Truc Minh Nguyen

10/11/2024

MA615

Strawberries

Intro:

This dataset is collected from the USDA on strawberries production data in the U.S. The goal of the analysis is to give some insight into these questions:

1. Where they are grown? By whom?
2. Are they really loaded with carcinogenic poisons?
3. Are they really good for your health? Bad for your health?
4. Are organic strawberries carriers of deadly diseases?
5. When I go to the market should I buy conventional or organic strawberries?
6. Do Strawberry farmers make money?
7. How do the strawberries I buy get to my market?

The dataset is split into two sub datasets: strawberry census and strawberry survey. The strawberry census dataset contains data about the acres production of the strawberries for organic and inorganic strawberries and profit amount. Strawberry survey datasets focus on the chemical usage and profit amount.

Strawberry Census:

Census data is isolated into organic and inorganic subsets. This is determined by piping and using the separate_wider_delim function to split the “Data Item” column into two columns, “strawberries” and category, we then use the drop function to drop columns that contains only one unique value. I also split the “Category Column” into “Measure” and “Bearing_type”. I used str_replace to clean up extra words, commas, ect. I used str_rename to rename columns to a more appropriate term.

Next, I created a new subset of Strawberry Census to filter out “Measure” into “Acres” and “Operations”. I split the acres ranges into columns of minimum acres and maximum acres. For example, before the change, an observation would have one column with values “0.1 to 0.9 acres”, now I’ve taken that character string and convertetd to a double. Then, I cleaned up and set minimum and maximum columns for the two values. I ordered the minimum values from least to greatest per state per “Measure”. I did this to fit a regression that would impute the missing values in the “Value” column of the strawberry census data. I noticed a pattern that the values of “Value” grows exponentially as the minimum acres increases. I decided to use a polynomial regression of Value ~ poly(Min Acres). What I realized after is that while it does make consistent prediction for the missing values of “Value”, the sum does not add up to the Total. I have not come up with a solution on how I would adjust for this error. I do think that using a polynomial regression might cause overfitting but since we are predicting values of a set range instead outside the range, I think it’s an appropriate method for this objective.

Strawberry Survey:

Sorry I did not get a chance to dig much into the strawberry survey data as I was spending time on strawberry census. However, the steps would be to first split the dataset into sub datasets based on categories of chemical, fertilizer, and production. This analysis would aim to figure out the motivation of using chemical and fertilizers. The prediction would be that it would increase harvest rate and quantity, leading to a higher profit margin.

Conclusion:

At this point in my analysis, I can answer that the primary states of focus are California and Florida as they are the leading producer of our nation's strawberries, accounting for more than 50%. They can be grown by local farmers or an operation of corporate farming. The strawberry business is profitable though I'm not sure how much of that profit is rewarded to farmers working in the field and how much is collected by corporate executives. The harvest size of the strawberries determines where they go in the market. The large ones are sold as fresh produce and the small ones are frozen produce. I believe the small ones would also be used in canned goods and processed foods such as pastries. I am not able to draw and inferences between organic vs. inorganic being the carrier of diseases and cannot advise what the user should buy at the grocery store. I hope to have a clearer insight once I dig more into the census data and the survey.