CS598 - Foundations of Data Curation Tan Dong Xiang

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

Progress report

Proposal feedback incorporated

After initially scoping down the project in the project proposal and planning phase, there were no more comments to be addressed and there was no need to further scope down the project from the final project proposal.

Challenges during data acquisition

This section will outline the challenges faced during the data acquisition process. There were no changes in the data sources from the initial project planning phase and this section will talk about the data acquisition process of each of the two data sources and the challenges faced while working with them.

NOAA Global Historical Climatology Network (GHCN)

Acquiring data from the GHCN website was much more complex than what was estimated during the project planning phase. In this project, the average temperature for the state of Iowa. However, data from GHCN is split at several levels and here are the levels and the process to get the target end state, average temperature of the state for each year.

1. **State**

A file, *ghcnd-states.txt*, will list out the state code of each state, this state code will then be used to filter out which stations belong to which state in *ghcnd-stations.csv*.

2. **Station**

By filtering *ghcnd-stations.csv* with the state code, a list of weather station IDs belonging to stations in Iowa is created. This is then used to download the appropriate reading data (in the form of .dly files) from the GHCN website with Python's request library. (e.g. https://www.ncei.noaa.gov/pub/data/ghcn/daily/all/AFM00040938.dly for the station with the ID AFM00040938)

3. **Reading**

After downloading all the readings from the stations in Iowa which contain the maximum and minimum temperature for each month, we can calculate the average temperature for the entire state for each year. Each .dly file contains rows of data that have fixed length columns. Each column are formatted as follows:

Variable	Columns	Type
ID	1–11	Character
YEAR	12–15	Integer
MONTH	16–17	Integer
ELEMENT	18–21	Character
VALUE1	22–26	Integer
MFLAG1	27–27	Character
QFLAG1	28–28	Character
SFLAG1	29–29	Character
VALUE2	30–34	Integer
MFLAG2	35–35	Character
QFLAG2	36–36	Character
SFLAG2	37–37	Character
.	.	.
.	.	.
.	.	.
VALUE31	262–266	Integer
MFLAG31	267–267	Character
QFLAG31	268–268	Character
SFLAG31	269–269	Character

Out of all the variables, the first five variables are needed, ID, YEAR, MONTH, ELEMENT and VALUE1. An example of a row with these variables will look like: AFM00040938202501TMAX21.

By pulling out the relevant substring for each row, we can get a pandas dataframe with the needed columns. Then, the dataframe is filtered and split into dataframes with the elements TMAX (temperature max) and TMIN (temperature minimum). The two dataframes are then joined according to their ID, YEAR, and MONTH column to get a dataframe of each station's TMAX and TMIN. This is used to calculate the average temperature for the station in the time frame (TAVG) which is then used to calculate the average temperature for the entire state for each year after merging the dataframes from all relevant stations.

USDA NASS Quick Stats (Crop Data)

Retrieving the crop data was easier than initially estimated. The quick stats platform (https://quickstats.nass.usda.gov/) was easy to navigate and the data was easily acquired. However, to programmatically download the data (i.e. in the jupyter notebook) an API key has to be requested. For this project, the data is downloaded beforehand to make testing and demonstration easier.

Other challenges identified

This section will discuss the challenges that are not related to the data acquisition process.

Portability of the data

Due to the size of data, it will take a long time to move and download the data to the other platforms, in this case, GitHub. To solve this problem, the jupyter notebook will download the GHCN data on the first run and use the downloaded data on subsequent runs on new machines. On top of that, there will be another branch that contains the GHCN data called “with-data” created.

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

Scope adjustments

There are no scope adjustments

Updated Timeline

Task	**Due date**	**Progress**
Proposal submission	9/15	Done
Project planning (Plan)	9/20	Done
Data acquisition (Aquire)	10/20	Done
Progress report	10/27	Done
Data cleaning & transformation (Process)	11/14	In-progress
Data integration & graph plotting (Analyze)	11/30	Not started
Data storage and automation (Preserve & publish/share)	12/3	Not started
Final Submission	12/10	Not started