Agriculture Crop Yield Analysis

Objective:

Analyze crop yield datasets to uncover environmental and agricultural factors that influence crop productivity across U.S. regions. The goal is to identify actionable insights to support strategic yield optimization efforts.

Skills:

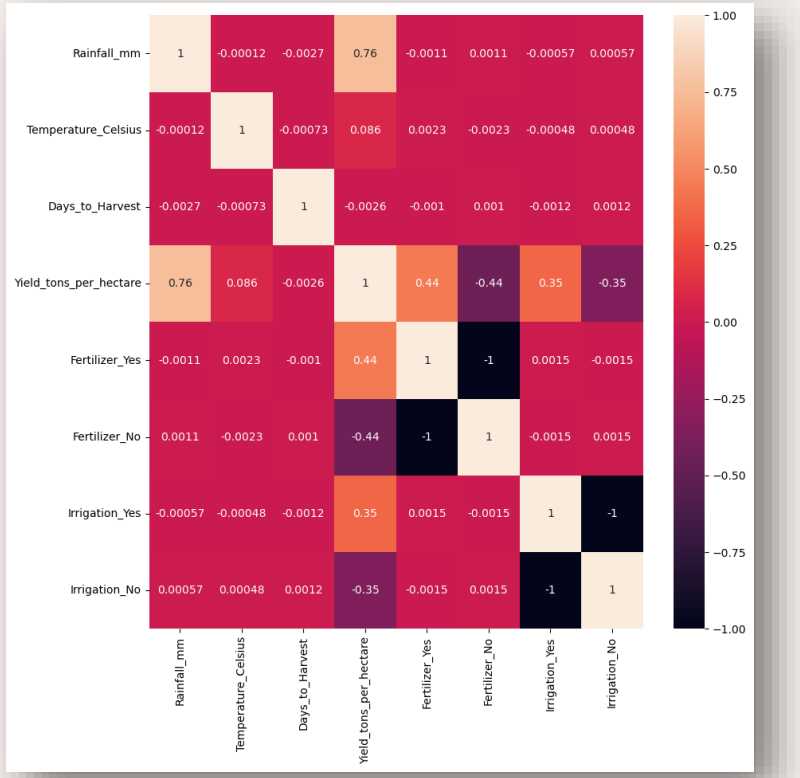
Data Cleaning & Preparation, EDA, Regression Modeling, Clustering Analysis, Geospatial Visualization, Time Series Analysis (stationarity testing/autocorrelation/Dickey-Fuller test), Decision Tree Modeling, Dashboard Development

Tools:

Project Flow

1. Data Collection	- Imported crop yield dataset - Sourced monthly yield data (‘08 – ’22) from NASDAQ WASDE
2. Data Cleaning & Preparation	- Performed data wrangling and consistency checks - Filtered & cleaned time-series data for a clean monthly analysis
3. Exploratory Data Analysis (EDA)	Explored variable distributions and correlations with yield
4. Geospatial Visualization	Mapped regional crop yields using a choropleth map
5. Regression Analysis	Built a linear model to predict yield based on key features
6. Clustering Analysis	Applied K-Means to segment regions and crop conditions
7. Decision Tree Modeling	Created a decision tree to identify top yield predictors
8. Dashboard Development	Built an interactive Tableau Dashboard showcasing results and recommendations

Exploratory Data Analysis



Rainfall (mm) and **Yield (tons/ha)** show a strong positive correlation **(r = 0.87)**

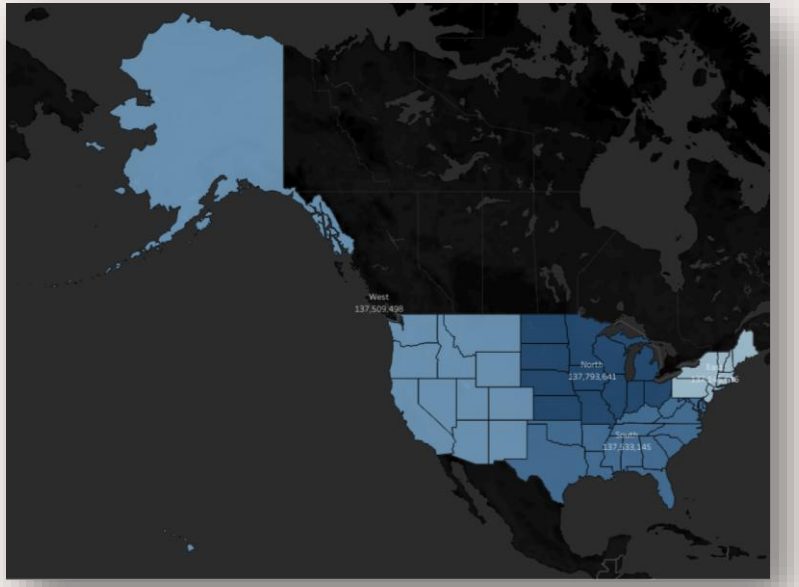
Null Hypothesis (H_0): There is no significant relationship between the rainfall a crop receives and its yield. I.E. increased rainfall does not result in a higher crop yield.

Alternative Hypothesis (H_1): There is a significant relationship between the amount of rainfall a crop receives and its yield. I.E. increased rainfall is associated with higher crop yields.

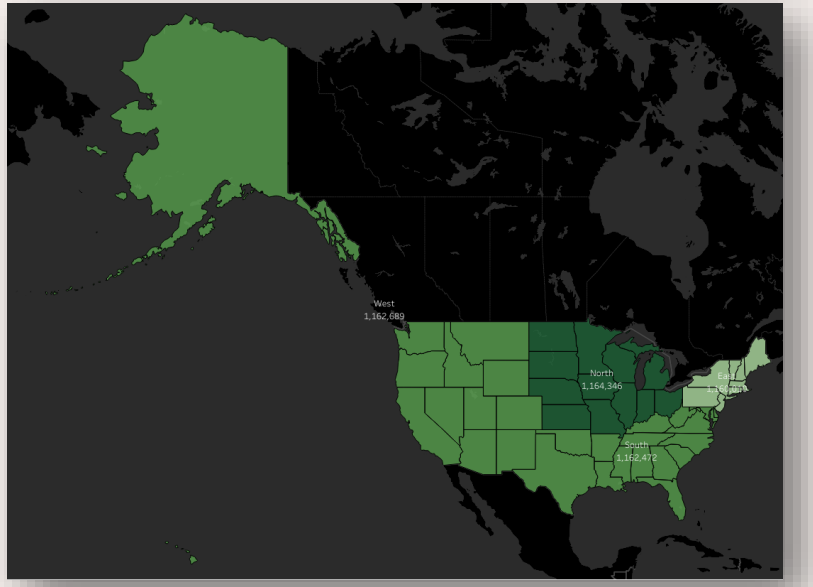
Geospatial Analysis

Regional Patterns in Rainfall and Crop Yield

Rainfall Choropleth



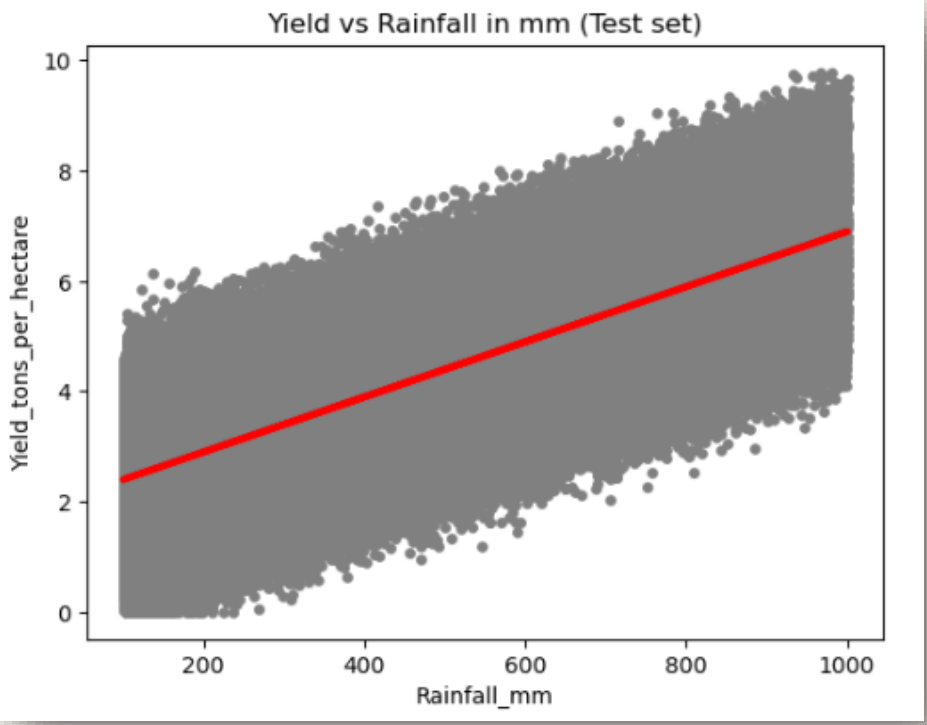
Yield Choropleth



- Regions with higher total rainfall also showed **greater crop yields**
- The **North** had the **highest rainfall and yield**, while **Eastern regions** trailed in both
- These maps visually reinforce the earlier finding that **rainfall is a key driver of yield**

Regression Modeling

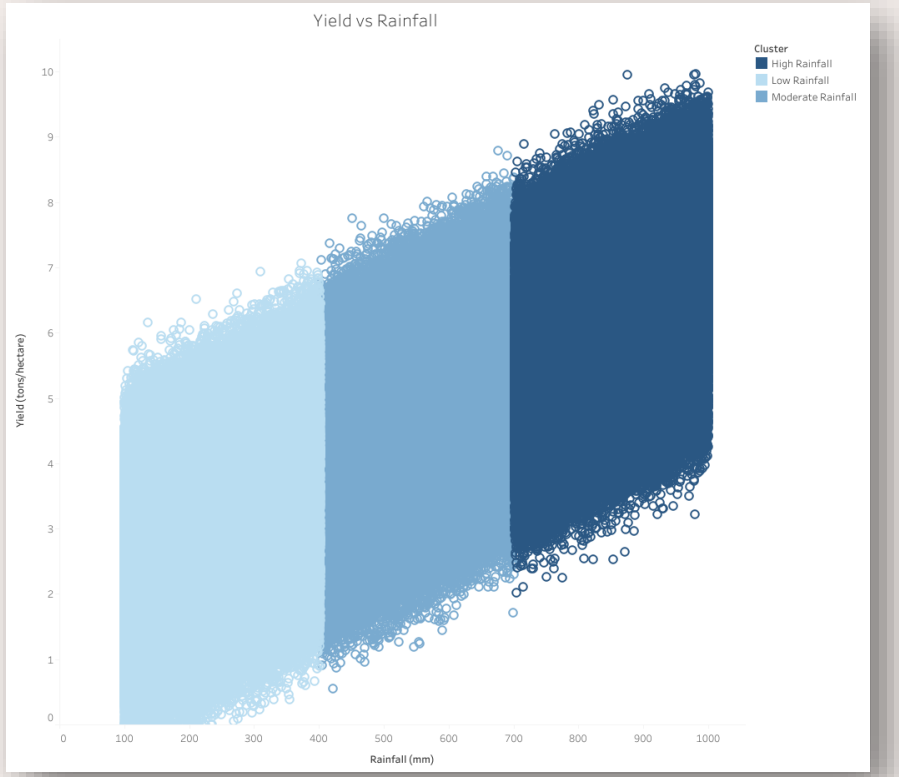
Predicting Crop Yield from Rainfall



- Yield vs Rainfall (Test Set)
  - Red line represents predicted linear relationship
  - Points show actual test data
- Model Performance
  - **R^2 Score**: 0.58
  - **Slope:** 0.005
  - **MSE:** 1.19
- Interpretation
  - Rainfall explains 58% of yield variation
  - Each additional mm of rainfall increases yield by 0.005 tons/hectare
- Rainfall has a **strong positive effect** on yield
- Confirms earlier EDA and geospatial patterns

Clustering Analysis

Yield Clusters by Rainfall Intensity



Objective: Segment crop yield observations using **K – Means Clustering** to uncover how yield levels group across different rainfall conditions

Interpretation:

High Rainfall Cluster

- Highest mean yield (6.15 tons/ha); shorter harvest time; optimal growing conditions

Moderate Rainfall Cluster

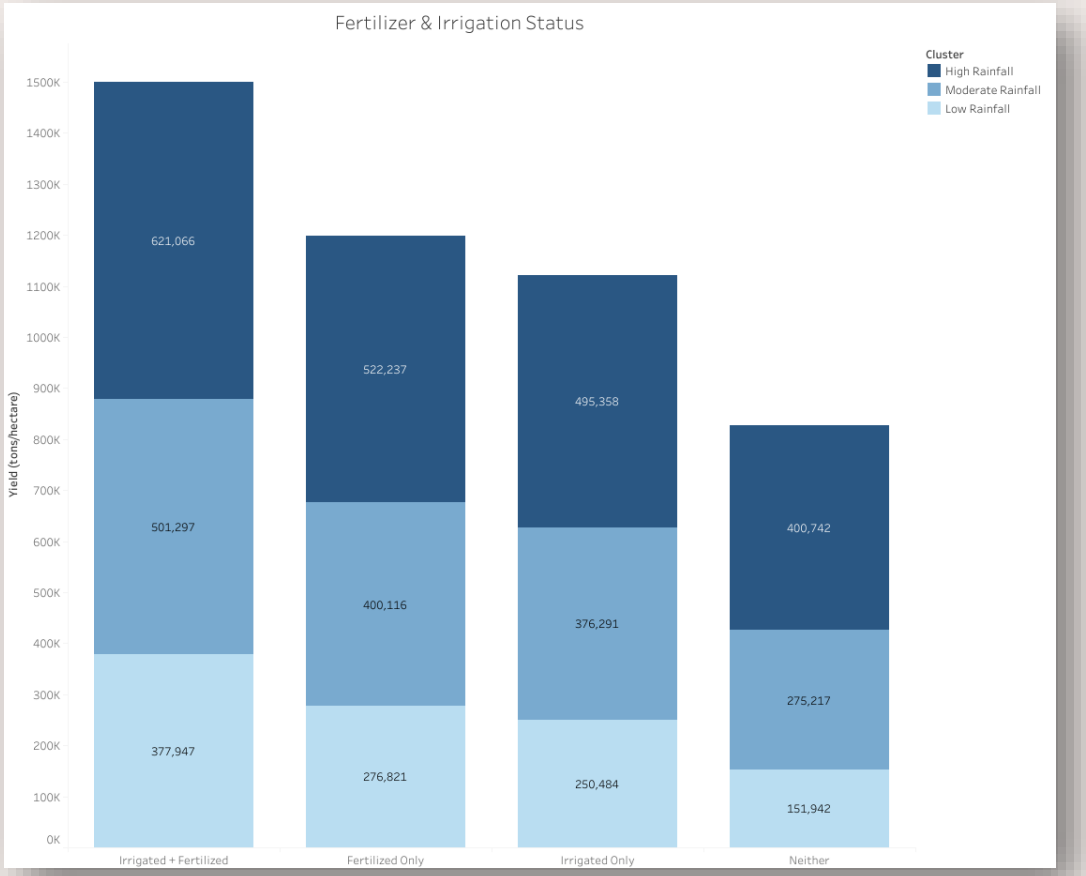
- Solid yield (4.68 tons/ha); balanced harvest period; consistent productivity

Low Rainfall Cluster

- Lowest yield (3.15 tons/ha) despite similar growing time; rainfall – limited regions

Decision Tree Modeling

Enhancing Yield Through Common Agricultural Practices



- After rainfall, **fertilizer use and irrigation** were identified as the next most influential variables using a **decision tree regression model**
- **Crops receiving both inputs consistently outperformed all others**, across all rainfall levels
- Crops without any intervention had the **lowest total yields**

Final Insights & Recommendations

Key Findings

- Hypothesis Test (Rainfall)
  - With a **p – value of 0**, the **null hypothesis is rejected**
  - There is a **significant positive relationship** between rainfall and yield
  - Rainfall alone explains **58% of the variance** in yield
- Irrigation & Fertilizer Impact
  - Fertilizer increases yield by **38%**
  - Irrigation increases yield by **30%**
  - Combined use increases yield by **43%**
- Other Variables
  - **Categorical variables** like weather, crop type, soil and temperature showed **no strong correlation** with yield

Recommendations

- If operating in a **high rainfall environment** (>700 mm), expect a median yield of **6.15 tons/hectare**
- If growing in **moderate or low rainfall regions → Adopt common agricultural practices** (fertilizer + irrigation) to **mitigate lower rainfall impacts**

Takeaways & Deliverables

What did I learn?

- How to combine exploratory, geospatial, and machine learning techniques to tell a cohesive story
- Clustering and decision trees helped validate and visualize complex relationships
- Real – world interventions (fertilizer and irrigation) are as important as modeling skills

What can I improve?

- Strengthen my overall machine learning skills, especially in model tuning, feature engineering, and evaluation techniques
- Build more robust multivariate regression models

Deliverables

- Interactive Tableau Dashboard
- GitHub Repository