCoRaHS precipitation-only networks.

3. Environment + API setup

I spent time resolving issues in:

- Conda environment activation
- PATH configuration for Python and packages
- Loading API keys from .env
- Ensuring consistent execution inside VS Code

These issues temporarily slowed the development of the retrieval pipeline, but everything is now functioning.

4. Potential future challenges

As the project progresses, I expect challenges in:

- Aggregating climate data across multiple stations per state
- Aligning monthly climate data with annual yield data
- Handling missing or inconsistent data from NOAA
- Designing proper statistical tests and visualizations

Despite these issues, the data retrieval foundation is now working and I am on track for the final steps of analysis.