Berries Project

Chenghao Meng

2020/10/11

1 Data Cleaning

1.1 Data Import

```
# Import necessary packages
pacman::p_load("tidyverse")
```

By downloading data from the National Agriculture Statistics Service (NASS) of United States Department of Agriculture(USDA), we have the dataset containinig information about three types of berries: Blueberries, Strawberries and Raspberries.

Since there are only 8 out 21 columns that are useful for further analysis, we will drop those first for simplicity of the dataset.

```
# Read in the file
dt <- read.csv(file="C:/Users/CH.Meng/Desktop/berries.csv",header=T)
berry_raw <- dt %>%
  select(Year,Period,State,Commodity,Data.Item,Domain,Domain.Category,Value)
head(berry_raw)
```

```
##   Year         Period      State   Commodity
## 1 2019 MARKETING YEAR CALIFORNIA BLUEBERRIES
## 2 2019 MARKETING YEAR CALIFORNIA BLUEBERRIES
## 3 2019 MARKETING YEAR CALIFORNIA BLUEBERRIES
## 4 2019 MARKETING YEAR CALIFORNIA RASPBERRIES
## 5 2019 MARKETING YEAR CALIFORNIA RASPBERRIES
## 6 2019 MARKETING YEAR CALIFORNIA RASPBERRIES
##                                                               Data.Item Domain
## 1              BLUEBERRIES, TAME - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 2 BLUEBERRIES, TAME, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 3   BLUEBERRIES, TAME, PROCESSING - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 4                    RASPBERRIES - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 5      RASPBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
## 6        RASPBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / LB  TOTAL
##   Domain.Category Value
## 1   NOT SPECIFIED  2.85
## 2   NOT SPECIFIED  3.56
## 3   NOT SPECIFIED  0.29
## 4   NOT SPECIFIED  2.69
## 5   NOT SPECIFIED   (D)
## 6   NOT SPECIFIED   (D)
```

1.2 Initial Screening of the Data

From the output above, we can notice that there are a lot of categorical varibles. However, `Value` is supposed to be a numeric varible according to the defination on the website.

By looking at the column of `Value` , many (D),(NA),(X) and (Z) appears to be the reason why this column is defined as categorical. So, we will replace those with NA.

```
berry_raw$Value <- as.numeric(berry_raw$Value)
# Replace (D),(NA),(X) and (Z) with NA
berry_raw[berry_raw =="(D)"] <- NA
berry_raw[berry_raw =="(NA)"] <- NA
berry_raw[berry_raw =="(X)"] <- NA
berry_raw[berry_raw =="(Z)"] <- NA
```

Since those irregular “NA”s have been replaced, a summary of the dataset should be made for further exploration of the data.

```
# Summary of berry_raw
summary(berry_raw)
```

```
##       Year         Period             State            Commodity
##  Min.   :2015   Length:13238       Length:13238       Length:13238
##  1st Qu.:2016   Class :character   Class :character   Class :character
##  Median :2017   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2017
##  3rd Qu.:2019
##  Max.   :2019
##
##   Data.Item            Domain          Domain.Category        Value
##  Length:13238       Length:13238       Length:13238       Min.   :  0.000
##  Class :character   Class :character   Class :character   1st Qu.:  0.550
##  Mode  :character   Mode  :character   Mode  :character   Median :  1.831
##                                                           Mean   : 49.564
##                                                           3rd Qu.: 26.000
##                                                           Max.   :960.000
##                                                           NA's   :8854
```

1.3 Further Data Cleaning on Strawberries

After finishing the initial screening of the dataset, we use the `filter` function to extract data of strawberries to conduct further study.

```
strawberry_raw <- berry_raw %>%
  filter(Commodity=="STRAWBERRIES")
# Summary of the dataset
summary(strawberry_raw)
```

```
##       Year         Period             State            Commodity
##  Min.   :2015   Length:3476        Length:3476        Length:3476
##  1st Qu.:2016   Class :character   Class :character   Class :character
##  Median :2018   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2017
##  3rd Qu.:2019
##  Max.   :2019
##
##   Data.Item            Domain          Domain.Category        Value
##  Length:3476        Length:3476        Length:3476        Min.   :  0.000
##  Class :character   Class :character   Class :character   1st Qu.:  0.307
##  Mode  :character   Mode  :character   Mode  :character   Median :  2.000
##                                                           Mean   : 63.618
##                                                           3rd Qu.: 37.000
##                                                           Max.   :960.000
##                                                           NA's   :2247
```

The summary of the strawberry dataset shows that there are 4958 NAs in the column `Value` . Since those observations does not contain much information, we choose to delete them.

1.3.1 Cleaning: `Data Item`

```
strawberry_raw2 <- strawberry_raw %>% drop_na()
```

```
item_pre <- strawberry_raw2$Data.Item
# Replace "-" with "," for the convenience of spliting
item <- gsub(" - ",",",item_pre)
```

Now, we use regular expression to extract the measurement and the type of the berry.

```
# Measurement of the strawberry
unit_stberry <- str_extract_all(item,"MEASURED.*[^./AVG]|ACRES.*")
# Delete the comma and space
unit_stberry <- str_replace(unit_stberry, ",","")
unit_stberry <- trimws(unit_stberry)
```

By looking at the original dataset, we find that there is only one strawberry type in the dataset, and we also extract them by using regular expression.

```
# Market Channel of the strawberry
market_stberry <- str_extract_all(item,"(FRESHMARKET)|(PROCESSING)")

col_market_stberry <- data.frame(Market.Channel=as.character(market_stberry))
col_market_stberry[col_market_stberry=="character(0)"] <- NA
```

1.3.2 Cleaning: `Domain Category`

Then, we will separate the chemical type and the detail of certain kind of chemical from the column `Domain Category` by using `separate` function in tidyverse package.

```
chemical_obj <- data.frame(strawberry_raw2$Domain.Category)
chemical_info <- separate(data=chemical_obj, col=colnames(chemical_obj), into = c("Chemical.Type", "C
hemical.Detail"), sep = ",")
head(chemical_info)
```

```
##   Chemical.Type                          Chemical.Detail
## 1 NOT SPECIFIED                                     <NA>
## 2 NOT SPECIFIED                                     <NA>
## 3 NOT SPECIFIED                                     <NA>
## 4 NOT SPECIFIED                                     <NA>
## 5 NOT SPECIFIED                                     <NA>
## 6      CHEMICAL  FUNGICIDE: (BORAX DECAHYDRATE = 11102)
```

1.4 Cleaned Dataset: Inforamtion of Strawberries

Now we have the final dataset for further exploration by using `select` and `mutate` function in tidyverse package.

```
stberry <- strawberry_raw2 %>%
  select(Year,State,Commodity,Value) %>%
  mutate(Unit=as.character(unit_stberry),chemical_info,col_market_stberry)
head(stberry)
```

```
##   Year        State    Commodity Value                   Unit Chemical.Type
## 1 2019   CALIFORNIA STRAWBERRIES 108.0    MEASURED IN $ / CWT NOT SPECIFIED
## 2 2019      FLORIDA STRAWBERRIES 152.0    MEASURED IN $ / CWT NOT SPECIFIED
## 3 2019 OTHER STATES STRAWBERRIES 129.0    MEASURED IN $ / CWT NOT SPECIFIED
## 4 2019 OTHER STATES STRAWBERRIES  52.8    MEASURED IN $ / CWT NOT SPECIFIED
## 5 2019   CALIFORNIA STRAWBERRIES 580.0 MEASURED IN CWT / ACRE NOT SPECIFIED
## 6 2019   CALIFORNIA STRAWBERRIES 300.0         MEASURED IN LB      CHEMICAL
##                            Chemical.Detail Market.Channel
## 1                                     <NA>           <NA>
## 2                                     <NA>           <NA>
## 3                                     <NA>           <NA>
## 4                                     <NA>     PROCESSING
## 5                                     <NA>           <NA>
## 6  FUNGICIDE: (BORAX DECAHYDRATE = 11102)           <NA>
```

2 Exploratory Data Analysis

2.1 Measurements of Strawberry

2.1.1 Count the Types of Measurement

After cleaning the data, we will first count the types for measurement of the strawberry.

```
# Summary of the measurement for strawberry
stberry_unit_sum <- stberry %>%
  group_by(Unit)%>%
  summarize(
    Count=n(),
    Mean.Value=round(mean(Value),2)
    )
```

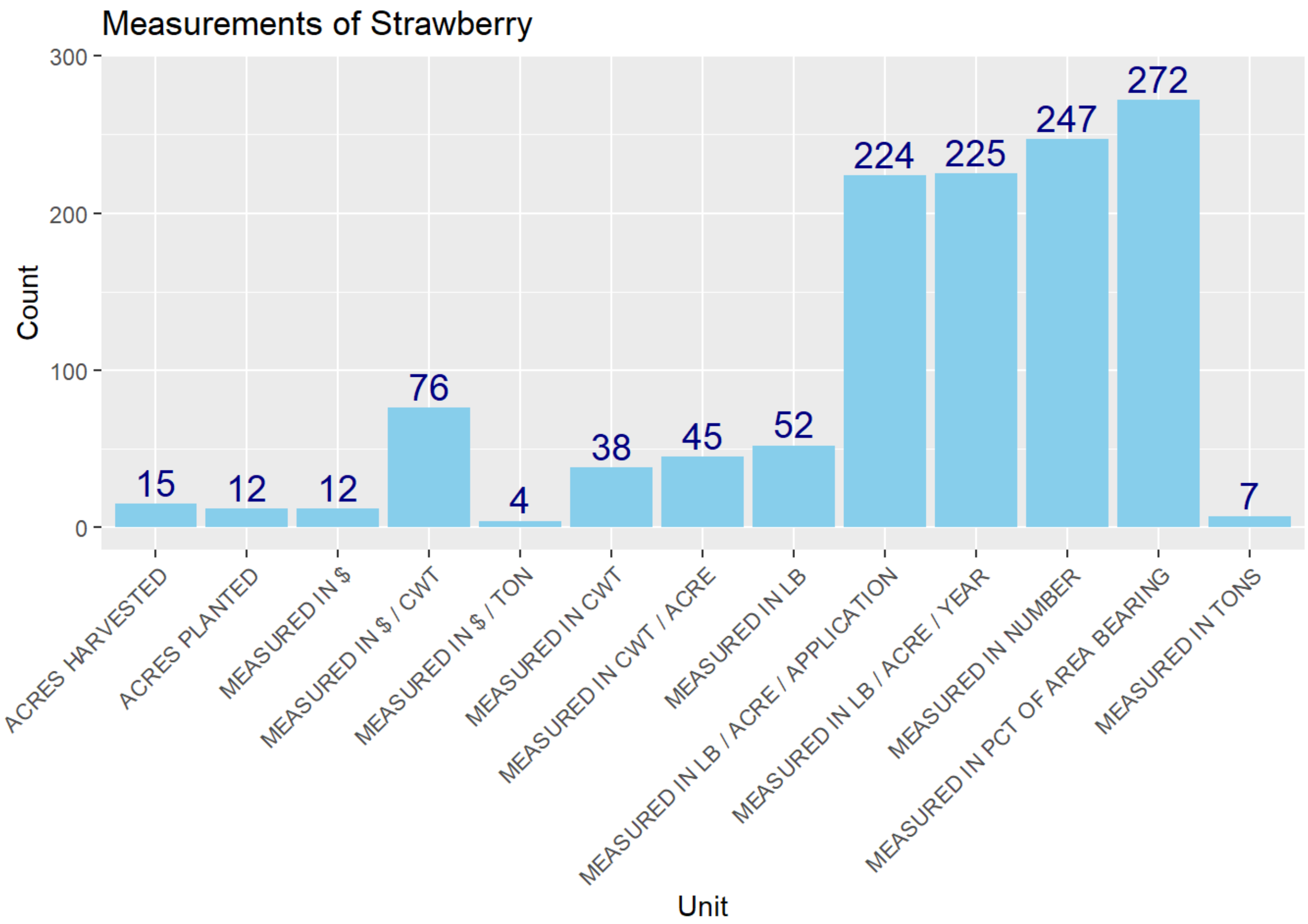
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
cat(paste("There are",length(stberry_unit_sum$Unit),"types of measurements for strawberry in the data
set."))
```

```
## There are 13 types of measurements for strawberry in the dataset.
```

The we will make a bar plot to identify the frequency of different measurements for strawberry.

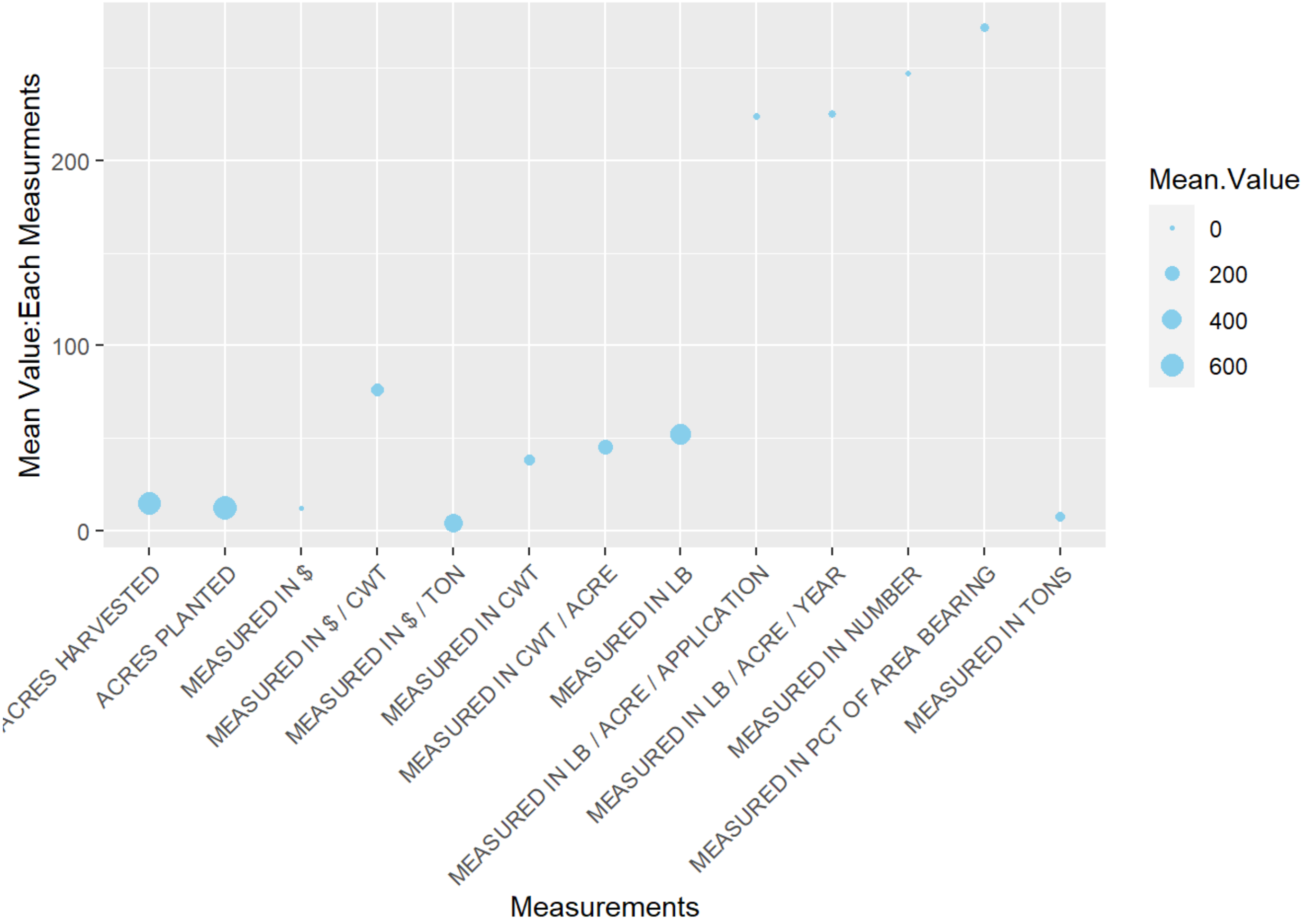
```
# Bar Plot: Measurement of stberry
ggplot(data=stberry_unit_sum,mapping=aes(x=Unit,y=Count))+
  geom_bar(stat='identity',fill="sky blue")+
  ggtitle("Measurements of Strawberry")+
  geom_text(aes(label=Count,y=Count+14),size=5,color="navy blue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



From the bar plot, we can see that *Measured in PCT of area bearing* is mostly used for 272 times, and *Measured in $/Ton* is leastly used for only 4 times in the strawberry dataset.

2.1.2 Plot the Value of Measurements

```
# Plot: mean value of the measurements
ggplot(data=stberry_unit_sum, mapping=aes(x = Unit, y= Count,size=Mean.Value)) +
  geom_point(shape=20,color="sky blue")+
  xlab("Measurements") +
  ylab("Mean Value:Each Measurments")+
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This scatterplot indicates that there are big variation between each measurements, which proved that the data cleaning in previous sections is very necessary to make different numbers comparible.

2.2 Plot Different Measurement

To make the plot of different measurement versus other varibles in the dataset, `group_by` function should be used to generate the data frame for further use.

```
# Group by "Unit"
stberry_unit_df <- stberry %>%
  group_by(Unit) %>%
  summarize(
    States=State,
    Years= Year,
    Count=n(),
    Values=Value
    )
```

```
## `summarise()` regrouping output by 'Unit' (override with `.groups` argument)
```

```
tail(stberry_unit_df)
```

```
## # A tibble: 6 x 5
## # Groups:   Unit [1]
##   Unit             States         Years Count Values
##   <chr>            <chr>          <int> <int>  <dbl>
## 1 MEASURED IN TONS NORTH CAROLINA  2018     7      0
## 2 MEASURED IN TONS FLORIDA         2018     7      0
## 3 MEASURED IN TONS NORTH CAROLINA  2018     7      0
## 4 MEASURED IN TONS NORTH CAROLINA  2017     7    149
## 5 MEASURED IN TONS NORTH CAROLINA  2017     7    150
## 6 MEASURED IN TONS FLORIDA         2016     7      0
```

2.2.1 Measurement: *Measured in Number*

To explore the status of value measured in Number, a data frame should to created for the convenience of `ggplot` function.

```
# Generate a dataframe of Measurement: Measured in Number
df_measur_in_number <- stberry_unit_df %>%
  filter(Unit=="MEASURED IN NUMBER") %>%
  group_by(States,Years) %>%
  summarise(Number_Total=sum(Values))
```

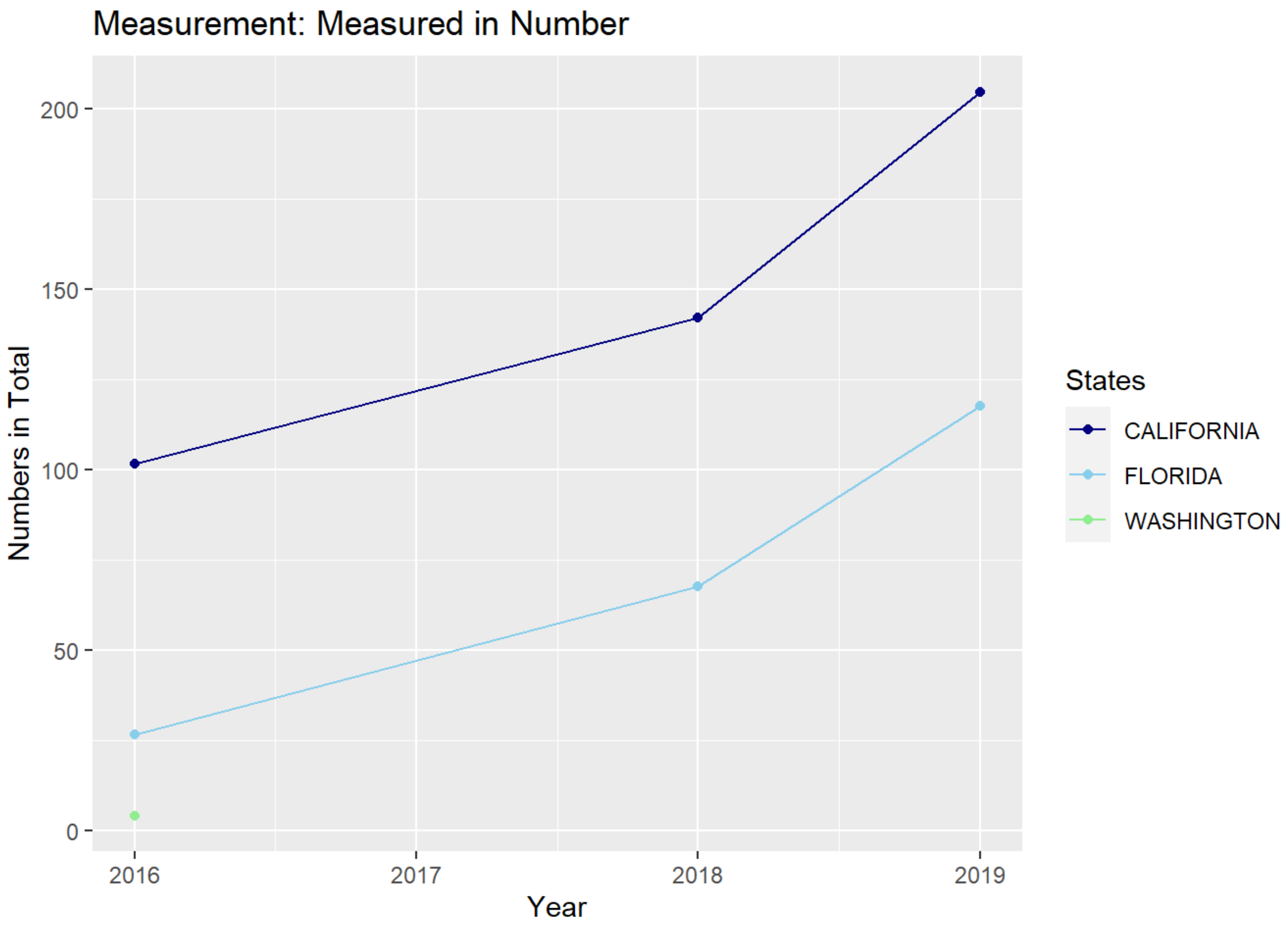
```
## `summarise()` regrouping output by 'States' (override with `.groups` argument)
```

```
df_measur_in_number
```

```
## # A tibble: 7 x 3
## # Groups:   States [3]
##   States     Years Number_Total
##   <chr>      <int>        <dbl>
## 1 CALIFORNIA  2016        102.
## 2 CALIFORNIA  2018        142.
## 3 CALIFORNIA  2019        205.
## 4 FLORIDA     2016         26.5
## 5 FLORIDA     2018         67.5
## 6 FLORIDA     2019        118.
## 7 WASHINGTON  2016          4.1
```

Now, We draw a plot of Total Numbers v.s. Year.

```
ggplot(data=df_measur_in_number)+
  geom_line(mapping=aes(x=Years,y=Number_Total,color=States))+
  geom_point(mapping=aes(x=Years,y=Number_Total,color=States))+
  scale_color_manual(values = c("navy blue","sky blue","light green"))+ # Change the color of the leg
end
  xlab("Year") + ylab("Numbers in Total") +
  ggtitle("Measurement: Measured in Number")
```



The plot above shows that the total number of strawberry in California and Florida keep growing from 2016 to 2019. Meanwhile, the total number of strawberry in California is larger than that in Florida.

2.2.2 Measurement: *Measured in LB*

Now, we can include the value measured by LB into a dataframe for the convenience of plotting.

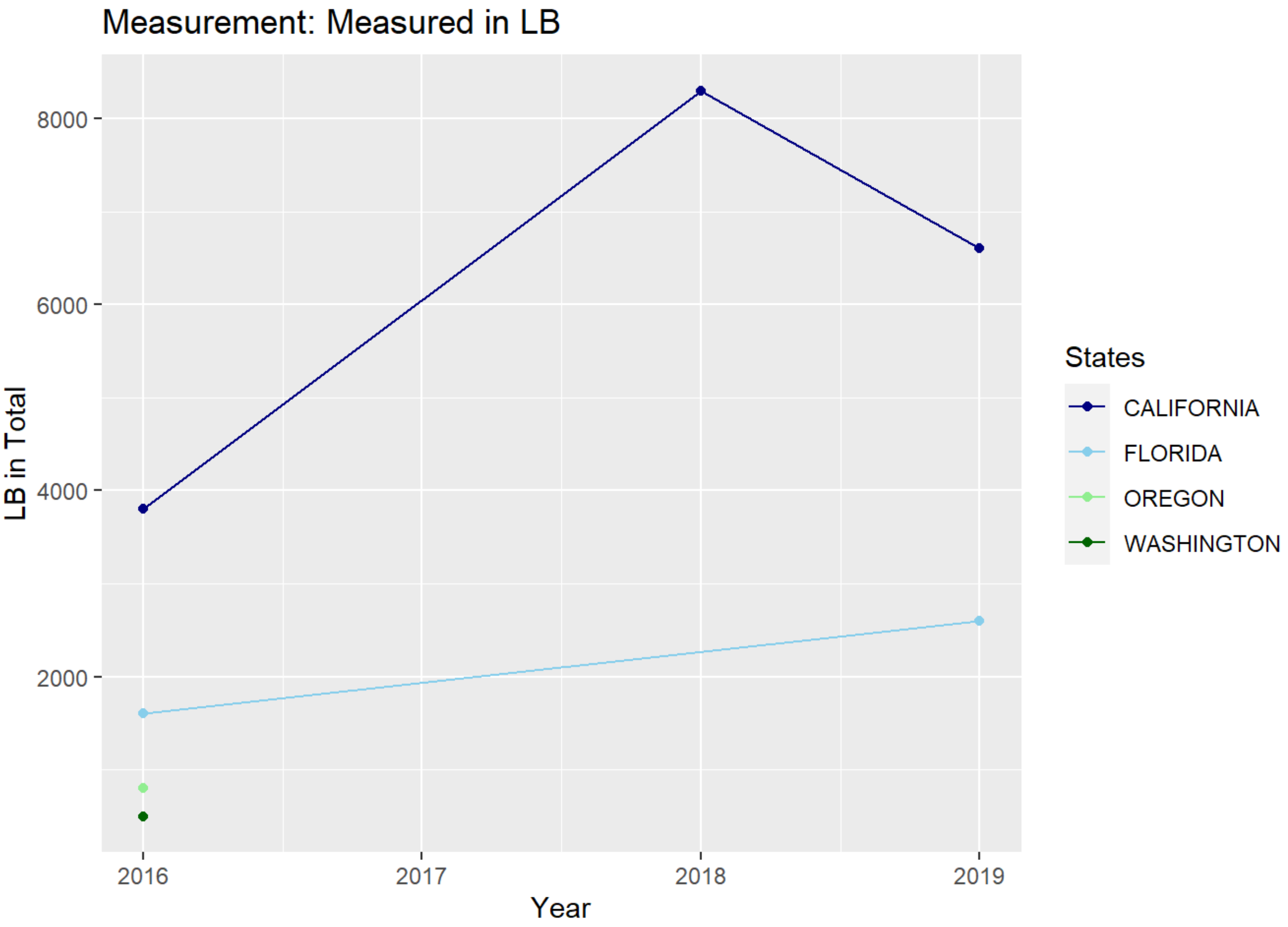
```
df_measur_in_lb <- stberry_unit_df %>%
  filter(Unit=="MEASURED IN LB") %>%
  group_by(States,Years) %>%
  summarise(LB_Total=sum(Values))
```

```
## `summarise()` regrouping output by 'States' (override with `.groups` argument)
```

```
df_measur_in_lb
```

```
## # A tibble: 7 x 3
## # Groups:   States [4]
##   States     Years LB_Total
##   <chr>      <int>    <dbl>
## 1 CALIFORNIA  2016     3800
## 2 CALIFORNIA  2018     8300
## 3 CALIFORNIA  2019     6600
## 4 FLORIDA     2016     1600
## 5 FLORIDA     2019     2600
## 6 OREGON      2016      800
## 7 WASHINGTON  2016      500
```

```
# Plot Total LB v.s. Year
ggplot(data=df_measur_in_lb)+
  geom_line(mapping=aes(x=Years,y=LB_Total,color=States))+
  geom_point(mapping=aes(x=Years,y=LB_Total,color=States))+
  scale_color_manual(values = c("navy blue","sky blue","light green","dark green"))+ # Change the col
or of the legend
  xlab("Year") + ylab("LB in Total") +
  ggtitle("Measurement: Measured in LB")
```



The plot shows that the state of California has the total weight measured by LB, but it expereinced a sharp drop in the year of 2019.

2.2.3 Measurement: `Measured in $/CWT`

Then, we will explore the price of the strawberry by adopting the same methods above.

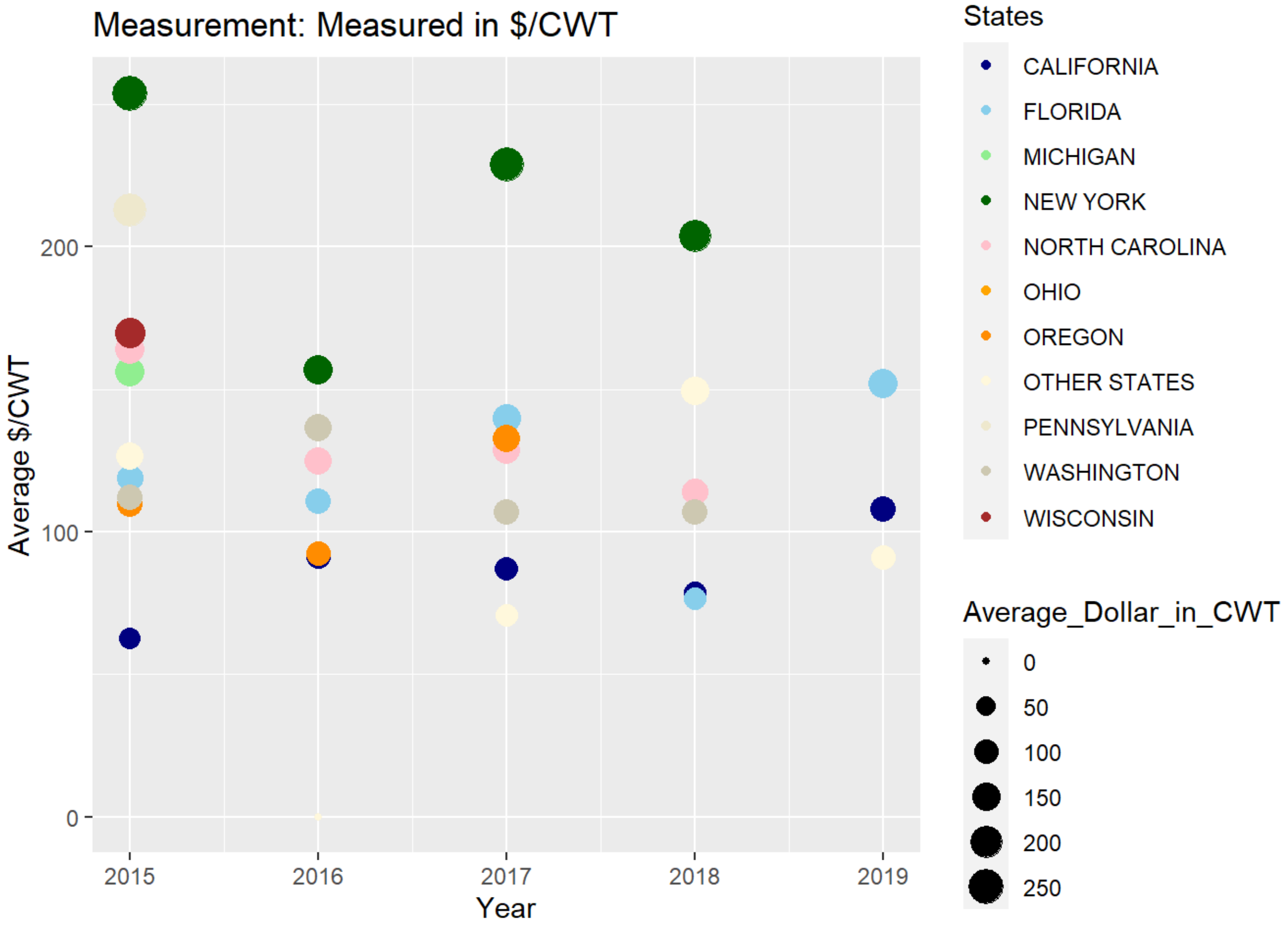
```
df_measur_price_cwt <- stberry_unit_df %>%
  filter(Unit=="MEASURED IN $ / CWT") %>%
  group_by(States,Years) %>%
  summarise(Average_Dollar_in_CWT=mean(Values))
```

```
## `summarise()` regrouping output by 'States' (override with `.groups` argument)
```

```
head(df_measur_price_cwt)
```

```
## # A tibble: 6 x 3
## # Groups:   States [2]
##   States     Years Average_Dollar_in_CWT
##   <chr>      <int>                 <dbl>
## 1 CALIFORNIA  2015                  62.5
## 2 CALIFORNIA  2016                  91.3
## 3 CALIFORNIA  2017                  87
## 4 CALIFORNIA  2018                  78.6
## 5 CALIFORNIA  2019                 108
## 6 FLORIDA     2015                 119
```

```
ggplot(data=df_measur_price_cwt)+
  geom_point(mapping=aes(x=Years,y=Average_Dollar_in_CWT,color=States,size=Average_Dollar_in_CWT))+
   scale_color_manual(values = c("navy blue","sky blue","light green","dark green","pink","orange","d
ark orange","cornsilk","cornsilk2","cornsilk3","brown"))+
  xlab("Year") + ylab("Average $/CWT") +
  ggtitle("Measurement: Measured in $/CWT")
```



From the plot, we can see that the average price in $/CWT in State of New York is always the highest among other states from 2015 to 2019.

Moreover, the average price of strawberry in state of California is realtively low compared to other states.

3 Recommendation

According to the analysis above, the state of California is the best place to buy strawberries with the advantages of the highest production and the lowest price compared to other states.

However, since not every state has values for all unit of measurements, this recommendation is not very solid. For example, the production information measured by *Numbers* and *LB* is not included in the dataset. This recommendation can be seen as a reference when choosing the place to purchase strawberries.

4 Reference

[1]Hadley Wickham, Romain François, Lionel Henry, Kirill Müller.(2020) dplyr: A Grammar of Data Manipulation, version 1.0.2

[2]Hadley Wickham.(2019) tidyverse: Easily Install and Load the ‘Tidyverse’, version 1.3.0

[3]Alboukadel Kassambara.(2020) ggpubr: ‘ggplot2’ Based Publication Ready Plots, version 0.4.0