CS598 - Foundations of Data Curation Tan Dong Xiang

End-to-end data curation workflow Final project report dxt2@illinois.edu

Project Report

Project Overview

The goal of this project is to demonstrate several key concepts of data curation by creating an end-to-end data curation workflow. For this purpose, this project will pull and merge two datasets: Temperature readings of weather stations from the NOAA Global Historical Climatology Network (GHCN) and corn yield data from USDA NASS Quick Stats. The two sets of data will be merged to find out if there are any correlations between the average temperature over the years and its effects on the corn yield. The final output of this project will come in the form of a static graph with the processed and merged dataset.

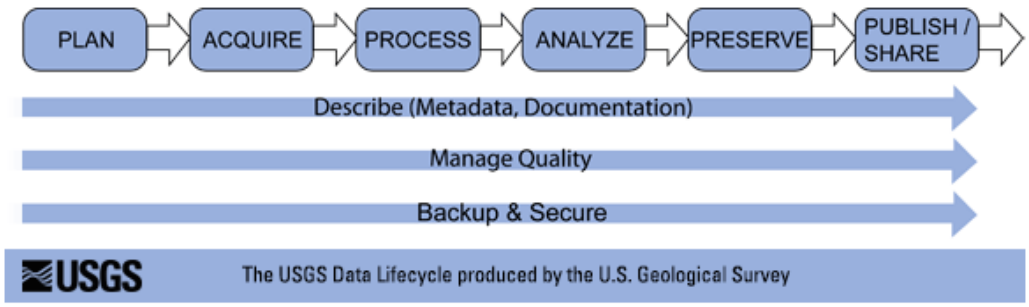
Report scope

The objective of this report is to document the work done for the End-to-end data curation workflow project for CS598. This report will include the data lifecycle for the project, a short description of the data used for this project, findings, challenges, as well as where to find the artifacts produced by the project. Other important parts of this project can be found in the GitHub repository documented in the “Artifacts” section of this report. This includes:

- How to setup and run the project can be found in the README.md
- Data dictionary (data structure and details) can be found in metadata/data_dictionary.md
- Metadata using the schema.org standard can be found in metadata/dataset.jsonld
- Provenance details can be found in metadata/provenance.md

Data lifecycle

The data as well as the project will be broken down into 6 main stages based on the USGS Science Data Lifecycle model (Faundeen, n.d.). This section will describe what happens in each stage of the data lifecycle for this project.



Primary Model Elements

Plan: At this stage, a plan will be developed for the project. This includes the plan on how data will be acquired and managed throughout the lifecycle of the project and the expected outputs from the project.

Acquire: In this stage, data will be gathered through different data sources. The two datasets were downloaded and scraped from their sources.

Process: Collected data were filtered for relevance, specifically for corn yield in Iowa (location) and temperature readings in Iowa in June, July, and August (time, location, remove unused data) The two filtered datasets will then be merged into a single pandas dataframe.

Analyze: To compare the datasets, they will be plotted onto a single graph.

Preserve: Metadata, data dictionary, and provenance information is documented and stored into the repository (in the metadata folder)

Publish/share: The notebook, along with the data is uploaded onto a public GitHub repository.

Cross-cutting activities

Describe: The metadata stored in the repository is stored and updated as the definitions and how data is processed and stored changes.

Manage Quality: Invalid and irrelevant data is removed before the final dataset.

Backup & Secure: Data can be downloaded at any time. There are no security requirements as the data is publicly accessible.

Data Description

This section will describe the two datasets used in the project, and some observations and decisions made while using these datasets.

NOAA Global Historical Climatology Network (GHCN)

This data source will provide historical data on changes in climate over time. Namely the temperature, precipitation, and other climate drivers affected by climate change. The data can be downloaded using the jupyter notebook itself.

Aside from the processing required to gather the minimum, maximum, and average temperature across the years, the data itself was also filtered for two more things:

1. Only keep the readings for the months June, July, and August. This is because corn is the most temperature sensitive.
2. Only keep the readings from stations in Iowa, which is one of the highest corn-producing states.

Step for data processing

1. Fetch Iowa station list from ghcnd-stations.txt (STATE = IA).
2. For each station, load its .dly file (or predownloaded DLV) and parse fixed-width fields: ID, YEAR, MONTH, ELEMENT (TMAX/TMIN).
3. For each TMAX/TMIN line, scan daily value slots (31 days, in their corresponding columns); ignoring invalid or empty fields which are denoted by "-9999". Then compute the monthly mean in degrees celsius.
4. Keep months June to August and for each station/month, calculate and keep monthly mean TMAX, TMIN, and TAVG (average of TMAX and TMIN).
5. Across stations, aggregate annual means: mean of TMAX, mean of TMIN, mean of TAVG per YEAR and round them to 1 decimal place.
6. Save columns: YEAR, Average Temperature (TAVG), Average TMAX, Average TMIN.

GHCN Data terms of use (Ethical, legal, or policy constraints)

Since this is a data set from the U.S. government, it is effectively public domain. However, proper citation and date of access are still expected. And according to the README of the data source (https://ncei.noaa.gov/pub/data/ghcn/daily/readme.txt), it is required to cite the dataset version and the Menne et al. 2012 paper.

USDA NASS Quick Stats (Crop Data)

The data from this data source includes the yields and production for major crops on a county, state, and national level. This will be used to represent the agricultural production outcomes (crop yield) over time. The data is available on USDA's quick stats API, but it would require a signup to get an API key.

To ease the testing process and ensure that it is available for anyone using the notebook, the data has been pre-downloaded and stored with the notebook. This is only possible due to its relatively small size. Similar to the GHCN dataset, this dataset also uses corn production in Iowa because corn is one of the most prominent crops in the USA with Iowa being one of the top producers of it. The chosen units used for this dataset is bushels per acre of land to make sure elements like the changes in size of farmland does not affect the results.

Step for data processing

1. Download QuickStats CSV with filters: State=IOWA, Commodity=CORN, Data Item=CORN, GRAIN - YIELD, Program=SURVEY (annual yields).
2. Keep annual records (Period = YEAR); drop forecast/duplicate period rows.
3. Convert Value to numeric (bushels/acre), select YEAR and yield.
4. Save columns: YEAR, Yield.

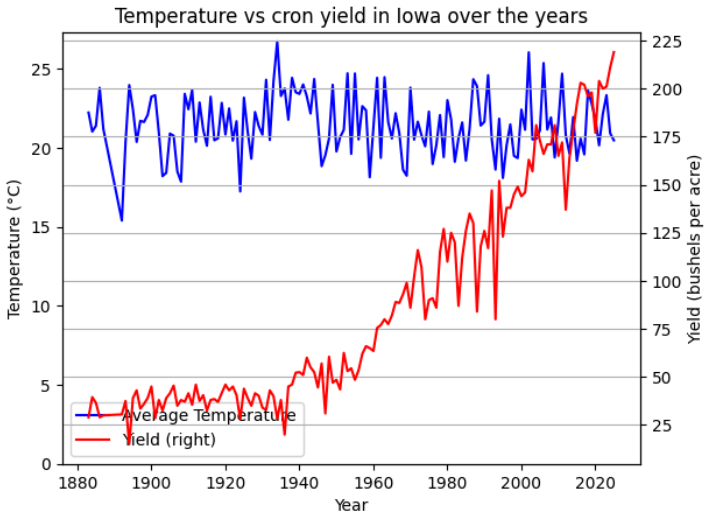
Crop Data terms of use (Ethical, legal, or policy constraints)

USDA NASS QuickStats data (corn CSV) are in the public domain. NASS requests acknowledgment such as “Source: USDA National Agricultural Statistics Service (QuickStats), accessed <date>” and that the accessors do not imply USDA endorsement or attempt to identify individual respondents (data are aggregated). No commercial-use restriction is imposed.

Findings

This section will highlight the conclusion from comparing the two datasets and conclude if temperature does affect corn yield in Iowa

Unfortunately, based on the processed data and graph plot based on it, there was no strong correlation between the temperature and corn yield. However, it can be observed that in the years when the temperature drastically deviates from the average temperature, the corn yield tends to be lower at the same time.



Further reading also suggests that there are other factors that can have a bigger impact on corn yield such as availability of water, fertilizers, and pests. (Zai et al., 2024)

Another observation is that over the years, corn yield (in bushels per acre) has been steadily increasing. This would mean that despite other factors, efficiency of farmland has been improving over the years. This can be attributed to advances in agricultural technology and science which helps crops adapt to different conditions. (Zai et al., 2024)

Challenges

This section will highlight the challenges faced in this project.

1. **Size of data**

   Specifically the GHCN dataset, due to its size and span across the many stations, downloading and processing the data was a time consuming process. There were also some issues uploading the data onto GitHub at the start.

2. **Complexity of the GHCN dataset**

   As discussed in the progress report of this project as well as the provenance documentation, processing the GHCN dataset took many steps and data had to be joined across multiple files. These files have an unconventional format which took some effort to convert to pandas dataframes.

3. **Domain knowledge**

   Because I had no knowledge of farming and little knowledge of the climate, it was hard to come up with a strong hypothesis that can yield conclusive results. For example, there might have been other factors that can have a large impact on corn yield like precipitation.

   There was also no easy way to know if a station in the GHCN dataset was truly relevant because the data collection stations might be far away from the farmland and the data collection does not represent the temperatures there.

Possible improvements

This section will discuss some of the pitfalls of the project and what can be done to improve it.

1. **Explore other factors (e.g. precipitation)**

   As discussed in the previous section, there might be other factors that can affect the corn yield. Since some of these data are readily available in the dataset or made publicly available by the US government, it is possible to explore if these factors have an impact on the corn yield and the drivers behind these factors changing over time and how it affects crop yield. This can make the final result more useful and potentially actionable to know what factors to prioritize when adapting agricultural techniques to them.

2. **Make the final product of data processing more interactive**

   Instead of a static graph, an interactive graph that allows users to zoom and show/hide other datasets will result in a tool that will be more useful for data analysis and finding patterns within the data.

3. **Mark major events and how it affects yield**

   Other than environmental factors, it is also important to consider events like world wars, scientific breakthroughs, and the introduction of new agricultural technologies/techniques. Factoring these events into the analysis allows us to get a clearer picture of how different environmental factors and world events affect crop yield.

Artifacts

Public GitHub repository: https://github.com/T-D-X/CS598

All work is contained mainly in *climate_crop_study.ipynb* this includes the downloading and downloading and reading of datafiles for the GHCN data source while the crop data is contained in *data/crop_data/raw_data/corn.csv.* Processed data after preprocessing will be stored in *data/crop_data/processed/crrop.csv* and *data/climate_data/processed/ghcnd-IA-data.csv*.

There are two main branches to take note of:

1. “master” branch which contains all the work done for the project minus the raw data
2. “with-data” branch which contains everything in the master branch + the GHCN data

The details on running the project can be found in the README.md of the repository:

https://github.com/T-D-X/CS598/blob/master/README.md

Bibliography

Faundeen, J. L. (n.d.). *The united states geological survey science data lifecycle model*.

Zai, F. H., McSharry, P. E., & Hamers, H. (2024). Impact of climate change and genetic development on Iowa corn yield. *Frontiers in Agronomy*, *6*. https://doi.org/10.3389/fagro.2024.1339410