Stawberries EDA: Final draft

MA615

2023-10-23

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.

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.

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~
```

```
[1] "No Columns to drop"
```

```
[1] TRUE
```

```
# echo: True

## This will be done in stages --

#####################################################
## split `Data Item` into "Fruit", "temp1","temp2","temp3"
```

```
## then test the columns created for numer of distinct values
## split the columns until you have columns of
## subjects, properties, values, and metrics (where metrics
## are the units defined for the values)

## In this case, the subject is State/Strawberries --
## strawberries grown reported by state.

## When using separate_wider_delim() when you don't know the
## number of columns the function will return,
## use the "too_many" and "too_few" parameters to set up
## the function.  Generally, setting both parameters
## to "error" will produce helpful error messages.

  strwb_census <- strwb_census |>
  separate_wider_delim(  cols = 'Data Item',
                         delim = ",",
                         names = c("Fruit",
                                 "temp1",
                                 "temp2",
                                 "temp3"),
                         too_many = "error",
                         too_few = "align_start"
                       )

## Test the columns for the number of distinct values.
## for example:
##
 a <- strwb_census |> distinct(Fruit)
## The Fruit column only has one value: STRAWBERRIES the
## subject under investigation.
##
## Remember - the value in single-value columns
## are often needed for Labels on tables and plots.
##
## Testing the temp1 column guides the next step.
# a <- strwb_census |> distinct(temp1)
## The "temp1" column has 4 distinct values
##
##    " ORGANIC - OPERATIONS WITH SALES"
##    " ORGANIC - PRODUCTION"
##    " ORGANIC - SALES"
```

```
##    " ORGANIC"
##
##  (Note the leading space in each string --
##       which is fixed below.)
##
##  You can see that this column needs to be split between
##  "organic" and the properties "OPERATIONS WITH SALES",
##  "PRODUCTION" and "SALES",
##    using " - " as the column delimiter.
##
##  The column "prop_acct" contains the properties,
##   which are are accounting metrics related to
##   strawberry growing operations.


###########################################
## split temp1 into crop_type, Prop_acct

strwb_census <- strwb_census |>
  separate_wider_delim(  cols = temp1,
                         delim = " - ",
                         names = c("crop_type",
                                   "prop_acct"),
                         too_many = "error",
                         too_few = "align_start"
                       )

## Once again, test the columns to plan your next step.
##
# a <- strwb_census |> distinct(crop_type)
## Column "crop_type' has single value  "organic"

# a <- strwb_census |> distinct(prop_acct)

##
## The stringss in the "prop_acct" column are row labels
## for values reported in the "Values" column.

##    "OPERATIONS WITH SALES"
##    "PRODUCTION"
##    "SALES"
##    "NA"
```

```r
## Note that the NA is in a row where the value
## is labeled in another column.
##

#############################################
## trim the strings
## you can see which columns contain string values that need
## to have leading or trailing spaces that need to be trimmed.


# glimpse(strwb_census)

strwb_census$crop_type <- str_trim(strwb_census$crop_type, side = "both")

strwb_census$temp2 <- str_trim(strwb_census$temp2, side = "both")

strwb_census$temp3 <- str_trim(strwb_census$temp3, side = "both")



#############################################
## split temp2 into market_type, measure

##
## The temp2 column requires a different logic.
##

## start by looking at the unique entries in the temp2 column.

# a <- strwb_census |> distinct(temp2)
#
# temp2
# 1  NA
# 2 " MEASURED IN CWT"
# 3 " MEASURED IN $"
# 4 " FRESH MARKET - OPERATIONS WITH SALES"
# 5 " FRESH MARKET - SALES"
# 6 " PROCESSING - OPERATIONS WITH SALES"
# 7 " PROCESSING - SALES"

## temp2 contains data for three separate columns
```

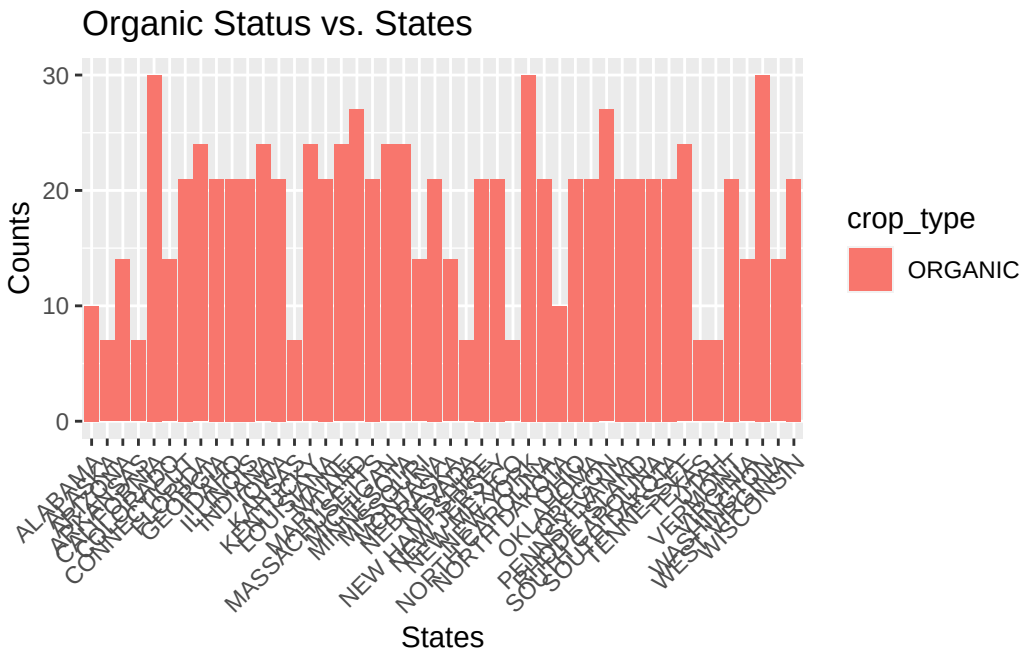
```
##
##   All Strawberries  (is this a Total?)
##   Fresh Market
##   Processing
##
##  To understand these labels see
##     "Strawberries: An Economic Assessment of the Feasibility
##      of Providing Multiple-Peril Crop Insurance",
##        prepared by Economic Research Service, USDA
##             for the Federal Crop Insurance Corporation
##                  October 31, 1994
##
```

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.

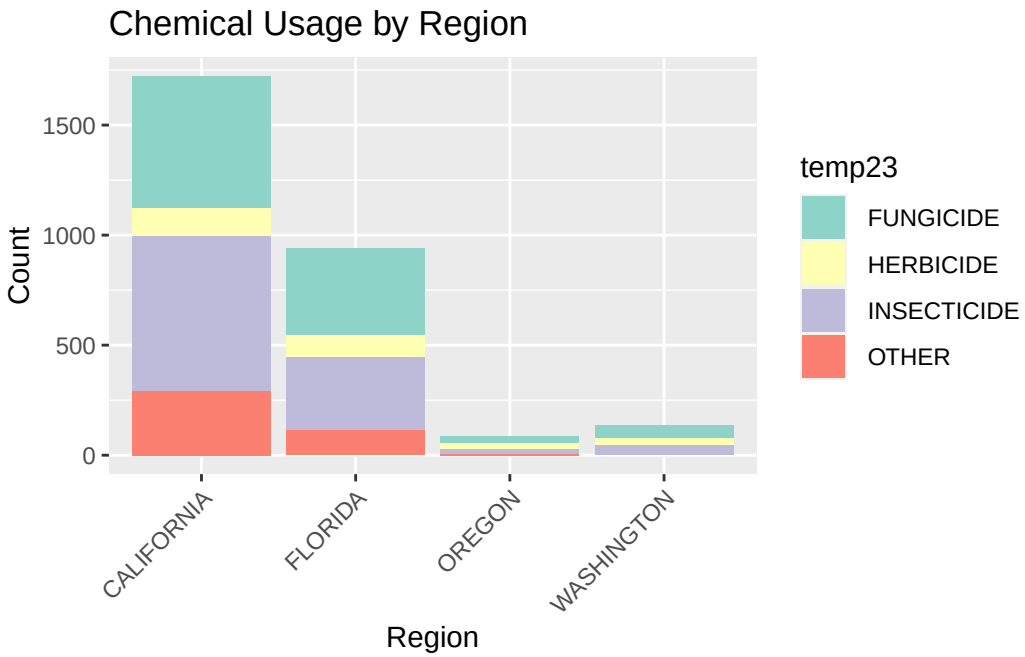
EDA

```
# EDA
ggplot(strwb_census, aes(x = State, fill = crop_type)) +
  geom_bar(position = position_dodge(width = 5)) +
  labs(title = "Organic Status vs. States",
       x = "States",
       y = "Counts") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Following Haviland work from the class after seperating the census program and survey program, the crop planted in organic way and the crop planted using chemicals are also devided at the same time. Up here is the histogram of the amount of strawberry sales that came from organic planting. There are states like Alaska which cannot implement organic farming for long period due to weather issue. But overall a lot of states have organic crops for sale.

```
ggplot(strwb_survey_chem, aes(x = State, fill = temp23)) +
  geom_bar(position = "stack") +
  labs(title = "Chemical Usage by Region",
       x = "Region",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```



On the other aspect, lots of farming cannot avoid using chemicals to kill those things which would be harmful to the crops. The chemicals used includes fungicide, herbicide, insecticide, etc. According to the colored histogram above, California seems to be the states using the most chemicals. California is also the state with most organic crops too based on the previous plot. Overall the fungicide and insecticide are used for the most than the other chemicals.

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