Stawberries: exploratory data analysis

Ruicheng Zhang

2023-10-16

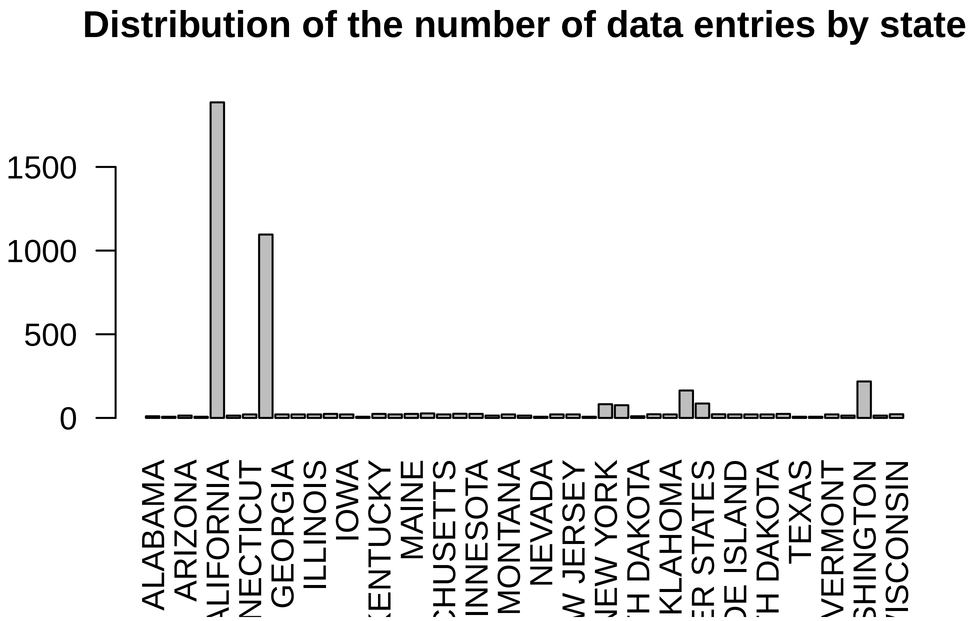
Initial questions

Is the data complete, what direction do we need to take our research, is there a relationship between the variables, and is this data source reliable?

Data acquisition and assessment

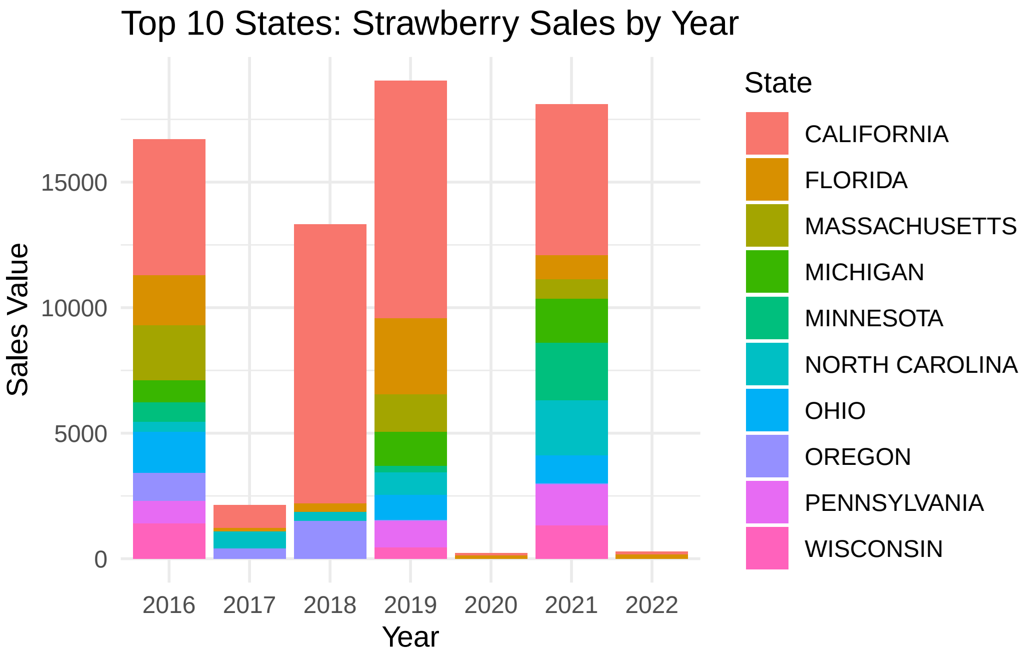
```
Rows: 4,314
Columns: 21
$ Program             <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "~
$ Year                <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 202~
$ Period              <chr> "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YE~
$ `Week Ending`       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Geo Level`         <chr> "STATE", "STATE", "STATE", "STATE", "STATE", "STATE~
$ State               <chr> "ALASKA", "ALASKA", "ALASKA", "ALASKA", "ALASKA", "~
$ `State ANSI`        <chr> "02", "02", "02", "02", "02", "02", "02", "06", "06~
$ `Ag District`       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Ag District Code` <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ County              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `County ANSI`       <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ `Zip Code`          <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Region              <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ watershed_code      <chr> "00000000", "00000000", "00000000", "00000000", "00~
$ Watershed           <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ Commodity           <chr> "STRAWBERRIES", "STRAWBERRIES", "STRAWBERRIES", "ST~
$ `Data Item`         <chr> "STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES", "S~
$ Domain              <chr> "ORGANIC STATUS", "ORGANIC STATUS", "ORGANIC STATUS~
$ `Domain Category`   <chr> "ORGANIC STATUS: (NOP USDA CERTIFIED)", "ORGANIC ST~
$ Value               <chr> "2", "(D)", "(D)", "(D)", "2", "(D)", "(D)", "142",~
$ `CV (%)`            <chr> "(H)", "(D)", "(D)", "(D)", "(H)", "(D)", "(D)", "1~
```

```
state <- table(strawberry$State)
barplot(state, main="Distribution of the number of data entries by state", las=2)
```



```
# Filtering non-numeric values in Value fields and converting them to numbers
strawberry$Value <- as.numeric(as.character(strawberry$Value), na.rm=F)

# Grouping by state and year and totaling strawberry sales
sales_by_state_year <- strawberry %>%
  group_by(State, Year) %>%
  summarise(Value = sum(Value, na.rm=TRUE), .groups='drop')

# Select the top 10 states with the highest sales
top_states <- sales_by_state_year %>%
  group_by(State) %>%
  summarise(Total = sum(Value), .groups='drop') %>%
  arrange(-Total) %>%
  head(10) %>%
  pull(State)

# Filtering data
filtered_data <- sales_by_state_year %>%
  filter(State %in% top_states)
```

```r
# Plotting stacked bar charts
ggplot(filtered_data, aes(x=as.factor(Year), y=Value, fill=State)) +
  geom_bar(stat="identity", position="stack") +
  labs(title="Top 10 States: Strawberry Sales by Year", x="Year", y="Sales Value") +
  theme_minimal()
```



Data cleaning and organization

```
Rows: 4,314
Columns: 10
$ Program           <chr> "CENSUS", "CENSUS", "CENSUS", "CENSUS", "CENSUS", "C~
$ Year              <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021~
$ Period            <chr> "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEAR", "YEA~
$ State             <chr> "ALASKA", "ALASKA", "ALASKA", "ALASKA", "ALASKA", "A~
$ `State ANSI`      <chr> "02", "02", "02", "02", "02", "02", "02", "06", "06"~
$ `Data Item`       <chr> "STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES", "ST~
$ Domain            <chr> "ORGANIC STATUS", "ORGANIC STATUS", "ORGANIC STATUS"~
$ `Domain Category` <chr> "ORGANIC STATUS: (NOP USDA CERTIFIED)", "ORGANIC STA~
$ Value             <dbl> 2, NA, NA, NA, 2, NA, NA, 142, NA, NA, NA, 141, NA, ~
$ `CV (%)`          <chr> "(H)", "(D)", "(D)", "(D)", "(H)", "(D)", "(D)", "19~
```

EDA

First,for the survey part of the data is processed by splitting the chemistry into two columns and removing meaningless variables.

```
stb_survey <- strwb_survey %>%
  filter(str_detect(`Data Item`, "MEASURED IN")) %>%
  mutate(`Data Item` = str_extract(`Data Item`, "(?<=MEASURED IN ).*"))
stb_survey <- stb_survey %>%
  mutate(
    Chemical = if_else(str_detect(`Domain Category`, "\\(.*=.*\\)"),
                       str_extract(`Domain Category`, "(?<=\\().*?(?=\\=)"),
                       NA_character_),
    Chemical_Code = if_else(str_detect(`Domain Category`, "\\(.*=.*\\)"),
                            str_extract(`Domain Category`, "(?<=\\=).*?(?=\\))"),
                            NA_character_)
  )


stb_survey <- subset(stb_survey, select = -Program)
stb_survey <- subset(stb_survey, select = -`Domain Category`)
```

Dealing with Missing Values, Outliers, and Duplicates

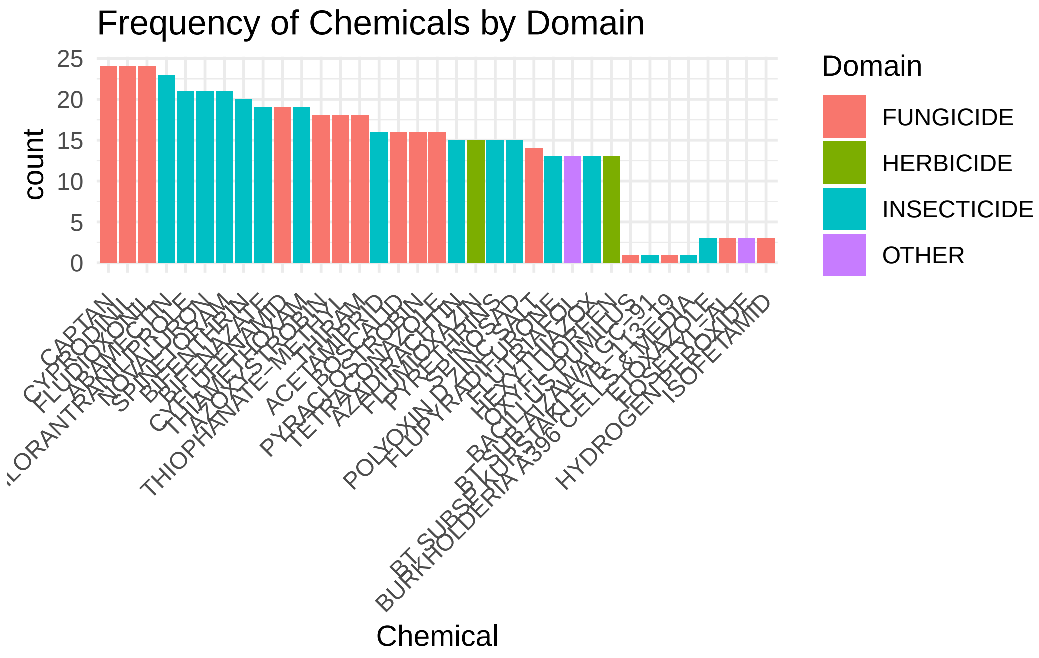
```
stb_survey <- stb_survey[, !sapply(stb_survey, function(col) all(is.na(col)))]


stb_survey <- stb_survey[!is.na(stb_survey$Value), ]


stb_survey <- stb_survey[stb_survey$State != "OTHER STATES", ]
```

```
strwb_census$`CV (%)`<- as.numeric(strwb_census$`CV (%)`)
strwb_census <- strwb_census %>%
  select(-Program,-`Period`,-Fruit,-crop_type,-Domain,-`Domain Category`)
```

Do the same for census

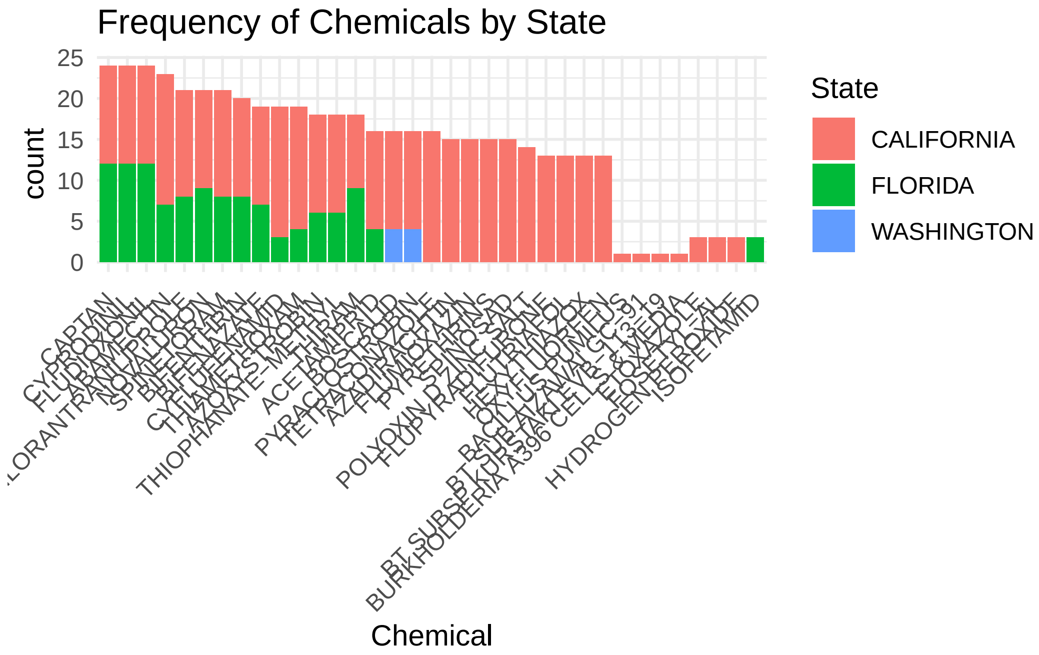
```
stb_survey$Domain <- gsub("CHEMICAL,", "", stb_survey$Domain)
stb_survey$Domain <- trimws(stb_survey$Domain)
```

```r
chemical_counts <- table(stb_survey$Chemical)
top_10_chemicals <- names(sort(chemical_counts, decreasing = TRUE)[1:27])
bottom_10_chemicals <- names(sort(chemical_counts)[1:8])
selected_chemicals <- c(top_10_chemicals, bottom_10_chemicals)


subset_stb_survey <- stb_survey[stb_survey$Chemical %in% selected_chemicals, ]


ggplot(subset_stb_survey, aes(x = Chemical, fill = Domain)) +
  geom_bar() +
  scale_x_discrete(limits = selected_chemicals) +
  labs(title = "Frequency of Chemicals by Domain") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

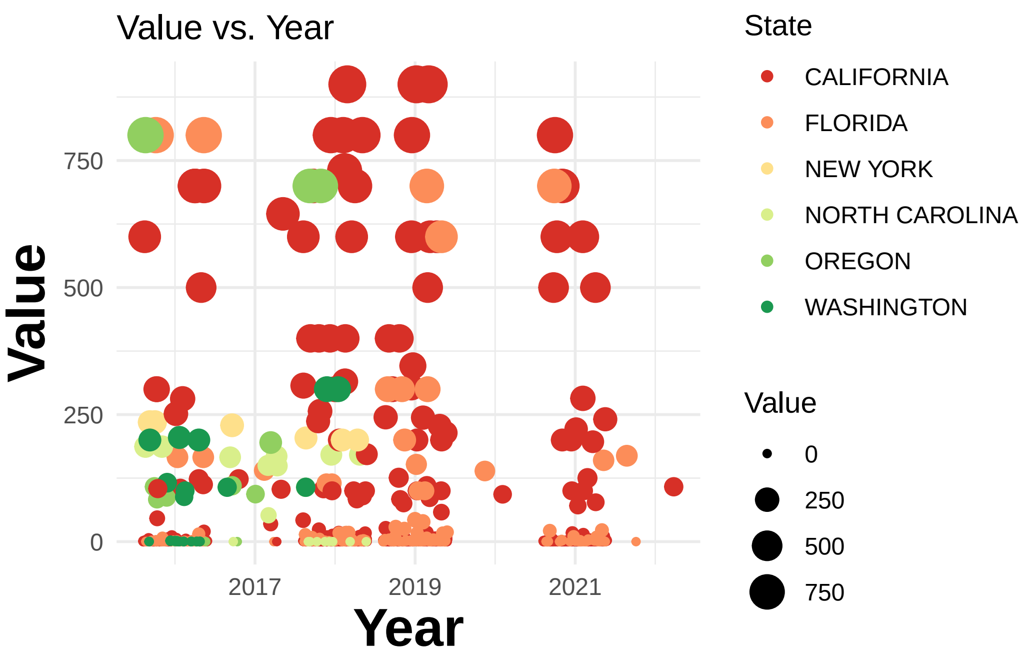


For the top ten chemicals, such as "MALATHION" and "2,4-D", they are mainly associated with the fields of "FIELD CROPS" and "FRUIT & TREE NUTS". For chemicals in the bottom ten frequencies, such as "CHLORPYRIFOS METHYL" and "DIAZINON", their frequencies are lower, but they are also associated with several domains. Some domains (e.g., "FRUIT & TREE NUTS" and "FIELD CROPS") occur in multiple chemicals, while others occur less frequently.

```r
ggplot(subset_stb_survey, aes(x = Chemical, fill = State)) +
  geom_bar() +
  scale_x_discrete(limits = selected_chemicals) +
  labs(title = "Frequency of Chemicals by State") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```r
ggplot(stb_survey) +
 aes(x = Year, y = Value, colour = State, size = Value) +
 geom_jitter() +
 scale_color_brewer(palette = "RdYlGn",
 direction = 1) +
 labs(title = "Value vs. Year") +
 theme_minimal() +
 theme(axis.title.y = element_text(size = 20L,
 face = "bold"), axis.title.x = element_text(size = 20L, face = "bold"))
```

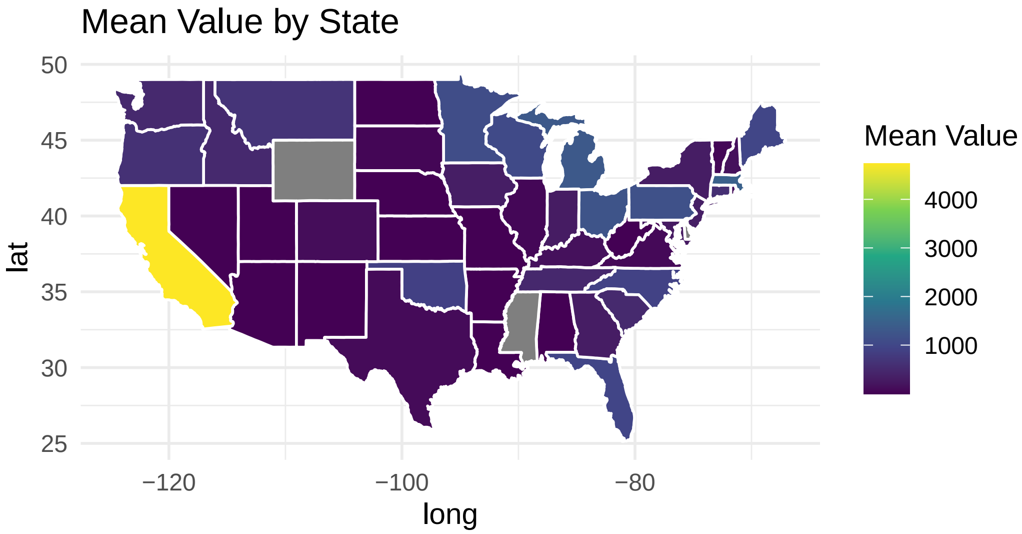


There are significant differences in Value across states. Some states (e.g., Florida and Washington, D.C.) have wider ranges of Value, indicating that the data are more variable in these states. Most states have a median Value in the lower range, but some have a higher median Value.

```
state_value_mean <- sales_by_state_year %>%
  group_by(State) %>%
  summarise(MeanValue = mean(Value, na.rm = TRUE))

capitalize_first <- function(string) {
  paste0(tolower(substr(string, 1, nchar(string))))
}

state_value_mean$State<-sapply(state_value_mean$State, capitalize_first)

library(maps)
```

```
Attaching package: 'maps'

The following object is masked from 'package:purrr':

    map
```

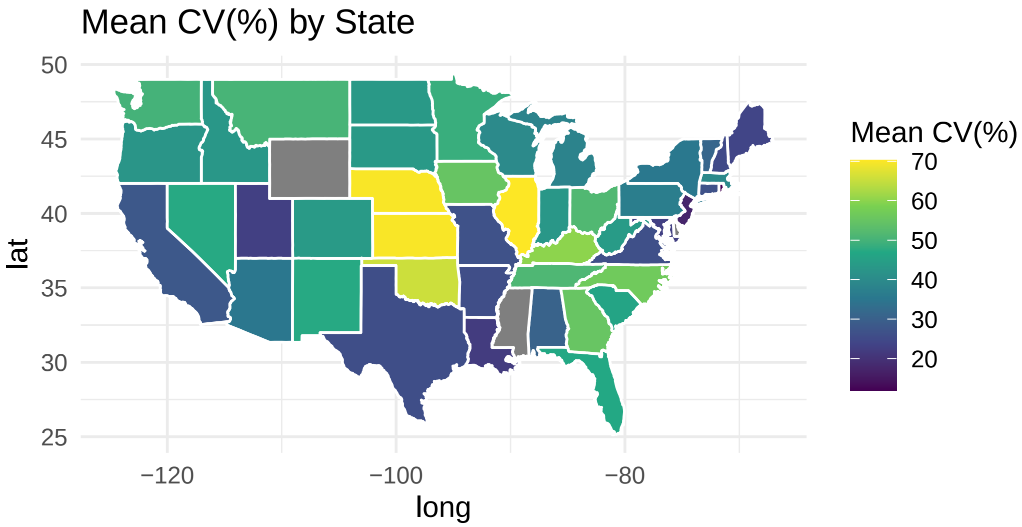
```
usa_map <- map_data("state")

state_value_mean$region <- state_value_mean$State

merged_data <- left_join(usa_map, state_value_mean, by = "region")


ggplot(data = merged_data, aes(x = long, y = lat, group = group, fill = MeanValue)) +
  geom_polygon(color = "white") +
  scale_fill_viridis_c(na.value = "grey50", name = "Mean Value") +
  labs(title = "Mean Value by State") +
  coord_fixed(1.3) +
  theme_minimal()
```



We further explore Value and the specific performance of each state, and with the help of the map we can clearly see the specifics of each state.

```
state_cv_mean <- strwb_census %>%
  group_by(State) %>%
  summarise(MeanCV = mean(`CV (%)`, na.rm = TRUE))
state_mean <- sales_by_state_year %>%
  group_by(State) %>%
  summarise(MeanValue = mean(Value, na.rm = TRUE))

merged_d <- left_join(state_cv_mean, state_mean, by = "State")
```

```
merged_d[is.na(merged_d)] <- 0
merged_d$State<-sapply(merged_d$State, capitalize_first)

merged_d$region <- merged_d$State

merg <- left_join(usa_map, merged_d, by = "region")

ggplot(data = merg, aes(x = long, y = lat, group = group,fill = MeanCV )) +
  geom_polygon(color = "white") +
  scale_fill_viridis_c(na.value = "grey50", name = "Mean CV(%)") +
  labs(title = "Mean CV(%) by State") +
  coord_fixed(1.3) +
  theme_minimal()
```



References

https://quickstats.nass.usda.gov/src/glossary.pdf https://quickstats.nass.usda.gov/param_define