Analysis of Agricultural Crop Yield

By: Vishnu Prasad and Shane Lindquist

Data Description

Dataset:

[“Agriculture Crop Yield” – Samuel Oti Attakorah](https://www.kaggle.com/datasets/samuelotiattakorah/agriculture-crop-yield/data)

This dataset consists of agricultural data, containing one million samples (rows). Each row represents an observation and can be used for modeling crop yields. The design of our dataset is conducive to building predictive models estimating crop yields in tons per hectare based on a variety of agricultural and environmental factors.

This dataset includes attributes such as:

**Region** – Geographical location; North, South, East, West

**Soil\_type** – Soil characteristics in which the crop is planted; Clay, Sandy, Loam, Silt, Peaty, Chalky

Crop – Type of crop grown; Wheat, Rice, Maize, Barley, Soybean, Cotton

**Rainfall\_mm** – Amount of rainfall in millimeters during crop growth period

**Temperature\_Celsius** – The average temp during crop growth period

**Fertilizer\_Used** – indicates if fertilizer was used Y/N

**Irrigation\_Used** – Indicates if irrigation was used Y/N

**Weather\_Condition** – Type of weather experienced during crop growth period; Sunny, Rainy, Cloudy

**Days\_to\_Harvest** – Days taken for crop to reach optimal harvesting growth

**Yield\_tons\_per\_hectare** – total crop yield produced per hectare

This dataset paints a detailed picture of how regional, environmental, and management factors interact to affect agricultural productivity. It offers an opportunity for exploring patterns in crop yield and potentially guiding decisions in farming practices or policy.

Our intended use and application of this data set is to train regression or machine learning models to forecast crop yield per hectare. Furthermore, we can study the impact of variables like rainfall, soil type, and fertilizers usage on yield. With one million records and a rich set of features we can make some robust predictions and explore how variables interact with each other.

|  |  |
| --- | --- |
| Rainfall\_mm | Temperature\_Celsius |
| Min.: 100.0 | Min.: 15.00 |
| 1st Qu.: 324.9 | 1st Qu.: 21.25 |
| Median: 550.1 | Median: 27.51 |
| Mean: 550.0 | Mean: 27.50 |
| 3rd Qu.: 774.7 | 3rd Qu.: 33.75 |
| Max.: 1000.0 | Max.: 40.00 |
|  |  |

**Conduct some preliminary descriptive statistical analysis for your project data. Select at least 2 numeric variables of interest, report the results of descriptive statistics using summary function. Comment on any observed findings.**

We can see that rainfall values are evenly spread from 100mm to 1000mm, with a symmetrical distribution around the mean/median of 550mm which makes this set great for modeling. The temperature distribution is also centered and balanced, ranging widely form 15°C to 40°C which gives us a good basis to test it’s effect on yield.

**Select at least 2 categorical (or group) variables of interest, report frequency table and contingency table. Comment on any observed findings.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Frequencies |  |  |  |  |  |
| **df\_list$Fertilizer\_Used** | |  |  |  |  |
| Type: Logical |  |  |  |  |  |
|  | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
| FALSE | 500060 | 50.01 | 50.01 | 50.01 | 50.01 |
| TRUE | 499940 | 49.99 | 100 | 49.99 | 100 |
| <NA> | 0 | 0 | 0 | 0 | 100 |
| Total | 1000000 | 100 | 100 | 100 | 100 |

From the fertilizer categorical frequency, we can see that the dataset is split evenly between plots that used fertilizer and those that didn’t. This balanced distribution is ideal for comparing the effects of fertilizer on crop yield.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Frequencies |  |  |  |  |  |
| **df\_list$Soil\_Type** |  |  |  |  |  |
| Type: Character |  |  |  |  |  |
|  | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
| Chalky | 166779 | 16.68 | 16.68 | 16.68 | 16.68 |
| Clay | 166352 | 16.64 | 33.31 | 16.64 | 33.31 |
| Loam | 166795 | 16.68 | 49.99 | 16.68 | 49.99 |
| Peaty | 166283 | 16.63 | 66.62 | 16.63 | 66.62 |
| Sandy | 167119 | 16.71 | 83.33 | 16.71 | 83.33 |
| Silt | 166672 | 16.67 | 100 | 16.67 | 100 |
| <NA> | 0 | 0 | 0 | 0 | 100 |
| Total | 1000000 | 100 | 100 | 100 | 100 |
|  |  |  |  |  |  |

The distribution of soil types is also very uniform, with each type taking up almost exactly 1/6th of the dataset. Neither fertilizer nor soil frequencies have any missing values.

A number of numbers and symbols

AI-generated content may be incorrect.

As we consider our contingency analysis, we can see that fertilizer usage is once again uniform across, meaning no soil type appears to be favored/avoided for fertilization. The balanced nature of this data makes it easier to isolate causal effects in our model. However, this also raises a question of the reliance of the dataset itself being real, the samples so far are extremely uniform suggesting manufactured data.

**Develop box plots, histograms and frequency tables/charts using the plotting features of R. Comment on any interesting finding.**

A graph of a chart

AI-generated content may be incorrect.

When comparing the Yield from fertilized and non-fertilized crops we can see that fertilizer increases both the median and overall yield. The IQR is wider with fertilizer, which means there is a higher upper range. This suggests greater productivity potential. Outliers exist in both groups, but more so on the low end without fertilizer. **Overall, we conclude that fertilizer use has a strong positive effect on yield.**

A chart of different soil types

AI-generated content may be incorrect.

From the Yield vs Soil Type boxplot, we can see that soil types have similar central yield tendencies. Slight differences do exist; for example, Loam and Sandy seem to have slightly higher medians. **Overall, our conclusion is that Soil type has a negligible effect on yield, with no soil types being significantly more productive.**

A graph of a number of numbers

AI-generated content may be incorrect.

A graph of a number of objects

AI-generated content may be incorrect.Regarding Histogram for yield distribution, the shape is bell – curved and centered around 4.5 – 5.0 tons per hectare. The distribution is normal and the yield values range from 2 – 7, with a few outliers on each end. The distribution of values are almost perfectly normally distributed, suggesting again this data is not from real world recordings.

Finally, the proportion of recorded crop types shows us the occurrence of any given crop type is almost precisely 1/6th of the recorded observances.

**Develop correlation plots for different combination of variables. Develop scatterplots to show these relationships.  Use faceting. Comment on any observed trends.**

A screenshot of a computer

AI-generated content may be incorrect.

Beginning with a correlation matrix, Rainfall is the most predictive numerical variable for yield. Temperature and Days\_to\_Harvest are very weakly/not at all correlated to yield.

A graph showing the temperature of celsius

AI-generated content may be incorrect.

In our scatterplot for Temperature vs Yield, it is very dense and mostly vertical. This shows that temperature varies widely at all yield levels. Though there is a small positive trend shown, it can be considered negligible and concluded temperature has no significant impact on yield.

A graph showing the height of a rain fall

AI-generated content may be incorrect.

In our plotting for Rainfall vs Yield, there is a clear positive trend. As rainfall increases, yield generally increases.

A graph of different types of paint

AI-generated content may be incorrect.A graph of different weather conditions

AI-generated content may be incorrect.

When testing for all combinations of other given variables with rainfall and their influence on yield, all six soil types show a similar rising pattern – yield that increases with rainfall. The slope appears nearly identical across all soil types from a visual standpoint. Our conclusion is that rainfall matters more than soil type in determining yield, and that soil type does not significantly modify the effect of rainfall on yield.

Finally in our scatterplot for Rainfall vs Yield by Region, there is a strong positive relationship visible across all regions. There are slight visual differences in slope steepness and yield spread, but overall trends are consistent. Overall, regional variation may exist but rainfall consistently boosts yield in every region.

Ultimately given all the data points we’ve considered, rainfall amount and fertilizer use are seemingly the most impactful predictors for crop yield.

**Test the Regression model(s) in R - choose more than 2 independent (predictor) variables and a dependent (response) variable of interest to predict. All chosen variables should be numerical variables. Report your findings and how you use the model(s) for prediction and answering business questions.**

To understand how environmental factors influence crop yield (in tons per hectare), our regression model gives us the tools for prediction and decision making.

Original Formula: Yield\_tons\_per\_hectare ~ Rainfall\_mm + Temperature\_Celsius + Days\_to\_Harvest

**Simplified Model Excluding Days of Harvest: Yield\_tons\_per\_hectare ~ Rainfall\_mm + Temperature\_Celsius**

Rainfall and temperature both had positive and highly significant impacts on yield, as indicated by their coefficients: each additional 1 mm of rainfall is associated with an increase of 0.004992 tons in crop yield, and each 1°C increase in temperature contributes approximately 0.02013 tons. In contrast, Days to Harvest showed a t-value close to 0 and a coefficient with a high probability of being near zero, suggesting little to no explanatory power. As a result, we removed this variable from the model without compromising accuracy. Our model’s adjusted R2  is .592, meaning that 59.2% of yield variance is explained by the model. Rainfall and Temperature are strong predictors of crop yield; Insight into these variables can guide:

1. Irrigation strategies to optimize water use based on forecasted rainfall
2. Crop selection based on regional temperature
3. Yield forecasting, critical for supply planning and market pricing.
4. Understanding yield risks under droughts and other abnormal conditions.

A screenshot of a computer screen

AI-generated content may be incorrect.As Days to Harvest does not significantly affect yield when Rainfall and Temperature are accounted for, this suggests that the timing of the harvest is not as critical as the conditions of the climate.

**Test the CART model(s) in R - choose more than 2 independent (predictor) variables and a categorical dependent (response) variable of interest to predict. Report your findings and how you use the model(s) for prediction and answering business questions.**

Our CART model predicts the type of crop grown based on environmental and agricultural inputs. Our dependent variable was crop yield and we took Rainfall, Temperature, Soil Type, Fertilizer used, and Irrigation used into consideration. The train/test split was 70%/30%.

Our model performed with:

* ~16.67% accuracy, comparable to random guessing.
* The kappa is ~0 indicating no meaningful predictive power
* The model appeared to have a prediction bias, overpredicting certain crops like rice and never predicting ones like soybean.

The top predictors influencing crop type:

* Rainfall\_mm – strongest predictor
* Temperature\_Celsius
* Soil\_Type
* Irrigation\_Used
* Fertilizer\_Used – very low influence

An example prediction given inputs:

**Input Conditions:**

* Rainfall: 850 mm
* Temperature: 23°C
* Soil Type: Loam
* Fertilizer Used: Yes
* Irrigation Used: Yes

**Predicted Crop:** Cotton

The main takeaway here is to prioritize rainfall and temperature as primary signals in planning crop models. This model can support a high-level assessment of an environment’s suitability for crops, early-stage decision making for agricultural planning, and understanding what environmental variable influence crop growth. This model does not generalize well due to potential class imbalance and is not suitable for high – stakes or large-scale investment without some enhancement.

Conclusion

All evidence points to a consistent conclusion: rainfall and temperature are the two most important drivers of crop yield in this dataset. Their predictive strength, independence, and visual clarity across multiple analyses highlight them as the core environmental levers for optimizing agricultural output.

Lessons learned and challenges faced:

* Dropping insignificant variable improved model clarity without hurting performance.
* Plots are an essential tool for communicating data.
* Overly deep trees are hard to interpret and generalize poorly.
* Extremely clean and unform class distribution and values in the data may limit some real-world usability.
* Our CART model heavily factored certain crops due to class imbalance.
* Initial CART trees were too dense/cluttered, adjustments had to be made.
* Despite tuning, CART could not outperform random chance, indicating a need for more complex models.

While the regression model provided valuable insights into yield factors, the CART model’s limitations illustrated the complexity of crop prediction tasks. Nonetheless statistical models combined with thoughtful visualization can guide agricultural decisions, such as estimating yield under weather patterns or identifying key variables that differentiate crop types. Proper data, model tuning, and interpretability are essential for transforming raw data into actionable business intelligence.

**Group member roles**:

Work on the final project was evenly split by both members; we edited and contributed equitably to each section of the project.