Strawberry EDA Report

Wuge Li

2023-10-16

Introduction:

The data set I used for this report is strawberry data.

The column names normally represent what this data will show you. So, first, Let's take a look at the column names of strawberry raw data

```
strawberry <- read.csv("strawberry.csv", header = T)
colnames(strawberry)
```

```
##  [1] "Program"          "Year"             "Period"           "Week.Ending"
##  [5] "Geo.Level"        "State"            "State.ANSI"       "Ag.District"
##  [9] "Ag.District.Code" "County"           "County.ANSI"      "Zip.Code"
## [13] "Region"           "watershed_code"   "Watershed"        "Commodity"
## [17] "Data.Item"        "Domain"           "Domain.Category"  "Value"
## [21] "CV...."
```

From the column name of the data we know, the information about Program (census or survey), year, period, state(origin of the strawberries) and other information(whether or not use chemical? Which chemical used when the strawberry grow up)

After watching the data of strawberry. I am curious about what influence the quality of strawberry? Year? Origin? Chemicals? Also, the relationship between the origin of strawberry produce and the market the strawberry sold. And whether the strawberry fresh of not.

Data cleaning and tidying:

Look at origin of the strawberry

```
state_all <- strawberry |>
  group_by(State) |>
  count()
state_all
```

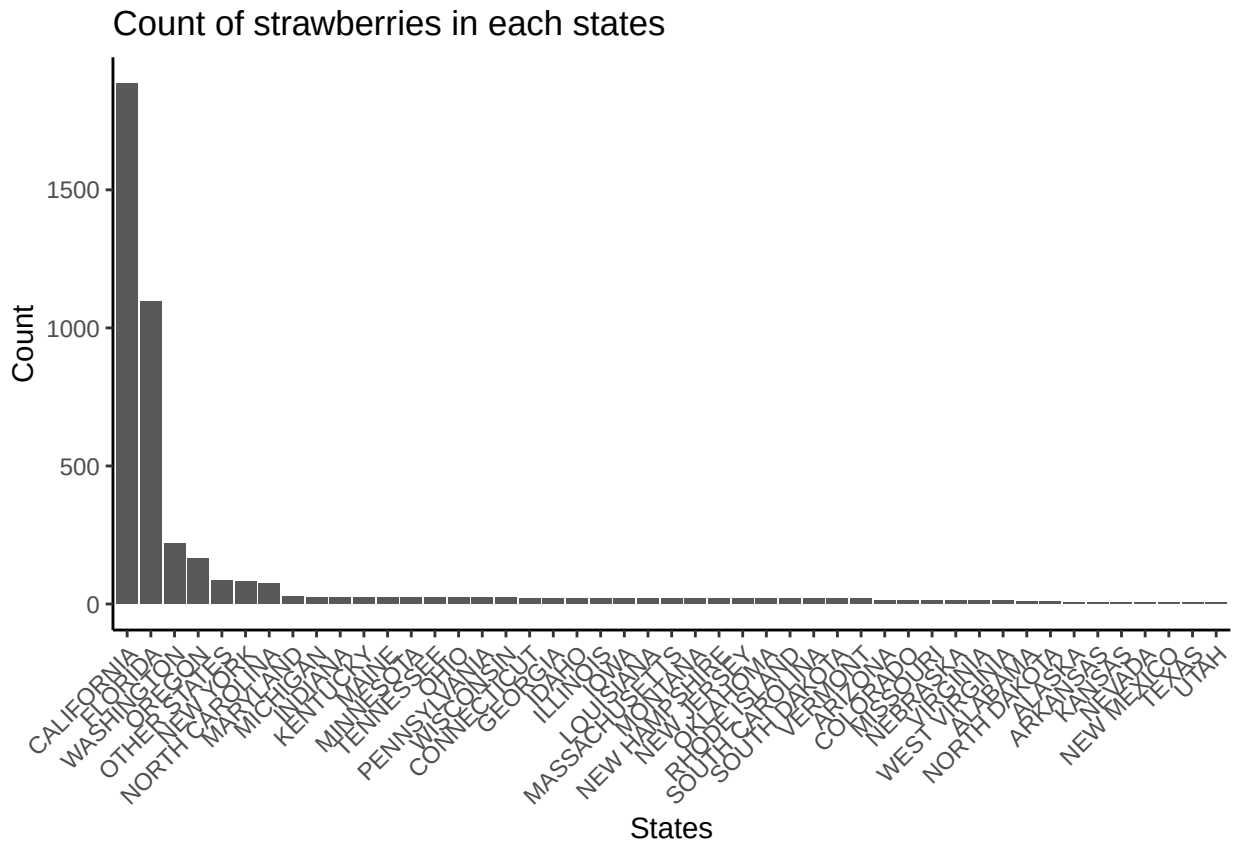
```
## # A tibble: 47 x 2
## # Groups:   State [47]
##    State           n
##    <chr>       <int>
##  1 ALABAMA        10
##  2 ALASKA          7
##  3 ARIZONA        14
##  4 ARKANSAS        7
##  5 CALIFORNIA   1886
```

```
##  6 COLORADO       14
##  7 CONNECTICUT    21
##  8 FLORIDA      1096
##  9 GEORGIA        21
## 10 IDAHO          21
## # i 37 more rows
```

```
state_max <- state_all$State[which(state_all$n ==  max(state_all$n))]
state_max
```

```
## [1] "CALIFORNIA"
```

```
ggplot(data = state_all, mapping = aes(reorder(State, desc(n)), n)) +
  geom_bar(stat = 'identity')+
  labs(title = "Count of strawberries in each states",
       x = "States",
       y = "Count") +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```



The graph above shows that the count of strawberries in each states by decreasing order

From the chart and graph above, we know that California contributes the most strawberry

We know that California contributes the most strawberry, then we concentrate on strawberry which origin is California

```
calif_census <- strawberry |> filter((State=="CALIFORNIA") & (Program=="CENSUS"))
calif_survey <- strawberry |> filter((State=="CALIFORNIA") & (Program=="SURVEY"))
nrow(calif_census)
```

```
## [1] 30
```

```
nrow(calif_survey)
```

```
## [1] 1856
```

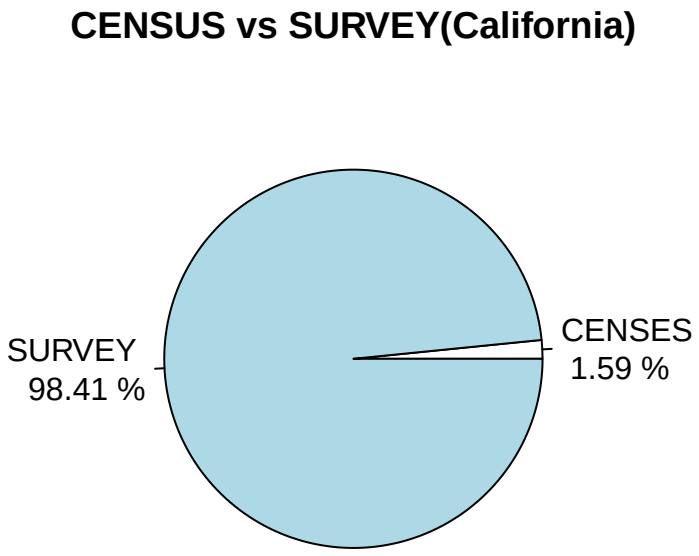
```
calif_cen_per <- nrow(calif_census)/(nrow(calif_survey)+nrow(calif_census))
calif_svy_per <- nrow(calif_survey)/(nrow(calif_survey)+nrow(calif_census))
calif_cen_per
```

```
## [1] 0.01590668
```

```
calif_svy_per
```

```
## [1] 0.9840933
```

```
slices <- c(calif_cen_per, calif_svy_per)
labels <- c("CENSES", "SURVEY")
calif_percentages <- round(c(calif_cen_per * 100, calif_svy_per * 100), digits = 2)
calif_labels_with_percentages <- paste(labels, "\n", calif_percentages, "%")
pie(slices, labels = calif_labels_with_percentages, main = "CENSUS vs SURVEY(California)")
```



From above, we know that there are 30 programs of 1886 are CENSUS, and 1856 of 1886 are SURVEY. CENSUS Program takes 1.59% of whole Programs and SURVEY Program takes 98.41% of whole Programs

Then I am curious about the number of CENSUS Programs and SURVEY Programs.

```
strwb_census <- strawberry |> filter(Program == "CENSUS")
strwb_survey <- strawberry |> filter(Program == "SURVEY")

nrow(strwb_census)
```

```
## [1] 864
```

```
nrow(strwb_survey)
```

```
## [1] 3450
```

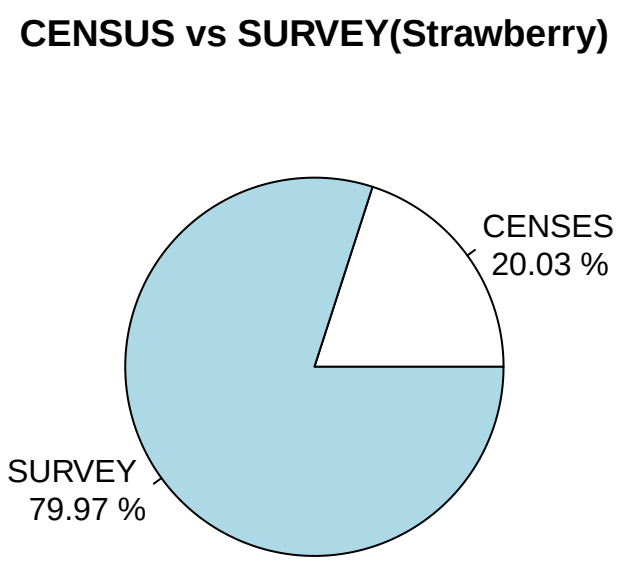
```
strwb_cen_per <- nrow(strwb_census)/(nrow(strwb_survey)+nrow(strwb_census))
strwb_svy_per <- nrow(strwb_survey)/(nrow(strwb_survey)+nrow(strwb_census))
strwb_cen_per
```

```
## [1] 0.2002782
```

```
strwb_svy_per
```

```
## [1] 0.7997218
```

```
slices <- c(strwb_cen_per, strwb_svy_per)
labels <- c("CENSES", "SURVEY")
strwb_percentages <- round(c(strwb_cen_per * 100, strwb_svy_per * 100), digits = 2)
strwb_labels_with_percentages <- paste(labels, "\n", strwb_percentages, "%")
pie(slices, labels = strwb_labels_with_percentages, main = "CENSUS vs SURVEY(Strawberry)")
```



However, when we looked at the programs of all strawberries, we found that CENSUS Program takes 20.03% and SURVEY Programs take 79.97% of the whole programs.

Conclusion:

Based on the information and tidy data above, we can know that California and Florida contributes much more strawberries than other states. SURVEY Programs contributes nearly 80% for the whole strawberries and contributes nearly 98% for California strawberries.