Understanding the Instruction Prompt 3.

Here is short summary of the whole promp3.pdf outlines a data analysis challenge focused on crop diversification, agricultural land, and production metrics in the United States. The challenge encourages investigating the historical and spatial relationships among these factors, considering regional variations and external influences such as urban expansion. Here's a breakdown of the key aspects:

Definition of a Farm: The document emphasizes that a farm is defined as any location generating over $1,000 from agricultural product sales within the Census year

Challenge Objective: Participants are tasked with examining how crop production patterns relate to land value, farm size, production output, and external factors like urban sprawl This involves analyzing historical trends and potentially building predictive models

Flexibility and Focus: The prompt is designed to be open-ended, allowing teams to focus on specific regions or crop types

Data Provided:Two datasets are provided: "land_use_farm_ops.csv": Contains information about land use practices (cropland, pastureland, woodland), farm operations, and economic indicators like crop insurance. "Prompt3_wide_3.0.csv": Focuses on crop production volume, the number of farms growing specific crops, and data on harvested and irrigated land

Essentially, this prompt invites an in-depth exploration of how crop choices, land use, and economic factors intersect within the US agricultural landscape. The provided datasets and guiding questions aim to uncover trends, patterns, and potential future implications within this domain.

Understanding the data set.

1. **land_use_farm_ops.csv** – The dataset captures a comprehensive overview of agricultural land use and management practices, focusing on cropland, pastureland, and woodland metrics, alongside farm operation details. It provides a robust foundation for analyzing land efficiency, risk management through crop insurance, and the economic health of farming operations across various regions.
2. **Prompt3_wide_3.0.csv** – The dataset focuses on measuring the extent of crop production and the number of farming operations associated with diverse crops, emphasizing both harvested and irrigated land. It serves as a valuable resource for analyzing trends in agricultural practices, crop yields, and resource allocation across different regions.
3. **GDP_price_index_Table.xlsx** The "GDP_price_index_Table.xlsx" dataset presents a comprehensive table detailing various price index metrics associated with Gross Domestic Product (GDP). Each column represents different dimensions of the GDP, such as specific economic categories or time periods. The dataset includes columns labeled with categories like consumption, investment, and government spending, each tracked over a series of years. The dataset also features unlabeled columns, possibly containing metadata or auxiliary data related to conversions or formatting.This table

enables a detailed analysis of how different components of the GDP's price index evolve over time, allowing for comparisons between multiple sectors and their contribution to economic changes.

4. **Sales_data_county.csv** – This dataset focuses on agricultural production and sales at the county level across various crop types and animal products. It provides detailed metrics related to the number of operations, total sales measured in dollars, and other economic indicators across multiple years. The dataset also captures regional agricultural trends over time, making it valuable for analyzing crop and livestock production, sales, and operational scale in different counties.

Possible questions to answer: Analyze the historical trends in crop diversification by the crop types and
sales grown in a county and how they relate to the economic valuation of
agricultural land.

Tools

Viz Hints:
To create your team’s final visualizations, it will be beneficial to map your values across the counties of the contiguous United States. A choropleth map is a type of map that uses color to represent data across geographic areas. The term comes from the Greek words choros (region) and plethos (multitude). Using a tool like Python Plotly's choropleth maps can enhance your visual analysis. For more information and examples, visit Plotly's choropleth map documentation and ways to create USA County Choropleth Maps in Python.

Ultimate goal:

Link crop diversification with the overall increase in sales of crop production and how this affects the economic growth of the region which is measured as the economic value of assets.

Stretch goal:
Fit a time series forecasting model to predict next year economic value

Context1: A wealthy farm man wants to open another farm business. Where and what crop should I invest into and how much is the investment of the land in terms of buying.

Context2: A local farmer would like to know which crop to grow for the next 5 years to optimize the revene output in tuheir land

MAP DIVISION

https://www.worldatlas.com/articles/the-officially-recognized-four-regions-and-nine-divisions-of-the-united-states.html

THINGS WE NEED TO DO:
1. Get the price the one we have in the TOTAL 5 CORPS
2. Top 5 prices
3. Top 5 profitable crops

Analysis on Crop Diversification, Agricultural Land, and Production Metrics

HDSI Agri Datathon 2024

Team Crop Pop

Samuel Damon	Frank O	Sai Akhil Rayapudi	Harish Narava
Umass Boston	Student	NEU	NEU

Abstract

This study examines the interplay between crop diversification, agricultural land use, and production metrics across the United States, drawing on historical data from the Census of Agriculture. Since 1974, farms have been defined by a sales threshold, ensuring comprehensive coverage of various crop types and regions. By focusing on key agricultural areas such as the Midwest, Southeast, and West, we explore how shifts in crop production have correlated with changes in farmland size, land value, and productivity. Additionally, we assess the role of external factors, including urban expansion, in influencing crop distribution and land use patterns. This analysis aims to provide insights into the evolving dynamics of U.S. agriculture and contribute to strategies for sustainable land management and agricultural development.

1. Introduction

The U.S. agricultural sector has experienced significant changes in crop diversification, land use, and production patterns over recent decades. These shifts have been influenced by a variety of factors, including market demands, technological advancements, and external pressures such as urban expansion. Understanding the dynamics of crop diversification is crucial for addressing key issues in agricultural sustainability and economic efficiency. This project seeks to analyze how crop diversification has evolved over time across various regions in the contiguous United States, focusing on the relationship between changing crop patterns, farm sizes, and land values.

By examining historical trends in crop production, we aim to uncover patterns where certain crops have become increasingly dominant while others have declined, reflecting shifts in farming practices and regional agricultural preferences. Additionally, we explore how changes in farm size, whether measured by acreage or revenue, correlate with these diversification patterns. This analysis will provide insights into the underlying economic and environmental factors driving these changes, contributing to future land use strategies and agricultural policies.

2. Data and Methods

The analysis relies on three primary datasets, “land_use_farm_ops.csv”, “Prompt3_wide_3.0.csv”, and “sales_data_county.csv” which provide a detailed view of land use, management practices, and crop production across different regions of the United States. These datasets capture metrics related to cropland, pastureland, woodland, farm operation details, harvested and irrigated land, and the number of farming operations by crop type.

2.1. Data Preprocessing

a) Missing Data Handling: The datasets were first examined for missing values, and any gaps were handled through

either imputation or exclusion, depending on their significance.

b) Data Cleaning: Special characters and formatting inconsistencies were resolved, particularly for county names, which were standardized using FIPS codes. Inflation adjustment was applied to economic data to account for changing dollar values over time.

c) Feature Engineering: New variables were derived from the original datasets, including farm revenue per acre and diversification indices, which quantify the variety of crops grown in each county or region.

2.2. Methodologies

a) Exploratory Data Analysis (EDA): Descriptive statistics and visualizations were used to identify historical trends and spatial patterns in crop diversification and farm size across the U.S. Choropleth maps were created using Python Plotly's choropleth function to visualize diversification trends at the county level.

b) Trend Analysis: Historical shifts in crop diversification were analyzed through time-series models to assess how crop production patterns have changed over time in different regions.

c) Correlation Analysis: The relationships between farm size (acreage and revenue), crop diversity, and land value were explored using correlation coefficients and regression models.

2.3.Predictive Modeling

A machine learning approach was applied to predict future trends in crop diversification and farm size. Models such as Random Forest and Gradient Boosting were employed to forecast these values, with features including historical diversification, land value, and external factors like urban expansion. The model performance was evaluated using cross-validation techniques to ensure generalizability.

2.4. Tools Used

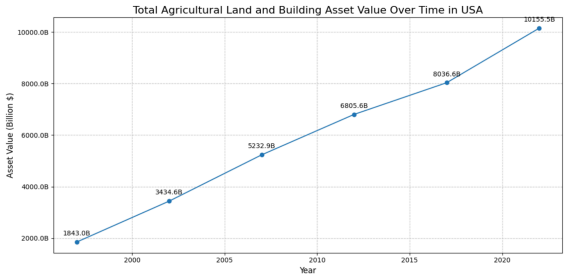
a) Python Libraries: Key libraries used for analysis include pandas for data manipulation, scikit-learn for machine learning, Plotly for visualization, and geopandas for geographical mapping.

b) Choropleth Mapping: Python Plotly was employed to create choropleth maps that visualize crop diversification across counties, helping to highlight spatial patterns and trends.

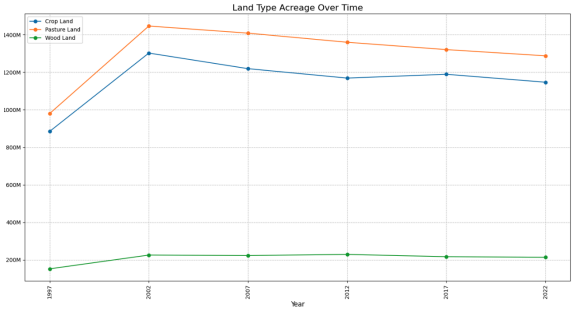
3. Results

We start at the country level and drill down to the state level.

First we observe there is a clear trend that the value of the land has increased over time. (Fig 1)

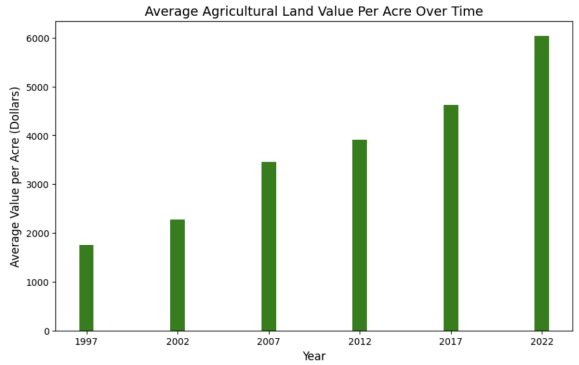


We compared that with the amount of land that is dedicated to agriculture. From 1997 till 2002 there was an increase, then constantly dropping for the next 20 years. In total we cannot see any significant increase over time (Fig 2)



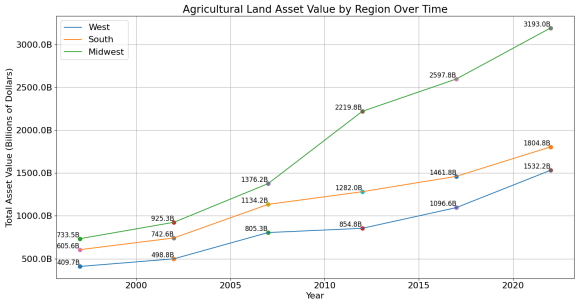
Hence, together with the previous result means that the value of the land is increasing. Further research, not covered in this study, could probably reveal the reason for that increase till 2002 and the constant later decrease for 20 years.

If we look into the price per acre the results look like this (Fig 3):



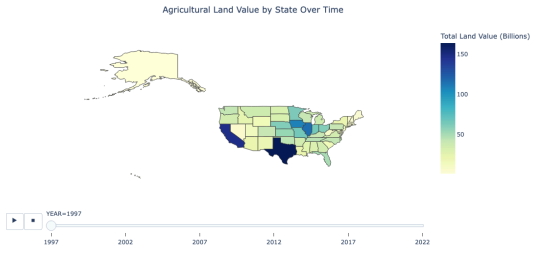
Which means that in the time of our observations (1997-2022) the price of an acre has increased from a bit less than $2000 till around $6000, a 30% increase in 25 years.

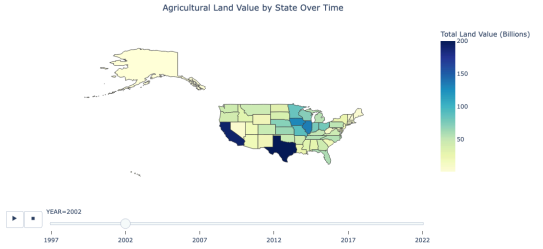
From the point of view of regions (Fig 4)

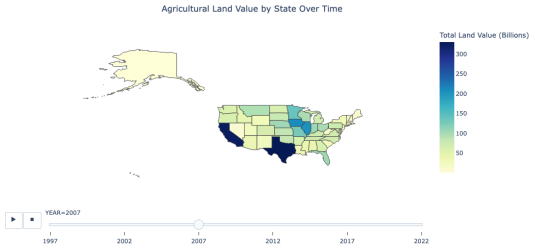


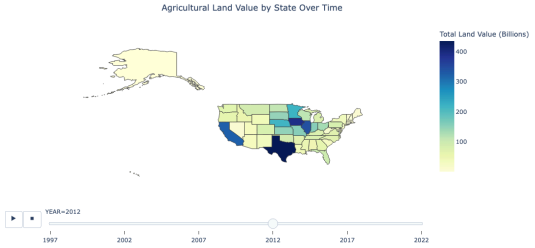
We can see that the midwest is leading this rank and double the South and West.

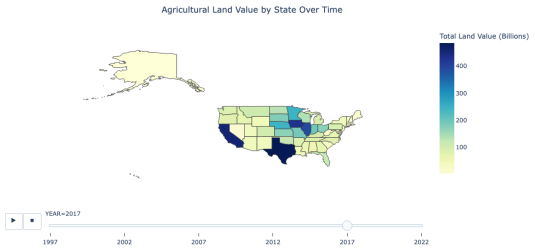
If we look into this by states, per year (Figs 5-11)

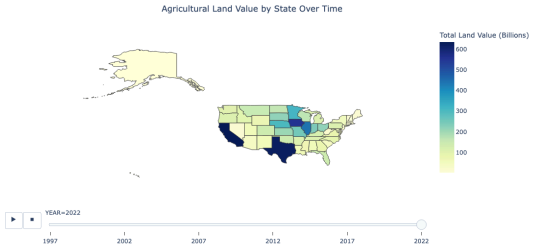






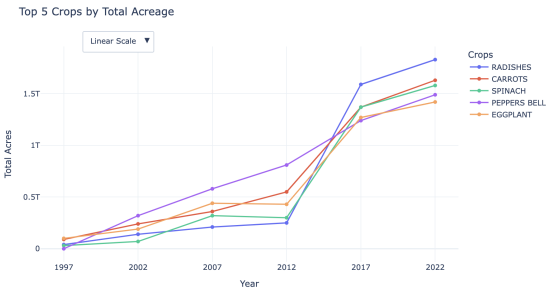




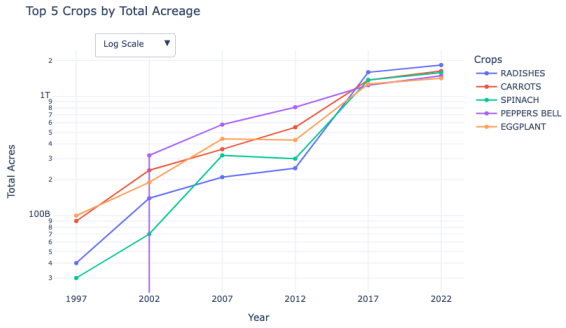


We can observe that Texas in the South has always been leading this rank. The value of land in California in the West and Iowa in the Midwest has significantly increased. The rest of the Midwest states also got an increase over the years but less noticeable compared with Iowa.

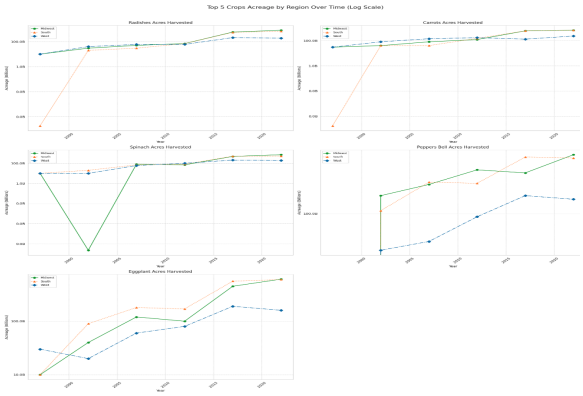
In terms of crops, the following (Fig 12)



Showcases how the top 5 most crops based on total acres within the states had increased significantly their acreage in the observed time. We can see that radish has suffered a change from ranking 5th into the list till coming 1st. This extreme increase happened in a period of 10 years, from 2012 till 2022.
We also displayed the previous values in the log scale (Fig 13)



so we could detail that there is no data available for peppers bell till year 2002. A further investigation could maybe give us insights of the reason behind that.
Unfortunately, there is no data in the sales dataset to explore a bit further all of this. We looked it at the region level and this is the breakout (Fig):



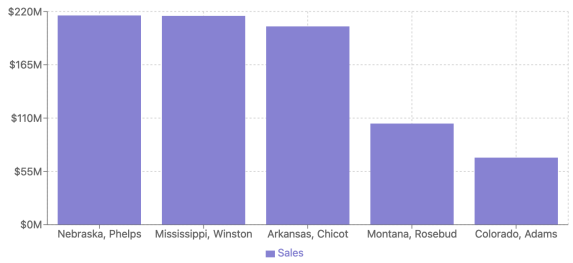
The top 5 total sales per year excluding animal-based products (Fig)

We can see that the best performance product is [], followed by []

If we look at this from the region level (Fig)

the top performance product is []

And if we go further into the county level (Fig)



Tell us that Nebraska, Phelps County ($216,104,000), Mississippi, Winston County ($215,388,000) and Arkansas, Chicot County ($204,748,000) are the most profitable counties to take into account when looking in the top sell non animal products

4. Conclusion

This analysis of crop diversification, agricultural land use, and production patterns across the

United States reveals significant trends and correlations that highlight the evolving nature of the agricultural sector. Our findings show that certain crops have become more dominant in specific regions, while others have diminished, reflecting both market dynamics and environmental changes. The correlation between shifts in farm size—measured by both acreage and revenue—and changes in crop diversity indicates that economic pressures and land value play a critical role in shaping farming practices.

The predictive models developed in this project suggest that future crop diversification trends will continue to be influenced by external factors such as urban expansion and economic fluctuations. The insights gained from this analysis have important implications for policymakers and farmers, as they navigate the complexities of land management and sustainable agricultural practices in the face of shifting market conditions.

While this project provides a comprehensive overview of historical trends and predictive models, there are limitations. The analysis could be enhanced by incorporating more granular data on climate change impacts, regional water usage, and crop-specific factors that were not fully explored in this study. Future research could also investigate the long-term effects of urbanization on rural land values and crop distribution, offering a more complete picture of the factors shaping the agricultural landscape.

Our findings contribute valuable knowledge to the field of agricultural economics and land management, helping to inform strategies for sustainable growth and efficient resource use in the future.

Please see this link to our Submission Video
Please see this link to our Google Colab Notebook1 and Google Colab Notebook2

References

US MAP DIVISION (worldatlas)

MORE_DOLLARS 11254
non-null with a total of 7521 NAs

This first one, FARM_OPERATIONS_AREA_OPERATED_MEASURED_IN_PCT_OF_TOTAL_LAND, we could drop it, as it doesn't seem to be adding much value.

For the second one, AG_LAND_INCL_BUILDINGS_OPERATIONS_WITH_ASSET_VALUE_WHERE_VALUE_10_000_000_OR_MORE_DOLLARS is part of a division in bins

APPENDIX

EDA

We are working on the land_use_farm_ops.csv. We are identifying all the similar/common data and plotting them in the graph.

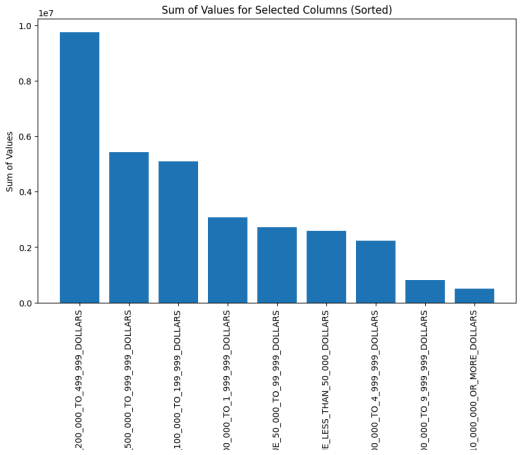
The data is 18775 rows and 75 columns

We found at least two columns that has many NAs:

1. FARM_OPERATIONS_AREA_OPERATED_MEASURED_IN_PCT_OF_TOTAL_LAND
   3129 non-null with a total of 15646 NAs

And

2. AG_LAND_INCL_BUILDINGS_OPERATIONS_WITH_ASSET_VALUE_WHERE_VALUE_10_000_000_OR_



We can observe that most of the data is in the range of 200000 to 499999 of total value.

In this subset of bins and for the rest of the division bins we found in the data, we decided that the value comes in adding them all, instead of having differentiated in bins so we proceeded to aggregate all the subsets bins together.