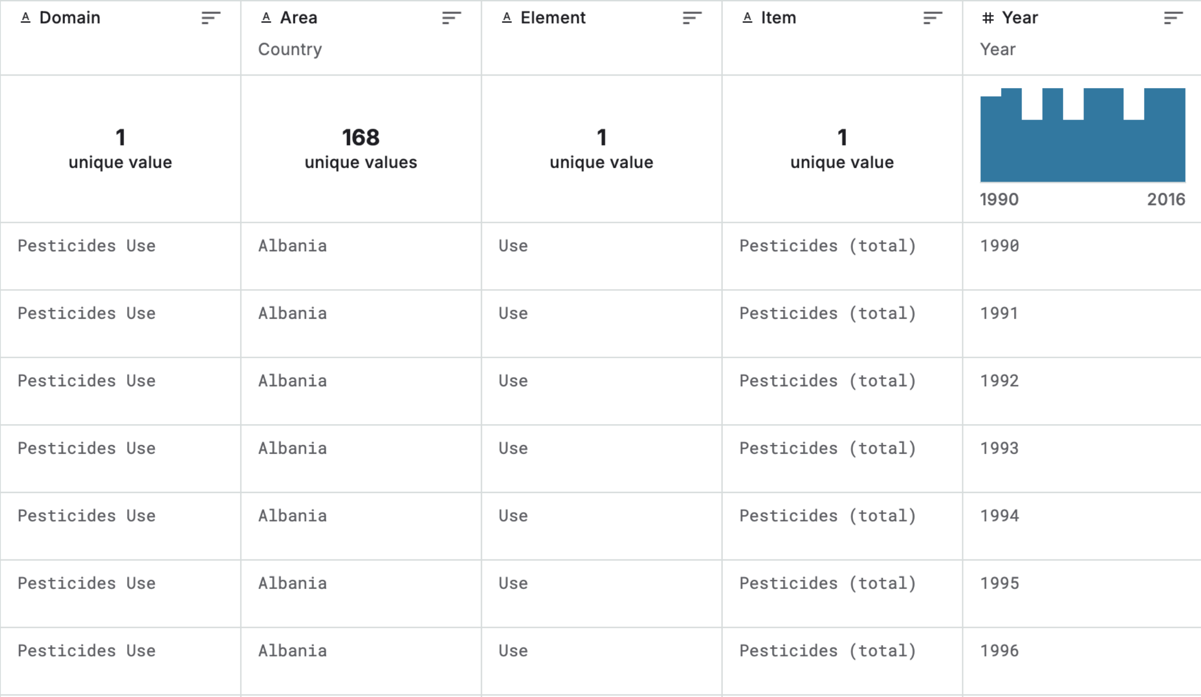
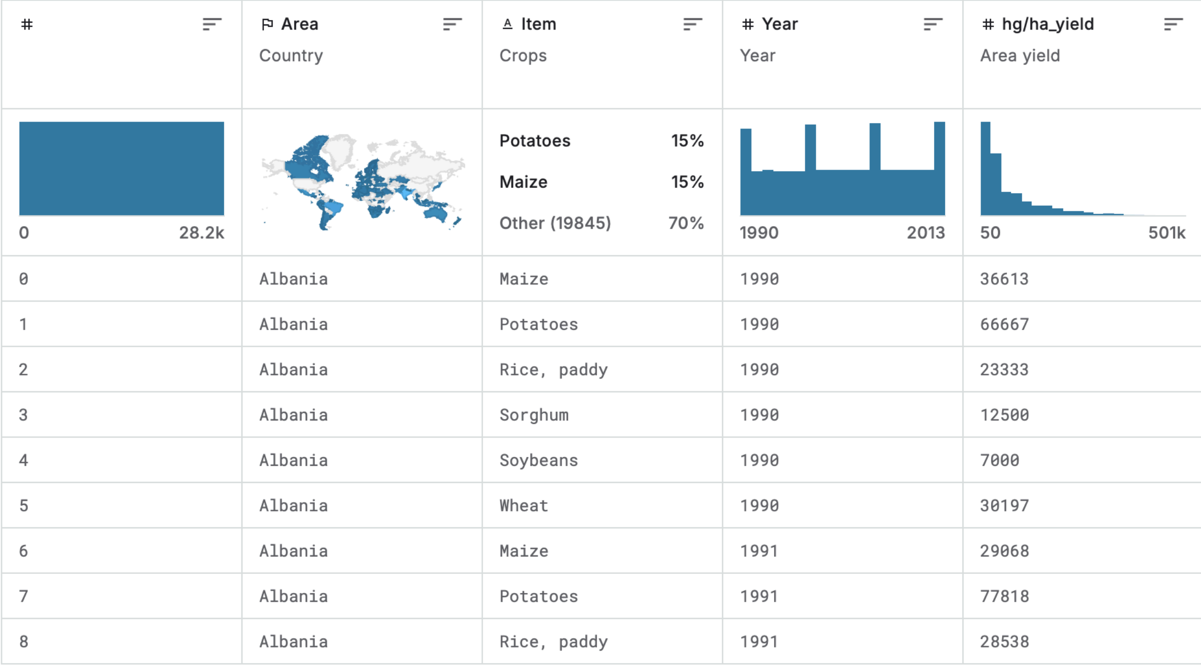
314 Project Proposal

1. Group Member Names:
    a. Evelyn Gonzalez Garcia
    b. Shoshana Stadlan
    c. Praneel Panchigar
    d. Chenxing Liu
2. Goals
    a. **Identify Key Factors**: We are interested in identifying the most important factors that affect crop yields. For example, which combination of factors (climate, pesticides, location) has the most significant impact on the yield of specific crops like maize, potatoes, or cassava.
    b. **Predict Crop Yields**: Using machine learning models, we aim to predict future crop yields based on historical data and environmental factors, such as rainfall, pesticides usage, and temperature. The models will help anticipate yields and guide decision-making in agriculture.
    c. **Explore Regional Trends**: Another objective is to identify regional differences in yield performance, helping us understand how different countries contribute to global crop production and how environmental factors vary by location.
3. Dataset
    a. The data we are using is from FAO (Food and Agriculture Organization). It includes historical data for the ten most consumed crops worldwide, covering the years 1961-2016. This dataset includes information like the country, crop type, year, and yield values. The link is: https://www.fao.org/faostat/en/#data. The Kaggle link is: https://www.kaggle.com/code/kushagranull/crop-yield-prediction/input?select=pesticides.csv.
    b. The dataset for our crop yield prediction project consists of multiple merged datasets representing crop yield, rainfall, pesticide usage, and temperature.
        i. **Crop Yield Data**: 56,717 rows and 12 columns
        ii. **Rainfall Data**: 6,727 rows and 3 columns.
        iii. **Pesticide Usage Data**: 4,349 rows and 3 columns.
        iv. **Temperature Data**: 71,311 rows and 3 columns.
        v. **Merged Dataset**: After merging all datasets based on Year and Area, the final dataset has 28,242 rows and 7 columns, including all four features mentioned above.
    c. No. The data can be directly accessed through the FAO website. Since we found this dataset on Kaggle, some preliminary data cleaning works have been done.

d.



Domain	Area Country	Element	Item	# Year Year
1 unique value	168 unique values	1 unique value	1 unique value	1990 2016
Pesticides Use	Albania	Use	Pesticides (total)	1990
Pesticides Use	Albania	Use	Pesticides (total)	1991
Pesticides Use	Albania	Use	Pesticides (total)	1992
Pesticides Use	Albania	Use	Pesticides (total)	1993
Pesticides Use	Albania	Use	Pesticides (total)	1994
Pesticides Use	Albania	Use	Pesticides (total)	1995
Pesticides Use	Albania	Use	Pesticides (total)	1996



#	Area Country	Item Crops	# Year Year	# hg/ha_yield Area yield
0 28.2k		Potatoes 15% Maize 15% Other (19845) 70%	1990 2013	50 501k
0	Albania	Maize	1990	36613
1	Albania	Potatoes	1990	66667
2	Albania	Rice, paddy	1990	23333
3	Albania	Sorghum	1990	12500
4	Albania	Soybeans	1990	7000
5	Albania	Wheat	1990	30197
6	Albania	Maize	1991	29068
7	Albania	Potatoes	1991	77818
8	Albania	Rice, paddy	1991	28538

4. Anticipated Challenges
   a. **Outliers**: Crop yields and environmental factors may vary widely across different regions and years. This could introduce outliers or extreme values that may skew the model’s predictions.
   b. **Categorical Data(Countries)**: Variables like countries have a lot of unique values, while converting these variables into a numeric format via one-hot encoding can lead to a very high-dimensional dataset, which might introduce computational challenges and the risk of overfitting.
   c. **Generalizability**: Since crop yield is influenced by many unpredictable factors like extreme weather or sudden changes in agricultural

practices, your model may face challenges in generalizing to new, unseen data or future conditions.