Raspberries

Alison Pedraza

Part of a project from 10/18/2020

Raspberries: Tiny, tart, and temptingly delicious. But before you pop them in your mouth, what chemicals have been used to get them to grow?

Project Summary

- **Purpose of project:**
  - Do exploratory data analysis to understand the kind and quantity of chemicals used to grow raspberries in the U.S.A.
- **Code Outline:**
  - Collected data from from the U.S. Department of Agriculture National Agricultural Statistical Service.
  - Cleaned and tidied up the data
  - Did Exploratory Data Anlysis on raspberry data to see what kind and how many chemicals are used to grow them.

Data Sources

The data was collected from the *U.S. Department of Agriculture National Agricultural Statistical Service*.

1. Data on the Chemical Agents used in California, Washington, and Oregon to grow strawberries and raspberries.
2. Data on the Total Production per year or raspberries in California, Washington, and Oregon.

Data on Chemical Usage:

The original data on chamical agents contained a number of variables which included: pollination used, yield produced, fertilizer used, as well as restricted and non-restricted use of chemical fungacide, herbacide, and insecticide. The data was stored online and downloaded as a .csv file.

Chemical Data: https://quickstats.nass.usda.gov/results/6E4F3616-5CDA-34DF-B0D1-46D4BE03321E (https://quickstats.nass.usda.gov/results/6E4F3616-5CDA-34DF-B0D1-46D4BE03321E)

Data on Raspberry production:

For total raspberry production by state, data was also collected from the U.S. Department of Agriculture National Agricultural Statistical Service. Please find links below:

For California: https://quickstats.nass.usda.gov/results/A82506DE-F3B2-3BDB-87D7-08B727A5C756 (https://quickstats.nass.usda.gov/results/A82506DE-F3B2-3BDB-87D7-08B727A5C756)

For Washington: https://quickstats.nass.usda.gov/#40159E97-BAE5-347A-B464-10DBC8AACBC5 (https://quickstats.nass.usda.gov/#40159E97-BAE5-347A-B464-10DBC8AACBC5)

For Oregon: https://quickstats.nass.usda.gov/results/8557227A-65F5-376E-8681-9A2D832003D8 (https://quickstats.nass.usda.gov/results/8557227A-65F5-376E-8681-9A2D832003D8)

For Symbol and Select Data Descriptions:

https://www.nass.usda.gov/Data_and_Statistics/Pre-Defined_Queries/ChemUseSymbolandDataItemDefinitions-fruit-vegetables.pdf (https://www.nass.usda.gov/Data_and_Statistics/Pre-Defined_Queries/ChemUseSymbolandDataItemDefinitions-fruit-vegetables.pdf)

The Data: Columns and Values

- Read in data from csv file.
- Dataframe name: rberry_data
- Get: column names, unique values for certain columns
- Get: Year range

```
##  [1] "Ag District"      "Ag District Code" "Commodity"        "County"
##  [5] "County ANSI"      "CV (%)"           "Data Item"        "Domain"
##  [9] "Domain Category"  "Geo Level"        "Period"           "Program"
## [13] "Region"           "State"            "State ANSI"       "Value"
## [17] "Watershed"        "watershed_code"   "Week Ending"      "Year"
## [21] "Zip Code"
```

```
## [1] 211
```

```
## [1] 1990 2019
```

Explanation of Chemical Usage Dataframe

- The crops (**Commodity**) identified in this dataframe are strawberries and raspberries.
- The dataframe shows the total amount of each agent (herbicide, fungacide, insecticide, fertilizer) listed used in pounds (**Value** column) by each crop producing state every year starting in 1990.
- Data spans over 29 years (1990 - 2019).

- **Domain Category** column shows the name of the chemical agents used. There are 211 unique chemical agents in the database.
- Below are the unique values from columns: Commodity, State, Data Item, Domain, Domain Category:

```
## [1] "RASPBERRIES"  "STRAWBERRIES"
```

```
##  [1] "CALIFORNIA"     "FLORIDA"        "WASHINGTON"     "OREGON"
##  [5] "MICHIGAN"       "NEW JERSEY"     "NEW YORK"       "NORTH CAROLINA"
##  [9] "PENNSYLVANIA"   "WISCONSIN"
```

```
## [1] "RASPBERRIES, BEARING - APPLICATIONS, MEASURED IN LB"
## [2] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB"
```

```
## [1] "CHEMICAL, FUNGICIDE"   "CHEMICAL, HERBICIDE"   "CHEMICAL, INSECTICIDE"
## [4] "FERTILIZER"
```

```
## [1] "CHEMICAL, FUNGICIDE: (AZOXYSTROBIN = 128810)"
## [2] "CHEMICAL, FUNGICIDE: (BACILLUS AMYLOLIQUEFACIENS STRAIN D747 = 16482)"
```

```
## [1] 1992 2019
```

```
## [1] 1990 2000
```

Data Cleaning and Scrubbing: Separate and Filter

Clean up was needed to filter out raspberries and details on chemicals used.

Data cleaning on the following collumns:

- Commodity - to filter out Raspberries
- Data Item
- Domain Category - various chemicals

Clean up consisted of:

- Separating values contained in the **Domain Category** column and created new columns of those values (**Substance, Agen, Agent Name**).
- The **Value** column:
    - Filtering out (D), (NA), (Z) entries from the **Value** column.

- Removing those values that were not numeric and would not help in the data analysis. Rows containing the following values were removed: (D), (NA), and (Z).

```
# Filter out Raspberries and separate/clean and arrange columns
rberry_data_bb <- rberry_data %>% filter(Commodity=="RASPBERRIES")
rberry_data_bb %<>%separate(`Data Item`, c("Berry", "Kind", "Units"), ",")
rberry_data_bb %<>% separate(`Domain Category`, c("Substance", "Agent", "Agent Name"))

# Removing values of (NA), (D), and (Z) in the Values column
rberry_data_bb <- rberry_data_bb%>%filter(Value != "(D)")
rberry_data_bb %<>% filter(Value != "(NA)")
rberry_data_bb %<>% filter(Value != "(Z)")

# The Value column entries were characters and had commas. Removed commas and changed characters
to numeric:
rberry_data_bb$Value = as.numeric(gsub("\\," , "", rberry_data_bb$Value))
```

Final Dataframe on Chemical Usage:

- The Final size of dataset: (number of rows): 522
    - After filtering for raspberries and removing rows with (NA), (D) and (Z), the number of rows were reduced from 897 entries to 522 entries.
- The number of unique chemical agents in final dataset: 58
- The Year range for the three states also reduced due to cleaning:

```
+ California years: 2019 - 2019 (narrowed from 1992 - 2019)

+ Washington years: 1991 - 2019

+ Oregon years: 1991 - 2019
```

Year	State	Commodity	Units	Substance	Agent	Agent Name	Value
2019	CALIFORNIA	RASPBERRIES	MEASURED IN LB	CHEMICAL	FUNGICIDE	BOSCALID	300
2019	CALIFORNIA	RASPBERRIES	MEASURED IN LB	CHEMICAL	FUNGICIDE	CYPRODINIL	900
2019	CALIFORNIA	RASPBERRIES	MEASURED IN LB	CHEMICAL	FUNGICIDE	FLUDIOXONIL	600

Year	State	Commodity	Units	Substance	Agent	Agent Name	Value
2019	CALIFORNIA	RASPBERRIES	MEASURED IN LB	CHEMICAL	FUNGICIDE	MYCLOBUTANIL	300
2019	CALIFORNIA	RASPBERRIES	MEASURED IN LB	CHEMICAL	FUNGICIDE	PYRACLOSTROBIN	200
2019	CALIFORNIA	RASPBERRIES	MEASURED IN LB	CHEMICAL	FUNGICIDE	TOTAL	24000

Other Dataframes: Production of Raspberries in California, Washington, and Oregon

- For total production of raspberries by year, dataframes for production from California, Washington, and Oregon were used.
- The 3 dataframes were merged vertically with rbind() since all had same column names.
- The dataframe shows the year and production total for each state.
- **Value** indicates the total production in pounds that year.
- Downloaded data and selected columns for final Production dataframe.

```
# Read in CSV files for Production Totals by State Dataframes
cali_prod <- read_csv("Cali_Rasp_Production.csv")
range(cali_prod$Year)
```

```
## [1] 2007 2019
```

```
wash_prod <- read_csv("Wash_Rasp_Production.csv")
range(wash_prod$Year)
```

```
## [1] 2012 2019
```

```
oreg_prod <- read_csv("Oreg_Rasp_Production.csv")
#range(oreg_prod$Year)
```

California: Years 2007-2019

Year	State	Commodity	Data Item	Value
2019	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	114500000

Year	State	Commodity	Data Item	Value
2018	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	135000000
2017	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	143280000
2016	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	169670000
2015	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	225300000
2014	CALIFORNIA	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	258900000

Washington: Years 2012-2019

Year	State	Commodity	Data Item	Value
2019	WASHINGTON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	81000000
2018	WASHINGTON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	86200000
2017	WASHINGTON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	78050000
2016	WASHINGTON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	80110000
2015	WASHINGTON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	74100000
2014	WASHINGTON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	69990000

Oregon: 2012-2017

Year	State	Commodity	Data Item	Value
2017	OREGON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	5400000
2016	OREGON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	8550000
2015	OREGON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	9390000
2014	OREGON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	8650000
2013	OREGON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	8000000
2012	OREGON	RASPBERRIES	RASPBERRIES - PRODUCTION, MEASURED IN LB	8750000

Creating of Additional Data Frames: Join, Group_by, Filter Operations

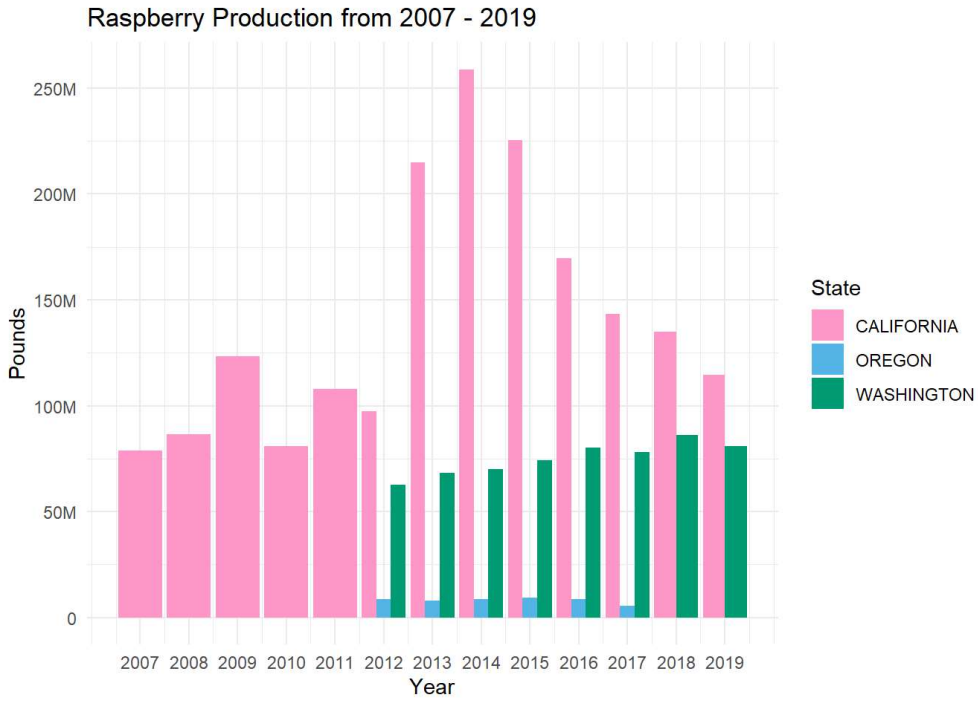
```
#Production Dataframes
# join the three separate dataframes on production to make one large production dataframe
prod_df <- rbind(cali_prod, wash_prod)  # 2007-2019
prod_df_new <- rbind(prod_df, oreg_prod)  # 2012-2019
prod_df_grouped <- prod_df_new%>%group_by(Year)%>%summarise(Total_Pounds = sum(Value))
prod_df_reduced <- prod_df_new%>%filter(Year>2011)

# Additional Rberry data frames
rberry_data_reduced <- rberry_data_final%>%filter(Year>2011)
oreg <-  rberry_data_final%>%filter(State == "OREGON")
rberry_data_grouped <- rberry_data_reduced%>%group_by(Year, State, Agent)%>%summarise(Total_Poun
ds = sum(Value))
rberry_data_grouped_big <- rberry_data_final%>%group_by(Year, State, Agent)%>%summarise(Total_Po
unds = sum(Value))
rberry_data_grouped2 <- rberry_data_grouped%>%group_by(Year, Agent)%>%summarise(Total_Pounds = s
um(Total_Pounds))

unique(prod_df_new$Year)
```

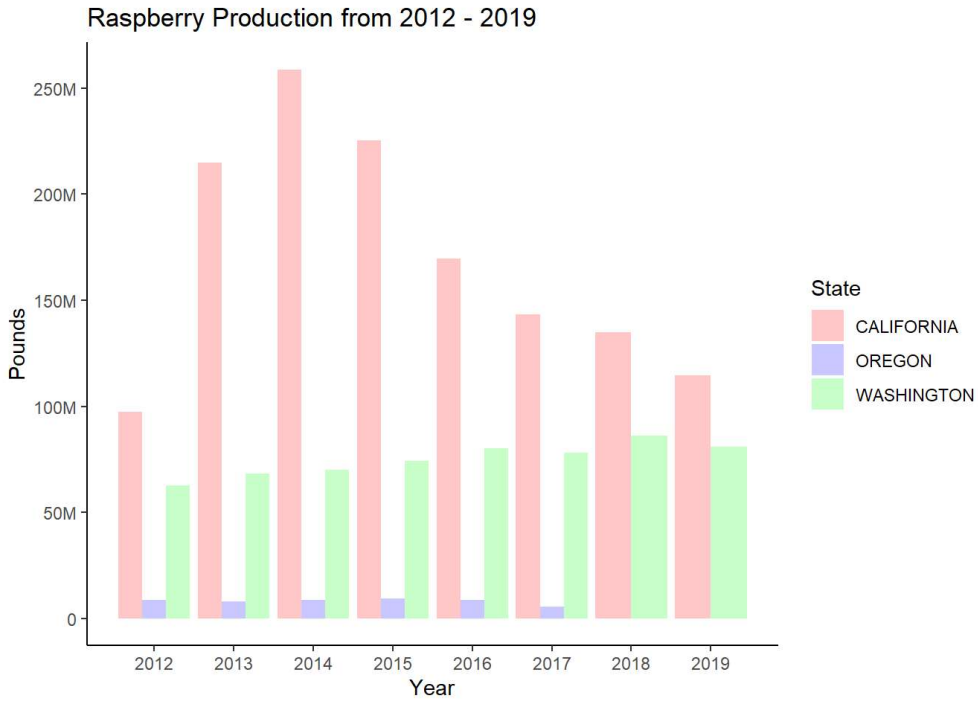
```
##  [1] 2019 2018 2017 2016 2015 2014 2013 2012 2011 2010 2009 2008 2007
```

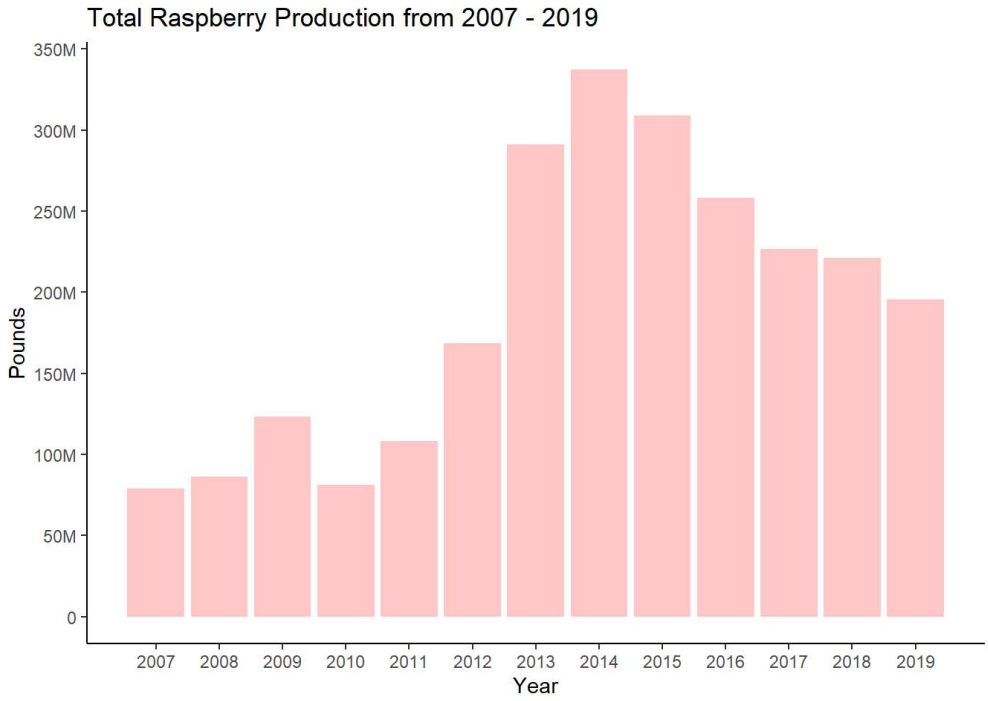
Plots

Raspberry Production from 2007 - 2011 and 2012 - 2019:

- Production data for all three states starts in 2012.
- For the years 2012-2019, California produced the most pounds of raspberries.
- California has reduced raspberry production over the years and Washington has increased its production of the commodity.
- There is no data on Oregon raspberry production for 2018 and 2019.
- Since there is not data on production for Oregon and Washington before 2012, we cannot say that California has always produced the most pounds of raspberries.

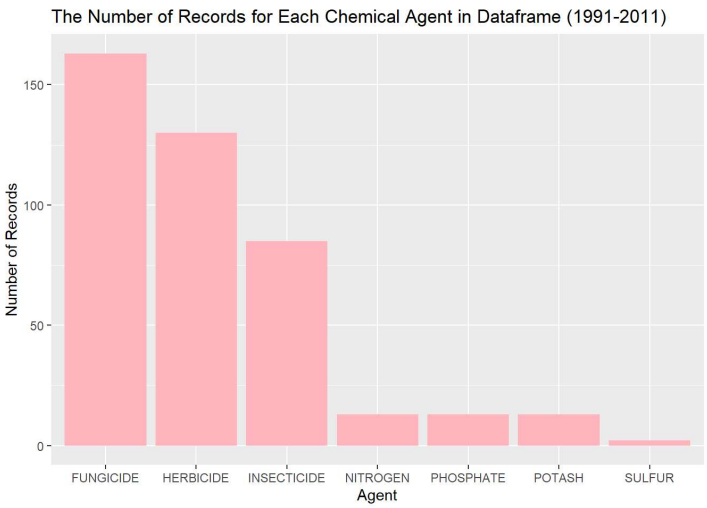

Raspberry Production from 2007 - 2019
250M
200M
150M
100M
50M
0
Pounds
2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
Year
State
CALIFORNIA
OREGON
WASHINGTON

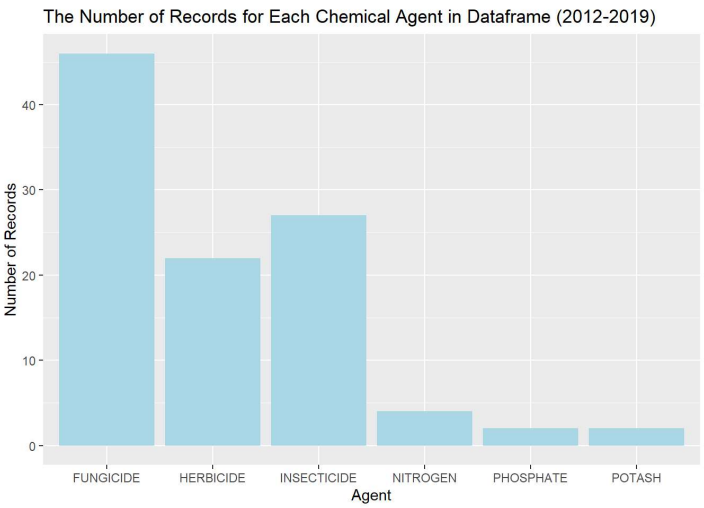

Raspberry Production from 2012 - 2019
250M
200M
150M
100M
50M
0
Pounds
2012 2013 2014 2015 2016 2017 2018 2019
Year
State
CALIFORNIA
OREGON
WASHINGTON



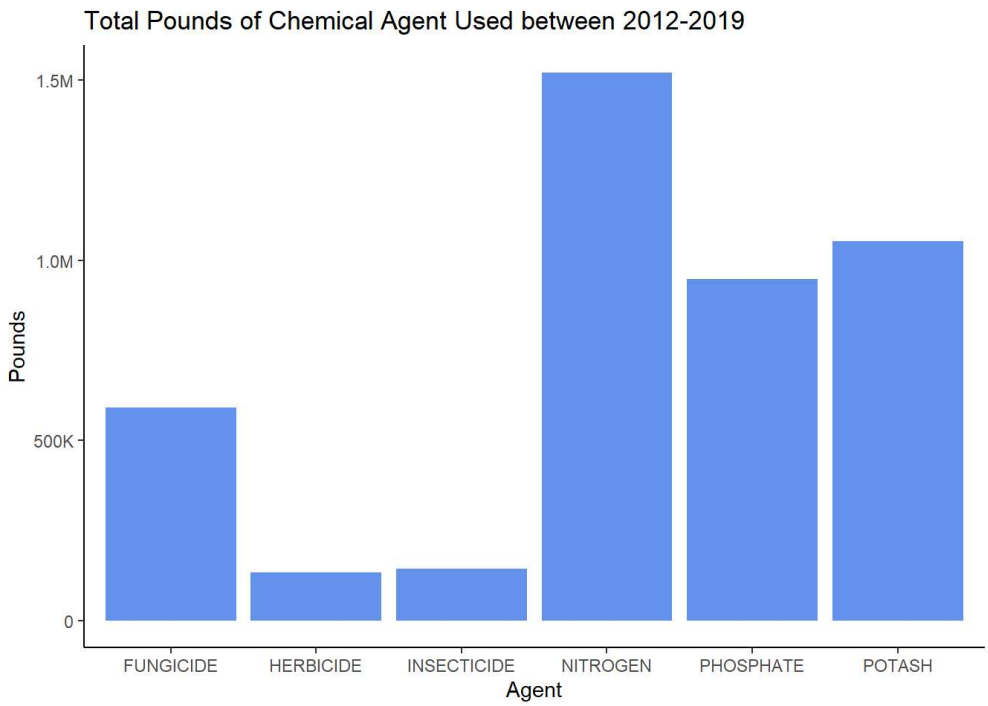
Chemical Agents

- Comparing the total number of records showing chemical agents in the dataset from 1991 - 2011 with the years 2012 - 2019.
- The amount of **herbicide** used before 2011 is much greater than after.
- The plots show that after 2011, there are slightly more instances of **insecticide** recorded and fewer instances of herbicide. Although we are comparing 20 years of records to 7 years of records, this difference is notable.





2012-2019: Total Pounds for Chemical Agents Used



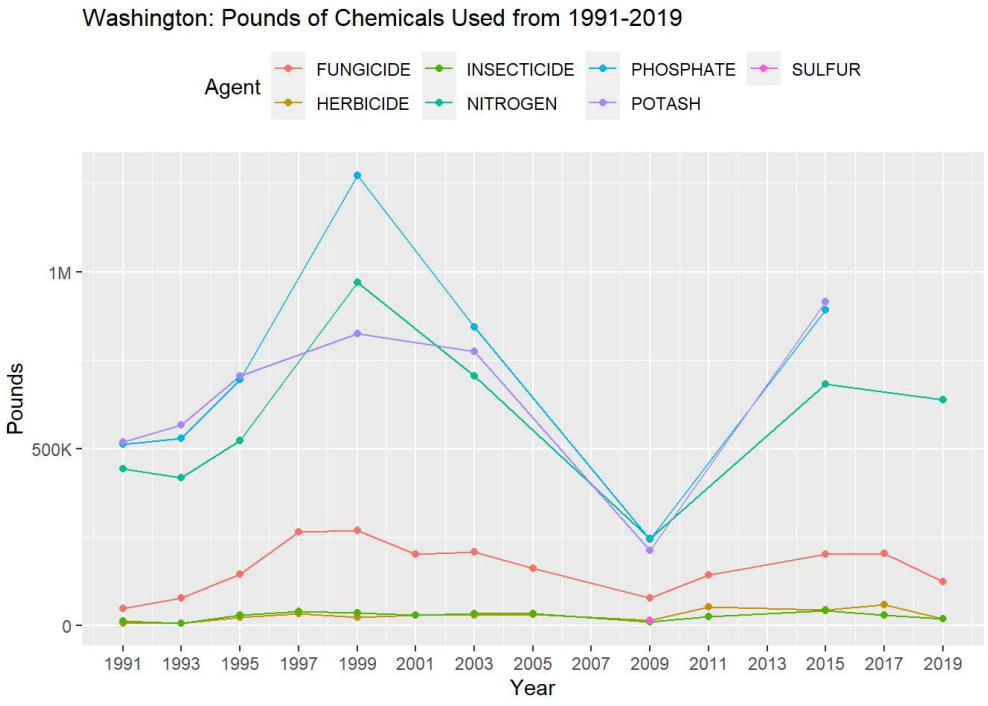
Analysis at State Level

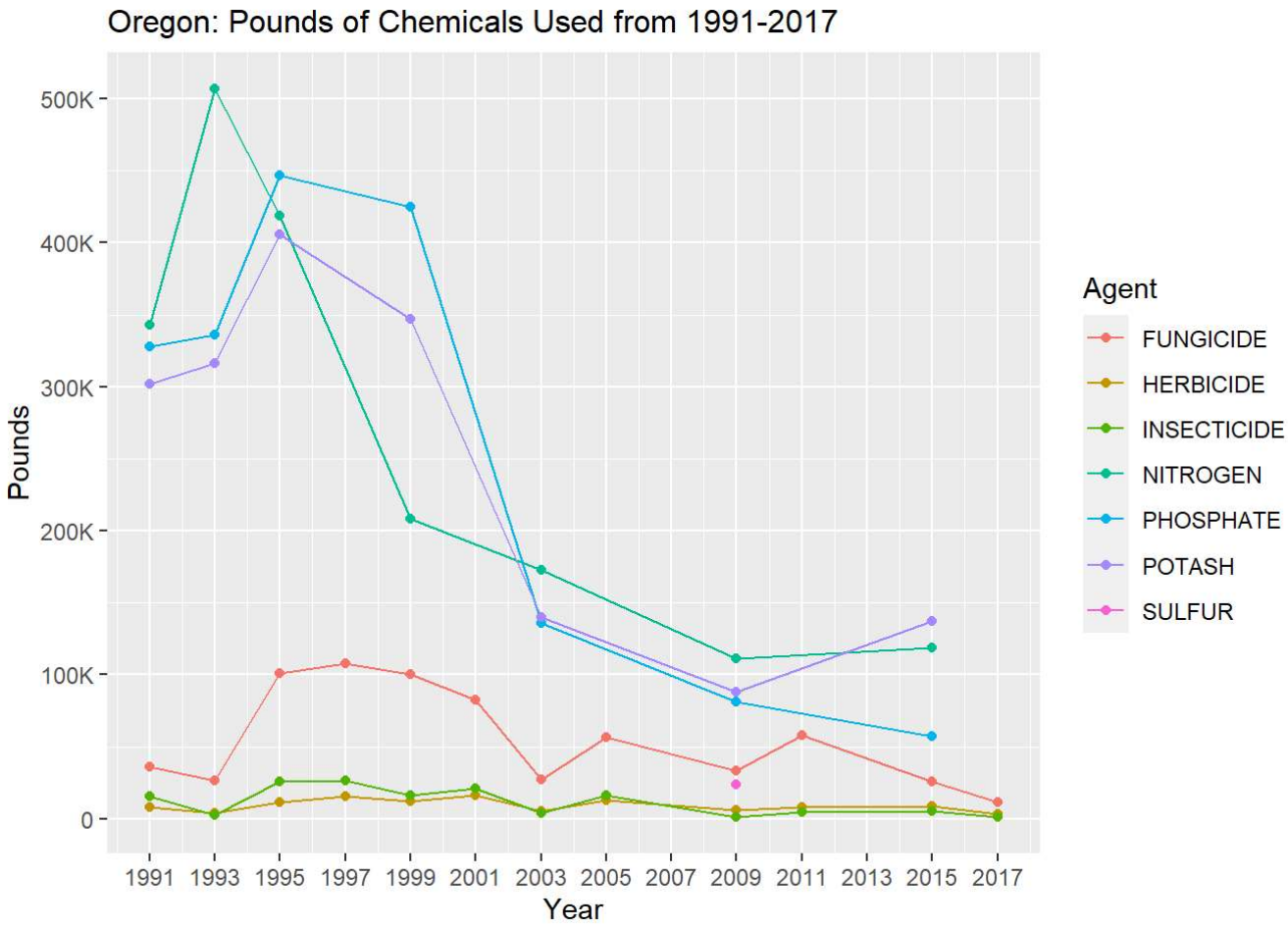
Create new dataframes: Group data by Year, State, and Agent (from 2011-2019)

```
# Create state level dta
wash_agents <- rberry_data_final %>%filter(State == "WASHINGTON")%>% group_by(Year, Agent, Stat
e)%>%
  summarise(Total_Pounds = sum(Value))
cali_agents <- rberry_data_final %>%filter(State == "CALIFORNIA")%>% group_by(Year, Agent, Stat
e)%>%
  summarise(Total_Pounds = sum(Value))
oreg_agents <- rberry_data_final %>%filter(State == "OREGON")%>% group_by(Year, Agent, State)%>%
  summarise(Total_Pounds = sum(Value))
```

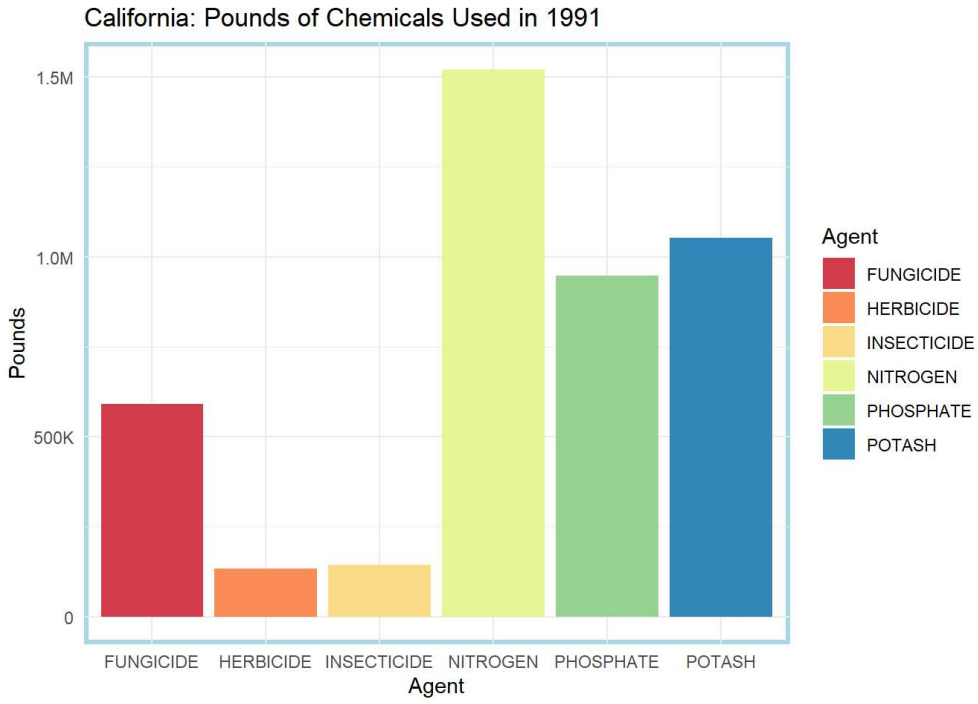
Pounds of Each Chemical Agent Used on a Yearly Basis

- Washington has used fewer chemicals in recent years than previously and only used sulfer once in 2009.
- Oregon has also used fewer chemicals in recent years than in previous years.
- Nitrogen, Potassium, and Phosphate seem to be the chemicals most used for all 3 states.



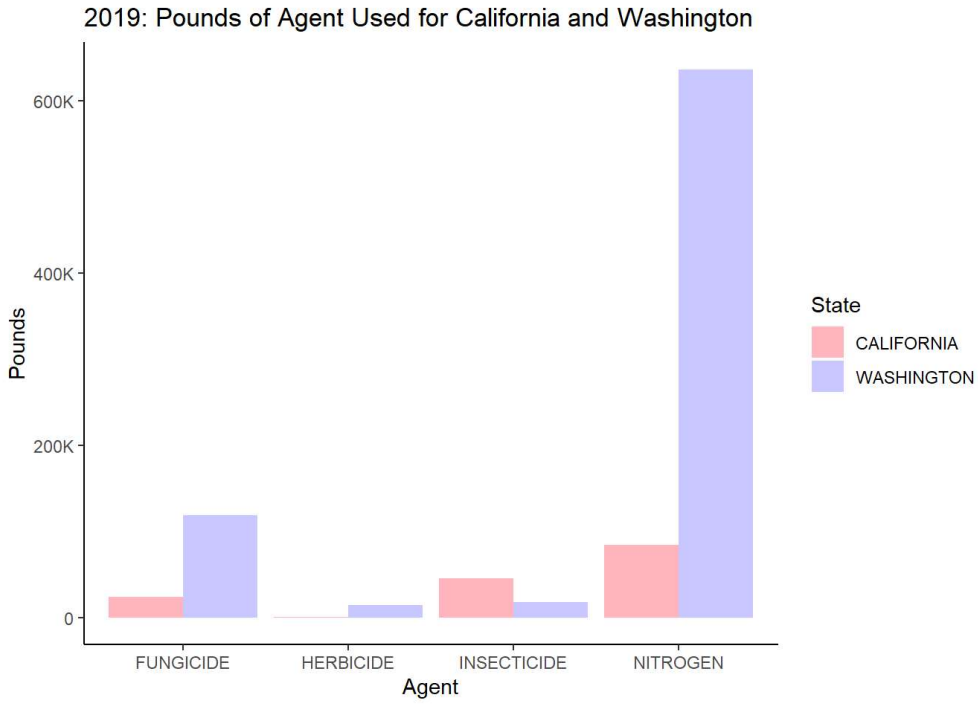


- For California, there is only chemical data for the year 2019:

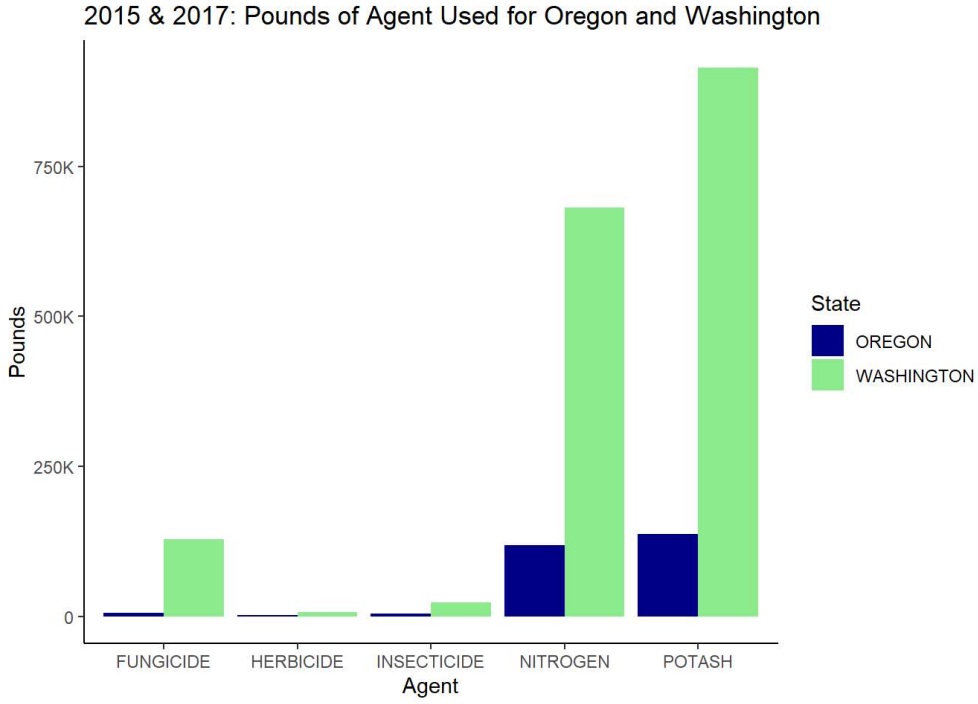


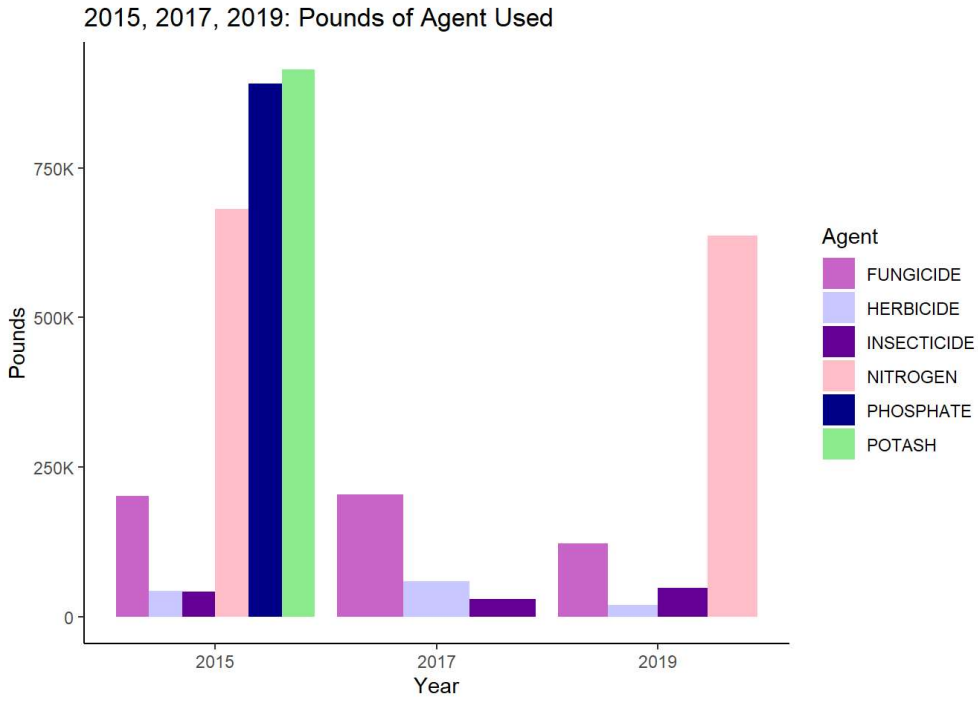
Comparing California and Washington in 2019

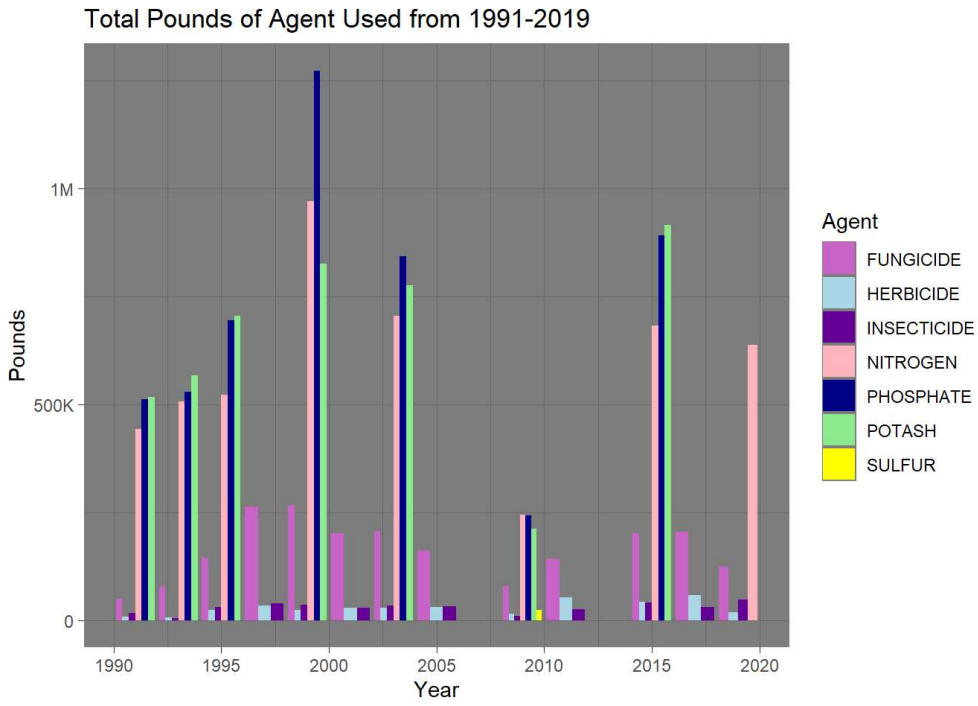
- **There is no data for Oregon in the year 2019.**



Comparing Oregon and Washington:







Some Additional Take-aways from the EDA

1. The pounds of insecticide used over the years has remained consistent for Washington and Oregon.
2. Washington has produced fewer raspberries overall but used more kinds of chemicals than California and Oregon. This may be because it needs more kinds of chemicals due to its environment.

3. Although California does not use as many kinds of chemicals, it does use more pounds of chemical than Washington and Oregon. This may be because it produces more pounds of the commodity.
4. Sulfur was used sparingly in the late 2000s. Normally very little is needed so this may be why there is very little use recorded.
5. This EDA analysis would have benefited from fewer entries of (NA), (D) and (Z) in the **Value** column resulting in more accurate plots.

Resources

The following resources were used to complete these slides and this project:

1. USDA National Agricultural Statistics Service
2. To learn how to remove the comma from the Value Column: Thank you to R for Excel Users. http://www.rforexcelusers.com/remove-currency-dollar-sign-r/ (http://www.rforexcelusers.com/remove-currency-dollar-sign-r/)
3. RMarkdown Cheatsheet
4. R for Data Science - for plot ideas.