Capstone 1 Proposal
Alyse Record

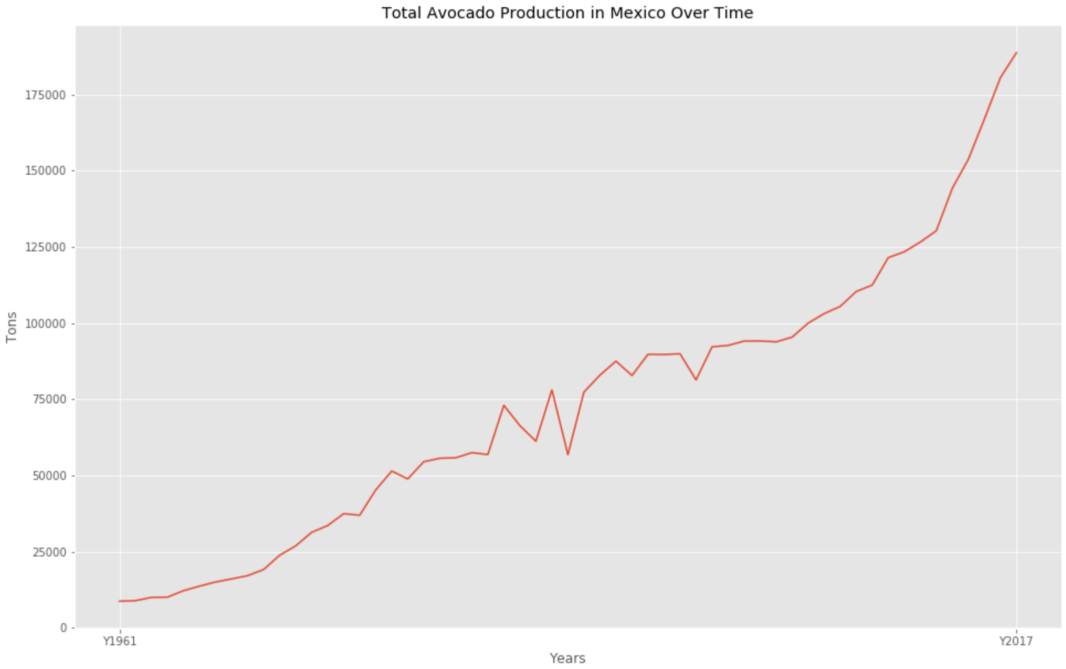
The Question

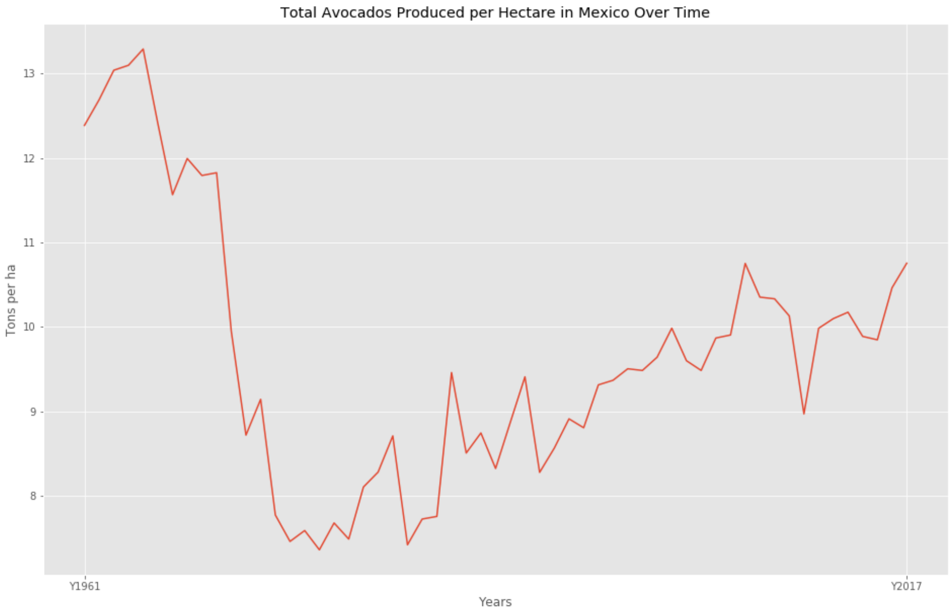
How does a crop production quantity per hectare correlate to the producer's price of the product? Which crops have the strongest correlation? Does it vary significantly by country? What crops should be produced in a given country for a maximum profit?

Description of the data

The data is obtained from the Food and Agriculture Organization of the United Nations. The FAO has acquired data from countries around the world in the areas of crop production, prices, trade, investment, etc. For this capstone I plan to focus on the crops, crops produced and producer prices datasets which are broken out by country, crop, and year.

Examples of the data were generated by reading the data into a Spark Dataframe, then using PySpark SQL to query the data, and finally plotting sample data using Matplotlib.





The Minimal Viable Product

Perform EDA on the datasets to:

- Identify which crops have the highest producer profit per hectare over time worldwide.
- For those crops, identify if the profit per hectare varies greatly per country.

The MVP +, MVP ++, etc.

- Predict which crops would be the best to produce for maximum profit in a given country.
- Plot the data on a map to compare visualize production rates and prices across countries.

Alternate Capstone Proposal

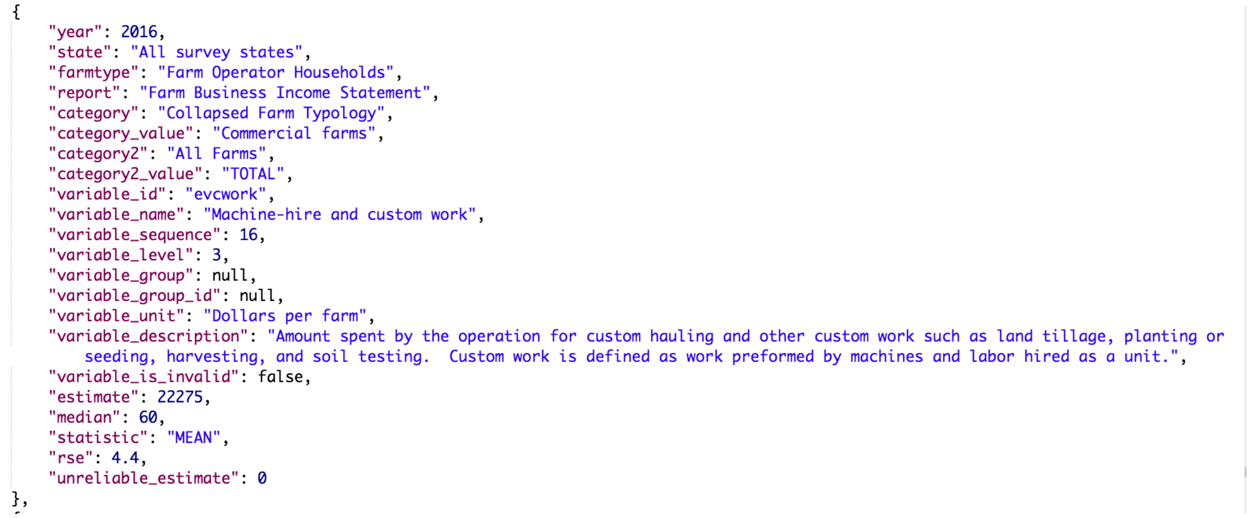
The Question

What is the distribution of large (commercial) farms to residential and intermediate (small) farms across different regions of the United States? Are there certain production specialties that are growing for small operations in some areas and not others?

Description of the data

The data will be obtained from the USADA ERS ARMS Data API:
https://www.ers.usda.gov/developer/data-apis/arms-data-api/

I am still working on how to get the Farm Business Balance Sheet and Structural Characteristics survey reports broken down in the the desired categories and variables, but here is an example of some of the data accessed via Postman.



```
{
    "year": 2016,
    "state": "All survey states",
    "farmtype": "Farm Operator Households",
    "report": "Farm Business Income Statement",
    "category": "Collapsed Farm Typology",
    "category_value": "Commercial farms",
    "category2": "All Farms",
    "category2_value": "TOTAL",
    "variable_id": "evcwork",
    "variable_name": "Machine-hire and custom work",
    "variable_sequence": 16,
    "variable_level": 3,
    "variable_group": null,
    "variable_group_id": null,
    "variable_unit": "Dollars per farm",
    "variable_description": "Amount spent by the operation for custom hauling and other custom work such as land tillage, planting or
        seeding, harvesting, and soil testing.  Custom work is defined as work preformed by machines and labor hired as a unit.",
    "variable_is_invalid": false,
    "estimate": 22275,
    "median": 60,
    "statistic": "MEAN",
    "rse": 4.4,
    "unreliable_estimate": 0
},
```

MVP

- Scrape the data from the USDA ERS ARMS API into a MongoDB.
- Put the data into a more usable format (ie. Pandas DF or Spark DF if too large)
- Create visualization of top product specialities across different regions of the United States.
- Create visuations how those product specialities are distributed among large and small farm operations.