Optimizing Investments on Agricultural Land

“Where should you invest your next acre? Let the data answer.”

Investing Strategy

AGENDA

Crop farming business plan

01 Introduction & Research Objectives

02 Data Collection & Preparation

03 Exploratory Data Analysis

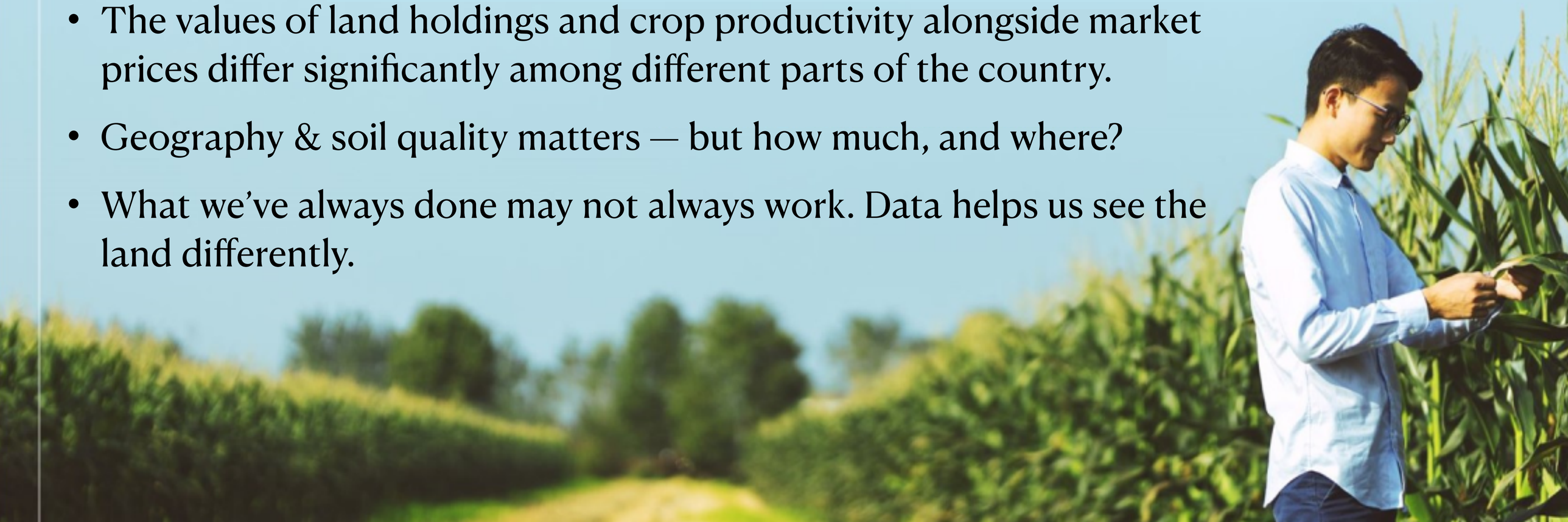
04 Modeling Strategy & Results

05 Limitations & Considerations

06 Conclusion

When you have data, why depend upon habits?

- Investments in agriculture are usually based on guesswork rather than data.
- The values of land holdings and crop productivity alongside market prices differ significantly among different parts of the country.
- Geography & soil quality matters — but how much, and where?
- What we've always done may not always work. Data helps us see the land differently.



Research questions

Soil Characteristics

What effects do soil characteristics have on the yield of crops in various geographical areas?

Historical Yield Data

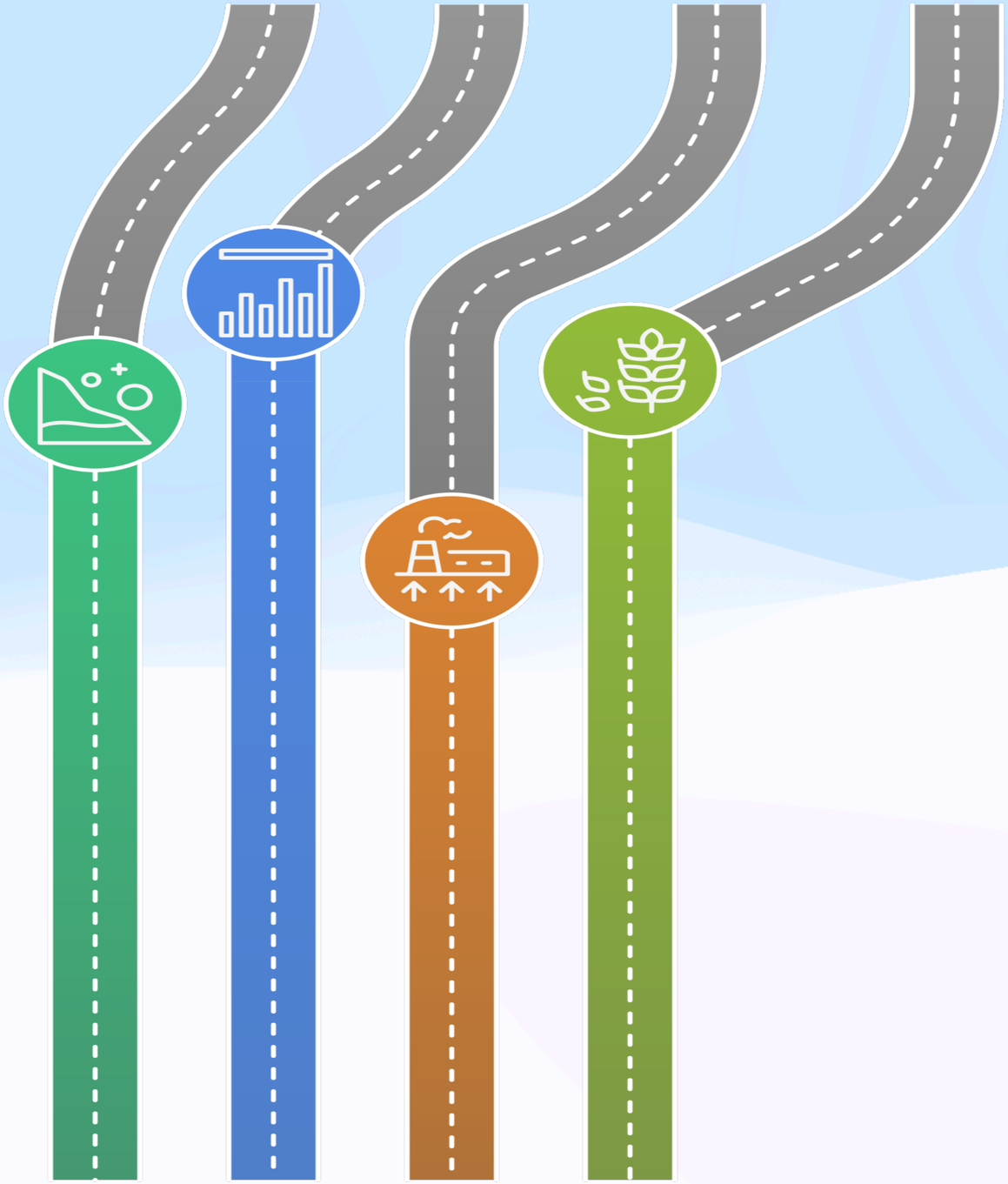
Over the last ten years, which states and counties have produced the best yields of various crop types?

Yield Influencing Factors

Which factors affect rising or falling yields and how have patterns in agricultural yields altered over time?

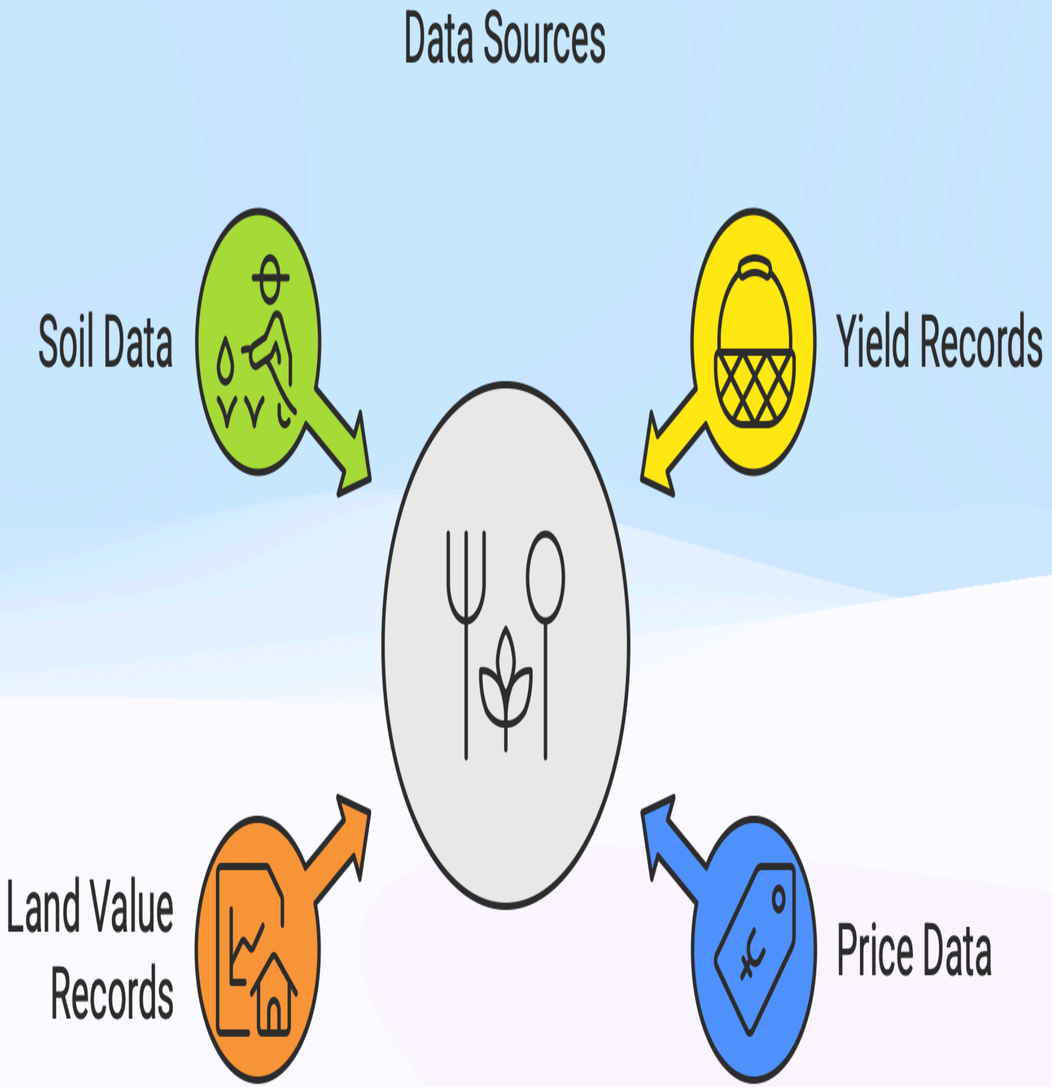
Crop Selection and Investment

In order to achieve maximum profitability in various farming sites. how can we best choose crops and make investments?



Data Collection

- There was no single, unified dataset.
- We collected **individual yield records** for 9 crops (Corn, Barley, Cotton, Oats, Peanuts, etc.) from USDA NASS.
- Then came the **price data** — separately tracked by crop and year.
- We sourced **land value records**, but only in periodic intervals.
- Finally, we gathered **soil data** — pH, depth, and texture — at the **county level**.



Data Cleaning & Preparation:

- **The units of various crops varied.**

  We modified all prices to USD/kg and yields to kg/acre.
- **Combining data from multiple sources**

  Using a composite key, it was unified: County + State + Crop Type + Year
- **Missing values existed in the soil dataset.**

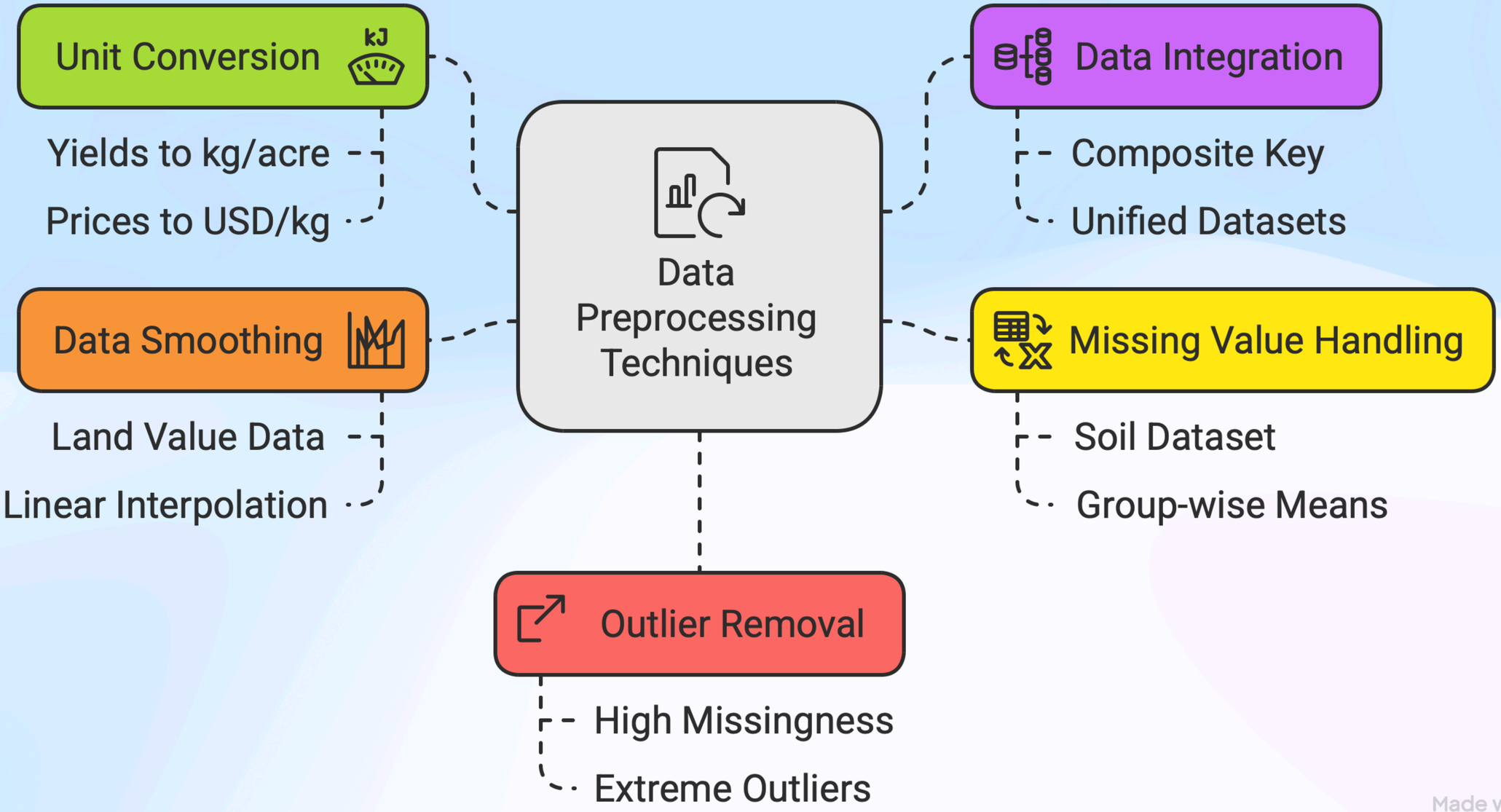
  Based on crop-soil clusters, imputed using group-wise means
- **Data on land values was periodic.**

  In order to estimate values in missing years, linear interpolation was used.

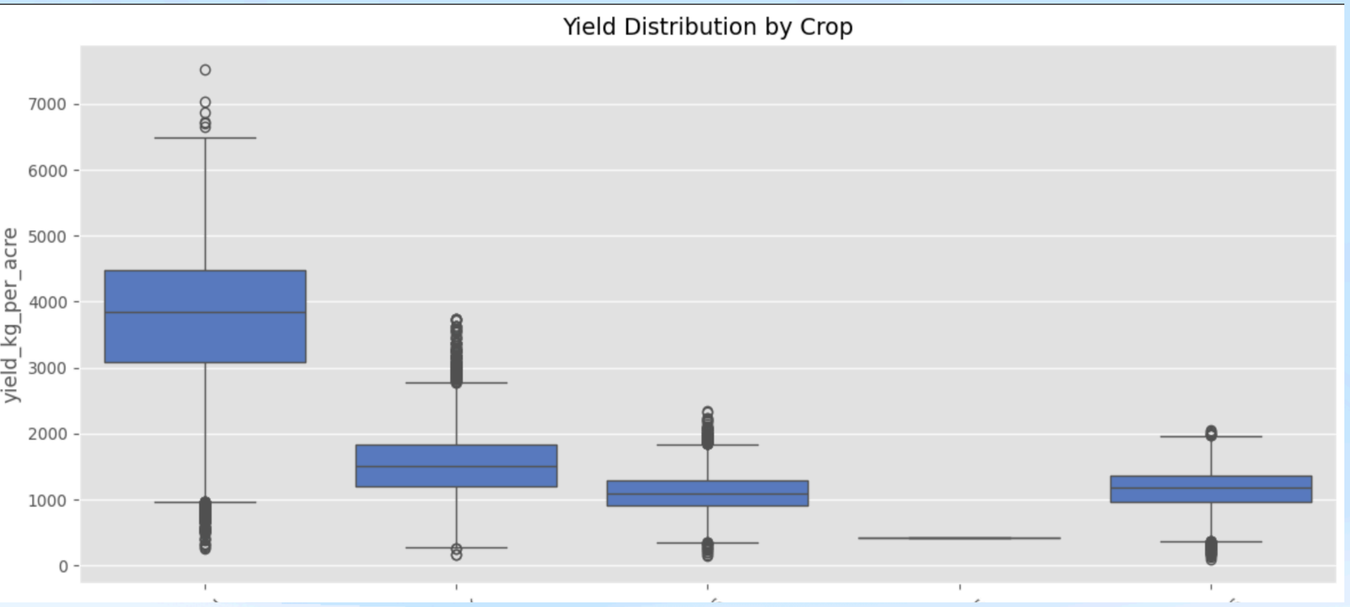
  Maintained growth patterns with no amplifying outliners
- **Crop types with extreme outliers or high missingness were eliminated.**

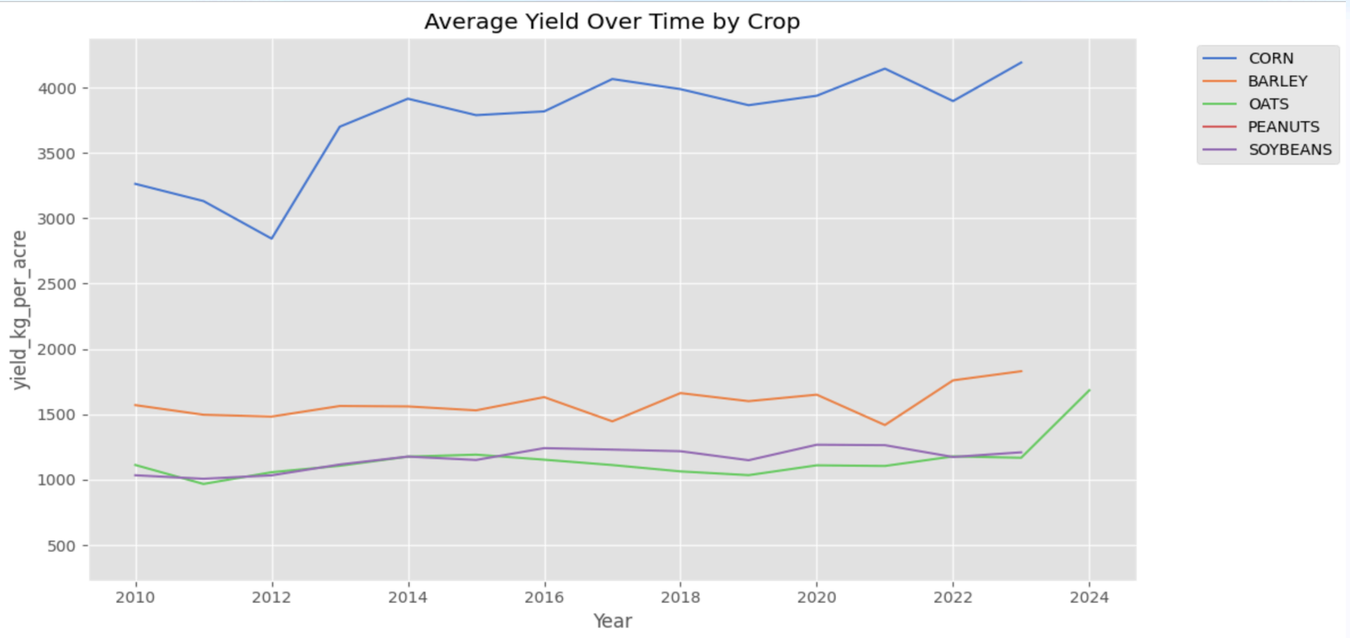
  Guaranteed dataset balance and model accuracy

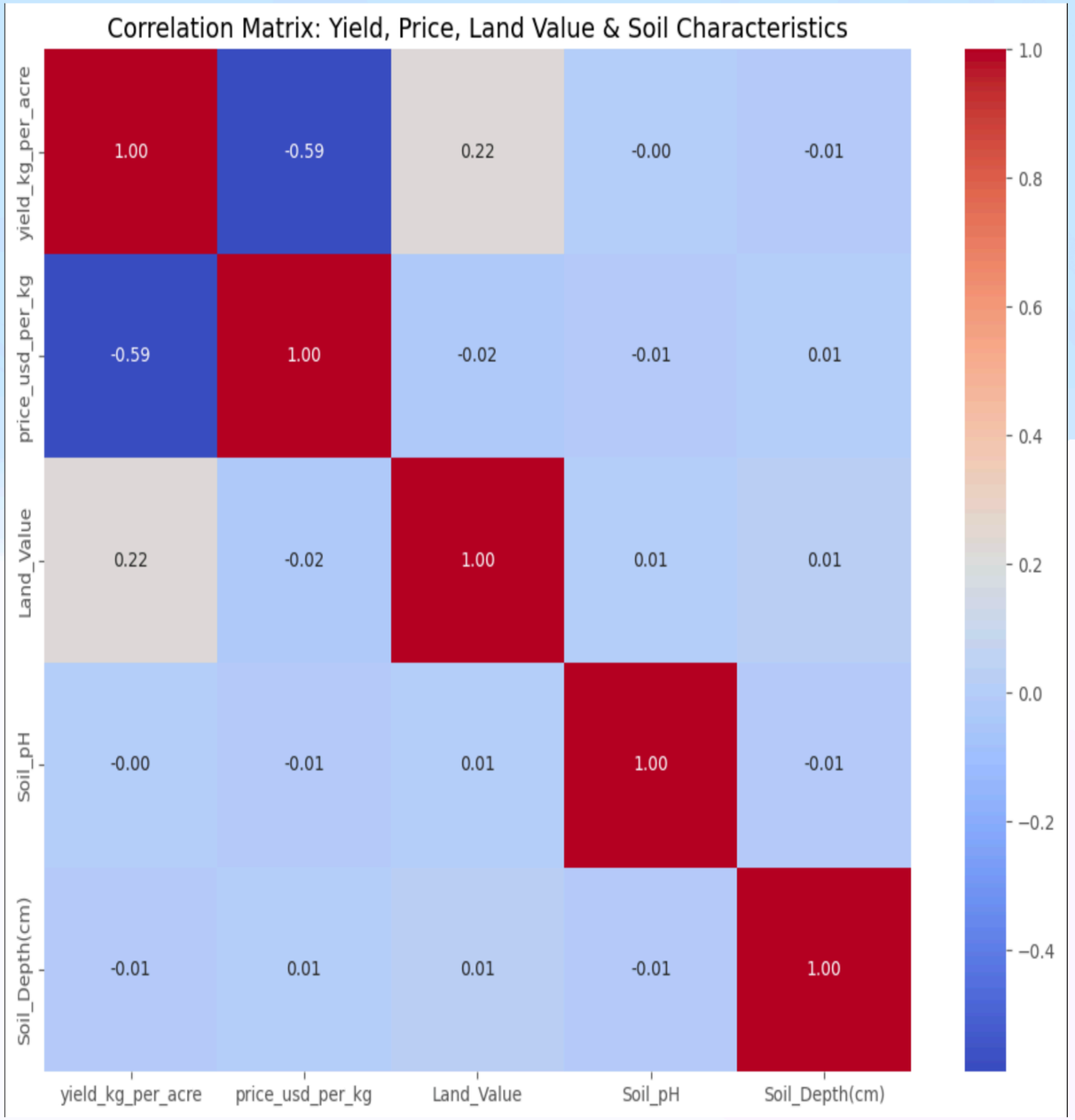
Data Preprocessing Techniques for Agricultural Analysis



Exploratory Data Analysis: Uncovering Patterns Across Soil, Yield & Value







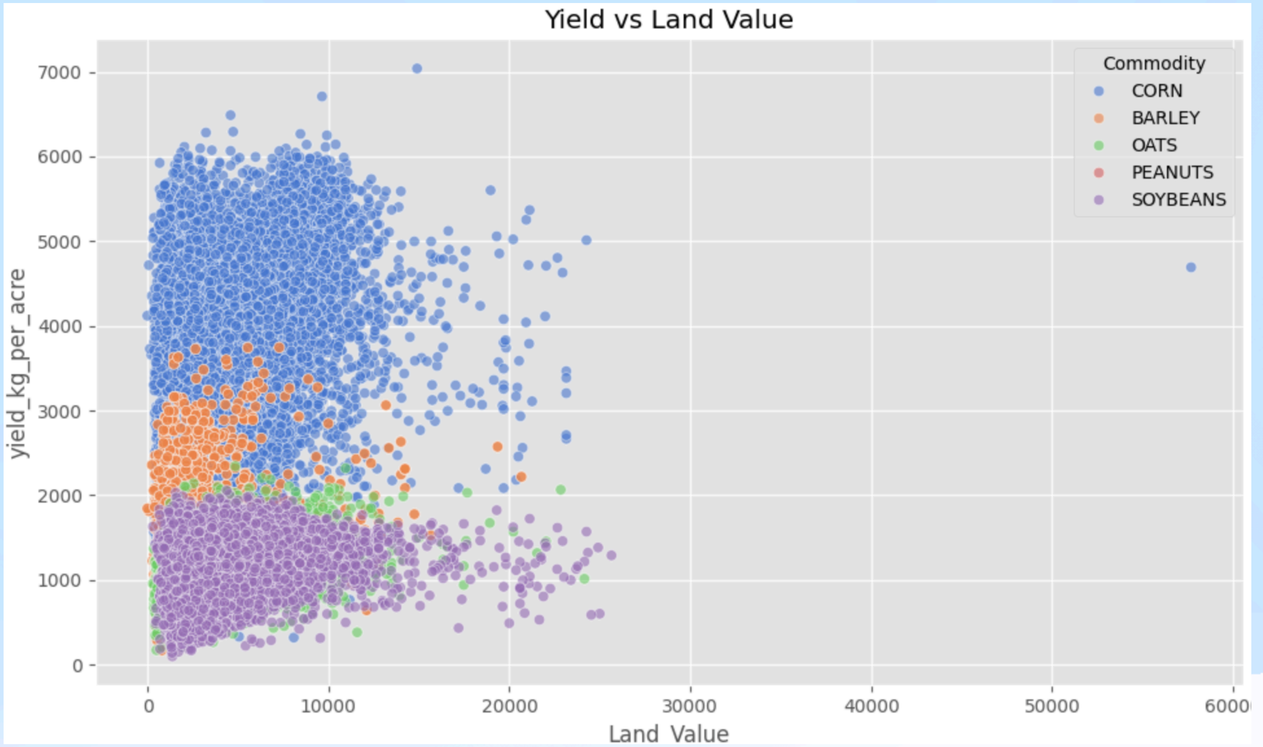
Understanding Yield & Land Patterns

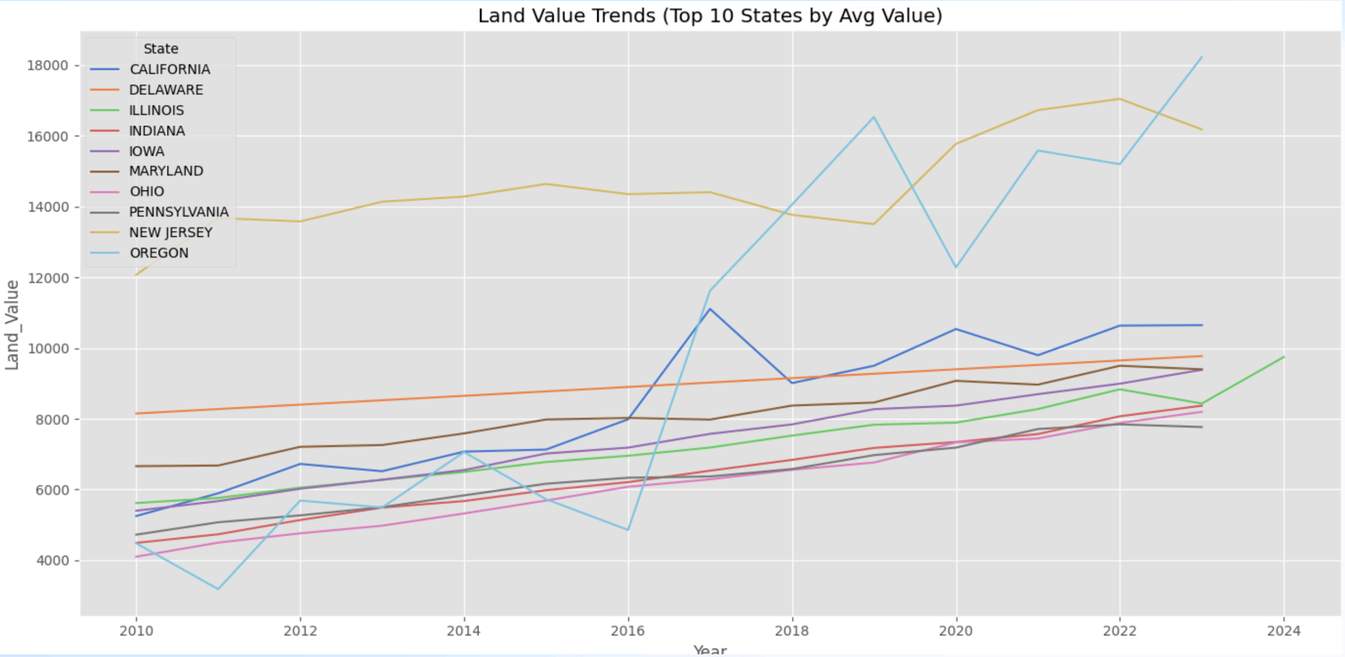
We began our analysis by exploring key patterns in crop yield, land value, and price behavior.

- Corn leads in yield, but also shows the highest variance -> risk + reward
- Oats & barley have lower median yields and are more stable
- Soil characteristics alone show weak linear correlation with yield
- hinting at non-linear relationships explored in modeling
- Yield Low as Price High (r = -0.59) → The traditional supply-demand relationship
- A high yield standard deviation (approximate 1,464 kg/acre) indicates significant productivity variation among counties
- The pH of the soil clusters at about 6.5, indicating that the majority of counties function within the agriculturally neutral range.

The most profitable land isn’t the most productive

- Shows that high-yield ≠ high-value land
- Yield doesn’t always mean value. Some counties have high yields, but they are on expensive land. Some are far more profitable however yield moderately.
- Many high-yield counties are located in areas with minimal land costs, making them perfect for profit-driven investments.



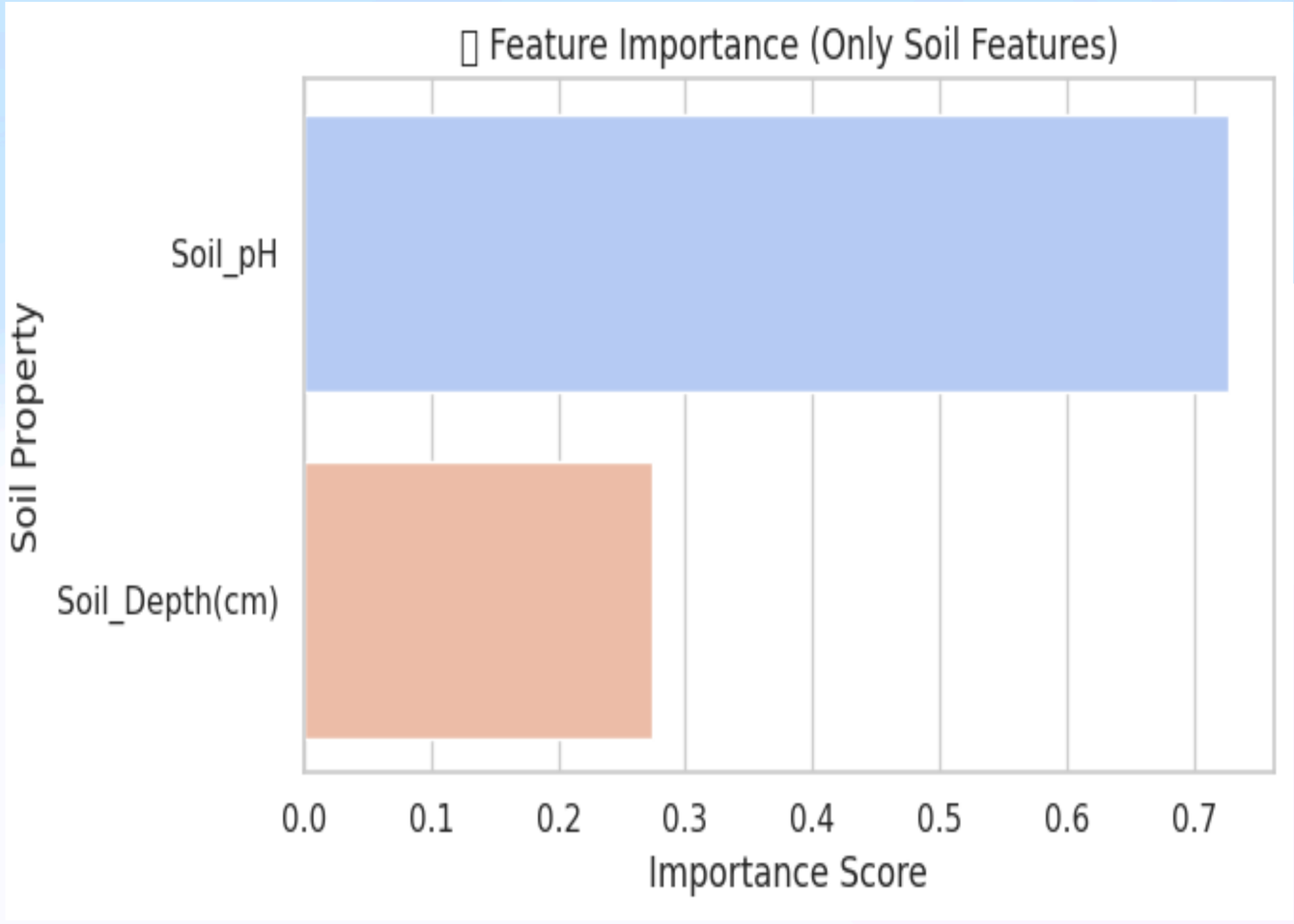


Landscape of Farmland

- California, New Jersey, Oregon top the land value charts.
- Land prices are steadily rising (2010–2024).
- In certain states, land prices are inflated by urban proximity, independent of agricultural productivity.
- The disconnect between cost and productivity prompted us to develop the Profitability Index, a more effective approach to determining where investment actually delivers returns.

Soil Characteristics Influence Crop Yield

- Soil pH – affects nutrient absorption
- Soil Depth – deeper soil generally supports stronger root growth
- Soil Texture – impacts water retention and drainage.
- Early analysis revealed a weak linear relationship between soil characteristics and yield. Yield patterns among counties could not be explained by soil alone.
- Even if the soil isn't ideal for a certain crop, farmers regularly use fertilizers & techniques to make it work, so poor soil doesn't always meant poor yield."



Prediction: How We Modeled Agricultural Yield

With the help of EDA's results, we moved toward applying machine learning to forecast agricultural yield in an effort to determine what actually influences productivity.

- Target Variable / Input Features :

  yield_kg_per_acre / Soil pH, Depth, Texture, State, County, Commodity

- Algorithms :

  Linear Regression (baseline – interpretable, limited)

  Random Forest Regressor (nonlinear, high accuracy)

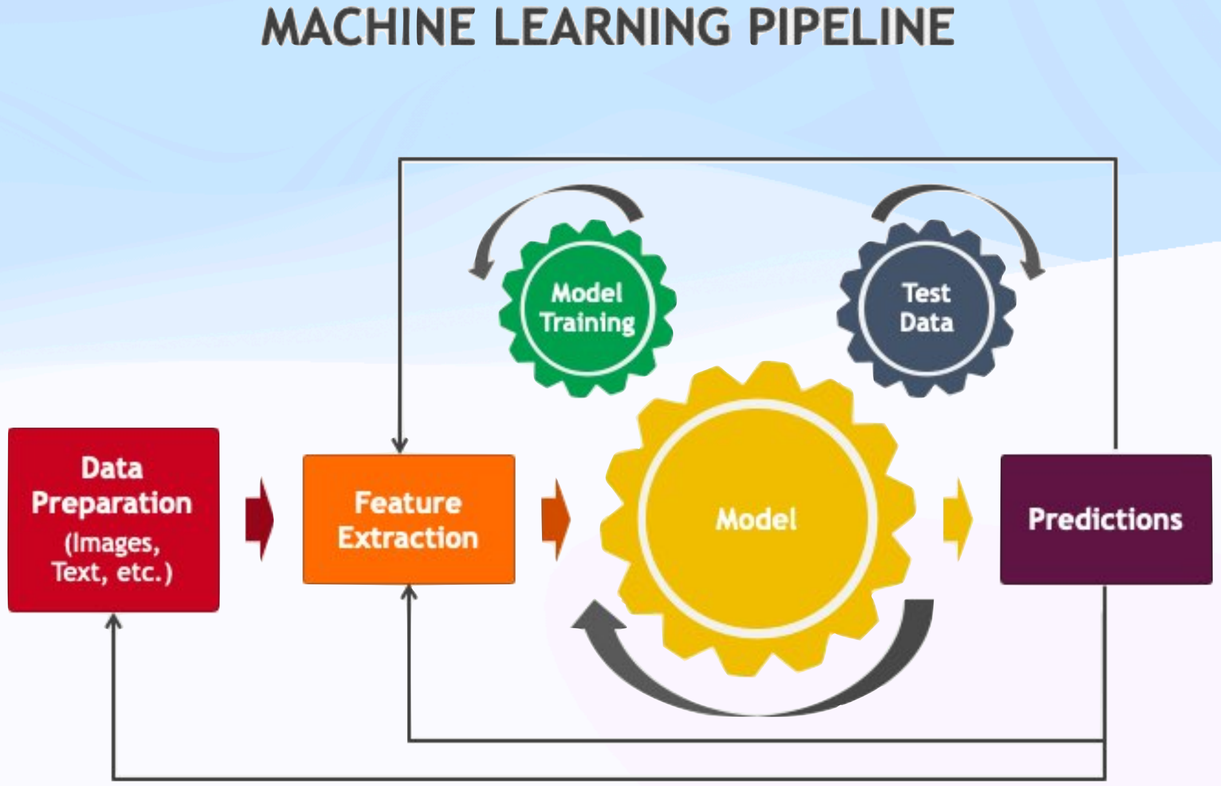
  XGBoost Regressor (advanced ensemble, refined performance)

- Preprocessing :

  Label encoding for categorical features (such as State, County, and Commodity)

  Imputed missing values before to training

  80% train / 20% test split for evaluation



Model Performance & Interpretation

- linear regression provided only a limited level of accuracy because it assumed linear, additive correlations.
- Non-linear and interaction effects were captured by Random Forest, which significantly enhanced performance.
- The most reliable and accurate predictions were generated by further refining those results using XGBoost.
- The wide difference between the ensemble models and linear regression is normal & expected. Nonlinear and contextual factors, such as market behavior, soil-crop interactions, and regional influences, affect agricultural yield and are not captured by linear models.

Model	R^2 Score	MAE	RMSE
Linear Regression	0.441	872.51	1090.39
Random Forest	0.911	290.93	434.85
XGBoost Regressor	0.916	292.52	421.95

Prediction to Investment

Our XGBoost model not only about accuracy.We utilize this as a prediction engine to support more intelligent crop and land investment strategies, compute profitability, & simulate yield outcomes.

Uses of our model.

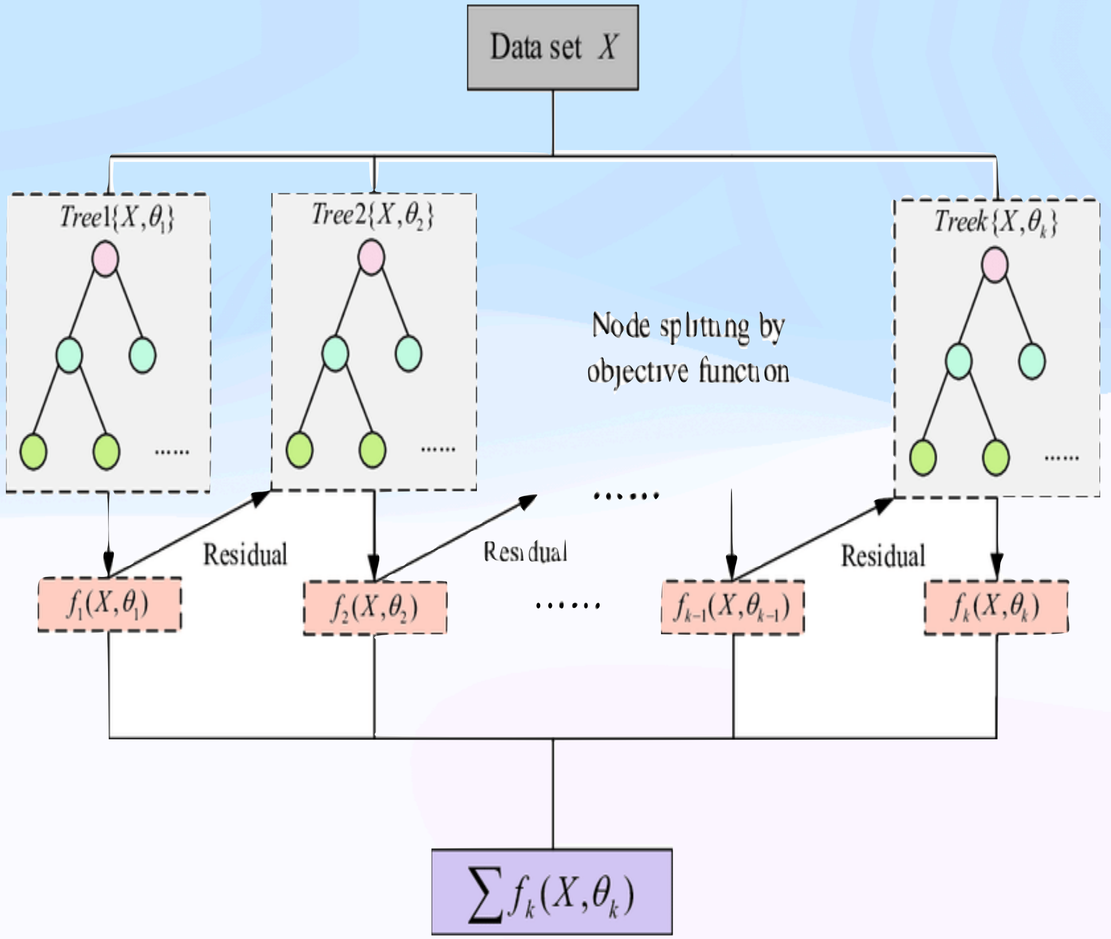
- The model can accurately predict county crop yields from merged soil, price and region datasets.
- We can calculate profitability index

  Profitability Index = (Yield × Price) / Land Value
- Compare crop scenarios across regions:

  “Barley in Montana” vs “Soybeans in Oklahoma”

  “Corn in California” vs “Peanuts in Virginia”

  can find the most profitable match of crop + location + market



The Profitability Index

- the Profitability Index (PI) for calculating the probable return on investments

  Profitability Index = (Predicted Yield × Price) / Land Value

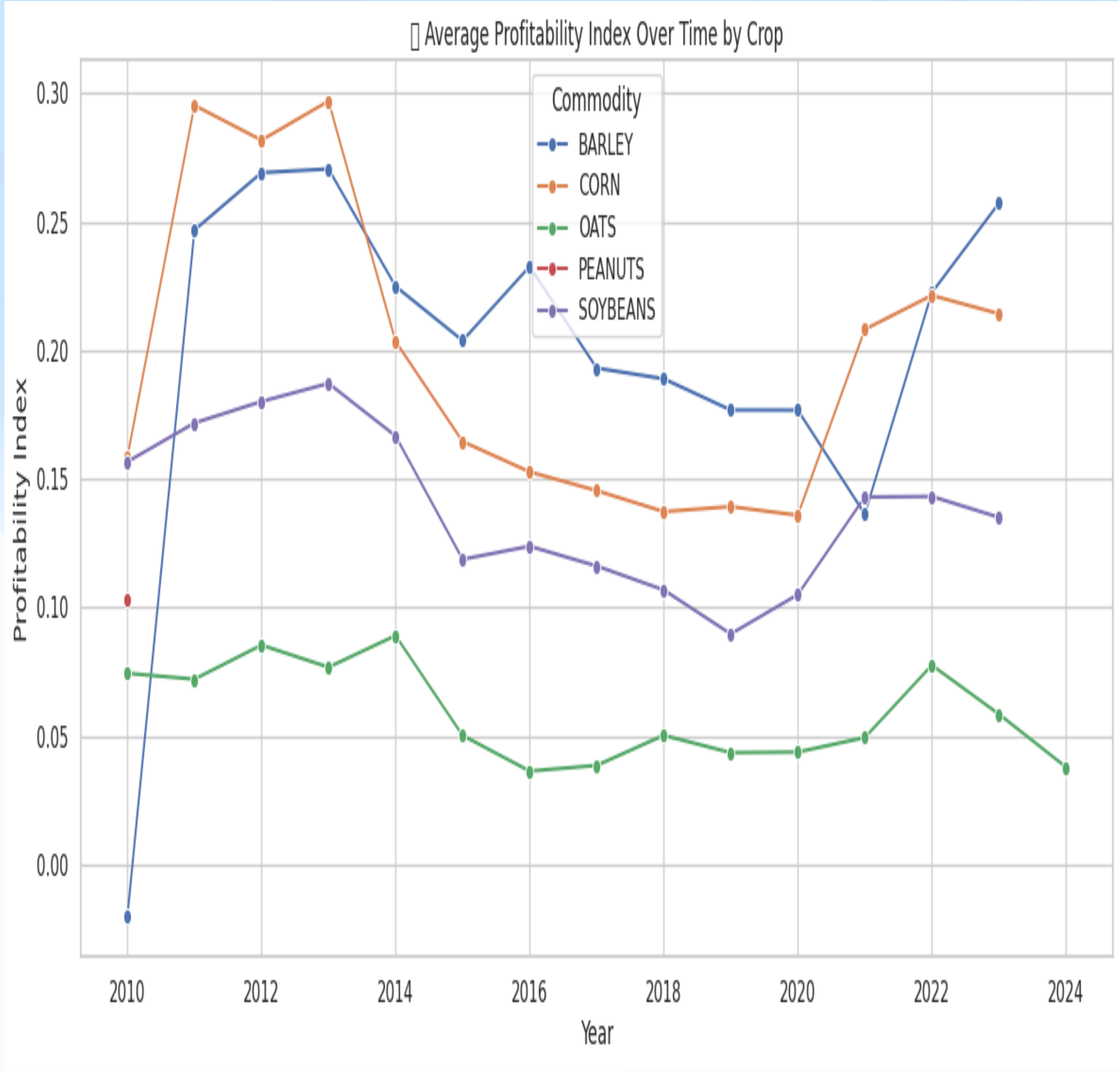
- It matters because Considerations for regional differences in land prices
- Allows for crop-location comparisons for improved decision-making;

  goes beyond productivity to concentrate on economic return per acre.

- Example

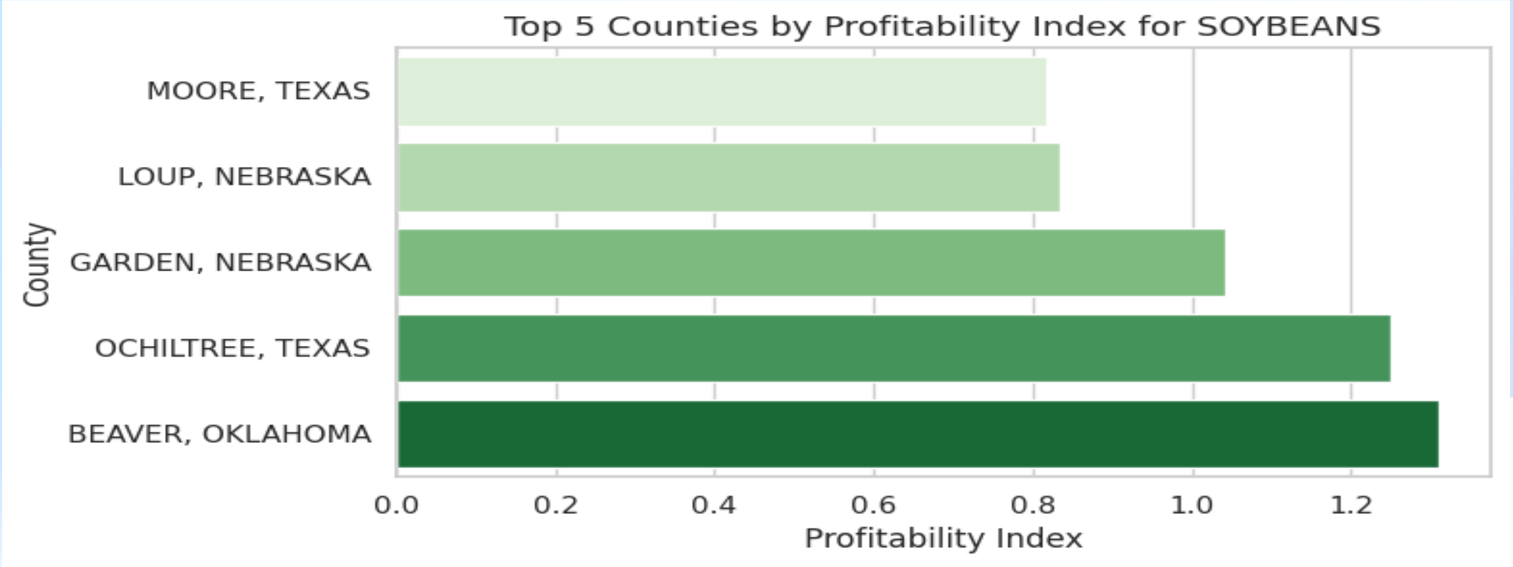
  County A: PI = 0.10 (high yield, high land cost)

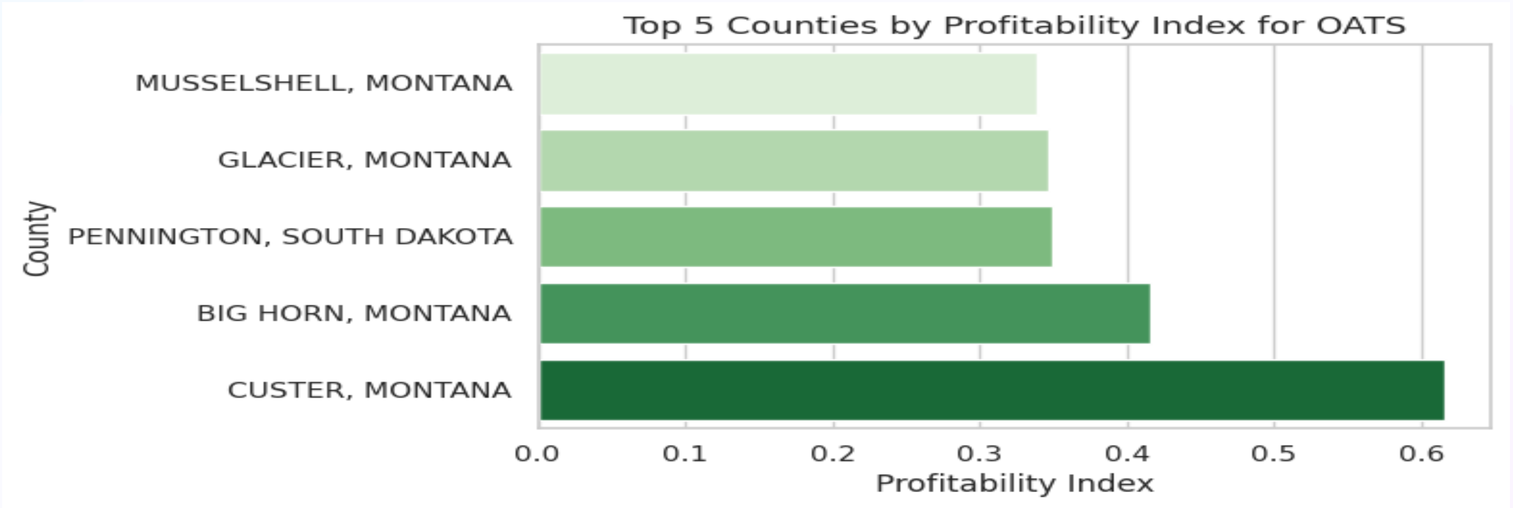
  County B: PI = 0.16 (lower yield, cheaper land)



Top Investment Zones – Profitable Crops by Region

- We calculated the Profitability Index for every county crop pair across the U.S.
- This enabled us to identify high-return areas that might not have the highest yields.
- In terms of return per dollar, several obscure counties performed better than well-known agricultural centers.
- Soybeans: 📍 Beaver (OK), Ochiltree (TX), Garden (NE) – Moderate yield, low cost → good ROI.
- Custer & Greensville *(VA)* – Lower overall PI, but stable for niche markets





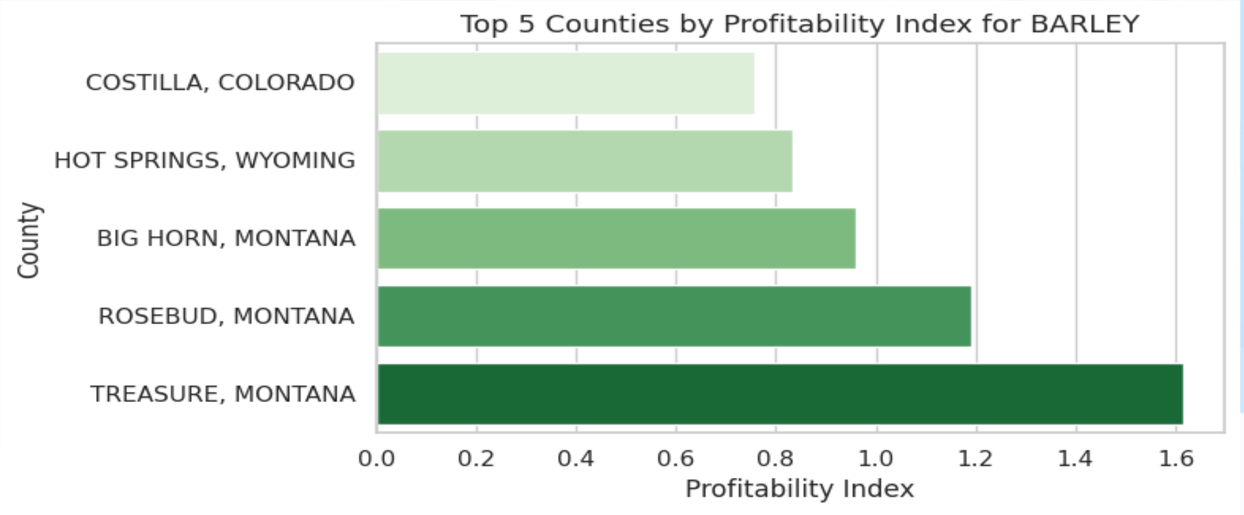
- Corn:

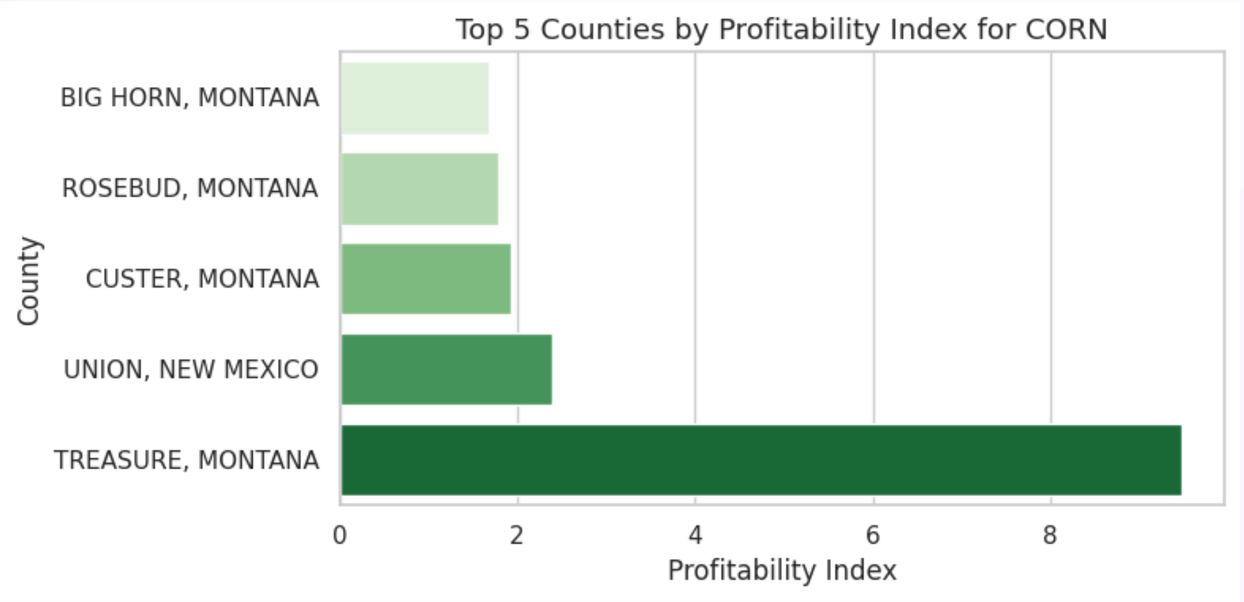
  📍 Treasure County, MT – PI = 9.48,Also strong in Union (NM), Custer (MT), Rosebud (MT)

- Barley:

  📍 *Big Horn & Rosebud (MT)* – Consistently high PI ,Ideal for cereal crop investments due to low land cost + decent yield

- These areas generate returns in addition to crops. Finding the land-crop combinations that optimize value per acre is made easier by the Profitability Index.





Limitations

- There is no weather or climate data included. The yield is extremely susceptible to frost, drought, and rainfall.
- Costs of labor, fertilizer, irrigation, transportation, Crop insurance or financial aid True profit = PI adjusted for actual expenses
- Limitations on Land Value

  Data was available only in intervals.To estimate the years that were missing, we utilized linear interpolation.No policy effects as well as market disruptions are mentioned.
- impact of the soil.

  Models predict low yield based only on soil characteristics.However, farmers use technology and fertilizers to improve poor soil.

Weather

Represents potential weather-related risks

Cost

Indicates financial risks and expenses

Policy

Highlights policy-related risks and regulations

Assumptions

Represents risks associated with underlying assumptions

Conclusion

- We organized an unorganized data—yield, price, land value, and soil—into something useful.
- We used machine learning, we didn’t just predict how much a crop would grow, but figured out where it actually makes sense to grow it.
- We have the ability to look at any county & any crop and have a good estimate of the expected yield with our model. More importantly, we combined this with land prices to create the Profitability Index, which is an even more effective tool.
- We learned that soil helps, but it’s not everything. Land cost and crop type matter more than we expected. And while our model doesn’t include things like weather or fertilizer costs, it’s still a strong first step toward smarter, data-backed farming
- In agriculture, yield is important — but profit is essential. With the right data and models, we don’t just grow more.


THANK YOU