Stawberries: exploratory data analysis

AUTHOR
Yuyang Sun

PUBLISHED
October 11, 2023

Assignment

Using our class discussions and this document as a starting point, produce an EDA report. The report should describe the data itself so that readers understand the data sources used in the report and how you cleaned and organized the data for analysis.

The sections below suggest how the report might be organized. The report should be succinct, communicating the information that you believe will be helpful to someone doing a fuller analysis of the data or using the data for model building. Implementation details should be included in commentary that is included in code.

Sections of the document as it was originally presented in class have been commented so that you can see them in the code.

Data acquisition and assessment

- Data sources
- Assumptions and motivations

Data cleaning and organization

Outline the approach taked to clean and organize the data.

Data cleaning !!

The strawberry data originally contains 26 columns but some columns are not necessary. Thus, we removed some of the columns and kept 10 columns that are essential for our further operations.

We got an idea of which state outputs the most strawberries by counting the strawberries. It turned out that California is the state.

The first column `Program` contains two categorical values: `CENSUS` and `SURVEY`. We filter them out and make them into two datasets: strwb_census and strwb_survey.

Now, we deal with the strwb_census and found two composite columns 'Data Item' and 'Data Category'.

First, we splited the 'Data Item' column into four parts: Fruit, temp1(type of fruit), temp2(market type), temp3(value in currency). Then we divided temp1 into two parts:crop_type and prop_acct.

We focus on temp2, we created a `Fresh Market` column by copying temp2 and remove entries with "MEASURED", "PROCESSING" and "NA" values. Further operations for `Fresh Market` are to remove the "FRESH" and refine the column by removing "FRESH MARKET -".

Next, we created a `Process Market` column which is very similar to `Fresh Market`. The procedure is basically the same.

Then, we removed all the "NA"'s in prop_acct, temp2, temp3 columns and combine temp2 and temp3 into a new column called "Metric" and remove "MEASURED IN". We rename `prop_acct` to "Total".

Then, we rearranged "Metric" before "Domain" Process Market" before "Metric" in strwb_census dataset.

The remain codes are correspond to values where we clean the values and possible footnotes. As there are some values with brackets(e.g. (D)), thus we treat those values as NA.

References

Material about strawberries

WHO says strawberries may not be so safe for you–2017March16

Pesticides + poison gases = cheap, year-round strawberries 2019March20

Multistate Outbreak of Hepatitis A Virus Infections Linked to Fresh Organic Strawberries-2022March5

Strawberry makes list of cancer-fighting foods-2023May31

Technical references

In their handbook "An introduction to data cleaning with R" by Edwin de Jonge and Mark van der Loo, de Jonge and van der Loo go into detail about specific data cleaning isssues and how to handle them in R.

"Problems, Methods, and Challenges in Comprehensive Data Cleansing" by Heiko Müller and Johann-Christoph Freytag is a good companion to the de Jonge and van der Loo handbook, offering additional insights.

Initial questions

- Initial questions about strawberries, the data, and about the work you are undertaking. Write these before you begin working.

What are the differences process market and fresh market? Which state yields the most volume of strawberries?

The data

Describe the source and original condition of the data: organization, problems with the data that needed to be addressed and so on. Cite data sources.

The data set for this assignment has been selected from: USDA_NASS
The data have been stored on NASS here: USDA_NASS_strawb_2023SEP19

Make relevant observations in the document and in your code about data. Add commentary to the code so that anthoer analysts could use or extend your code.

Discusss missing data, inclding how you handled it. Be careful to point out where NA's are being produced during processing and are not data missing in the original data.

Where it is relevant, include information of how you have organized the data for analysis. It might, for example, be helpful to know that there is both agricultural census data and survey data. It might be helpful to discuss data that appears to be redundant between these two sources.

Make sure you include details in your discussion and in your code about other data and information you used in your work. Cite sources and provide detail that would allow another analyst to reproduce your work.

```
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…
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             <chr> "2", "(D)", "(D)", "(D)", "2", "(D)", "(D)", "142", …
$ `CV (%)`          <chr> "(H)", "(D)", "(D)", "(D)", "(H)", "(D)", "(D)", "19…
```

```r
summary(strawberry)
```

```
   Program               Year         Period             State          
 Length:4314        Min.   :2016   Length:4314        Length:4314       
 Class :character   1st Qu.:2016   Class :character   Class :character  
 Mode  :character   Median :2018   Mode  :character   Mode  :character  
                    Mean   :2018                                        
                    3rd Qu.:2019                                        
                    Max.   :2022                                        
  State ANSI         Data Item            Domain          Domain Category   
 Length:4314        Length:4314        Length:4314        Length:4314       
 Class :character   Class :character   Class :character   Class :character  
```

```
Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
   Value              CV (%)
Length:4314        Length:4314
Class :character   Class :character
Mode  :character   Mode  :character
```

```
# Define functions to detect specific patterns
has_colon <- function(column) {
  return(any(str_detect(column, ":")))
}

has_comma_and_dash <- function(column) {
  return(any(str_detect(column, ",")) && any(str_detect(column, "-")))
}

# Apply the functions across all columns
colon_cols <- sapply(strawberry, has_colon)
comma_dash_cols <- sapply(strawberry, has_comma_and_dash)

# Filter and list the composite columns
composite_colnames <- c(names(colon_cols[colon_cols]), names(comma_dash_cols[comma_dash_cols]))

composite_colnames
```

```
[1] NA                "Domain Category" NA                NA
[5] "Data Item"       "Domain Category" NA
```

```
# domain category and data item are composite columns
```

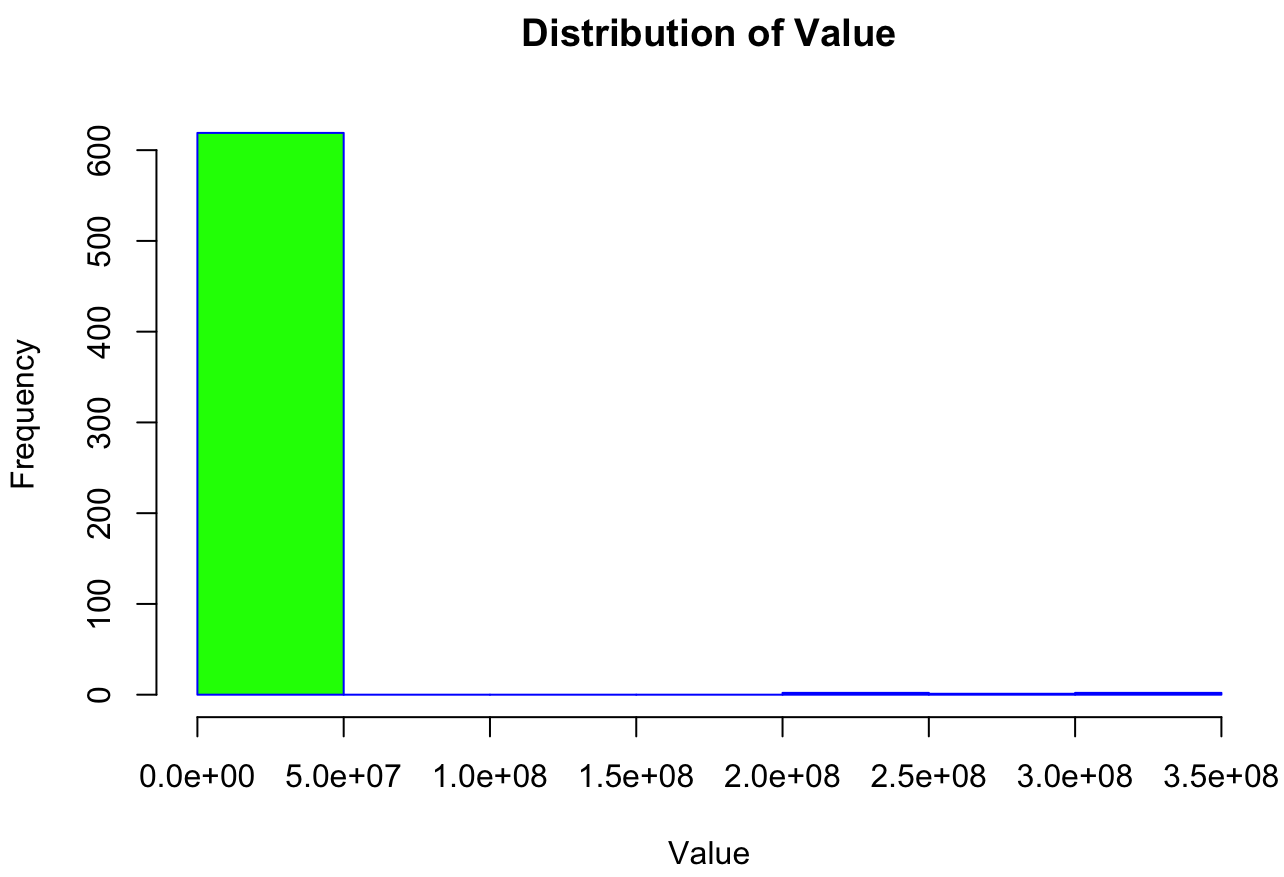
```
# paste(census_col[c(6, 8)])
```

```
# paste(survey_col[c(6,7,8)])
```

EDA

Once the data has been cleaned and organized, you must conduct your own EDA. Be sure to include a discussion of your analysis of the chemical information, including citations for data and other information you have used. Visualizations should play a key role in your analysis. Plots should be labeled and captioned.

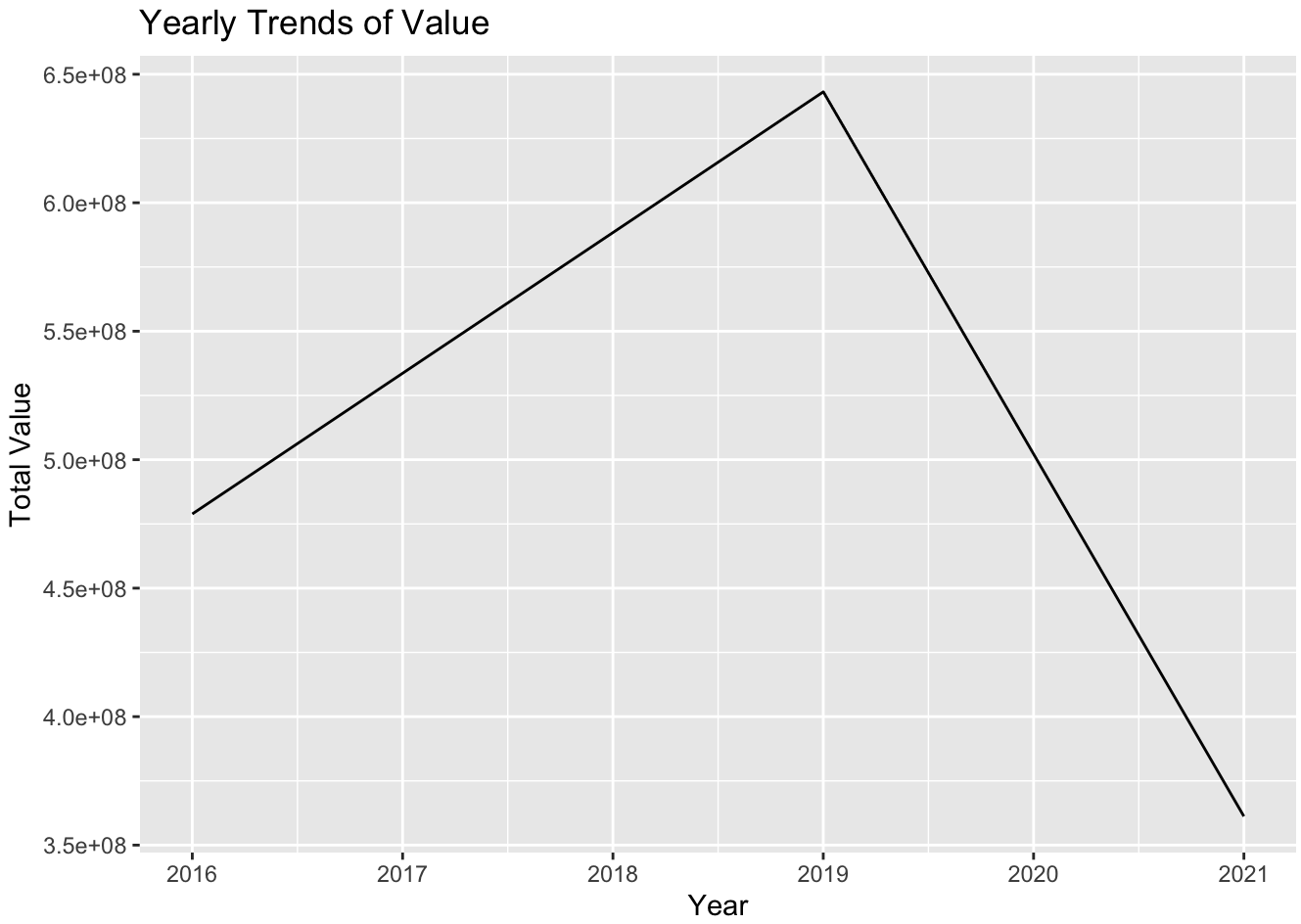
```
hist(c, main="Distribution of Value", xlab="Value", border="blue", col="green")
```




```
## yield trend
# Aggregate data by Year
yearly_totals <- aggregate(c ~ Year, data=strwb_census, sum, na.rm = TRUE)

# Plot
ggplot(yearly_totals, aes(x=as.numeric(Year), y=c)) +
```

```
  geom_line(group=1) +
  labs(title="Yearly Trends of Value", x="Year", y="Total Value")
```



Based on the graph, we notice that the yeild of strawberries was not always constant. Starting from 2016, there was a consistent increase in strawberry production, reaching its peak in 2019. This growth phase suggests favorable conditions for strawberry farming. After the peak in 2019, there was a sharp decline in production, reaching its lowest point in 2021. This decline could be attributed to several potential factors:

Climatic Events: Unfavorable weather conditions, such as droughts, excessive rain, or unseasonal temperatures, can adversely affect strawberry yields.

Pests or Diseases: An outbreak of pests or diseases can drastically reduce yields. Economic or Social Factors: Changes in demand, trade policies, or shifts in consumer preferences can influence farming decisions. Pesitides that might contain toxic chemicals that reduce yields.

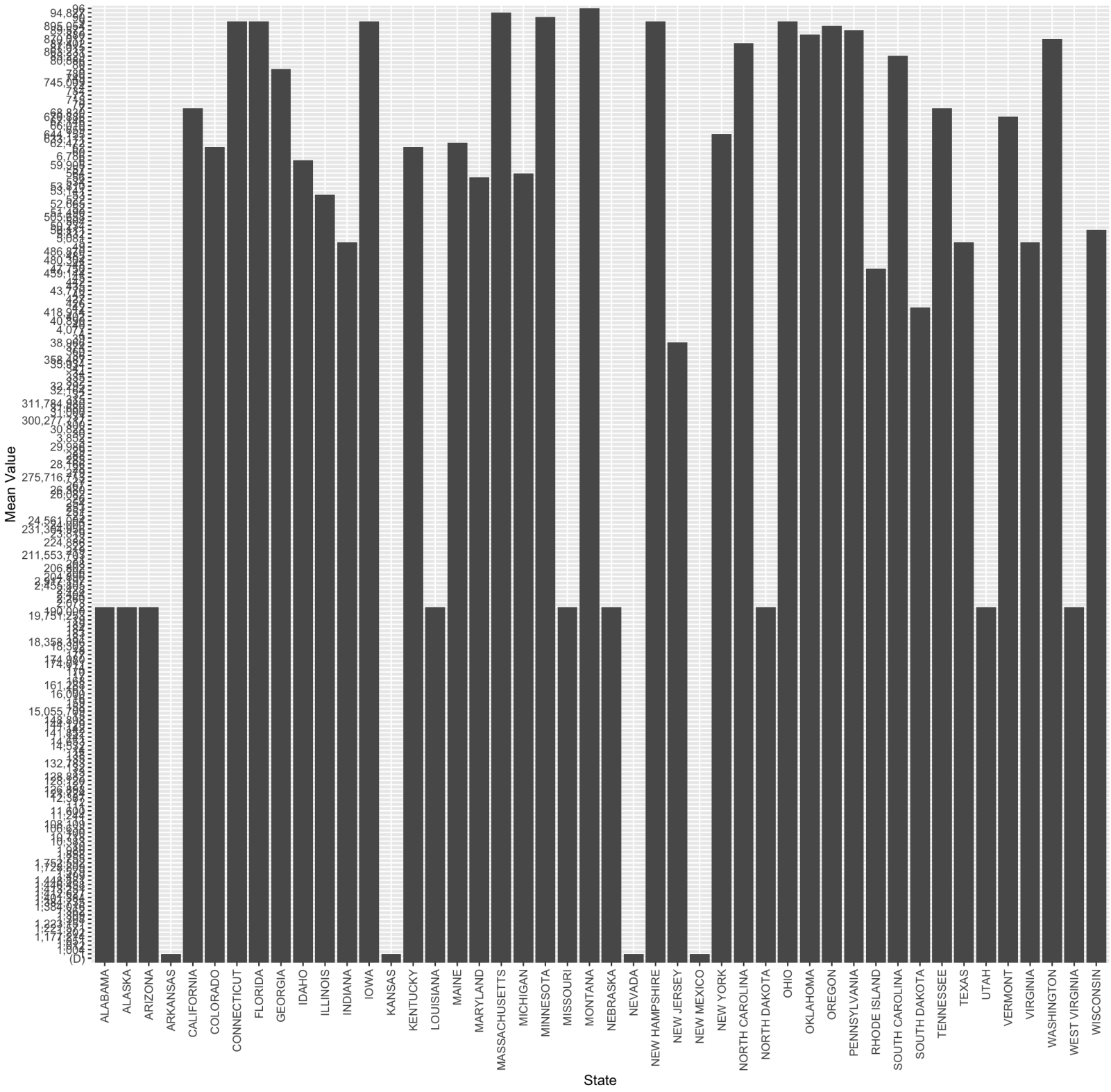
```r
# correlation
correlation <- cor(strwb_census$Year, c, use="complete.obs")

print(correlation)
```

```
[1] -0.002375557
```

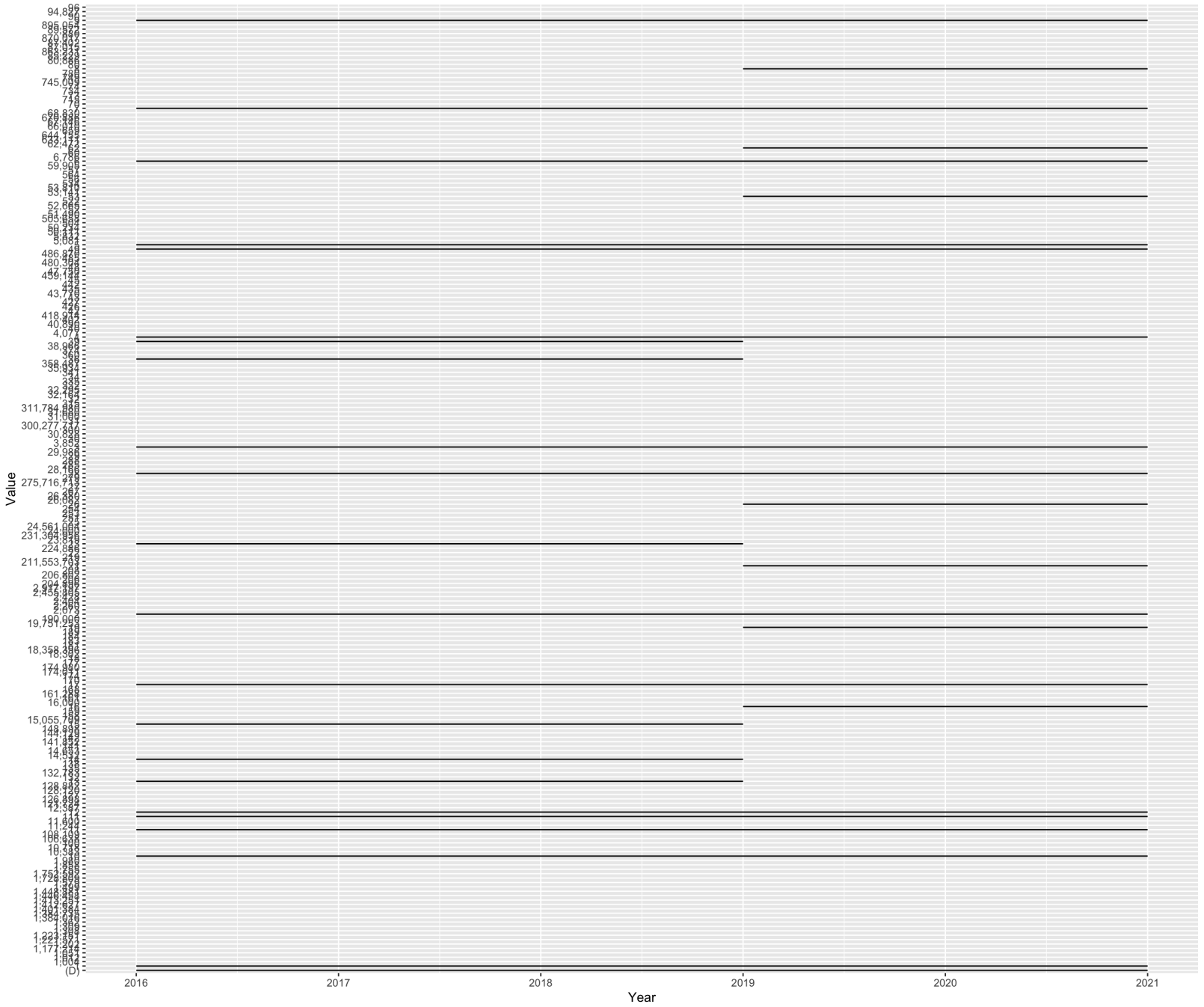
We notice that the year do not seem to have any correlations with the values.

```r
ggplot(strwb_census, aes(x=State, y=Value)) +
    stat_summary(fun=mean, geom="bar") +
    labs(title="Mean Strawberry Value by State", x="State", y="Mean Value") +
    theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```


Mean Strawberry Value by State
Mean Value
State
ALABAMA
ALASKA
ARIZONA
ARKANSAS
CALIFORNIA
COLORADO
CONNECTICUT
FLORIDA
GEORGIA
IDAHO
ILLINOIS
INDIANA
IOWA
KANSAS
KENTUCKY
LOUISIANA
MAINE
MARYLAND
MASSACHUSETTS
MICHIGAN
MINNESOTA
MISSOURI
MONTANA
NEBRASKA
NEVADA
NEW HAMPSHIRE
NEW JERSEY
NEW MEXICO
NEW YORK
NORTH CAROLINA
NORTH DAKOTA
OHIO
OKLAHOMA
OREGON
PENNSYLVANIA
RHODE ISLAND
SOUTH CAROLINA
SOUTH DAKOTA
TENNESSEE
TEXAS
UTAH
VERMONT
VIRGINIA
WASHINGTON
WEST VIRGINIA
WISCONSIN

```r
ggplot(strwb_census, aes(x=Year, y=Value)) +
    geom_line() +
    stat_smooth(method="loess") +
    labs(title="Time Series Analysis of Strawberry Value", x="Year", y="Value")
```

```
`geom_smooth()` using formula = 'y ~ x'
```


Time Series Analysis of Strawberry Value
Value
Year
2016
2017
2018
2019
2020
2021

These references have been left in the document to help while you are writing. Cite those you use and drop the rest from the final document.

NASS help

Quick Stats Glossary

Quick Stats Column Definitions

stats by subject

for EPA number lookup epa numbers

Active Pesticide Product Registration Informational Listing

pc number input pesticide chemical search

toxic chemical dashboard

ACToR – Aggregated Computational Toxicology Resource

comptox dashboard

pubChem

The EPA PC (Pesticide Chemical) Code is a unique chemical code number assigned by the EPA to a particular pesticide active ingredient, inert ingredient or mixture of active ingredients.

Investigating toxic pesticides

start here with chem PC code

step 2 to get label (with warnings) for products using the chemical

International Chemical safety cards

Pesticide Product and Label System

Search by Chemical

CompTox Chemicals Dashboard

Active Pesticide Product Registration Informational Listing

OSHA chemical database

Pesticide Ingredients

NPIC Product Research Online (NPRO)

Databases for Chemical Information

Pesticide Active Ingredients

TSCA Chemical Substance Inventory

glyphosate