eda report

AUTHOR
Huajin Xin

PUBLISHED
October 23, 2023

Data Source

The data set for this assignment has been selected from: [USDA_NASS](https://quickstats.nass.usda.gov)
 The data have been stored on NASS here: [USDA_NASS_strawb_2023SEP19](https://quickstats.nass.usda.gov/results/45FBC825-B104-38E2-9802-839F5F3C7036)

Data Cleaning

Here is the view of raw data:

```
Rows: 4,314
Columns: 11
$ ...1                <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 1…
$ 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…
```

Split the data by the column Program: Census and Survey and clean them seperately. And we first check for Census dataframe: splite the column 'Data Item' into reasonable columns, and clean the value of column 'Value' and 'CV(%)': For cleaning the number value, we delete all the commas and make every string value like "(D)" into NA.

```
Rows: 864
Columns: 9
$ ...1          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17…
$ Year          <dbl> 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 2021, 202…
$ State         <chr> "ALASKA", "ALASKA", "ALASKA", "ALASKA", "ALASKA", "ALASKA…
$ `State ANSI`  <chr> "02", "02", "02", "02", "02", "02", "02", "06", "06", "06…
$ Value         <dbl> 2, NA, NA, NA, 2, NA, NA, 142, 1413251, 3117849988, 141262…
$ `CV (%)`      <dbl> NA, NA, NA, NA, NA, NA, NA, 19.2, 51.6, 46.0, 51.7, 20.4,…
$ Type          <chr> "", "", "", "", " FRESH MARKET", " FRESH MARKET", " FRESH…
$ Condition     <chr> "SALES", "PRODUCTION", "SALES", "SALES", "SALES", "SALES"…
$ Metric        <chr> "", " CWT", " $", " CWT", "", " $", " CWT", "", " CWT", "…
```

then we check for Survey dataframe, clean it with the same way.

```
Rows: 3,450
Columns: 12
$ ...1                <dbl> 865, 866, 867, 868, 869, 870, 871, 872, 873, 874, 87…
$ Program             <chr> "SURVEY", "SURVEY", "SURVEY", "SURVEY", "SURVEY", "S…
$ Year                <dbl> 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022, 2022…
$ Period              <chr> "MARKETING YEAR", "MARKETING YEAR", "MARKETING YEAR"…
$ State               <chr> "CALIFORNIA", "CALIFORNIA", "CALIFORNIA", "FLORIDA",…
$ `State ANSI`        <chr> "06", "06", "06", "12", "12", "12", NA, NA, NA, "06"…
$ Domain              <chr> "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL", "TOTAL"…
$ `Domain Category`   <chr> "NOT SPECIFIED", "NOT SPECIFIED", "NOT SPECIFIED", "…
$ Value               <dbl> 1.08000e+02, NA, NA, 1.69000e+02, NA, NA, 0.00000e+0…
$ Product             <chr> "", " FRESH MARKET", " PROCESSING", "", " FRESH MARK…
$ Type                <chr> "PRICE RECEIVED", "PRICE RECEIVED", "PRICE RECEIVED"…
$ Metric              <chr> " $ / CWT", " $ / CWT", " $ / CWT", " $ / CWT", " $ …
```

EDA

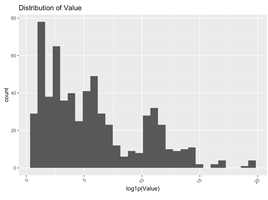
We first do some EDA for CENSUS part:

1.distribution for the log(Value).

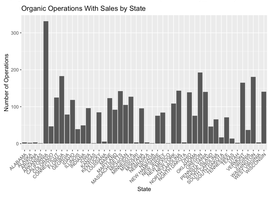
Because there are lots of NA values, we ignore them. And the distribution is large so we use log.

```
library(ggplot2)
data_to_plot <- strawberry_census[!is.na(strawberry_census$Value), ]

ggplot(data_to_plot, aes(x = log1p(Value))) +
  geom_histogram() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of Value")
```



2.Organic Operations With Sales by State (strawberry_census)



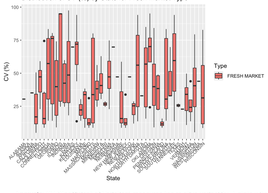
- **Analysis**: This bar chart would show the number of organic operations with sales for each state. States with taller bars have a higher number of operations. Observing which states have the highest and lowest counts can give insights into where organic strawberry farming is most prevalent and where it might be emerging or less common. The highest is California which means California has most prevalent strawberry farming. And for states like Alaska, Kansas, and Nevada definitely less common for strawberry farming.

3.Distribution of CV (%) by State (strawberry_census) for those with Fresh Market

```
library(ggplot2)

# Filter data for rows with Type = "Fresh Market"
data_to_plot2 <- strawberry_census[strawberry_census$Type == " FRESH MARKET", ]

ggplot(data_to_plot2, aes(x = State, y = `CV (%)`, fill = Type)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of CV (%) by State for Fresh Market Type")
```



- **Analysis**: CV (Coefficient of Variation) measures the relative variability. A state with a higher CV would have a higher relative variability in its data. If the CV is too high, it might indicate inconsistencies or potential issues in the data collection process, like Illinois. Conversely, a very low CV across many states might suggest that the data is too uniform and could be worth verifying for accuracy, like Rhode Island.

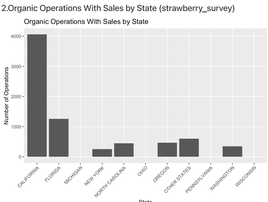
The there are some EDA for Survey Part:

1.distribution for Type of product condition.

```
ggplot(strawberry_survey, aes(x = State, fill = Type)) +
  geom_bar(position = "dodge") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Distribution of type of product condition by states", y = "Frequency")
```

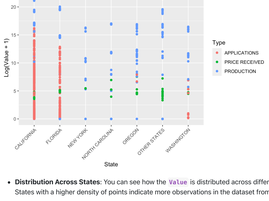
As you can see in the plot that most product in California and Florida are going for application but much less data recorded for price received.

2.Organic Operations With Sales by State (strawberry_survey)



3. Scatter Plot of Log Value by State with variation of types

```
data_to_plot3 <- strawberry_survey[!is.na(strawberry_survey$Value), ]
ggplot(data_to_plot3, aes(x=State, y=log1p(Value), color=Type)) +
    geom_point() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
    labs(title = "Scatter Plot of Log Value by State, Colored by Type", y = "Log(Value
```



- **Distribution Across States:** You can see how the `Value` is distributed across different states. States with a higher density of points indicate more observations in the dataset from that state.
- **Variation by Type**: The different colors allow you to see if certain types have consistently higher or lower values across states.
- **Outliers**: Any points that lie far from the general cluster of points for a state might indicate outliers or unique observations.
- **State Comparison**: You can compare states to see which ones have higher or lower values on average and how much variability there is within each state.