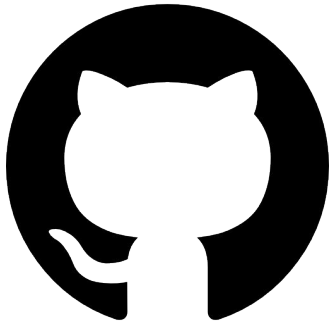
CS418 - Introduction to Data Science Course Project

An Examination of the Causal Relationship Between Rainfall and Crop Prices: A Focus on Wheat and Corn.

Data Bandits	Email address	Github
Marcin Kowalik	mkowa2@uic.edu	mkowa2
Filippo Corna	fcorn@uic.edu	FilippoCorna
Federico Mansutti	fmans@uic.edu	FedericoMansutti
Marco Conti	mcont46@uic.edu	C0NN
Patrick Poggi	ppogg@uic.edu	PatrickPoggi



https://github.com/mkowa2/Data_Bandits_CS418

Problems

Idea: We aim to investigate how meteorological patterns affect the yields of corn and wheat, particularly during critical growth phases like pollination, and how these changes lead to price fluctuations in U.S. markets.

Why It Matters? Corn and wheat are essential crops in food production and industrial use. Fluctuations in yield caused by weather conditions impact commodity prices, affecting farmers, supply chains, and futures markets. In commodity futures trading, significant changes in crop supply due to weather influence the contracts bought and sold, leading to price volatility and financial risk.

Hypotheses:

- Low rainfall during this period will reduce yields, resulting in increased prices and futures contracts.
- Excessive rainfall may delay planting, disrupting markets and complicating future price predictions
- Extreme heat, late frosts, and shifts in rainfall timing will also play critical roles in influencing crop quality and yield fluctuations
- Since temperatures show seasonal patterns, these might be reflected as significant patterns in price trends

Datasets

APIs:

- **USDA API:** Provides official data from the United States Department of Agriculture, offers **historical series of prices**, production, and corn stocks
- **Quandl API:** Offers data on corn futures from the **Chicago Board of Trade (CBOT)**, provides daily closing prices, volumes, and open interest
- **Visual Crossing API:** Offers **historical weather** data based on location.

Prices are time series and are pulled with API requests as JSON files. Data collection is typically not too time consuming since API queries tend to be quick.

DATASETS:

1) **USDA NASS QuickStats:** Comprehensive **database** of agricultural statistics in the United States, includes historical corn and wheat prices at the national and **state level,** free download in CSV format. Data is available year by year and shows total harvested (numerical, acres), total value sold (numerical, $) and damaged. For example, the dataset of corn produced, in $ per state, has 3284 rows.
2) **Weather Dataset Kaggle (US): Weather pattern and rainfall datasets** from 25 000 stations in the US. The dataset has he following categories: Date (YYYY-MM-DD) from 1992 to 2021., Maximum temperature(tenths of degrees C), Minimum temperature(tenths of degrees C), water evaporation and precipitations (tenths of mm).

Might be time consuming and require aggregation of data due to the vast number of possible dataset combinations on **USDA NASS QuickStats.**

Solution

The focus for this project will be the “Corn Belt” and surrounding states including, Colorado, Kansas, and North Dakota. We are looking for short and long term **relationships** between corn price and weather. This research could be further expand to other crops.
After the data cleaning phase, and a data exploration phase to better define our hypothesis, we will use a combination of different technique to address the problem:

- **Regression** to predict corn prices according to weather changes:
  - Least Square algorithms
  - Neural Networks
  - LSTMs for temporal relationships on time series
  - Other Machine Learning techniques
- **Hypothesis testing** to check for long or short term correlation between corn price and weather
- **Classification** to classify different crop types according to their reaction to the weather:
  - Logistic regression
  - K-Nearest Neighbors
  - Support Vector Machines

The result is finding the relationships between prices and weather and how one affects another.

ROLES and COORDINATION

Phase 1: getting the data (deadline: October 18th)

- Data collection: Marco, Filippo, Federico
- Data cleaning:Patrick, Marcin

Phase 2: analyze the data (deadline: November 6th)

- Exploratory data analysis: Filippo, Marcin, Federico
- Data visualization: Marco, Patrick

Phase 3: draw information from data (deadline: November 27th)

- Hypothesis testing: Patrick, Filippo
- ML analysis: Marco, Marcin, Federico

Phase 4: conclusion (deadline: November 29th)

- Reflection and next steps: Patrick, Filippo, Marco, Marcin, Federico