Midterm Project EDA

Jin Yan, Tao Guo, Jiaqi Sun

2022-11-09

Cleaning data

Separate the column of Data Item

```
strawb <- read_xlsx("strawberries-2022oct30-a.xlsx",col_names = TRUE)
## separate the Data Item
strawb2 <- strawb %>% separate(col=`Data Item`,
                               into = c("Strawberries", "items"),
                               sep = "-",
                               fill = "right")
strawb2 %<>% separate(col = `Strawberries`, into = c("Strawberries", "type", "type2"),
                      sep = ",", fill = "right")
strawb3 <- strawb2 %>% separate(col = `items`, into = c("sale type", "units"),
                                sep = ",", fill = "right")

strawb4 <- strawb3 %>% select(-c(4, 8:13, 15, 17))
```

1. Build organic and non organic tibble

```
Domain_organic1 <- grep("organic",
                        strawb4$Domain,
                        ignore.case = T)

org_rows1 <- Domain_organic1
strawb_organic <- strawb4 %>% slice(org_rows1, preserve = FALSE)

strawb_non_organic <- strawb4 %>% filter(!row_number() %in% org_rows1)
```

```
strawb_non_organic %<>% pivot_wider(names_from = `units`, values_from = `Value`)
strawb_organic %<>% pivot_wider(names_from = `units`, values_from = `Value`)
```

2. From non organic, separate the chemical tibble

```
##  [1] 312 313 314 315 381 382 383 384 507 508 509 510 557 558 559
```

```
strawb_chem <- strawb_non_organic %>% slice(chem_rows, preserve = FALSE)
```

Poison chemical Carbendazim, Bifenthrin, methyl bromide, 1,3-dichloropropene, chloropicrin, Telone

Searching the Poison chemical mentioned in reading

```
#empty
df_carbendazim <- grep("carbendazim",
                       strawb_chem$`Domain Category`, ignore.case = T)
df_Bifenthrin <- grep("Bifenthrin",
                       strawb_chem$`Domain Category`, ignore.case = T)
df_methyl_bromide <- grep("methyl bromide",
                      strawb_chem$`Domain Category`, ignore.case = T)
#empty
df_1_3_dichloropropene <- grep("1,3-dichloropropene",
                          strawb_chem$`Domain Category`,
                          ignore.case = T)
df_chloropicrin <- grep("chloropicrin",
                                strawb_chem$`Domain Category`, ignore.case = T)
## empty
df_Telone <- grep("Telone",
                       strawb_chem$`Domain Category`,
                       ignore.case = T)
```

Tibble for posion chemicals

The carbendazim, 1_3_dichloropropene, and Telone did not find in table

```
Bifenthrin <- strawb_chem[df_Bifenthrin,]
methyl_bromide <- strawb_chem[df_methyl_bromide,]
dichloropropene <- strawb_chem[df_chloropicrin,]
```

Total Posion chemicals by State

```
Posion_chem <- strawb_chem[c(df_Bifenthrin,df_methyl_bromide,df_chloropicrin ),]
```

```
LB <- ifelse(Posion_chem$` MEASURED IN LB`[1:16]=="(D)",0, Posion_chem$` MEASURED IN LB`[1:16])
LB <- as.numeric(LB)
Posion_chem$` MEASURED IN LB` <- LB
Posion_chem %>% group_by(State) %>% summarise(LB_sum = sum(` MEASURED IN LB`))%>% kable(caption =
      "Total Posion Chemicals Using by State", digits = 3,
      format.args = list(big.mark = ",")) %>%
  kable_minimal(full_width = F)
```

total chemical using by State

Table 1: Total Posion Chemicals Using by State

State	LB_sum
CALIFORNIA	20,192,300
FLORIDA	1,100
OREGON	0

Table 2: Total Chemicals Using by State

State	LB_sum
CALIFORNIA	91,367,900
FLORIDA	3,308,400
OREGON	3,500

```
LB_total <- ifelse(strawb_chem$` MEASURED IN LB`[1:728]=="(D)"|
                     strawb_chem$` MEASURED IN LB`[1:728]=="(NA)"|
                     strawb_chem$` MEASURED IN LB`[1:728]=="(Z)",0,
                   strawb_chem$` MEASURED IN LB`[1:728])
strawb_chem$` MEASURED IN LB` <- as.numeric(LB_total)

strawb_chem%>% group_by(State) %>% summarise(LB_sum = sum(` MEASURED IN LB`))%>%
  kable(caption = "Total Chemicals Using by State",
digits = 3, format.args = list(big.mark = ",")) %>%
kable_minimal(full_width = F)
```

Percentage poison chemicals by State

```
total_rows <- grep("(TOTAL)",
strawb_chem$`Domain Category`, ignore.case = T)
fertlizer_rows <- grep("FERTILIZER",strawb_chem$`Domain Category`, ignore.case = T)
chem_total <- strawb_chem[c(total_rows,fertlizer_rows),]
```

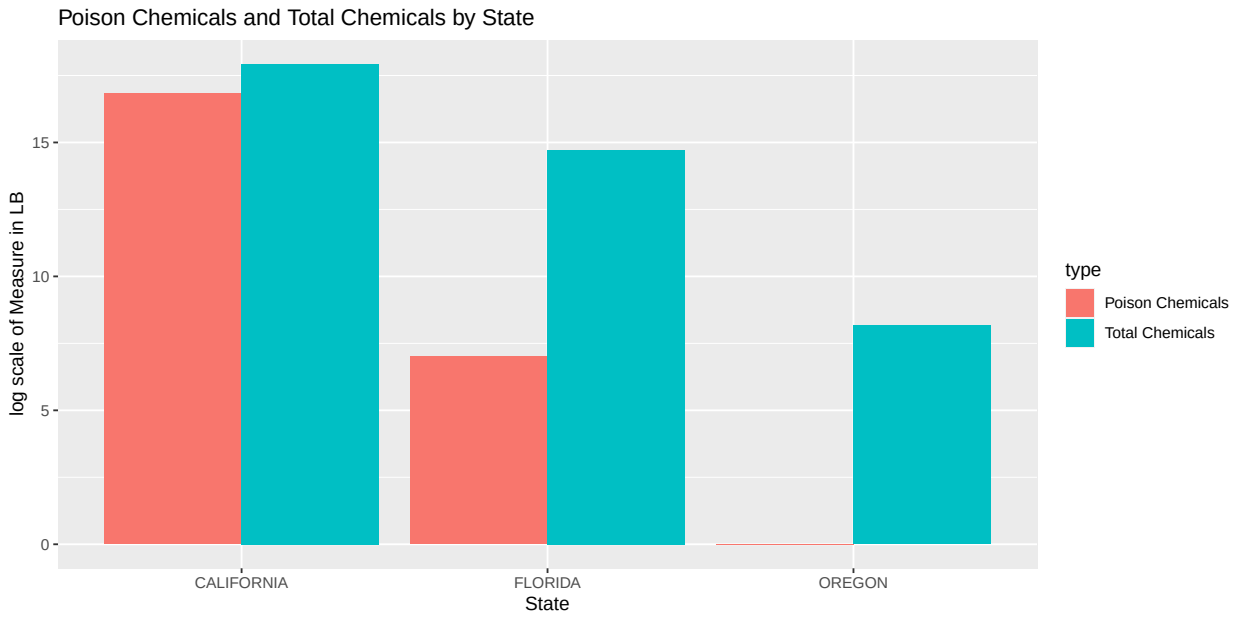
```
Posion_state <- Posion_chem %>% group_by(State) %>% summarise(LB_sum = sum(` MEASURED IN LB`))
chem_state <- chem_total%>% group_by(State) %>% summarise(LB_sum = sum(` MEASURED IN LB`))

Posion_state$Total_LB <- chem_state$LB_sum

percent_posion_state <- Posion_state %>%
  mutate(Percent_posion = LB_sum/Total_LB)
percent_posion_state %>%
  kable(caption = "Percentage Poison Chemicals by State",
      digits = 3, format.args = list(big.mark = ",")) %>%
  kable_minimal(full_width = F)
```

```
percent_posion_state1 <- percent_posion_state[-4] %>%
  pivot_longer(c(LB_sum,Total_LB),
  names_to = "type",values_to = "value")
```

Table 3: Percentage Poison Chemicals by State

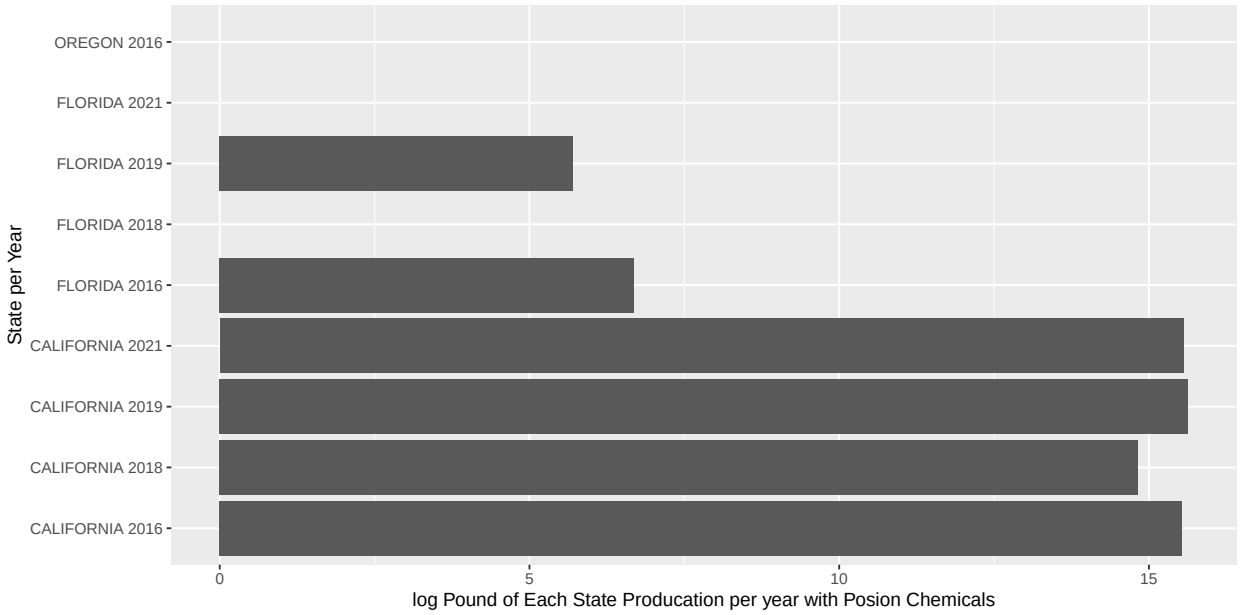
State	LB_sum	Total_LB	Percent_posion
CALIFORNIA	20,192,300	61,733,100	0.327
FLORIDA	1,100	2,466,300	0.000
OREGON	0	3,500	0.000

```
ggplot(percent_posion_state1, aes(fill=type, y=log(value+1), x=State)) +
    geom_bar(position="dodge", stat="identity")+ylab("log scale of Measure in LB")+
  ggtitle("Poison Chemicals and Total Chemicals by State")+
  scale_fill_discrete(labels = c("Poison Chemicals","Total Chemicals"))
```



From this plot, The California high weight of poison chemicals than other State, the Oregon almost did not use poison chemicals. Although the Florida use the poison chemicals, they only use about 0.45%. From percent_posion_state table, over 30 percent chemicals using in CA is poison chemicals which is risky for human health.

```
Posion_state_per_years <- Posion_chem %>%
  group_by(State,Year) %>%
  summarise(LB_sum = sum(` MEASURED IN LB`))
Posion_state_per_years$`State per Year` <-
  paste(Posion_state_per_years$State,
        Posion_state_per_years$Year, sep = " ")

ggplot(Posion_state_per_years, aes(x = `State per Year`, y = log(LB_sum)))+
geom_bar(stat = "identity")+
coord_flip()+ylab("log Pound of Each State Producation per year with Posion Chemicals")
```



In this Plot, from 2016 to 2021, the poison chemicals application in California still maintain at very high levels. Compared to California, the poison chemicals application in Florida is much lower than California and in 2021 the Florida did not use poison chemicals at all. Meanwhile, I find the quantity of California strawberry is extremely higher than others States, based on two plots, so in next I would like find the proportion of strawberry quantity by State. ## Strawberry Proportion by State

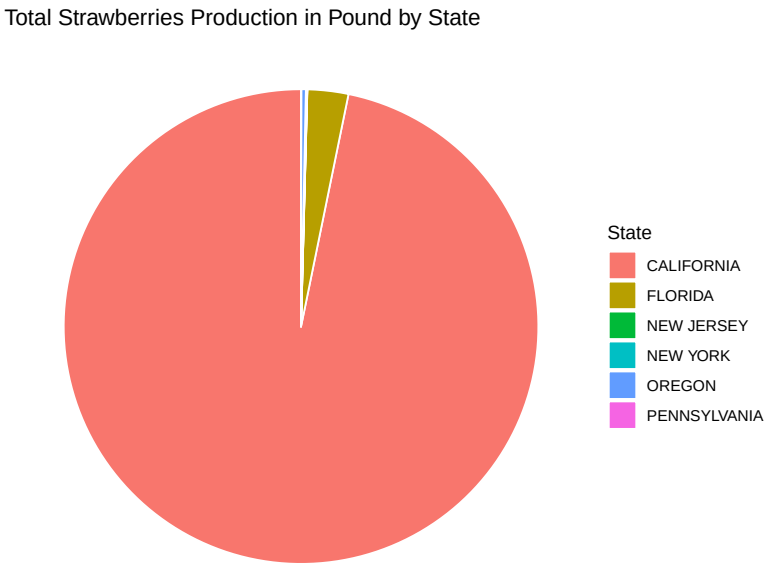
```
strawb5 <- strawb4 %>% pivot_wider(names_from = `units`, values_from = `Value`)
CWT <- ifelse(strawb5$` MEASURED IN CWT`=="(D)",0,strawb5$` MEASURED IN CWT`)
CWT_LB <- as.numeric(CWT)*100
strawb5$` MEASURED IN LB`[1:54]<- CWT_LB
LB_all <- ifelse(strawb5$` MEASURED IN LB`=="(D)"|strawb5$` MEASURED IN LB`=="(Z)"|
         strawb5$` MEASURED IN LB`=="(NA)",0,strawb5$` MEASURED IN LB`)
strawb5$` MEASURED IN LB` <- as.numeric(LB_all)
fresh_rows <-  grep("FRESH MARKET", strawb5$type2, ignore.case = T)
procss_rows <-  grep("PROCESSING", strawb5$type2, ignore.case = T)
strawb5 <- strawb5[-c(fresh_rows,procss_rows),]
total_rows1 <- grep("(TOTAL)",
strawb5$`Domain Category`, ignore.case = T)
fertlizer_rows <- grep("FERTILIZER",strawb5$`Domain Category`, ignore.case = T)
organic_rows1 <- grep("ORGANIC",strawb5$type, ignore.case = T)
strawb5 <- strawb5[c(total_rows1,fertlizer_rows,organic_rows1),]
strawb5 %>% group_by(State) %>%
  summarise(LB = sum(` MEASURED IN LB`, na.rm = TRUE))%>%
  kable(caption = "Total Strawberry Production by State in Pound",
        digits = 3, format.args = list(big.mark = ",")) %>%
  kable_minimal(full_width = F)
```

```
total_state_LB <- strawb5 %>% group_by(State) %>%
  summarise(LB = sum(` MEASURED IN LB`, na.rm = TRUE))
ggplot(total_state_LB, aes(x="", y=LB, fill=State)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0)+
  theme_void()+
```

Table 4: Total Strawberry Production by State in Pound

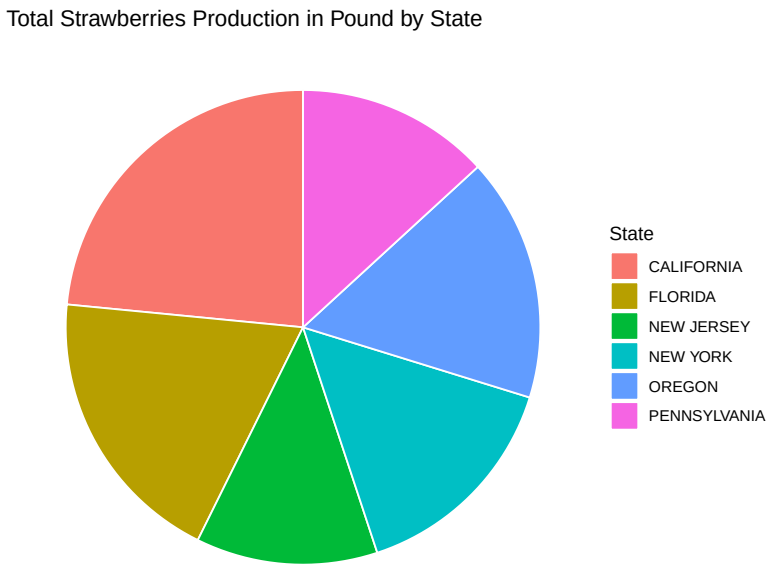
State	LB
CALIFORNIA	344,780,500
FLORIDA	9,910,000
NEW JERSEY	31,900
NEW YORK	327,200
OREGON	1,094,800
PENNSYLVANIA	62,600

```
  ggtitle("Total Strawberries Production in Pound by State")
```



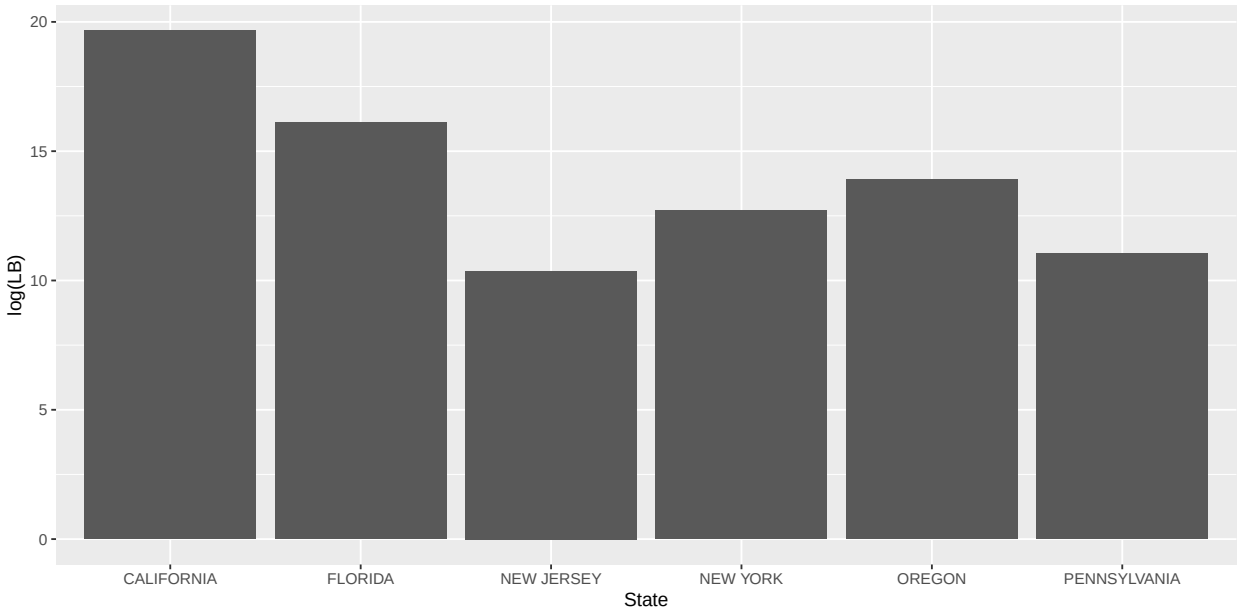
This Pi plot responds to my previous opinion, in this data sets, the California produce over 90% Strawberry. From my percentage poison chemicals by State, there are 20192300 LB strawberry may harm human health. Therefore, there are at least 6% strawberry produced in California, which are harmful.

```
ggplot(total_state_LB, aes(x="", y=log(LB), fill=State)) +
  geom_bar(stat="identity", width=1, color="white") +
  coord_polar("y", start=0)+theme_void()+
  ggtitle("Total Strawberries Production in Pound by State")
```



This last figure, I log the value of production of strawberry in order to find the proportion of other State without California. This Pi chart show the Florida production is relatively higher than other states. Others production is closes with each others beside California.

```
ggplot(total_state_LB, aes(x=State, y=log(LB))) +
  geom_bar(stat = "identity") + theme(legend.position="none")
```



Safe Chemicals

chemicals commonly used, which are safe ones: "ACIBENZOLAR-S-METHYL" - 12 results, " ACETAMIPRID " - 16 results, " ACEQUINOCYL " - 10 results

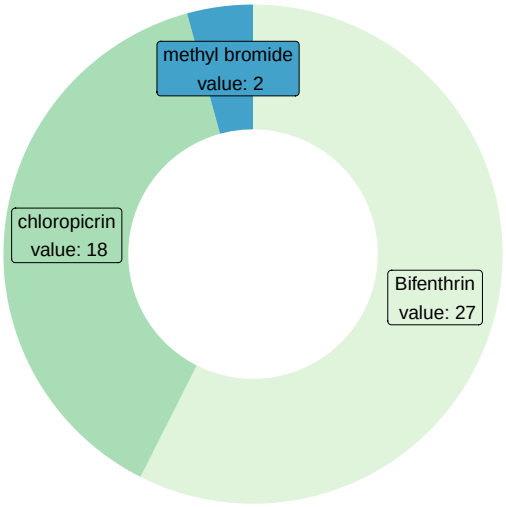
Percentage of poison chemicals

chemicals commonly used according to "Shopper's Guide to Pesticides in Produce™" which are deadly possion: "Bifenthrin"-27 results, "chloropicrin"-18 results and "methyl bromide"-2 results

```
# load library
library(ggplot2)

# Create test data.
hazardous_chems <- data.frame(
  category=c("Bifenthrin", "chloropicrin", "methyl bromide"),
  count=c(27, 18, 2)
)

# Compute percentages
hazardous_chems$fraction <- hazardous_chems$count / sum(hazardous_chems$count)

# Compute the cumulative percentages (top of each rectangle)
hazardous_chems$ymax <- cumsum(hazardous_chems$fraction)

# Compute the bottom of each rectangle
hazardous_chems$ymin <- c(0, head(hazardous_chems$ymax, n=-1))

# Compute label position
hazardous_chems$labelPosition <- (hazardous_chems$ymax + hazardous_chems$ymin) / 2

# Compute a good label
hazardous_chems$label <- paste0(hazardous_chems$category, "\n value: ", hazardous_chems$count)

# Make the plot
ggplot(hazardous_chems, aes(ymax=ymax, ymin=ymin, xmax=4, xmin=3, fill=category)) +
  geom_rect() +
  geom_label( x=3.5, aes(y=labelPosition, label=label), size=4) +
  scale_fill_brewer(palette=4) +
  coord_polar(theta="y") +
  xlim(c(2, 4)) +
  theme_void() +
  theme(legend.position = "none")
```

Table 5: Table Poison used Per Acre: raw form

Year	State	Per_acre_ys
2,016	CALIFORNIA	5,591,800
2,016	FLORIDA	800
2,018	CALIFORNIA	2,750,600
2,019	CALIFORNIA	6,153,900
2,019	FLORIDA	300
2,021	CALIFORNIA	5,696,000



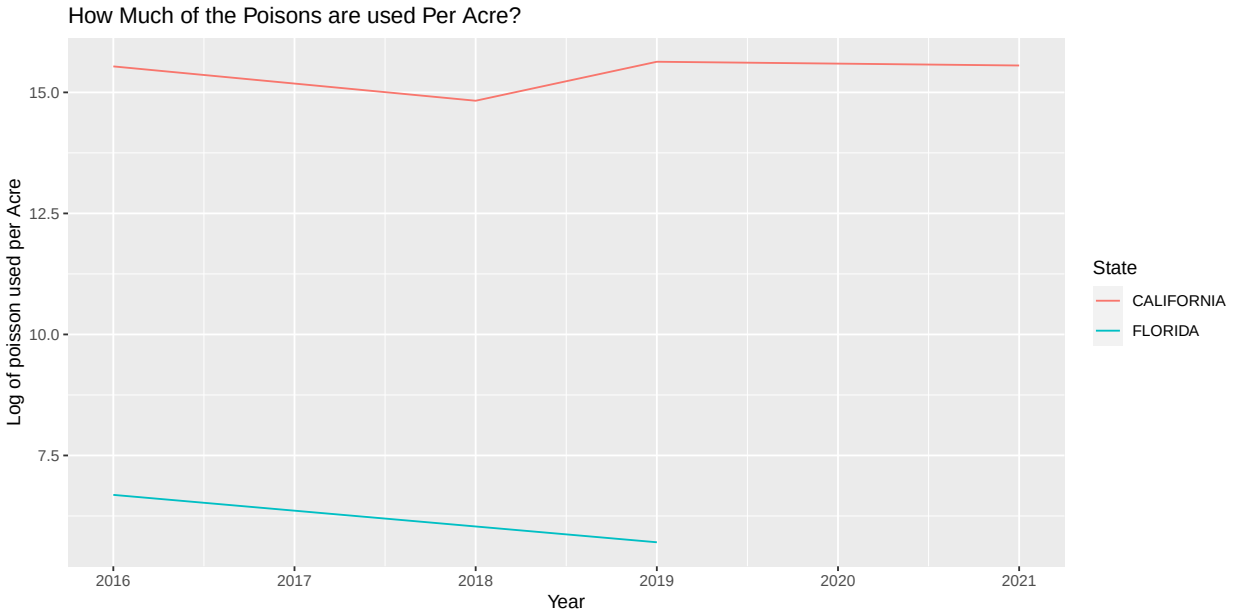
The donuts graph shows the percentage each poison takes in all, with Bifen the most, and chloropicrin second, Methyl bromide least.

How Much of the Poisons are used Per Acre?

```
library(ggplot2)

pois_all <- rbind(bifen_pa,bifen_papa,bifen_papy,
chloropicrin_pa, chloropicrin_papa, chloropicrin_papy)
pois_all %>%
  group_by(Year, State) %>%
  filter(`per Acre` != 0) %>%
   summarise(`Per_acre_ys` = sum(as.numeric(`per Acre`))) -> pois_all_new
```

```
pois_all_new %>%
  kable(caption = "<b>Table Poison used Per Acre: raw form</b>",
        digits = 3, format.args = list(big.mark = ",")) %>%
  kable_minimal(full_width = F)
```

```
ggplot(pois_all_new, aes(x = Year, y = log(`Per_acre_ys`+1), colour = State)) +
  geom_line() +
  labs(x = "Year",
  y = "Log of poisson used per Acre",
  title ="How Much of the Poisons are used Per Acre?")
```



we know from the graph below that California has been using very high level of poisson per acre with no plans to decrease them, while Florida hasn't been using much per acre and keep decrease the amount of poison.

We would suggest California to decrease the amount of poison using per acre and suggest customers buy strawberries from States other than California.

How Much of the Poisons are used Per Acre Per Application?

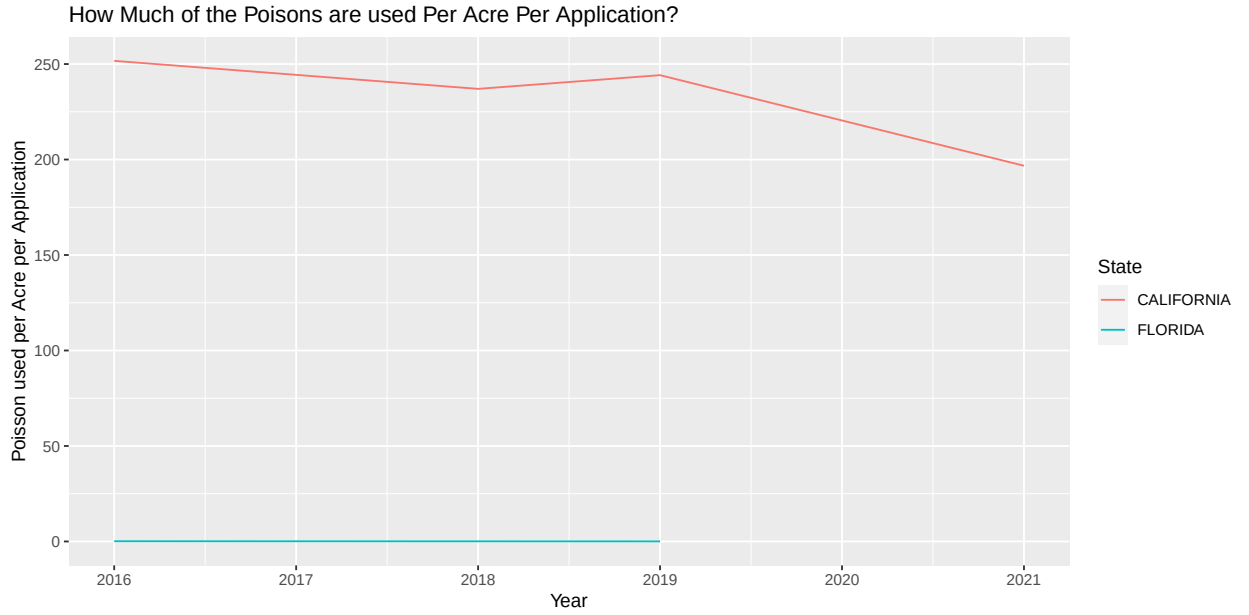
```
pois_all <- rbind(bifen_pa,bifen_papa,bifen_papy,
                  chloropicrin_pa, chloropicrin_papa,
                  chloropicrin_papy)
pois_all %>%
  group_by(Year, State) %>%
  filter(`Average per Acre per Application` != 0) %>%
   summarise(`Per_acre_pa_ys` = sum(as.numeric(`Average per Acre per Application`))) -> pois_all_papa
```

```
pois_all_papa %>% kable(
  caption =
"Table Poison used Per Acre Per Application: raw form",
digits = 3,
format.args = list(big.mark = ",")) %>%
kable_minimal(full_width = F)
```

Table 6: Table Poison used Per Acre Per Application: raw form

Year	State	Per_acre_pa_ys
2,016	CALIFORNIA	251.666
2,016	FLORIDA	0.117
2,018	CALIFORNIA	236.999
2,019	CALIFORNIA	244.175
2,019	FLORIDA	0.050
2,021	CALIFORNIA	196.738

```
ggplot(pois_all_papa, aes(x = Year, y = `Per_acre_pa_ys`, colour = State)) +
  geom_line() +
  labs(x = "Year",
  y = "Poisson used per Acre per Application",
  title ="How Much of the Poisons are used Per Acre Per Application?")
```



we know from the graph below that California has been using very high level of poison per acre per application but has to decrease them as the time going, while Florida's using of poisson is nearly 0 per acre per application.

We would suggest California to decrease the amount of poison using per acre per application and suggest customers buy strawberries from States other than California.

How Much of the Poisons are used Per Acre Per Year?

```
pois_all <- rbind(bifen_pa,bifen_papa,bifen_papy,
chloropicrin_pa, chloropicrin_papa, chloropicrin_papy)
pois_all %>%
  group_by(Year, State) %>%
  filter(`Average per Acre per Year` != 0) %>%
   summarise(`Per_acre_papy_ys` = sum(as.numeric(`Average per Acre per Year`))) -> pois_all_papy
```

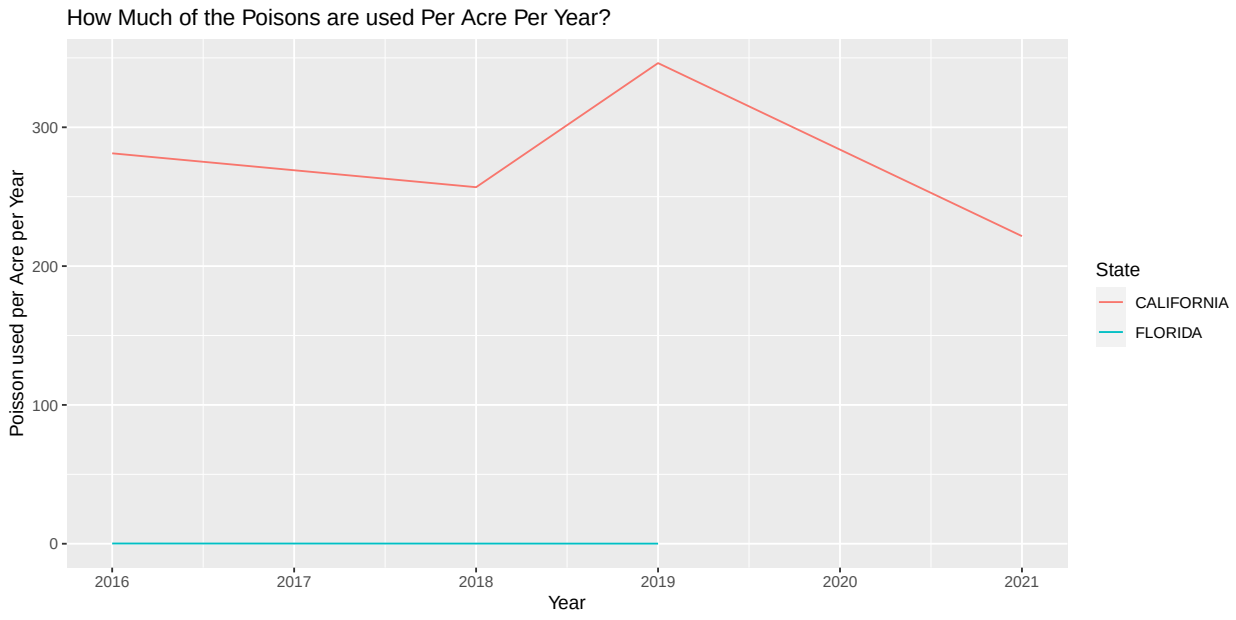
Table 7: Table Poison used Per Acre Per Year: raw form

Year	State	Per_acre_papy_ys
2,016	CALIFORNIA	281.253
2,016	FLORIDA	0.162
2,018	CALIFORNIA	256.850
2,019	CALIFORNIA	346.239
2,019	FLORIDA	0.077
2,021	CALIFORNIA	221.590

```
## `summarise()` has grouped output by 'Year'. You can override using the
## `.groups` argument.
```

```
pois_all_papy %>%
  kable(caption = "<b>Table Poison used Per Acre Per Year: raw form</b>",
        digits = 3, format.args = list(big.mark = ",")) %>%
  kable_minimal(full_width = F)
```

```
ggplot(pois_all_papy, aes(x = Year, y = `Per_acre_papy_ys`, colour = State)) +
  geom_line() +
  labs(x = "Year",
  y = "Poisson used per Acre per Year",
  title ="How Much of the Poisons are used Per Acre Per Year?")
```



we know from the graph below that California has been using very high level of poison per acre per year but has to decrease them as the time going, while Florida's using of poisson is nearly 0 per acre per year.

We would suggest California to decrease the amount of poison using per acre per year and suggest customers buy strawberries from States other than California.

Improvement data sets

Firstly, too many columns contained NA should be removed, which are useless. Secondly, many different variables contained in same column, which should be separated. Thirdly, total sales value and others are mixed, which are confused to calculation. Fourthly, the sales units are not recorded in same measurement method, which are difficult to calculate.

CV problem

```
population_mean=231304956
CV=0.137
SD=population_mean*CV
ci_upper<-population_mean+1.96*SD
ci_lower<-population_mean-1.96*SD
print(ci_upper)
```

```
## [1] 293414963
```

```
print(ci_lower)
```

```
## [1] 169194949
```

Due to the organic strawberries were collected by census, we can initiative to collect the data of CV and many of other details, but as for the non-organic strawberries, the method of collect is survey(the self-reported), the method lack the enough details for CV and other complete data, also, if ues we change the method collect of survey to census, there are much of tima and money we need to cost, it is not worthy,so there is no CV data for non-organic strawberries. As to the usage of CV, we through the critical value and SD, we can calculate the confidence interval for organic strawberries: Margin of error(parameter) = Critical value*standard deviation for population; population mean= 231304956,CV=13.7%,SD=mean*CV