Project Report For CS661: BIG DATA VISUAL ANALYTICS
2024-2025 Semester II
Project Title: AgriViz: Analyzing Farmer Queries with Data

Team members: Siddharth Kalra, Chitwan Goel
Geetika, Vikas Yadav, Divyansh, Shrey Bansal
Aniket Suhas Borkar, Kartik Anant Kulkarni

Member emails: siddharthk21@iitk.ac.in, chitwang21@iitk.ac.in, geetika21@iitk.ac.in, vikasy21@iitk.ac.in, divyansh21@iitk.ac.in, shreyb21@iitk.ac.in, aniketsb21@iitk.ac.in, kartik21@iitk.ac.in
IIT Kanpur

1 Introduction

Agriculture is a vital pillar of India's economy, employing 45% of the workforce and contributing nearly 18% to the GDP. However, the sector continues to face challenges such as climatic variability, market volatility, pest outbreaks, and complex policy frameworks. Limited access to reliable information exacerbates these issues, with many farmers relying on informal sources, leading to widespread misinformation, poor decision-making, and financial losses. To bridge this gap, the Government of India established the Kisan Call Centers (KCCs), providing toll-free expert advice to farmers. Although these centers have amassed a large repository of farmer queries and responses, the unstructured nature of the data restricts its effective utilization for large-scale analysis and policymaking. Structuring this information and integrating it with historical crop price data presents an opportunity to derive actionable insights into regional concerns and agricultural market dynamics.

The objective of this project, **AgriViz**, is to develop a scalable framework for cleaning, structuring, and visualizing Kisan Call Center query data alongside crop price datasets. Leveraging Natural Language Processing (NLP) techniques, clustering algorithms, and interactive visual analytics, the project aims to identify emerging agricultural issues, regional patterns of concern, and price trends. The resulting insights are intended to support policymakers, agribusiness stakeholders, and farmer advocacy groups in making informed, data-driven decisions to promote agricultural sustainability and economic resilience.

2 Tasks

The following key tasks were undertaken as part of this project to analyze and derive insights from the Kisan Call Center (KCC) query data and Crop Price data:

1. **Geospatial Analysis:** Analyze the volume of queries and crop price data across different states to identify regions with higher or lower engagement. Further drill down into districts within selected states to uncover localized patterns of concern, as well as regions with significant crop price fluctuations or disparities.

2. **Temporal Analysis:** Examine how the number of queries and crop price trends vary over time to detect seasonal trends, spikes during specific events, and long-term changes in both farmer concerns and crop price fluctuations. This analysis will help identify periods of price instability and shifts in farmer priorities.

3. **Query Type Distribution:** Investigate the types of queries being made, identifying which categories are more frequent and which are less common, to understand the dominant themes among farmers.

4. **Semantic and Linguistic Analysis:** Analyze the keywords used in queries through word clouds, identifying frequently occurring terms and observing how they vary by region or state.

5. **Inter-query Relationship Analysis:** Study the temporal correlation between different queries to identify co-occurring concerns and evolving patterns of related agricultural issues.

6. **Crop-wise Correlation Analysis:** Explore whether queries related to specific crops exhibit interdependence, helping to uncover relationships or shared challenges between crop categories.

7. **Monthly and Sectoral Breakdown:** Analyze the monthly distribution of queries and categorize them based on agricultural sectors (e.g., pests, fertilizers, weather) to understand seasonal and topical shifts in farmer queries.

8. **Crop Price and Price Variation Analysis:** Analyze the fluctuations in crop prices across different regions. This will help identify areas with high price volatility, seasonality in pricing, and the impact of market dynamics or external factors on crop prices.

9. **Modal Price Prediction:** Analyze the price time series and use a SARIMAX Model for prediction over the horizon of 1 year. Use the the model for decomposition of the seasonal and trend components.

3 Proposed Solution

To address the challenges outlined, we developed a modular and interactive dashboard-based framework that enables the exploration of Kisan Call Center (KCC) data and crop price data from multiple analytical perspectives. Our solution integrates geospatial analysis, temporal trends, semantic clustering, and linguistic pattern mining to provide a comprehensive view of farmer queries and crop price fluctuations across India. We separate the dashboard into two panes for insights from both the datasets.

3.1 KCC Query Analysis window:

This window consists of the following interactive visualizations. Due to the large size of the dataset ($\sim$ 16 GB), we first performed data preprocessing to enable efficient analysis. Irrelevant columns (such as categorical fields incorrectly populated with numerical values or excessive NaN entries) and inconsistent rows (such as mismatches in state or district names with geospatial references) were removed. After preprocessing, the dataset size was reduced to $\sim$ 5 GB and subsequently compressed to $\sim$ 1 GB for storage. To ensure high-speed and responsive visualizations, we implemented parallel processing for data loading and graph rendering. Initially, the graphs were updated serially, resulting in significant latency. By parallelizing the data-fetching processes across all visualizations, we achieved up to a 5x improvement in rendering speed, enabling smooth and interactive exploration even with large-scale datasets.

1. Spatial Distribution (State and District Level):
We implemented interactive choropleth maps to visualize the geographic distribution of queries. Each state is colored based on the volume of queries received, making it easy to identify regions with higher agricultural concerns. Upon selecting a state, the map drills down to display district-level data, enabling finer-grained insights into regional variations. This visual interface serves as the entry point for navigating the dataset and identifying spatial hotspots of farmer issues.

2. Temporal Trends in Query Volume:
To understand how the frequency of queries evolves over time, we constructed a dynamic time series plot. The graph displays daily counts of incoming queries, helping identify seasonal patterns, spikes (e.g., due to pest outbreaks or climatic events), and gradual trends. The plot is interactive, equipped with a range slider that allows users to focus on specific periods for in-depth temporal exploration.

3. Query Type Distribution and Clustering:
Given the diversity of queries, we used BERT-based embeddings to capture the semantic meaning of each query and applied agglomerative clustering to group them into coherent categories. These clusters were then labeled using TF-IDF-based keyword extraction. To visualize their evolution, we used a stream graph, where each colored stream represents a query cluster and its frequency over time. This helped us uncover dominant and emerging types of agricultural concerns. Users can interact with the graph to zoom in on specific clusters or time periods and observe changes in query composition.

4. Semantic Patterns via Word Clouds:
To further understand the language and key themes present in farmer queries, we generated word clouds for each state. These visualizations highlight the most commonly used words, allowing for quick identification of recurring issues. By comparing word clouds across states, we gained insights into how agricultural concerns and terminologies vary regionally, providing a linguistic lens on the dataset.

5. Temporal Relationships Between Queries:
To understand how different types of queries are interrelated over time, we constructed a node-link graph where each node represents a cluster of semantically similar queries. Edges between nodes were drawn based on the strength of temporal correlation between their query frequencies. This visualization revealed patterns of co-occurring concerns, helping us identify clusters that tend to emerge together during specific agricultural events or seasons.

6. Crop-wise Correlation Analysis:
Similar to the query cluster graph, we created a node graph for crops. The top-n most frequently mentioned crops were selected, and their respective time series of query volumes were analyzed. Each node represents a crop, and edges indicate correlation in their query trends. This helped uncover potential interdependencies between crops — such

as shared growing seasons, susceptibility to similar pests, or market-related linkages — providing valuable insights for crop management and agricultural planning.

7. Monthly and Sectoral Breakdown:
To examine seasonal trends and thematic focus, we visualized the monthly distribution of queries categorized by agricultural sectors such as pests, weather, fertilizer, etc. A multiline plot was generated where each line represents a sector and shows its query volume across the twelve months. This visualization helps uncover seasonal peaks in specific concerns (e.g., pest-related queries in monsoon months), allowing for timely intervention and resource planning. Additionally, the total query volume per month provides an overview of the overall demand for support throughout the year.

3.2 Crop Price Analysis window:

In this window, we developed an interactive dashboard to explore and analyze the crop price data of India across spatial and temporal dimensions. The application consists of multiple coordinated plots that update dynamically based on user interaction, offering a comprehensive view of crop price behavior. Below, we outline the key components of the interface and their functionalities:

1. Global Crop Selector:
A dropdown menu allows users to select a crop of interest. Upon selection, all plots and analyses update automatically to reflect the selected crop. This feature facilitates seamless exploration across different crops without reloading or reconfiguring the visualizations.

2. Spatial Distribution Map:
An interactive map of India displays a scatter plot where each point represents a district. The color of each point encodes either the *mean price* or the *price variance* of the crop in that district, averaged over the available time period. A dropdown allows users to toggle between viewing the mean price and the price variance, enabling flexible exploration of both the general price levels and their stability across time. This spatial view is motivated by the need to identify geographic patterns, regional disparities, and price volatility in crop pricing.

3. Temporal Price Evolution:
A time series plot illustrates the average price of the selected crop across all districts over time. To better understand underlying patterns, we apply a *Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (SARIMAX)* model to decompose the series into its *trend* and *seasonality* components. Future price predictions based on the fitted model are also visualized. This analytical layer helps users grasp both historical dynamics and plausible future behavior of crop prices.

4. Varietal Price Comparison:
A histogram compares the prices of the ten most popular varieties of the selected crop. Here, the price for each variety is averaged over both districts and time. This comparative visualization supports varietal-level decision-making, highlighting which varieties tend to command higher or lower prices in the market.

Each component of this window is designed to offer a distinct but complementary perspective on the crop price data, enabling users to move fluidly between spatial, temporal, and varietal analyses. The combination of interactive elements and statistical modeling aims to make the exploration both intuitive and insightful.

Together, these components form a unified visual analytics framework, allowing users to seamlessly transition between spatial, temporal, categorical, and semantic perspectives on agricultural queries, as well as spacial, temporal and varietal crop prices. This holistic approach enables stakeholders to detect region-specific issues, monitor trends over time, and derive actionable insights for agricultural policy and intervention, as well as make the query answering process for queries related to crop prices much more efficient and robust.

4 Results and Insights

4.1 KCC Query Analysis Window

The interactive visualizations developed for the KCC Query Analysis window provided several key insights into farmer concerns across India:

1. **Regional Hotspots of Agricultural Concerns:** The spatial distribution maps revealed distinct geographic clusters with high query volumes, indicating regions where farmers face greater agricultural challenges. Drill-down to district-level data allowed for finer localization of issues.

   (a) The states having larger population and area like Uttar Pradesh show the most number of queries.

2. **Variation in types of queries across States:** State-wise word clouds illustrated how the focus of agricultural concerns vary geographically. For instance,

   (a) Certain pests or diseases were found to be more prominent in specific regions based on keyword prominence.

   (b) The crops which are prominent in a region were shown to be highlighted in the wordcloud for the states of that particular region like coconut is prominently visible in Kerala and Wheat is dominating in Haryana.

   (c) Weather is a dominating word in the word cloud for most of the states except those where rainfall is generally stable, for example, Kerala.

3. **Interrelated Query Themes:** The node-link graphs based on temporal correlations between clusters uncovered patterns of co-occurring concerns, such as weather-related queries frequently correlating with pest management queries, highlighting the compound nature of agricultural issues. Some of such examples are:

   (a) For Maharashtra, and many other states, 'Magur culture', 'water storage' and 'freshwater fishery' are quite interrelated whereas these are unrelated with 'whether varieties'

4. **Crop Inter dependencies:** The crop-wise correlation analysis exposed crops whose growing cycles or market behavior are closely linked, providing hints at regional agricultural planning and suggesting opportunities for better resource allocation. Some of such examples are:

   (a) For the state of Rajasthan, tomato and chillies are more related with themselves as they are with say lemon.

5. **Sectoral Seasonality:** Monthly and sectoral breakdowns revealed the cyclical nature of different types of concerns:

   (a) The number of queries in the Agricultural sector for most of the states tends to increase in monsoon months of June and July, which is generally the sowing season for many crops.

Overall, the KCC Query Analysis window transformed a large and complex dataset into an intuitive, multi-faceted exploration environment, enabling discovery of patterns and relationships that would be difficult to detect through raw data inspection alone.

4.2 Crop Price Analysis Window

The Crop Price Analysis window offered a comprehensive exploration of crop price behaviors, leading to several actionable insights:

1. **Spatial Price Variations:** The interactive map revealed significant spatial disparities in mean crop prices across districts. Certain regions consistently showed higher or lower prices, suggesting differences in local supply-demand conditions, infrastructure, or market access. For example,

   (a) For example, for crop wheat the mean prices vary from 3794 Kollam, Kerala to 1903 in Agra, UP.

2. **Detection of Price Instability Zones:** By switching to the price variance view, users could identify districts where crop prices fluctuated heavily over time. These volatility hotspots could indicate areas vulnerable to market shocks or inconsistent supply chains.

   (a) Variance can vary from 1475 in Shimoga, Karnataka to 225 in Panipat.

3. **Historical Trends and Seasonality:** The time series plots combined with SARIMAX decomposition highlighted clear seasonal patterns for many crops, such as price rises during off-season months. This understanding is crucial for timing interventions and planning storage or transport. For example,

   (a) Apple shows high seasonality in price, resulting in greater magnitude of he seasonal graph values, whether crops such as wheat have almost no seasonality.

4. **Forecasting Future Prices:** The SARIMAX-based future projections provided an initial estimate of where prices are expected to move, offering valuable information for farmers, traders, and policymakers to make informed decisions.

5. **Varietal Price Differentiation:** The varietal histogram allowed users to quickly identify which crop varieties tend to command higher prices on average, supporting variety selection decisions for farmers aiming to maximize profits. For example,

   (a) For Apple, varieties such as American and Kashmir are significantly more expensive than other varieties such as Red Gold.

By making the data accessible through multiple coordinated views, the Crop Price Analysis window enabled users to extract rich, actionable insights about crop market behavior that would be extremely difficult to derive from tabular data or static reports alone.

5 Conclusion:

In this project, we developed **AgriViz**, an interactive, modular framework to visualize and analyze large-scale agricultural data from Kisan Call Center (KCC) queries and crop price datasets across India. By integrating geospatial analysis, temporal trend detection, semantic clustering, and linguistic pattern mining into intuitive dashboards, we enabled seamless exploration of complex, high-volume data.

For farmer queries, our system revealed critical spatial and temporal patterns, uncovered emerging agricultural concerns, and highlighted linguistic and regional variations that would otherwise remain hidden in unstructured data. For crop price analysis, our tools exposed regional price disparities, zones of price instability, seasonal trends, and varietal price differentiation, offering deep market insights.

The interactive, multi-pane design allowed users to transition smoothly between regional, temporal, and semantic perspectives, fostering a holistic understanding of the agricultural landscape. Insights derived from these visualizations have significant potential to support policymakers, agribusinesses, and farmer support initiatives by facilitating informed, data-driven decision-making.

Overall, **AgriViz** demonstrates the power of combining big data analytics with visual storytelling to address real-world agricultural challenges, paving the way for more resilient and informed agricultural systems in India.

6. Link to source code:

GitHub Link

References

[1] Plotly Technologies Inc. Dash: A python framework for building analytical web applications, 2017. Accessed: 2025-04-27.

[2] Plotly Technologies Inc. Plotly: Collaborative data science, 2015. Accessed: 2025-04-27.