Assignment

2024-10-07

Introduction-

This report presents the exploration, cleaning, and organization of the USDA strawberry dataset. The goal is to clean, organize, and explore the strawberry data, prepare the data for analysis

Assignment

Step 1- Data Exploration:

Before cleaning any data, we need to understand the content and structure of the data. This involves looking at the types of variables, missing values, and general trends in the dataset.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

First, we checked the head of the dataset. We few the first few rows of the dataset.

```
strawberry_data <- read.csv("strawberries25_v3.csv")
head(strawberry_data)
```

```
##   Program Year Period Week.Ending Geo.Level   State State.ANSI Ag.District
## 1  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 2  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 3  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 4  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 5  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
## 6  CENSUS 2022   YEAR          NA    COUNTY ALABAMA          1  BLACK BELT
##   Ag.District.Code  County County.ANSI Zip.Code Region watershed_code Watershed
## 1               40 BULLOCK          11       NA     NA              0        NA
## 2               40 BULLOCK          11       NA     NA              0        NA
```

```
## 3               40 BULLOCK          11       NA     NA              0        NA
## 4               40 BULLOCK          11       NA     NA              0        NA
## 5               40 BULLOCK          11       NA     NA              0        NA
## 6               40 BULLOCK          11       NA     NA              0        NA
##      Commodity                                      Data.Item Domain
## 1 STRAWBERRIES                    STRAWBERRIES - ACRES BEARING  TOTAL
## 2 STRAWBERRIES                      STRAWBERRIES - ACRES GROWN  TOTAL
## 3 STRAWBERRIES                STRAWBERRIES - ACRES NON-BEARING  TOTAL
## 4 STRAWBERRIES     STRAWBERRIES - OPERATIONS WITH AREA BEARING  TOTAL
## 5 STRAWBERRIES       STRAWBERRIES - OPERATIONS WITH AREA GROWN  TOTAL
## 6 STRAWBERRIES STRAWBERRIES - OPERATIONS WITH AREA NON-BEARING  TOTAL
##   Domain.Category Value CV....
## 1   NOT SPECIFIED   (D)    (D)
## 2   NOT SPECIFIED     3   15.7
## 3   NOT SPECIFIED   (D)    (D)
## 4   NOT SPECIFIED     1    (L)
## 5   NOT SPECIFIED     6   52.7
## 6   NOT SPECIFIED     5   47.6
```

The dataset has several columns representing different attributes like Program, Year, Geo Level, and State. Some columns, such as Week.Ending, contain missing values (NA).

Next, we explore the structure of the dataset to understand the data types of each column

```
str(strawberry_data)
```

```
## 'data.frame':    12669 obs. of  21 variables:
##  $ Program         : chr  "CENSUS" "CENSUS" "CENSUS" "CENSUS" ...
##  $ Year            : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022 ...
##  $ Period          : chr  "YEAR" "YEAR" "YEAR" "YEAR" ...
##  $ Week.Ending     : logi  NA NA NA NA NA NA ...
##  $ Geo.Level       : chr  "COUNTY" "COUNTY" "COUNTY" "COUNTY" ...
##  $ State           : chr  "ALABAMA" "ALABAMA" "ALABAMA" "ALABAMA" ...
##  $ State.ANSI      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Ag.District     : chr  "BLACK BELT" "BLACK BELT" "BLACK BELT" "BLACK BELT" ...
##  $ Ag.District.Code: int  40 40 40 40 40 40 40 40 40 40 ...
##  $ County          : chr  "BULLOCK" "BULLOCK" "BULLOCK" "BULLOCK" ...
##  $ County.ANSI     : int  11 11 11 11 11 11 101 101 101 101 ...
##  $ Zip.Code        : logi  NA NA NA NA NA NA ...
##  $ Region          : logi  NA NA NA NA NA NA ...
##  $ watershed_code  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Watershed       : logi  NA NA NA NA NA NA ...
##  $ Commodity       : chr  "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" "STRAWBERRIES" ...
##  $ Data.Item       : chr  "STRAWBERRIES - ACRES BEARING" "STRAWBERRIES - ACRES GROWN" "STRAWBERRIES -
##  $ Domain          : chr  "TOTAL" "TOTAL" "TOTAL" "TOTAL" ...
##  $ Domain.Category : chr  "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" "NOT SPECIFIED" ...
##  $ Value           : chr  " (D)" "3" " (D)" "1" ...
##  $ CV....          : chr  "(D)" "15.7" "(D)" "(L)" ...
```

The dataset contains 12,669 observations and 21 variables. The columns include both character and numeric data types, as shown below:

Program: chr (character) Year: int (integer) Week.Ending: logi (logical), contains only NA values. Geo.Level: chr (character) State: chr (character) State ANSI: int (integer) Ag.District: chr (character) Ag.District.Code:

int (integer) The column Week.Ending contains entirely missing data, and other columns like Zip.Code and Watershed also contain many missing values.

Next, we generate a summary of its contents to detect any unusual patterns, missing values, or outliers.

```
summary(strawberry_data)
```

```
##    Program               Year         Period          Week.Ending
##  Length:12669       Min.   :2018   Length:12669       Mode:logical
##  Class :character   1st Qu.:2021   Class :character   NA's:12669
##  Mode  :character   Median :2022   Mode  :character
##                     Mean   :2021
##                     3rd Qu.:2022
##                     Max.   :2024
##
##   Geo.Level            State             State.ANSI    Ag.District
##  Length:12669       Length:12669       Min.   : 1.00   Length:12669
##  Class :character   Class :character   1st Qu.: 9.00   Class :character
##  Mode  :character   Mode  :character   Median :21.00   Mode  :character
##                                        Mean   :24.43
##                                        3rd Qu.:39.00
##                                        Max.   :56.00
##                                        NA's   :264
##  Ag.District.Code    County           County.ANSI     Zip.Code
##  Min.   :10.00    Length:12669       Min.   :  1.00   Mode:logical
##  1st Qu.:20.00    Class :character   1st Qu.: 29.00   NA's:12669
##  Median :50.00    Mode  :character   Median : 69.00
##  Mean   :46.18                       Mean   : 83.82
##  3rd Qu.:62.00                       3rd Qu.:119.00
##  Max.   :96.00                       Max.   :810.00
##  NA's   :5359                        NA's   :5385
##   Region        watershed_code Watershed       Commodity
##  Mode:logical   Min.   :0      Mode:logical   Length:12669
##  NA's:12669     1st Qu.:0      NA's:12669     Class :character
##                 Median :0                     Mode  :character
##                 Mean   :0
##                 3rd Qu.:0
##                 Max.   :0
##
##   Data.Item            Domain          Domain.Category       Value
##  Length:12669       Length:12669       Length:12669       Length:12669
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     CV....
##  Length:12669
##  Class :character
##  Mode  :character
##
##
##
```

##

The summary shows the range of values for numerical columns like Year, State.ANSI, and Ag.District.Code. It also highlights the number of missing values in columns like Zip.Code, Region, Watershed, and Value. These columns either contain a significant amount of missing data or contain non-numeric placeholder values like (D) or (L) in the Value and CV.... columns.

Step 2 - Data Cleaning

Based on Data Exploration, we can now proceed to Data Cleaning. We can now drop irrelevant columns with mostly missing data, replace inconsistent placeholder values with NA, split columns containing multiple pieces of information into separate columns.

Dropping irrelevant columns- Columns like as Week.Ending, Zip.Code, Region, Watershed.Code, and Watershed contain mostly missing data, so we can drop these columns

```
cleaned_data <- strawberry_data %>%
  select(-c(Week.Ending, Zip.Code, Region, watershed_code, Watershed))
```

Handling Missing and Inconsistent Data-

The dataset contains placeholder values such as (D), (L), and (Z) in the Value and CV.... columns, representing missing data. We replace these placeholders with NA to handle them appropriately in future analysis.

```
cleaned_data <- cleaned_data %>%
  mutate(across(where(is.character), ~na_if(., "(D)"))) %>%
  mutate(across(where(is.character), ~na_if(., "(L)"))) %>%
  mutate(across(where(is.character), ~na_if(., "(Z)")))
```

```
unique(cleaned_data$Data.Item)
```

```
##  [1] "STRAWBERRIES - ACRES BEARING"
##  [2] "STRAWBERRIES - ACRES GROWN"
##  [3] "STRAWBERRIES - ACRES NON-BEARING"
##  [4] "STRAWBERRIES - OPERATIONS WITH AREA BEARING"
##  [5] "STRAWBERRIES - OPERATIONS WITH AREA GROWN"
##  [6] "STRAWBERRIES - OPERATIONS WITH AREA NON-BEARING"
##  [7] "STRAWBERRIES, ORGANIC - ACRES HARVESTED"
##  [8] "STRAWBERRIES, ORGANIC - OPERATIONS WITH AREA HARVESTED"
##  [9] "STRAWBERRIES, ORGANIC - OPERATIONS WITH SALES"
## [10] "STRAWBERRIES, ORGANIC - PRODUCTION, MEASURED IN CWT"
## [11] "STRAWBERRIES, ORGANIC - SALES, MEASURED IN $"
## [12] "STRAWBERRIES, ORGANIC - SALES, MEASURED IN CWT"
## [13] "STRAWBERRIES, ORGANIC, FRESH MARKET - OPERATIONS WITH SALES"
## [14] "STRAWBERRIES, ORGANIC, FRESH MARKET - SALES, MEASURED IN $"
## [15] "STRAWBERRIES, ORGANIC, FRESH MARKET - SALES, MEASURED IN CWT"
## [16] "STRAWBERRIES, ORGANIC, PROCESSING - OPERATIONS WITH SALES"
## [17] "STRAWBERRIES, ORGANIC, PROCESSING - SALES, MEASURED IN $"
## [18] "STRAWBERRIES, ORGANIC, PROCESSING - SALES, MEASURED IN CWT"
## [19] "STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, ADJUSTED BASE, MEASURED IN $ / CWT"
## [20] "STRAWBERRIES, PROCESSING - PRICE RECEIVED, ADJUSTED BASE, MEASURED IN $ / TON"
```

```
## [21] "STRAWBERRIES - PRICE RECEIVED, MEASURED IN $ / CWT"
## [22] "STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, MEASURED IN $ / CWT"
## [23] "STRAWBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / CWT"
## [24] "STRAWBERRIES - ACRES HARVESTED"
## [25] "STRAWBERRIES - ACRES PLANTED"
## [26] "STRAWBERRIES - PRODUCTION, MEASURED IN $"
## [27] "STRAWBERRIES - PRODUCTION, MEASURED IN CWT"
## [28] "STRAWBERRIES - PRODUCTION, MEASURED IN TONS"
## [29] "STRAWBERRIES - YIELD, MEASURED IN CWT / ACRE"
## [30] "STRAWBERRIES - YIELD, MEASURED IN TONS / ACRE"
## [31] "STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, 10 YEAR AVG FOR PARITY PURPOSES, MEASURED IN $ / C
## [32] "STRAWBERRIES, FRESH MARKET - PRICE RECEIVED, 10 YEAR AVG, MEASURED IN $ / CWT"
## [33] "STRAWBERRIES, FRESH MARKET - PRODUCTION, MEASURED IN $"
## [34] "STRAWBERRIES, FRESH MARKET, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [35] "STRAWBERRIES, NOT SOLD - PRODUCTION, MEASURED IN CWT"
## [36] "STRAWBERRIES, PROCESSING - PRICE RECEIVED, 10 YEAR AVG FOR PARITY PURPOSES, MEASURED IN $ / TO
## [37] "STRAWBERRIES, PROCESSING - PRICE RECEIVED, 10 YEAR AVG, MEASURED IN $ / TON"
## [38] "STRAWBERRIES, PROCESSING - PRODUCTION, MEASURED IN $"
## [39] "STRAWBERRIES, PROCESSING, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [40] "STRAWBERRIES, UTILIZED - PRODUCTION, MEASURED IN CWT"
## [41] "STRAWBERRIES, UTILIZED - PRODUCTION, MEASURED IN TONS"
## [42] "STRAWBERRIES - APPLICATIONS, MEASURED IN LB"
## [43] "STRAWBERRIES - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG"
## [44] "STRAWBERRIES - APPLICATIONS, MEASURED IN LB / ACRE / YEAR, AVG"
## [45] "STRAWBERRIES - APPLICATIONS, MEASURED IN NUMBER, AVG"
## [46] "STRAWBERRIES - TREATED, MEASURED IN PCT OF AREA BEARING, AVG"
## [47] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB"
## [48] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / APPLICATION, AVG"
## [49] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN LB / ACRE / YEAR, AVG"
## [50] "STRAWBERRIES, BEARING - APPLICATIONS, MEASURED IN NUMBER, AVG"
## [51] "STRAWBERRIES, BEARING - TREATED, MEASURED IN PCT OF AREA BEARING, AVG"
## [52] "STRAWBERRIES, PROCESSING - PRICE RECEIVED, MEASURED IN $ / TON"
## [53] "STRAWBERRIES, PROCESSING, UTILIZED - PRODUCTION, MEASURED IN TONS"
```

```
colnames(cleaned_data)
```

```
##  [1] "Program"          "Year"             "Period"           "Geo.Level"
##  [5] "State"            "State.ANSI"       "Ag.District"      "Ag.District.Code"
##  [9] "County"           "County.ANSI"      "Commodity"        "Data.Item"
## [13] "Domain"           "Domain.Category"  "Value"            "CV...."
```

```
head(cleaned_data)
```

```
##   Program Year Period Geo.Level   State State.ANSI Ag.District Ag.District.Code
## 1  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 2  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 3  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 4  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 5  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
## 6  CENSUS 2022   YEAR    COUNTY ALABAMA          1  BLACK BELT               40
##    County County.ANSI    Commodity
## 1 BULLOCK          11 STRAWBERRIES
## 2 BULLOCK          11 STRAWBERRIES
```

```
## 3 BULLOCK          11 STRAWBERRIES
## 4 BULLOCK          11 STRAWBERRIES
## 5 BULLOCK          11 STRAWBERRIES
## 6 BULLOCK          11 STRAWBERRIES
##                                        Data.Item Domain Domain.Category Value
## 1                     STRAWBERRIES - ACRES BEARING  TOTAL   NOT SPECIFIED   (D)
## 2                       STRAWBERRIES - ACRES GROWN  TOTAL   NOT SPECIFIED     3
## 3                 STRAWBERRIES - ACRES NON-BEARING  TOTAL   NOT SPECIFIED   (D)
## 4     STRAWBERRIES - OPERATIONS WITH AREA BEARING  TOTAL   NOT SPECIFIED     1
## 5       STRAWBERRIES - OPERATIONS WITH AREA GROWN  TOTAL   NOT SPECIFIED     6
## 6 STRAWBERRIES - OPERATIONS WITH AREA NON-BEARING  TOTAL   NOT SPECIFIED     5
##   CV....
## 1   <NA>
## 2   15.7
## 3   <NA>
## 4   <NA>
## 5   52.7
## 6   47.6
```

There seems to be no chemical data, so we shall proceed without it.

```
summary(cleaned_data$Value)
```

```
##    Length     Class      Mode
##     12669 character character
```

```
#converting value column to numeric
non_numeric_values <- cleaned_data$Value[!grepl("^\\d+$", cleaned_data$Value)]
unique(non_numeric_values)
```

```
##   [1] " (D)"          " (Z)"          "13,347"        "13,586"
##   [5] "3,099"         "1,790"         "1,797"         "15,684"
##   [9] "11,969"        "12,027"        "10,989"        "11,224"
##  [13] "1,161"         "3,195"         "46,265"        "2,514"
##  [17] "4,231"         "3,396"         "10,146"        "70,709"
##  [21] "1,199"         "3,729"         "46,931"        "2,758"
##  [25] "4,441"         "3,942"         "10,461"        "73,462"
##  [29] "2,753"         "4,690"         "1,812"         "7,596"
##  [33] "5,508"         "1,884"         "8,491"         "1,270"
##  [37] "2,130"         "35,019"        "1,699"         "8,154"
##  [41] "46,093"        "35,457"        "1,736"         "8,323"
##  [45] "46,813"        "1,414"         "1,401"         "13,413"
##  [49] "9,938"         "1,459"         "1,478"         "13,663"
##  [53] "1,012"         "1,073"         "1,664"         "1,853"
##  [57] "1,011"         "5,301"         "1,495,299"     "335,964,420"
##  [61] "1,494,673"     "1,483,234"     "11,440"        "4,228"
##  [65] "1,413,251"     "311,784,980"   "1,412,627"     "1,401,384"
##  [69] "11,244"        "40,890"        "67,146"        "18,358,396"
##  [73] "11,600"        "31,680"        "16,000"        "30,828"
##  [77] "174,980"       "53,810"        "94,827"        "174,011"
##  [81] "418,914"       "80,886"        "52,665"        "28,166"
##  [85] "1,755"         "633,111"       "190,000"       "106,638"
##  [89] "1,858"         "1,362"         "895,054"       "148,898"
```

```
##  [93] "1,202"          "480,304"       "1,579"          "870,017"
##  [97] "863,231"        "1,499"         "6,786"          "128,120"
## [101] "5,158"          "1,461,988"     "320,793,584"    "1,461,266"
## [105] "294,996,931"    "1,250,324"     "25,796,653"     "210,942"
## [109] "4,022"          "1,384,735"     "300,277,717"    "1,384,016"
## [113] "275,716,713"    "1,177,214"     "24,561,004"     "206,802"
## [117] "31,000"         "32,164"        "59,905"         "15,055,709"
## [121] "1,309"          "26,380"        "128,882"        "32,295"
## [125] "12,387"         "62,472"        "50,234"         "144,129"
## [129] "29,986"         "18,302"        "24,000"         "2,260"
## [133] "644,155"        "87,402"        "68,830"         "161,288"
## [137] "5,081"          "1,728,809"     "505,658"        "1,004"
## [141] "1,223,151"      "4,077"         "89,572"         "85,229"
## [145] "35,934"         "1,051"         "486,870"        "2,428"
## [149] "745,009"        "2,404"         "50,111"         "10.9"
## [153] "4.04"           "43.8"          "56,800"         "57,300"
## [157] "3,398,943,000"  "27,560,000"    "1,378,000"      "24.26"
## [161] "10.8"           "3,173,579,000" "22,382,200"     "33,400"
## [165] "44.4"           "4.05"          "225,364,000"    "5,144,400"
## [169] "27,526,600"     "1,376,330"     "42,700"         "43,100"
## [173] " (NA)"          "2,965,387,000" "24,600,000"     "3,300"
## [177] "2,800"          "6,600"         "603,100"        "30,300"
## [181] "8,600"          "22,400"        "14,600"         "7,100"
## [185] "7,200"          "10,800"        "4,000"          "2,300"
## [189] "10,600"         "8,300"         "3,600"          "1,258,100"
## [193] "1,300"          "269,500"       "2,338,800"      "3,900"
## [197] "71,400"         "89,700"        "12,800"         "28,700"
## [201] "4,600"          "2,000"         "1,700"          "7,600"
## [205] "6,200"          "5,000"         "5,400"          "3,100"
## [209] "19,400"         "7,900"         "5,600"          "3,800"
## [213] "231,600"        "11,299,000"    "1,642,600"      "15,611,900"
## [217] "393,000"        "216,000"       "0.234"          "0.042"
## [221] "0.354"          "1.693"         "0.316"          "0.659"
## [225] "0.211"          "0.119"         "0.161"          "0.5"
## [229] "0.105"          "0.271"         "0.061"          "0.112"
## [233] "0.167"          "0.658"         "0.095"          "5.471"
## [237] "0.036"          "0.695"         "2.201"          "0.078"
## [241] "0.923"          "0.019"         "0.299"          "0.129"
## [245] "0.03"           "0.116"         "0.063"          "0.117"
## [249] "0.178"          "0.26"          "0.087"          "0.182"
## [253] "0.186"          "0.457"         "1.807"          "0.169"
## [257] "0.077"          "0.071"         "0.094"          "0.062"
## [261] "188.284"        "111.391"       "0.114"          "0.326"
## [265] "0.093"          "0.885"         "15.932"         "0.872"
## [269] "0.753"          "0.599"         "0.389"          "0.294"
## [273] "0.727"          "0.546"         "0.49"           "0.109"
## [277] "0.174"          "0.355"         "1.036"          "0.193"
## [281] "38.761"         "0.059"         "1.667"          "8.873"
## [285] "0.39"           "0.315"         "5.892"          "0.025"
## [289] "0.65"           "0.181"         "1.094"          "0.158"
## [293] "0.241"          "0.397"         "0.327"          "0.197"
## [297] "0.317"          "0.227"         "0.496"          "2.398"
## [301] "0.276"          "0.232"         "0.125"          "0.173"
## [305] "0.092"          "442.413"       "131.091"        "0.218"
```

```
## [309] "1.4"           "2.2"           "2.5"           "9.4"
## [313] "2.8"           "1.1"           "3.3"           "1.8"
## [317] "1.5"           "5.2"           "1.6"           "2.1"
## [321] "7.1"           "2.4"           "4.1"           "6.4"
## [325] "1.3"           "3.8"           "3.1"           "2.3"
## [329] "1.7"           "1.2"           "1.9"           "12.6"
## [333] "8.6"           "10.5"          "24,600"        "24,575,400"
## [337] "14,100"        "14,200"        "433,556,000"   "2,960,000"
## [341] "144,000"       "9,700"         "6,100"         "112,100"
## [345] "302,700"       "9,900"         "1,500"         "12,300"
## [349] "5,100"         "283,000"       "52,000"        "538,000"
## [353] "0.151"         "2.012"         "0.244"         "0.152"
## [357] "2.156"         "0.031"         "0.144"         "0.054"
## [361] "0.058"         "0.214"         "10.509"        "0.756"
## [365] "0.47"          "12.456"        "0.033"         "0.179"
## [369] "0.136"         "0.06"          "0.18"          "0.064"
## [373] "5.8"           "3.6"           "7.6"           "8,800"
## [377] "2,951,200"     "27.1"          "56,300"        "56,400"
## [381] "3,259,100,000" "28,520,000"    "1,426,000"     "25.33"
## [385] "3,112,100,000" "23,077,600"    "25,700"        "43.5"
## [389] "147,000,000"   "5,416,700"     "28,494,300"    "1,424,715"
## [393] "43,500"        "43,600"        "2,781,768,000" "25,700,000"
## [397] "25,674,300"    "477,332,000"   "2,820,000"     "52,500"
## [401] "3,583,960,000" "27,930,000"    "1,396,500"     "26.6"
## [405] "3,253,959,000" "22,697,900"    "75,300"        "44.1"
## [409] "330,001,000"   "5,156,800"     "27,854,700"    "1,392,735"
## [413] "40,200"        "3,132,279,000" "25,100,000"    "1,900"
## [417] "2,100"         "253,600"       "11,300"        "16,700"
## [421] "7,500"         "1,800"         "2,400"         "4,100"
## [425] "3,400"         "1,200"         "591,200"       "9,400"
## [429] "96,300"        "1,056,400"     "6,300"         "16,900"
## [433] "7,400"         "1,100"         "5,500"         "2,500"
## [437] "2,600"         "29,100"        "2,900"         "5,800"
## [441] "2,200"         "116,700"       "5,692,600"     "848,600"
## [445] "1,040,400"     "3,000"         "7,602,900"     "0.237"
## [449] "0.017"         "0.372"         "1.662"         "0.023"
## [453] "0.333"         "0.627"         "0.222"         "0.483"
## [457] "0.279"         "0.098"         "2.152"         "0.113"
## [461] "0.711"         "3.779"         "0.038"         "0.707"
## [465] "2.144"         "0.124"         "0.251"         "0.084"
## [469] "0.301"         "1.327"         "0.018"         "0.305"
## [473] "0.138"         "0.029"         "0.505"         "0.123"
## [477] "0.066"         "0.233"         "0.106"         "0.088"
## [481] "0.513"         "2.027"         "0.907"         "0.779"
## [485] "2.222"         "0.041"         "0.076"         "0.252"
## [489] "0.065"         "196.615"       "70.726"        "2.372"
## [493] "240.829"       "0.491"         "0.199"         "0.021"
## [497] "0.506"         "14.256"        "0.591"         "0.974"
## [501] "0.394"         "0.254"         "0.202"         "0.566"
## [505] "0.35"          "0.183"         "10.775"        "0.14"
## [509] "0.23"          "0.989"         "0.139"         "17.216"
## [513] "0.056"         "0.815"         "5.029"         "0.249"
## [517] "0.461"         "0.099"         "0.388"         "1.357"
## [521] "0.453"         "0.184"         "0.048"         "0.191"
```

```
## [525] "0.107"         "0.145"         "0.283"         "0.121"
## [529] "0.16"          "0.245"         "0.253"         "0.602"
## [533] "2.443"         "0.219"         "1.076"         "1.159"
## [537] "2.57"          "0.168"         "0.067"         "0.126"
## [541] "0.274"         "221.399"       "77.467"        "4.631"
## [545] "281.989"       "0.958"         "0.266"         "4.6"
## [549] "25,024,700"    "451,681,000"   "2,830,000"     "135,100"
## [553] "4,500"         "142,400"       "303,200"       "2.025"
## [557] "0.189"         "2.38"          "0.072"         "0.069"
## [561] "21.339"        "0.816"         "0.652"         "22.509"
## [565] "0.037"         "0.187"         "3.4"           "9.5"
## [569] "2.6"           "97.1"          "38.1"          "93.1"
## [573] "49,300"        "49,700"        "2,591,759,000" "26,740,000"
## [577] "1,337,000"     "27.12"         "2,404,152,000" "21,756,400"
## [581] "55,800"        "187,607,000"   "4,927,800"     "26,684,200"
## [585] "1,334,210"     "37,600"        "38,000"        "2,267,175,000"
## [589] "24,400,000"    "48,800"        "24,351,200"    "11,700"
## [593] "324,584,000"   "2,340,000"     "7,000"         "2,333,000"
## [597] "52.8"          "45,400"        "45,800"        "2,749,961,000"
## [601] "23,980,000"    "1,199,000"     "26.41"         "2,501,482,000"
## [605] "19,251,000"    "21,300"        "248,479,000"   "4,707,700"
## [609] "23,958,700"    "1,197,935"     "34,900"        "35,300"
## [613] "2,342,148,000" "21,300,000"    "283,500"       "16,600"
## [617] "15,000"        "11,100"        "34,200"        "9,300"
## [621] "1,400"         "1,600"         "4,800"         "686,000"
## [625] "90,900"        "1,233,500"     "4,200"         "10,200"
## [629] "17,900"        "7,300"         "5,200"         "6,000"
## [633] "3,500"         "56,700"        "12,700"        "51,600"
## [637] "1,000"         "212,600"       "6,146,600"     "1,045,700"
## [641] "254,000"       "244,000"       "7,698,900"     "668,000"
## [645] "493,000"       "430,000"       "57,000"        "0.205"
## [649] "0.482"         "0.337"         "1.739"         "0.32"
## [653] "0.623"         "0.213"         "3.997"         "0.493"
## [657] "0.303"         "0.073"         "2.381"         "0.51"
## [661] "0.097"         "3.656"         "0.039"         "0.669"
## [665] "2.219"         "0.024"         "0.089"         "0.313"
## [669] "1.202"         "0.297"         "0.034"         "0.111"
## [673] "0.135"         "0.286"         "0.103"         "0.085"
## [677] "0.177"         "0.515"         "1.991"         "0.904"
## [681] "5.151"         "0.478"         "0.044"         "0.068"
## [685] "0.243"         "244.066"       "83.557"        "214.011"
## [689] "228.063"       "0.447"         "0.54"          "0.051"
## [693] "0.508"         "10.456"        "0.765"         "0.25"
## [697] "1.163"         "0.224"         "0.361"         "18.136"
## [701] "1.215"         "0.281"         "0.625"         "0.149"
## [705] "2.908"         "0.883"         "0.264"         "26.646"
## [709] "1.041"         "5.064"         "0.235"         "0.55"
## [713] "0.074"         "0.217"         "0.428"         "1.84"
## [717] "0.04"          "0.091"         "0.978"         "0.346"
## [721] "0.581"         "0.531"         "0.128"         "0.399"
## [725] "0.269"         "0.793"         "5.571"         "1.665"
## [729] "0.334"         "7.924"         "0.192"         "0.884"
## [733] "0.108"         "0.307"         "0.137"         "0.164"
## [737] "345.893"       "125.71"        "0.188"         "244.72"
```

```
## [741] "303.425"       "0.231"         "4.5"           "2.7"
## [745] "7.3"           "3.2"           "4.3"           "3.5"
## [749] "9.9"           "11.3"          "14.3"          "8.2"
## [753] "21,278,700"    "10,500"        "407,813,000"   "2,680,000"
## [757] "12,000"        "146,100"       "2,700"         "54,800"
## [761] "286,900"       "125,900"       "320,000"       "86,000"
## [765] "389,000"       "0.858"         "2.098"         "0.206"
## [769] "1.34"          "2.036"         "0.045"         "0.141"
## [773] "0.05"          "4.123"         "16.856"        "0.642"
## [777] "4.888"         "10.166"        "0.127"         "0.663"
## [781] "0.212"         "4.8"           "4.7"           "28.7"
## [785] "18.6"          "25.2"          "93.5"          "43.7"
## [789] "90.9"          "42.2"          "49,320"        "49,800"
## [793] "2,472,447,000" "26,480,000"    "2,242,706,000" "21,196,100"
## [797] "25,100"        "229,741,000"   "5,258,800"     "284,740"
## [801] "26,454,900"    "36,300"        "2,179,066,000" "24,000,000"
## [805] "94,800"        "4,900"         "15,200"        "298,200"
## [809] "20,100"        "466,900"       "8,000"         "90,500"
## [813] "134,400"       "2,749,500"     "3,586,800"     "10,676,000"
## [817] "5,745,000"     "10,583,000"    "2,637,000"     "0.358"
## [821] "1.878"         "0.022"         "0.59"          "0.17"
## [825] "0.488"         "0.104"         "2.425"         "0.171"
## [829] "0.628"         "0.096"         "4.865"         "0.636"
## [833] "2.078"         "0.353"         "1.364"         "0.118"
## [837] "0.032"         "0.082"         "0.172"         "1.799"
## [841] "0.176"         "0.891"         "6.34"          "236.901"
## [845] "83.195"        "0.226"         "0.298"         "0.565"
## [849] "6.452"         "0.026"         "1.005"         "0.143"
## [853] "0.402"         "3.52"          "0.221"         "0.762"
## [857] "12.88"         "0.055"         "0.827"         "4.285"
## [861] "0.148"         "0.396"         "1.944"         "0.027"
## [865] "0.364"         "0.165"         "0.645"         "0.163"
## [869] "0.24"          "0.63"          "3.062"         "1.091"
## [873] "10.762"        "0.083"         "0.319"         "0.086"
## [877] "256.687"       "85.17"         "16.9"          "11.8"
## [881] "13.2"          "10.2"          "1,964,352,000" "18,888,000"
## [885] "214,714,000"   "5,088,000"     "276,358"       "23,976,000"
## [889] "243,800,000"   "2,120,000"     "87,400"        "22,500"
## [893] "128,400"       "102,700"       "136,000"       "42,000"
## [897] "173,000"       "1.858"         "0.33"          "1.632"
## [901] "9.193"         "0.997"         "0.501"         "6.049"
## [905] "4.9"           "3.7"           "17.6"          "6,065,000"
## [909] "30,000"        "29,800"        "22,230,000"    "130,000"
## [913] "11,687,000"    "110,000"       "109,300"       "12,324,000"
## [917] "58,100"        "15,027,000"    "170,800"       "9,599,000"
## [921] "90,000"        "89,800"        "49,920"        "50,600"
## [925] "3,371,461,000" "31,764,900"    "636.3"         "3,154,923,000"
## [929] "26,040,000"    "30,100"        "216,538,000"   "5,694,800"
## [933] "31,734,800"    "36,400"        "36,500"        "3,030,953,000"
## [937] "28,938,000"    "2,829,210,000" "23,381,900"    "28,900"
## [941] "201,743,000"   "5,527,200"     "28,909,100"    "10,000"
## [945] "292,050,000"   "2,475,000"     "6,229,000"     "30,800"
## [949] "30,600"        "21,375,000"    "125,000"       "12,288,000"
## [953] "14,795,000"    "167,600"       "9,167,000"     "86,100"
```

```
## [957] "85,800"
```

```
cleaned_data$Value <- as.numeric(gsub("[^0-9.]", "", cleaned_data$Value))

# Verify the structure and summary of the Value column
str(cleaned_data$Value)
```

```
##  num [1:12669] NA 3 NA 1 6 5 NA NA 2 2 ...
```

```
summary(cleaned_data$Value)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 0.000e+00 2.000e+00 4.000e+00 1.123e+07 2.100e+01 3.584e+09      4744
```

Based on the above, we see that there are a lot of NA values in the Value column. So now we will handle those missing values-

```
# Remove rows with NA in Value column
cleaned_data <- cleaned_data %>%
  filter(!is.na(Value))

# Replacing NA with the median value of the column
cleaned_data$Value[is.na(cleaned_data$Value)] <- median(cleaned_data$Value, na.rm = TRUE)
```

Now that our data is cleaned, we can move onto the next step.

Step 3- Exploratory Data Analysis

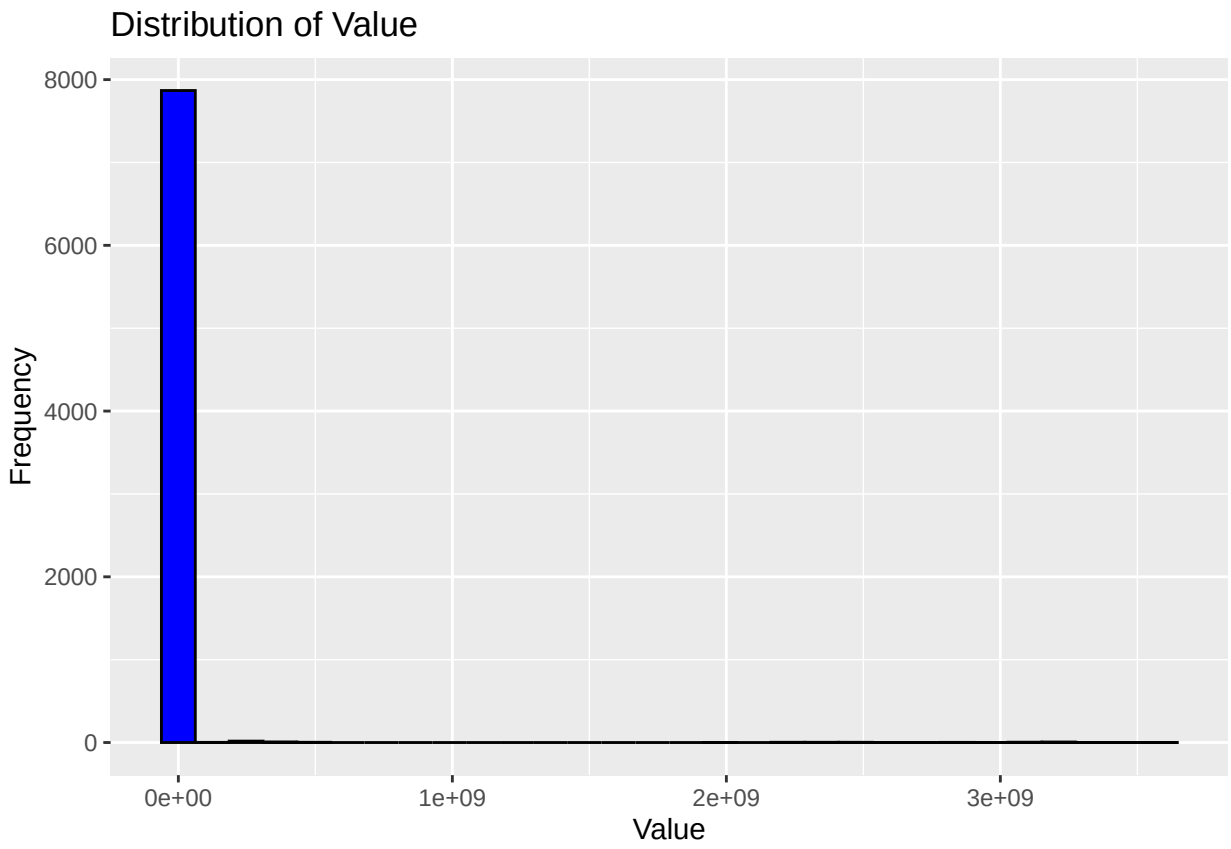
Summary statistics grouped by state, county, or year-

```
cleaned_data %>%
  group_by(State) %>%
  summarise(
    Min_Value = min(Value, na.rm = TRUE),
    Max_Value = max(Value, na.rm = TRUE),
    Mean_Value = mean(Value, na.rm = TRUE),
    Total_Value = sum(Value, na.rm = TRUE)
  )
```

```
## # A tibble: 52 x 5
##    State       Min_Value  Max_Value  Mean_Value  Total_Value
##    <chr>           <dbl>      <dbl>       <dbl>        <dbl>
##  1 ALABAMA         1            171       10.5         1076
##  2 ALASKA          1             50       10.5          251
##  3 ARIZONA         1             24        4.14         116
##  4 ARKANSAS        1            128        9.51         732
##  5 CALIFORNIA      0.017 3132279000 16059544.   25341960135.
##  6 COLORADO        1          31000      893.         62535
##  7 CONNECTICUT     1          40890     2355.        148344
##  8 DELAWARE        1             27        7.92         103
##  9 FLORIDA         0      477332000  6918550.    3279392684.
## 10 GEORGIA         1          11600      169.         28581
## # i 42 more rows
```

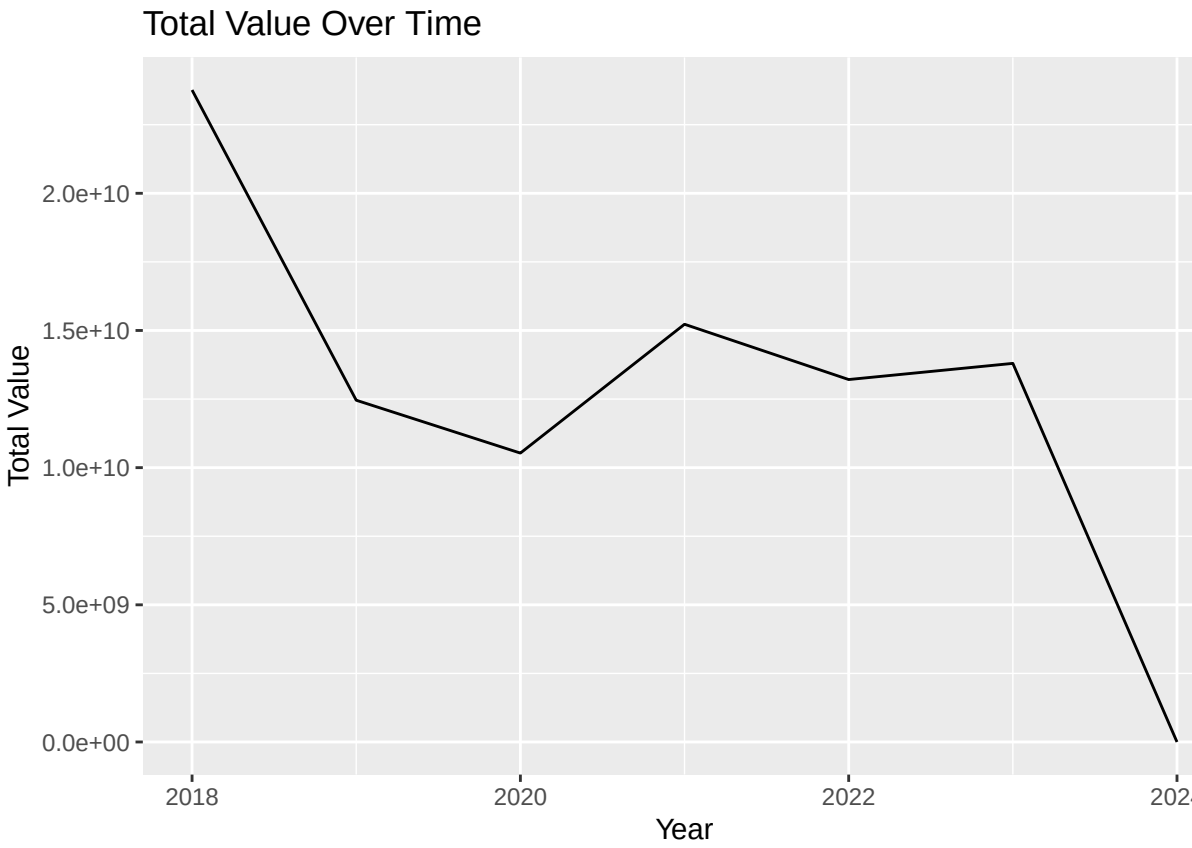
Visualising the data- Distribution Plot: We plot the distribution of the cleaned Value column to see its spread.

```
ggplot(cleaned_data, aes(x = Value)) +
  geom_histogram(bins = 30, fill = "blue", color = "black") +
  labs(title = "Distribution of Value", x = "Value", y = "Frequency")
```



he histogram demonstrates that the majority of the Value data is clustered near zero, with a frequency peak around very low values. The distribution is highly right-skewed, indicating the presence of a small number of extremely large values (likely outliers) that push the tail of the distribution far beyond the rest of the data. This kind of distribution suggests that most strawberry-producing regions contribute relatively small amounts to the total value, while a few regions dominate with very large contributions.

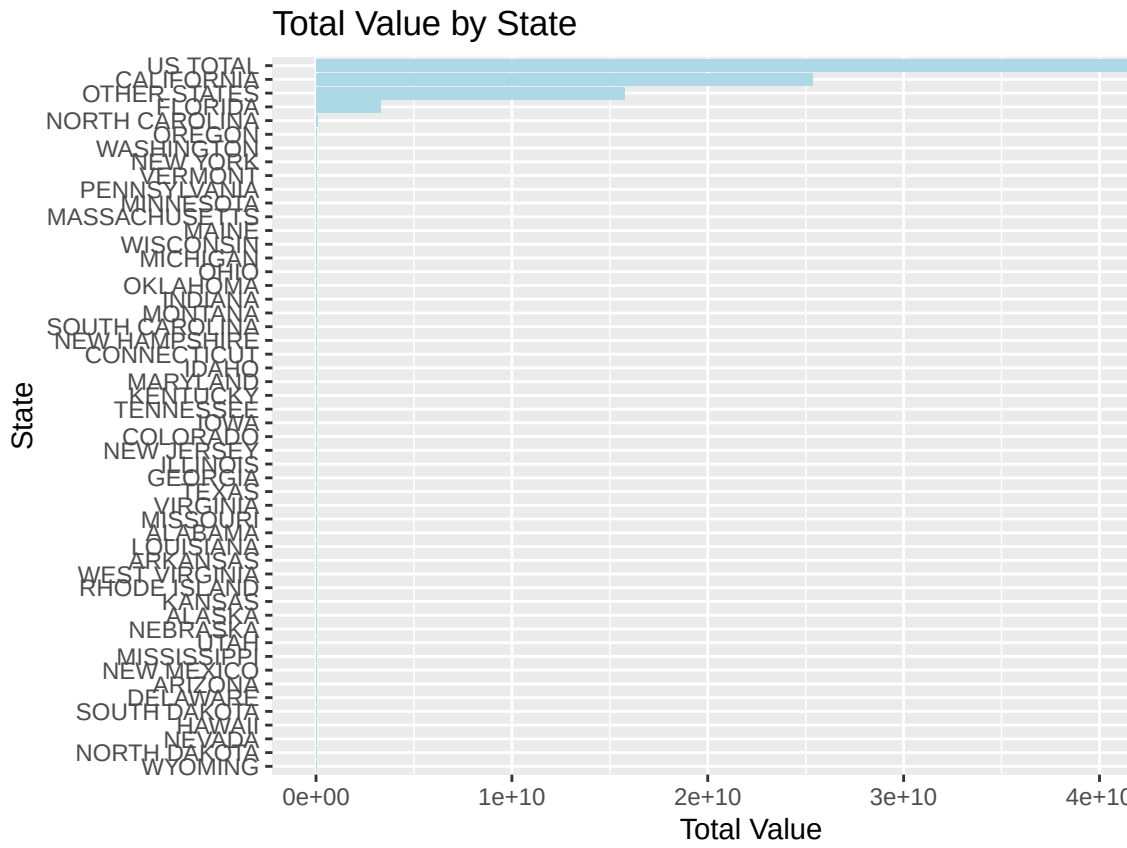
```
cleaned_data %>%
  group_by(Year) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE)) %>%
  ggplot(aes(x = Year, y = Total_Value)) +
  geom_line() +
  labs(title = "Total Value Over Time", x = "Year", y = "Total Value")
```



Time Series Plot-

The total value of strawberry production has experienced notable fluctuations from 2018 to 2024. There is a sharp decline between 2018 and 2020, possibly due to external factors like market demand, agricultural conditions, or economic downturns. From 2020 to 2022, the industry shows signs of recovery, with a sharp increase in total value. However, this growth is short-lived, as the value experiences another significant drop by 2024. These trends suggest the influence of external variables such as economic conditions, weather events, or shifts in production practices that could be explored in further analysis.

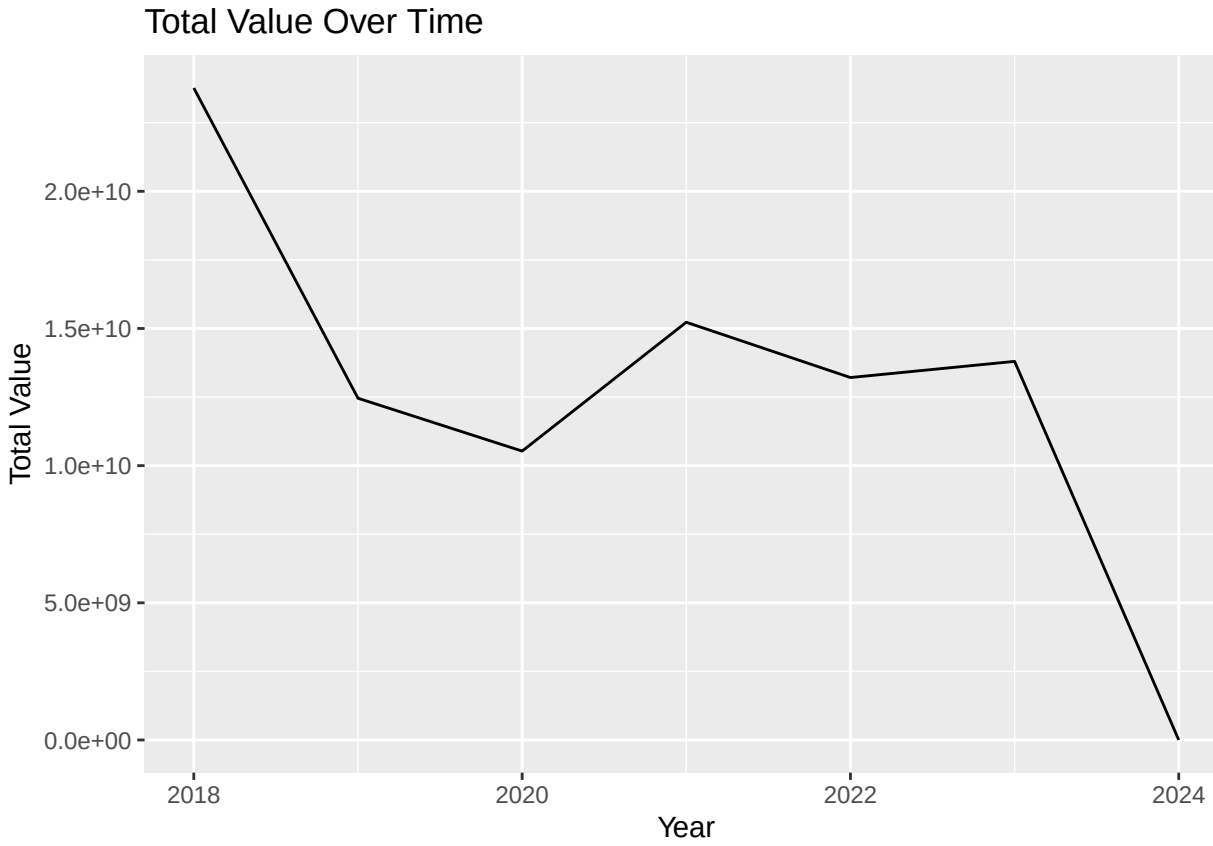
```
# Group by state and plot total value by state
state_summary <- cleaned_data %>%
  group_by(State) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE)) %>%
  arrange(desc(Total_Value))


# Bar plot of total value by state
ggplot(state_summary, aes(x = reorder(State, Total_Value), y = Total_Value)) +
  geom_bar(stat = "identity", fill = "lightblue") +
  coord_flip() +  # Flip for easier readability
  labs(title = "Total Value by State", x = "State", y = "Total Value")
```



Geographical analysis-

The bar plot highlights the dominance of California in strawberry production, as it contributes the largest total value by far, reaching levels much higher than any other state. Pennsylvania and North Carolina, while significant contributors, are a distant second and third, respectively. Most other states, including states like Washington, Oregon, and Florida, have much lower total values. This indicates that strawberry production is highly concentrated in a few regions, with California being the primary producer. This concentration may be due to favorable growing conditions, established infrastructure, or larger-scale farming operations in California compared to other states.

```
cleaned_data %>%
  group_by(Year) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE)) %>%
  ggplot(aes(x = Year, y = Total_Value)) +
  geom_line() +
  labs(title = "Total Value Over Time", x = "Year", y = "Total Value")
```



Trend Analysis-

The time series plot reveals considerable variation in the total value of strawberry production from 2018 to 2024. The total value drops between 2018 and 2020, which may be due to factors like changes in market demand, weather conditions, or production challenges. A sharp rise between 2020 and 2022 suggests a recovery phase, possibly spurred by favorable conditions or increased market demand. However, this growth is short-lived, as the total value declines steeply by 2024. These trends suggest that strawberry production or sales are sensitive to external influences, such as economic conditions, climate change, or policy shifts.

Conclusion-

I think I got a bit confused with the assignment in general. I couldn't find any chemical column in the data so I just went on and did some data cleaning and visualisations that felt suitable.