Data Analysis for Strawberries

Minqi Li

10/20/2020

1 Introduction

1.1 Overview

These data were collected from the USDA database selector: https://quickstats.nass.usda.gov. We select each variable and get the dataset. This dataset is about the information of three kinds of berries: blueberries, strawberries and raspberries. This report only analyzes the data of strawberries.

1.2 Outline

The outline of this report is as follows. Firstly, I read and clean the data. Next, I visualize and explore the data. Finally, I interpret my model and discuss the conclusions.

2 Data Cleaning

First of all, I read the data and the data is complicate. Then we follow these steps to clean the data.

- Remove the columns with only one unique value.
- Separate the data of strawberries into an individual table.
- Separate the column of data item, domain, domain category into the column of type, production, Avg, Measures, Materials, Chemical.
- Remove redundant columns and rearrange all the columns.
- Deal with the value column to transform (NA), (D), (Z) into 0 and convert character data to numeric.

3 Visualization

3.1 Explore the Relationship between State and Value

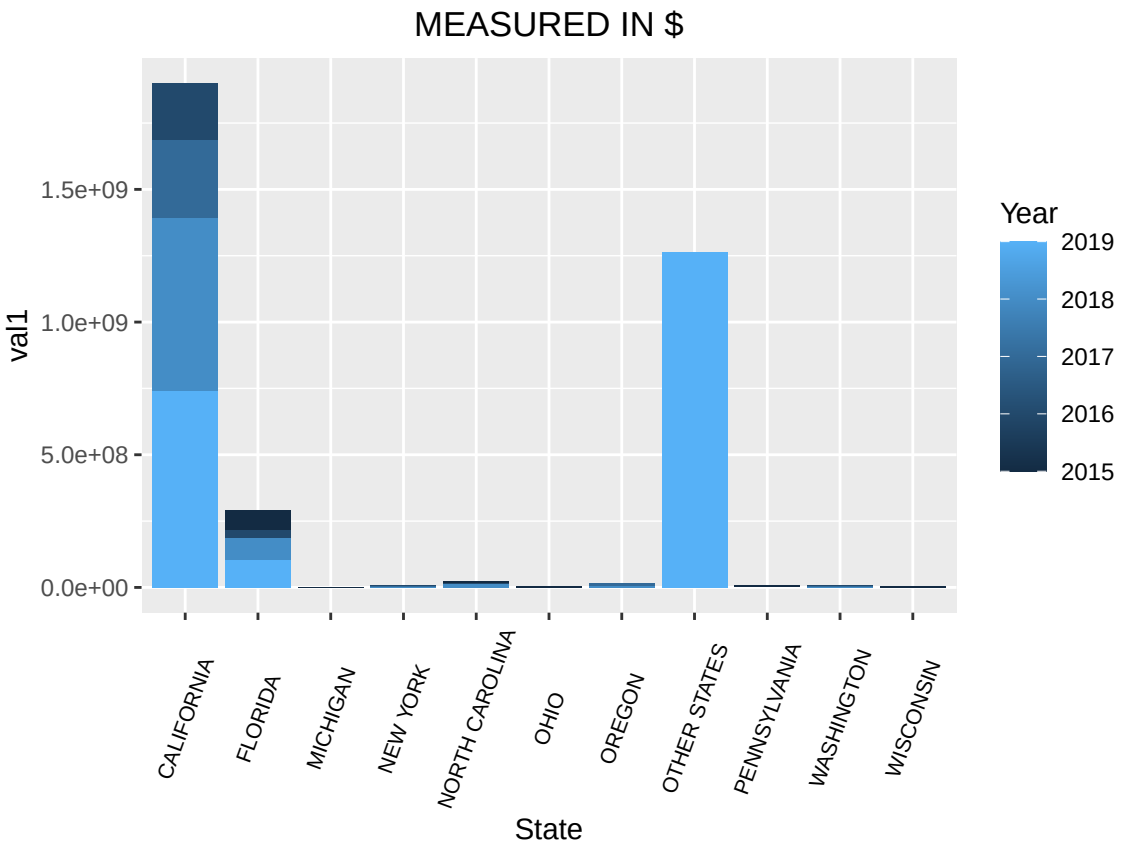

Figure 1: Plot the histogram with the value measured in $ for each state and each year.

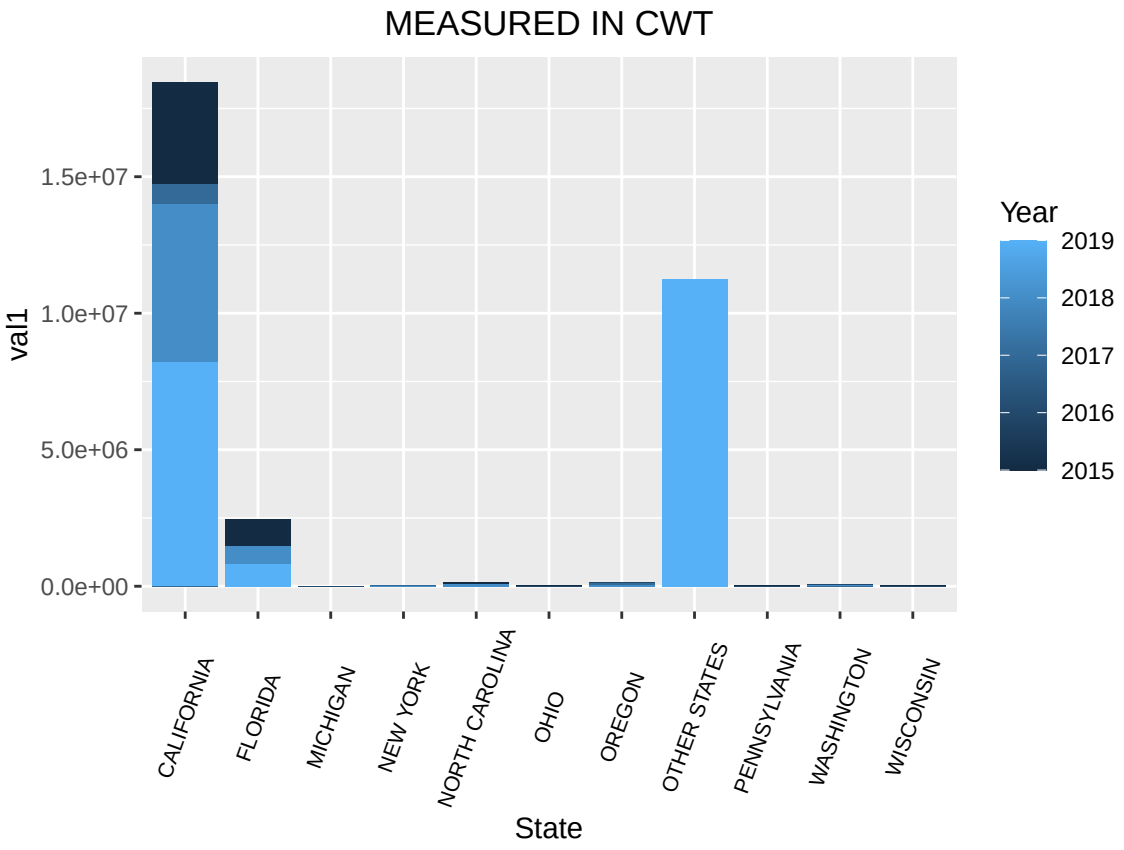

Figure 2: Plot the histogram with the value measured in CWT for each state and each year.

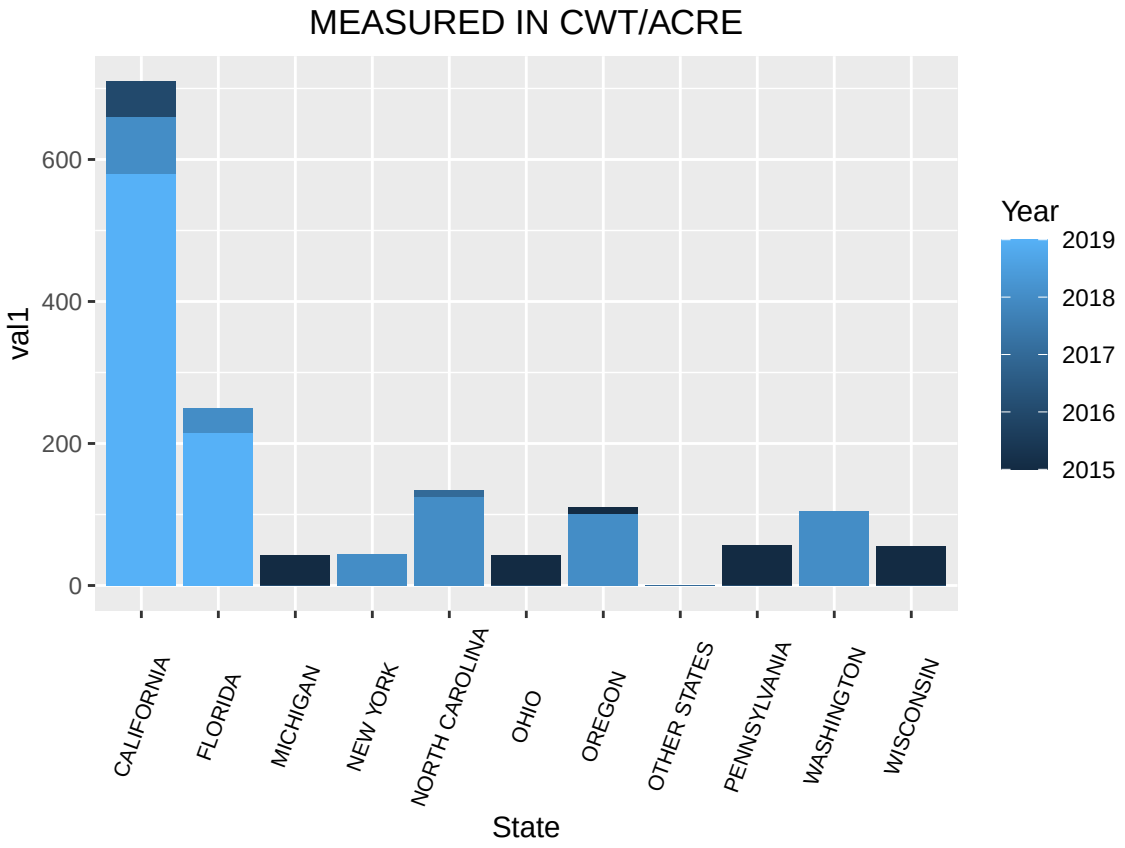

Figure 3: Plot the histogram with the value measured in CWT/ACRE for each state and each year.

3.2 Explore the Relationship between Chemical and Value

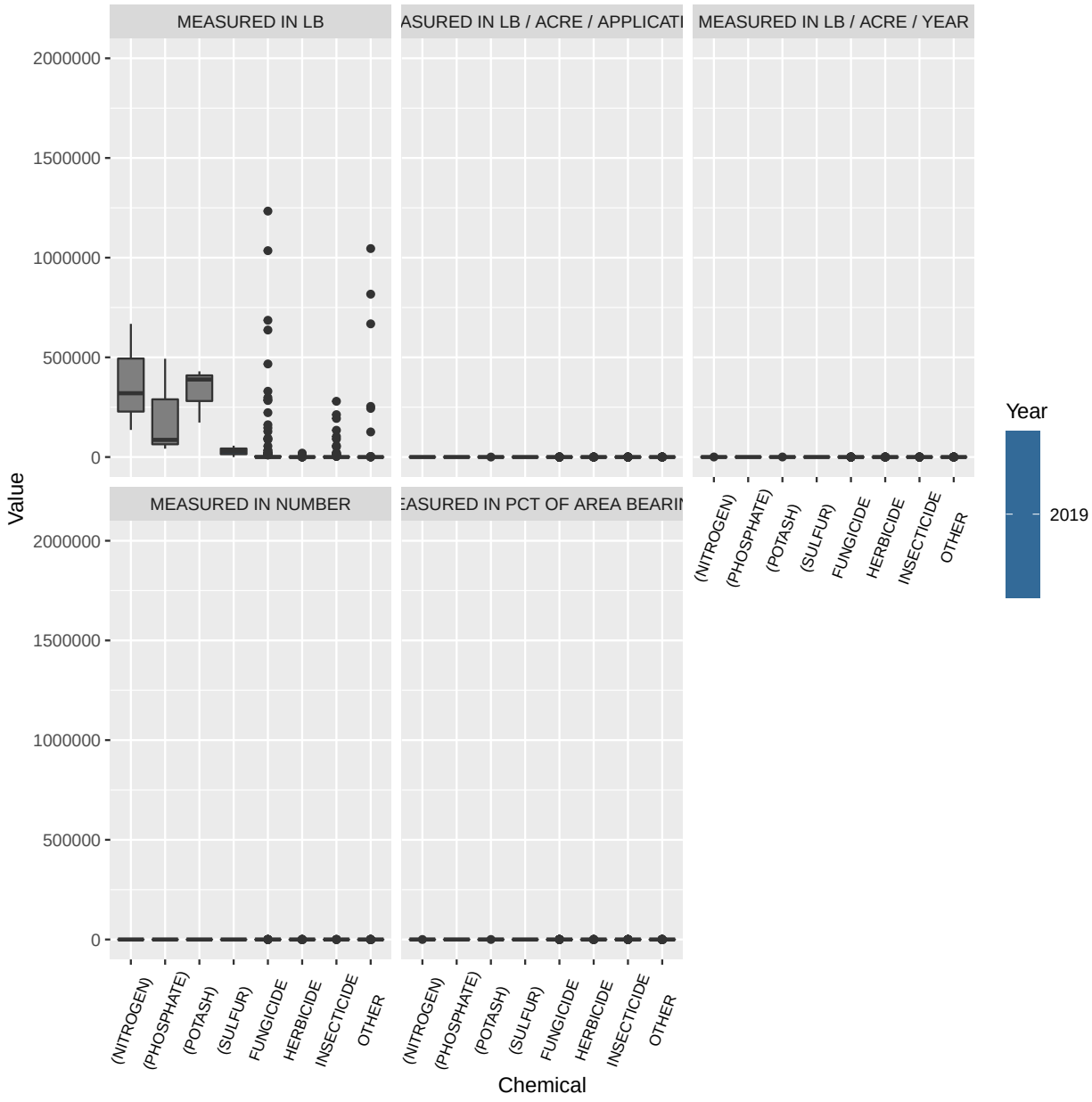

Figure 4: Boxplots separated for each measures.

3.3 Explore the Relationship between Type and Value

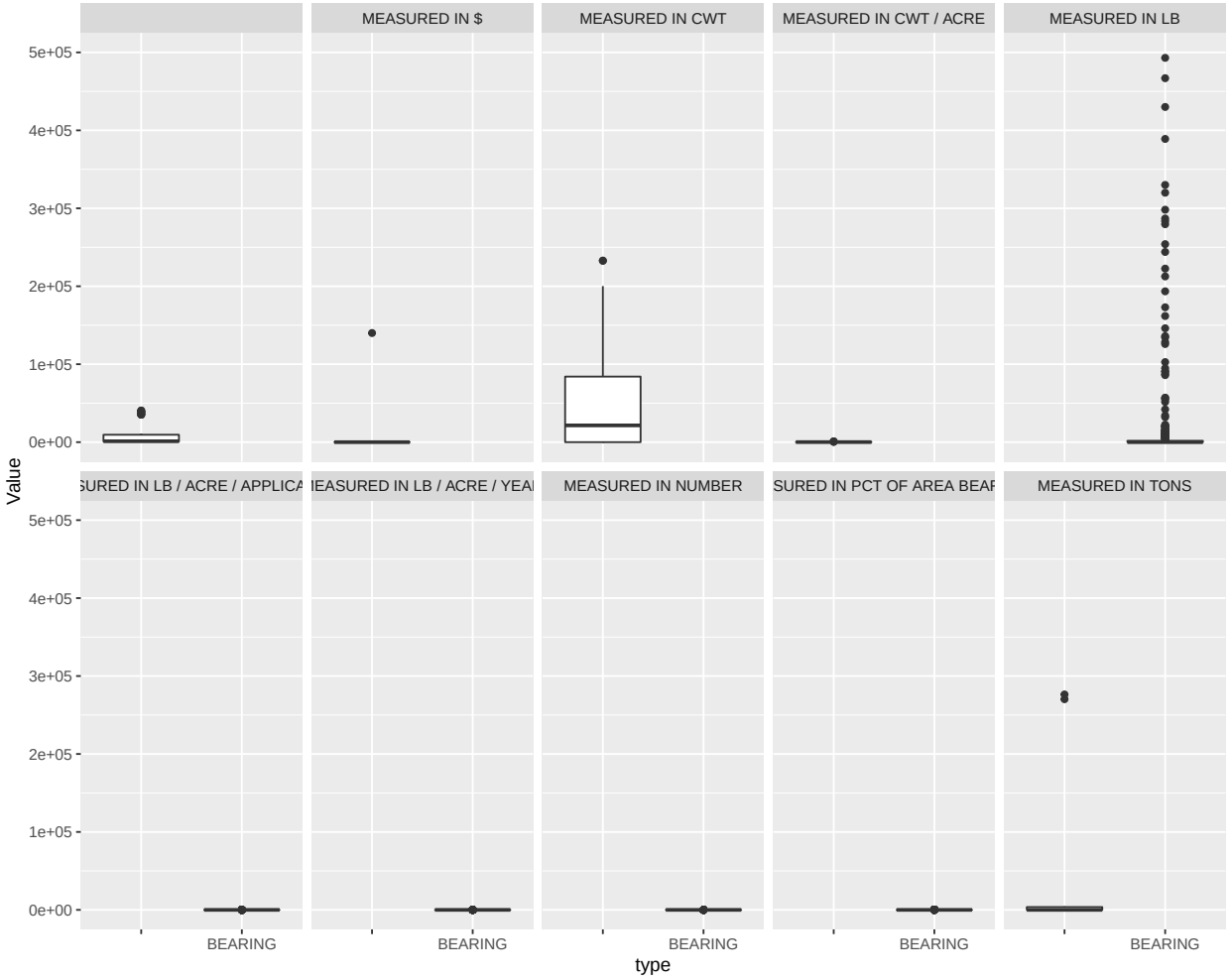

Figure 5: Boxplots separated for each measures.

4 Discussion

- From the figure 1, 2 and 3, we can conclude that California's strawberries have the highest price, the largest weight and the biggest density which increase by year, and Florida's strawberries take the second place. Therefore, California and Florida are more suitable for the growth of strawberries than other states in United states.
- Based on the figure 4, potash fertilizer has the biggest influence on facilitating the growth of strawberries, and nitrogen fertilizer and phosphate fertilizer take the second and third place respectively.
- From the figure 5, we can conclude that adding chemical materials during growing process has little influence on the growth of strawberries.

5 Weakness

Because of time constraints, I cannot clean data in more details such as transforming the value of (D) and (NA) into the mean of all the value with each measures. Besides, I should do more data analysis such PCA.

References

1. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for "Grid" Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra
2. Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.2.1. https://CRAN.R-project.org/package=kableExtra
3. Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

4. I appreciate Lin Zhou and Jinzhe Zhang's help.