Data Source

The data source is synthetically generated for the purpose of practicing data analysis and machine learning models, particularly those focusing on predicting crop yield.

I am choosing this data set because I have a bachelor's degree in Food Science with ten years of industry experience, so understanding what factors contribute to crop yield is an interest to me. I am interested in pursuing a job for the State of Colorado or Federal Government in any sector, so analyzing this information could provide an insight into the type of work I could be doing.

Column Name	**Description**
Region	The geographical region where the crop is grown (North, East, South, West)
Soil_Type	The type of soil in which the crop is planted (Clay, Sandy, Loam, Silt, Peaty, Chalky)
Crop	The type of crop grown (Wheat, Rice, Maize, Barley, Soybean, Cotton)
Rainfall_mm	The amount of rainfall received in millimeters during the crop growth period)
Temperature_Celsius	The average temperature during the crop growth period, measured in degrees Celsius.
Fertilizer_Used	Indicates whether fertilizer was applied (True = Yes, False = No)
Irrigation_Used	Indicates whether irrigation was used during the crop growth period (True = Yes, False = No).
Weather_Condition	Indicates weather condition during the growing season (Sunny, Rainy, Cloudy)
Days_to_Harvest	The number of days taken for the crop to be harvested after planting.
Yield_tons_per_hectare	The total crop yield produced, measured in tons per hectare.

Data Profile

Variables	Time-variant/ -invariant	Structured/ Unstructured	Qualitative/ Quantitative	Qualitative: Nominal/ Ordinal Quantitative: Discrete/ Continuous
Region	time-invariant	structured	qualitative	nominal
Soil_Type	time-invariant	structured	qualitative	nominal
Crop	time-invariant	structured	qualitative	nominal
Rainfall_mm	time-variant	unstructured	quantitative	continuous
Temperature_ Celsius	time_variant	unstructured	quantitative	continuous
Fertilizer_Used	time-invariant	structured	qualitative	nominal
Irrigation_Used	time-invariant	structured	qualitative	nominal
Weather_ Condition	time_variant	structured	qualitative	nominal
Days_to_Harvest	time_variant	unstructured	quantitative	discrete
Yield_tons_per_ hectare	time-invariant	unstructured	quantitative	continuous

Consistency Checks were performed to ensure no missing values, duplicates, mixed-type date or abnormal outliers. The only issue was 241 records with negative yields. I converted those negative values to 0 because it is not possible to have a negative crop yield.

Limitations

A limitation for this dataset is that it is synthetically generated meaning these are not actual values from crop yields across the United States. Even though the data is synthesized, it is based on commonly understood agricultural factors that influence crop yield.

Another limitation is that the regions are not more granular in terms of state locations and only listed as North, West, South. To mitigate this, I will be referencing the US Census Regions of the United States as follows.

West

- Arizona
- Colorado
- Idaho
- New Mexico
- Montana
- Utah
- Nevada
- Wyoming
- Alaska
- California
- Hawaii
- Oregon
- Washington

South

- Delaware
- District of Columbia
- Florida
- Georgia
- Maryland
- North Carolina
- South Carolina
- Virginia
- West Virginia
- Alabama
- Kentucky
- Mississippi
- Tennessee
- Arkansas
- Louisiana
- Oklahoma
- Texas

North

- Connecticut
- Maine
- Massachusetts
- New Hampshire
- Rhode Island
- Vermont
- New Jersey
- New York
- Pennsylvania

North

- Indiana
- Illinois
- Michigan
- Ohio
- Wisconsin
- Iowa
- Kansas
- Minnesota
- Missouri
- Nebraska
- North Dakota
- South Data

https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Ethical Considerations

There is a chance of collection bias in this data with it being synthetically generated. The data might not capture the nuances, randomness or noise found in real-world data. For example, with the hurricanes happening on the east and south coast, there would be a much higher rainfall amount than is probably reflected in the dataset. Another collection bias is the yield could be too evenly distributed and not reflecting the complexity of the real world, leading to an unrealistic distribution.

Sample bias can be happening in the dataset where the data that was generated may not translate directly to the real-world data from lack of diversity. For example, the southern region may not have enough sandy soil data even though the southern region contains a lot of land off the coast.

Another limitation with the dataset could be reduced interpretability. It could be harder to draw meaningful insights from synthetic data, especially since the data collection was not documented.

Jeffrey Frankenfeld
Task 6.1

Key Questions

1. Which variables have the strongest correlation with the crop yield?
2. Is there an optimal range of rainfall that maximizes crop yield?
    a. Rainfall_mm
3. Are specific temperature ranges more favorable for certain crops?
    a. Temperature_Celsius
    b. Crops
4. Does the use of fertilizers improve yield for specific soil types?
    a. Fertilizer_Used
    b. Soil_Types
5. Are yields significantly higher for crops that are irrigated vs non-irrigated crops?
    a. Irrigation_Used
    b. Crop
6. Is there a trend where longer growing periods lead to higher productivity? Or does yield plateau after a certain number of days?
    a. Days_to_Harvest
7. What variables (soil type, weather condition) contribute to each regions growth success?
    a. Region
8. How does crop yield vary across different regions?
    a. Region
9. Are certain soil types more productive in specific regions?
    a. Soil_Type
    b. Region
10. What is the relationship between soil type and success of different crops?
    a. Crop
    b. Soil_Type
11. Are certain weather conditions associated with higher or lower crop yields?
    a. Weather_Condition