Report

by Yu Du

1. The Introduction The project is designed to figure out whether the chemical applied to strawberries in states has an effect on the yield of strawberries in different locations.
2. Exploratory Data Analysis 1). Acquire and read the data: The data were stored online and then downloaded as a CSV file, and eight columns of twenty one columns which contain meaningful data are remained.
3. Data Cleaning:

2a) The analysis is focused on strawberries, so the data containing the "STRAWBERRIES" commodity and "YEAR" time period are selected.

Year	Period	State	Commodity	Data Item	Domain	Domain Category	Value
2019	YEAR	CALIFORNIA	STRAWBERRIES	STRAWBERRIES - ACRES HARVESTED	TOTAL	NOT SPECIFIED	35,400
2019	YEAR	CALIFORNIA	STRAWBERRIES	STRAWBERRIES - ACRES PLANTED	TOTAL	NOT SPECIFIED	36,000
2019	YEAR	CALIFORNIA	STRAWBERRIES	STRAWBERRIES - PRODUCTION, MEASURED IN $	TOTAL	NOT SPECIFIED	2,221,320,000
2019	YEAR	CALIFORNIA	STRAWBERRIES	STRAWBERRIES - PRODUCTION, MEASURED IN CWT	TOTAL	NOT SPECIFIED	20,500,000
2019	YEAR	CALIFORNIA	STRAWBERRIES	STRAWBERRIES - YIELD, MEASURED IN CWT / ACRE	TOTAL	NOT SPECIFIED	580

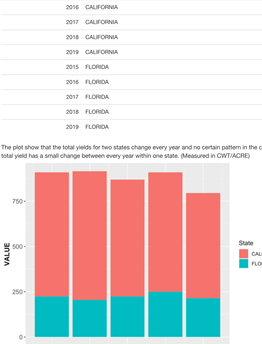
2b).Dividing some columns with mess data into more columns, deleting unuseful colunmns, and combining these columns later to make a more organized dataset.

Year	State	type	production	Avg	Measures	Materials	Chemical	Value
2019	CALIFORNIA		ACRES HARVESTED					35,400
2019	CALIFORNIA		ACRES PLANTED					36,000
2019	CALIFORNIA		PRODUCTION		MEASURED IN $			2,221,320,000
2019	CALIFORNIA		PRODUCTION		MEASURED IN CWT			20,500,000
2019	CALIFORNIA		YIELD		MEASURED IN CWT / ACRE			580

2c). Selecting the data with rows containing the real value in the last columns, therefore the remaining rows can be analyzed. Finalized Data Cleaning Process (before Starting the analysis on variables).

Year	State	type	production	Avg	Measures	Materials	Chemical	Value
2019	CALIFORNIA		ACRES HARVESTED					35,400
2019	CALIFORNIA		ACRES PLANTED					36,000
2019	CALIFORNIA		PRODUCTION		MEASURED IN $			2,221,320,000
2019	CALIFORNIA		PRODUCTION		MEASURED IN CWT			20,500,000
2019	CALIFORNIA		YIELD		MEASURED IN CWT / ACRE			580

3.Selecting subset from the data to analyze on California and Florida. 3a).Plotting the data for total yields in all states from 2015 to 2019

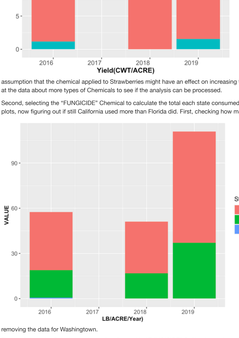


Basic Idea: try to find the relationship between chemicals and the total yield.

3b).Focusing on two main states producing strawberries in the U.S: California and Florida. The data below shows the total YIELD for California and Florida from year 2015 to 2019: The yield is measured in CWT/ACRE, therefore the total yield of California is more than twice as much as Florida has in each year .

	Year	State	total_yield
	2015	CALIFORNIA	685
	2016	CALIFORNIA	710
	2017	CALIFORNIA	645
	2018	CALIFORNIA	660
	2019	CALIFORNIA	580
	2015	FLORIDA	225
	2016	FLORIDA	205
	2017	FLORIDA	225
	2018	FLORIDA	250
	2019	FLORIDA	215

The plot show that the total yields for two states change every year and no certain pattern in the change. However, noticed from the plots, the total yield has a small change between every year within one state. (Measured in CWT/ACRE)



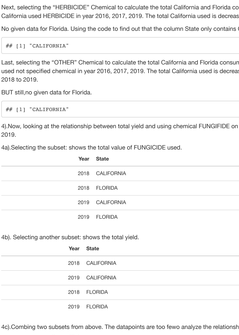
3b): Choosing only one measure and one type of chemical in the data: Looking for whether the chemical has a large effect on yield for California and Florida. Find the difference between the total values states consumed on planting strawberries.Here, choosing only the "MEASURE IN LB/ACRE/YEAR" measure.

First, selecting the "INSECTICIDE" Chemical to calculate the total each state consumed in each year.

The reason I choose to analyze on California and Florida: Some states' are not given in the data, you can see later in the process of analysis. As here, INSECTICIDE only shows for California and Florida. Besides, California and Florida are the top two strawberry producing states within the U.S.

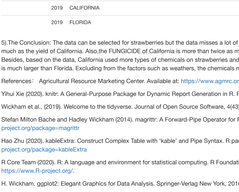
	Year	State	total
	2016	CALIFORNIA	19.953
	2016	FLORIDA	1.174
	2018	CALIFORNIA	19.431
	2019	CALIFORNIA	23.401
	2019	FLORIDA	1.541

The Result from calculation shows that California and Florida used INSECTICIDE in planting strawberries in year 2016,2018,2019.No given data for year 2015 and 2017 and no given data about Florida in year 2018. Noticed that, in each year, the total INSECTICIDE California used on planting strawberries is much larger than the total Florida used. The difference between each year within the same state is similar.



Based on the steps above, raising an assumption that the chemical applied to Strawberries might have an effect on increasing the yields in California and Florida, and starting to look at the data about more types of Chemicals to see if the analysis can be processed.

Second, selecting the "FUNGICIDE" Chemical to calculate the total each state consumed in each year. Based on the previous calculation and plots, now figuring out if still California used more than Florida did. First, checking how many states containing the data for "FUNGICIDE".



Showing 3 states. Therefore, removing the data for Washingtown.

The result from the Calculation shows that the total in "FUNGICIDE" California used on planting strawberries is much larger than the total Florida used, same as the result from calculating "INSECTICIDE". However, from year 2016 to 2019(noticing no rows of Year 2017 in the data), the difference between two states become larger. Besides, Florida used almost the same total value on year 2018, 2019, but the total value of California makes a big change from year 2016 to 2019.

The total FUNGICIDE of California is more than twice as much as Florida has in each year, which seems to correspond with the observation on the total yields between two states analyzed above.

Next, selecting the "HERBICIDE" Chemical to calculate the total California and Florida consumed in each year. The result shows that only California used HERBICIDE in year 2016, 2017, 2019. The total California used is decreasing.

No given data for Florida. Using the code to find out that the column State only contains California:

```
## [1] "CALIFORNIA"
```

Last, selecting the "OTHER" Chemical to calculate the total California and Florida consumed in each year. The result shows that only California used not specified chemical in year 2016, 2017, 2019. The total California used is decreasing from year 2016 to 2018, then is increasing a lot from 2018 to 2019.

BUT still,no given data for Florida.

```
## [1] "CALIFORNIA"
```

4).Now, looking at the relationship between total yield and using chemical FUNGIFIDE on strawberries in California and Florida from year 2018 to 2019.

4a).Selecting the subset: shows the total value of FUNGICIDE used.

	Year	State	total1
	2018	CALIFORNIA	34.316
	2016	FLORIDA	16.740
	2019	CALIFORNIA	73.855
	2019	FLORIDA	37.103

4b). Selecting another subset: shows the total yield.

	Year	State	total_yield
	2018	CALIFORNIA	660
	2019	CALIFORNIA	580
	2018	FLORIDA	250
	2019	FLORIDA	215

4c).Combing two subsets from above. The datapoints are too fewo analyze the relationship:

	Year	State	total1	total_yield
	2018	CALIFORNIA	34.316	660
	2016	FLORIDA	16.740	250
	2019	CALIFORNIA	73.855	580
	2019	FLORIDA	37.103	215

5).The Conclusion: The data can be selected for strawberries but the data misses a lot of information. The yield of California is more than twice as much as the yield of California. Also,the FUNGICIDE of California is more than twice as much as the FUNGICIDE had been used in Florida. Besides, based on the data, California used more types of chemicals on strawberries and the total value of all chemicals measured in lb/acre/year is much larger than Florida. Excluding from the factors such as weathers, the chemicals might help strawberries to yield more in two areas.

References:

Agricultural Resource Marketing Center. Available at: https://www.agmrc.org/commodities-products/fruits/strawberries

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Stefan Milton Bache and Hadley Wickham (2014). magrittr: A Forward-Pipe Operator for R. R package version 1.5. https://CRAN.R-project.org/package=magrittr

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

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.