Strawberry Data Analysis

Donghao Xue

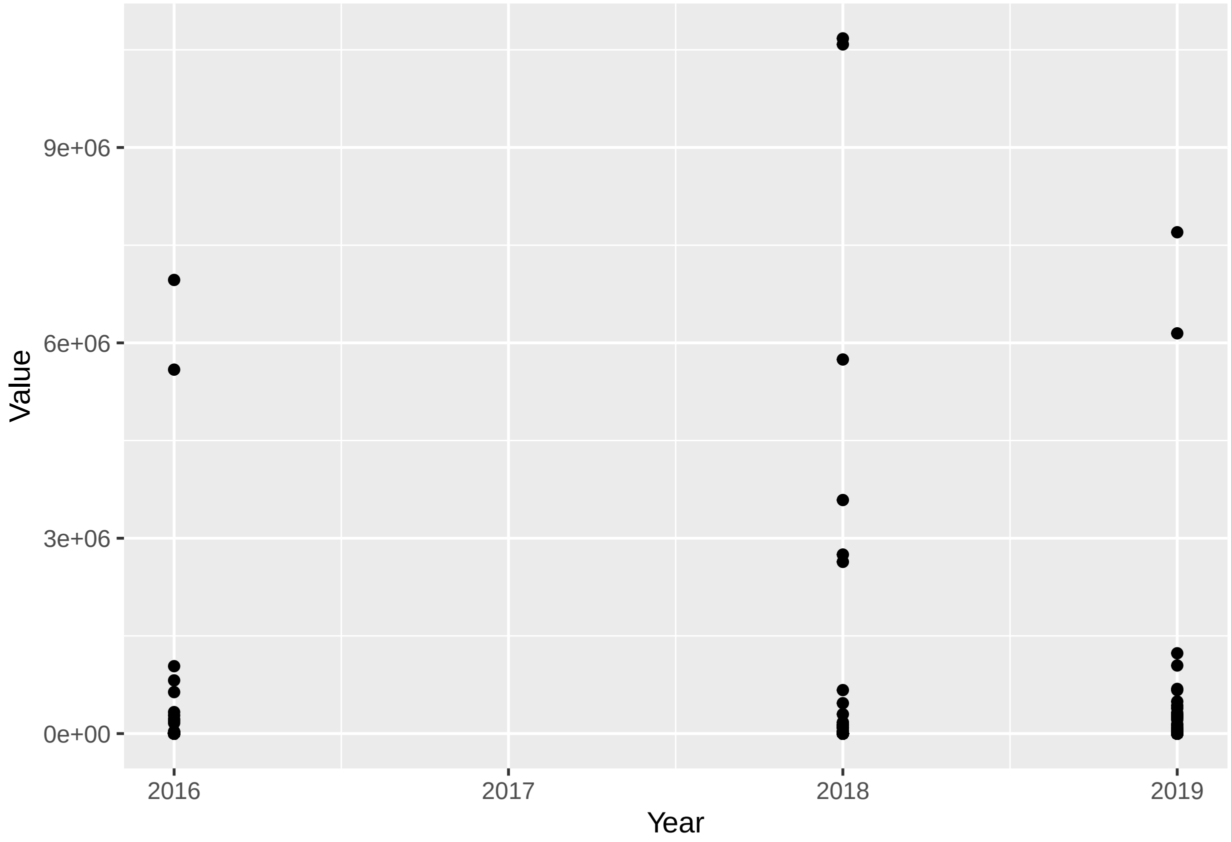
10/21/2020

Introduction

The data contains information about one type of red berry which is strawberry. I plan to discuss the distribution of the column of value and its relationship with other variables.

Analysis

```r
# organize data
sberry = read.csv("sberry.csv")

aa = str_extract_all(sberry$Value, "[0-9]|\\.")
bb = c()

for(i in 1:2271){
  bb[i]= as.numeric(paste(unlist(aa[i]), collapse = ""))

}
sberry %<>% mutate(Value = bb)

#plot of values for each year

ggplot(sberry, aes(x = Year, y = Value)) + geom_point()
```



We can see that the highest value occurs in 2018. Most values are clustered around the base line.

```
mean (sberry$Value[which(sberry$Value>0)])
```

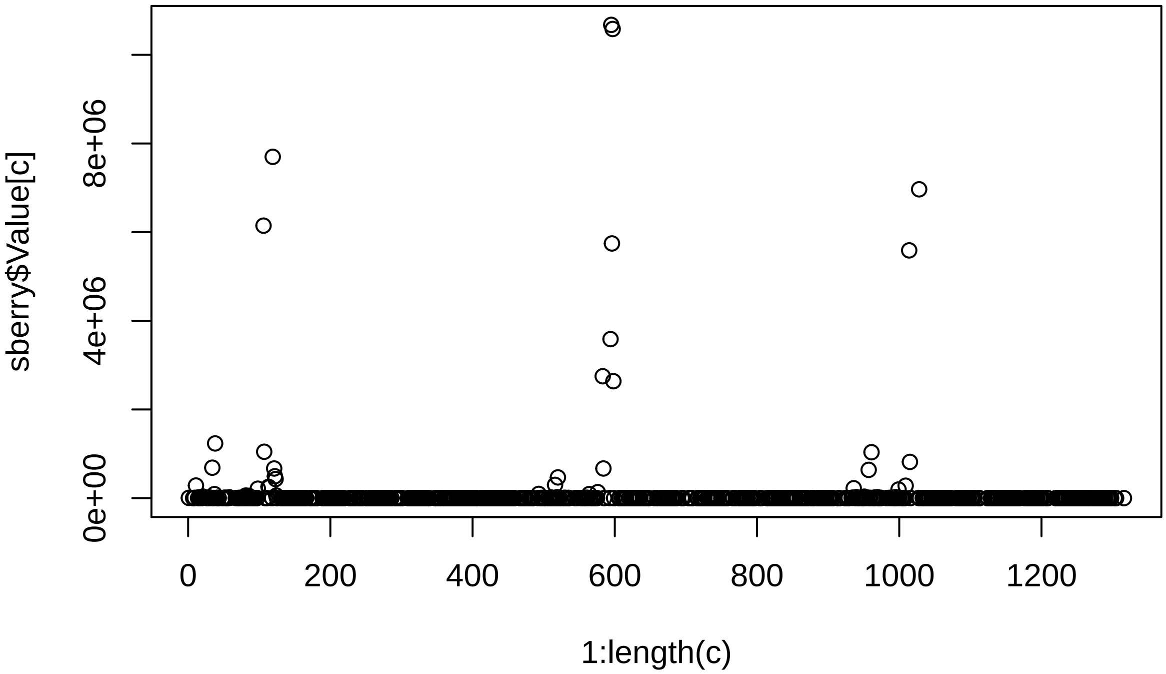
```
## [1] 81059.66
```

```
median(sberry$Value[which(sberry$Value>0)])
```

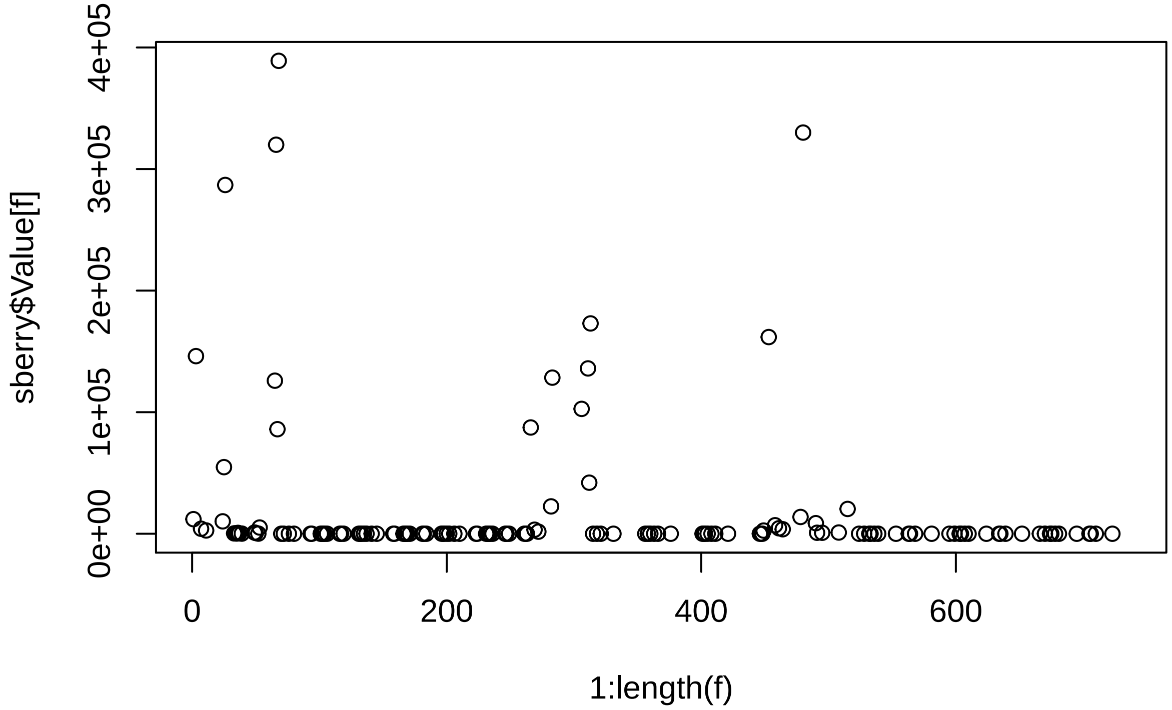
```
## [1] 1.5
```

The average value is about 81000. The median is 1.5, so in normal situation, the value should be around 1.5.

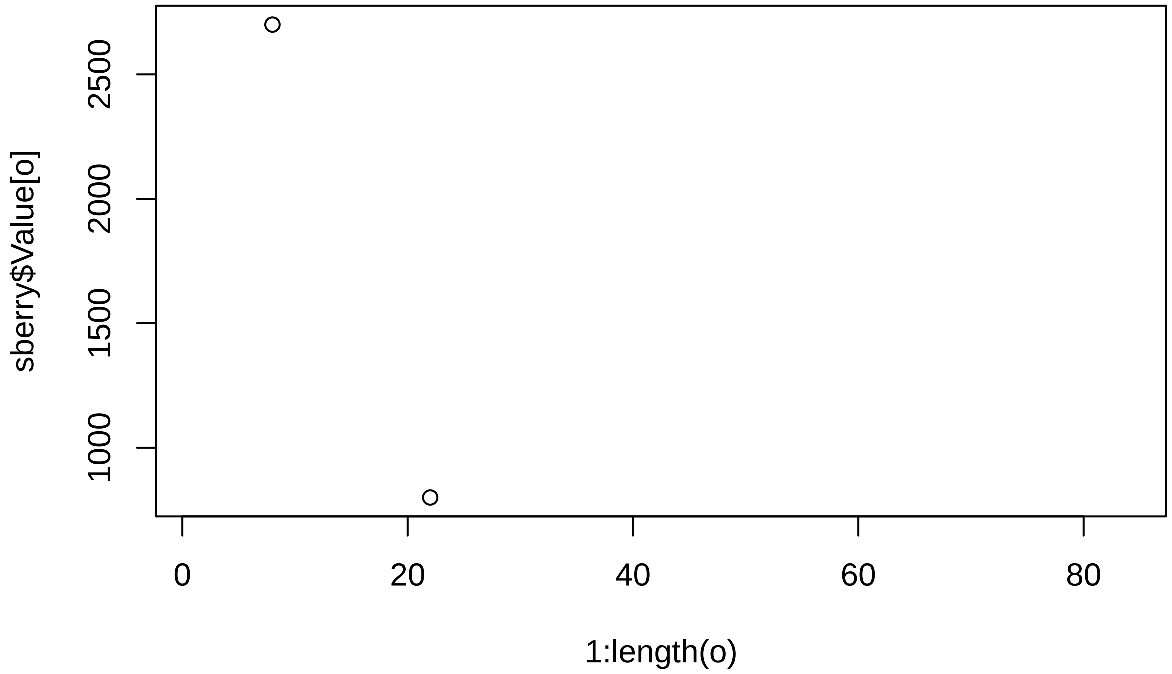
```
#The values of strawberry in different states.

#California
c = which(sberry$State == "CALIFORNIA")
plot(1:length(c), sberry$Value[c])
```



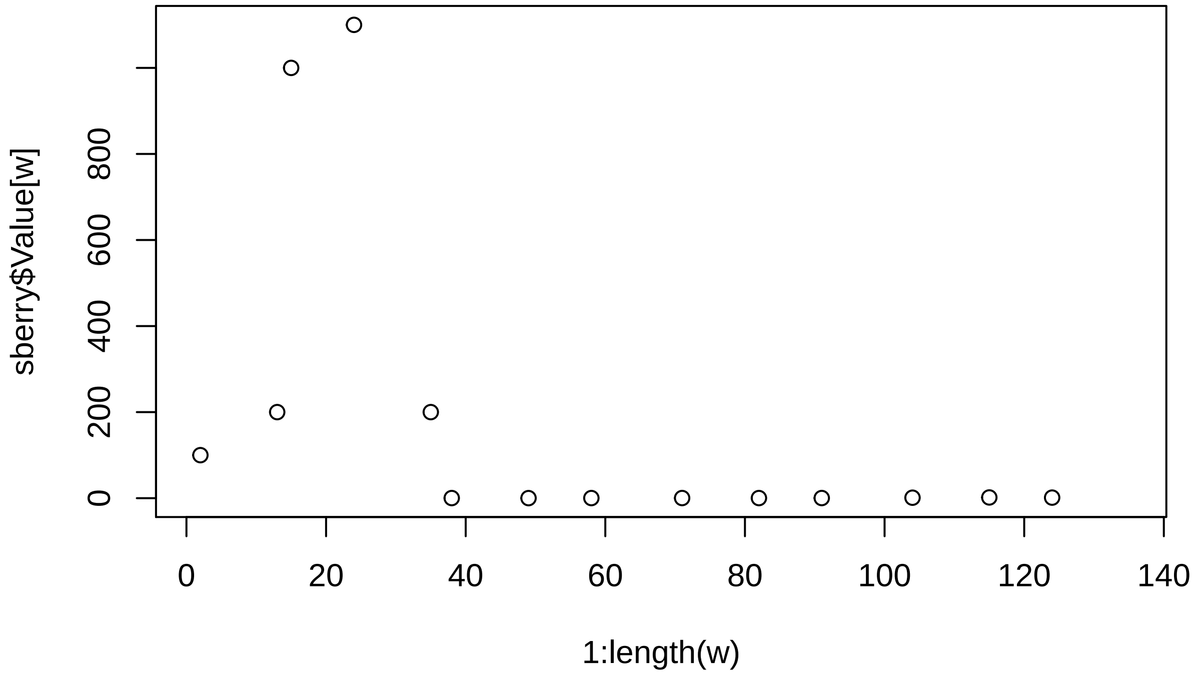
```
#Florida
f = which(sberry$State == "FLORIDA")
plot(1:length(f), sberry$Value[f])
```



```
#Oregon
o = which(sberry$State == "OREGON")
plot(1:length(o), sberry$Value[o])
```



```r
#Washington
w = which(sberry$State == "WASHINGTON")
plot(1:length(w), sberry$Value[w])
```



From the plots we can see that the state that produces most strawberries is California.

```
t1 = sberry$Value[c]
mean(t1[which(t1>0)])
```

```
## [1] 94392.93
```

```
t2 = sberry$Value[f]
mean(t2[which(t2>0)])
```

```
## [1] 18485.91
```

```
t3 = sberry$Value[o]
mean(t3[which(t3>0)])
```

```
## [1] 1750
```

```
t4 = sberry$Value[w]
mean(t4[which(t4>0)])
```

```
## [1] 186.1179
```

Strawberries in California have the highest average value.

Conclusion

From the dataset we can roughly idendify the distribution of the values of strawberries. Since there are lots of missing values, it needs more data to get an accurate result.

Reference

USDA database selector: https://quickstats.nass.usda.gov">https://quickstats.nass.usda.gov