Chemical EDA

Suheng Yao

2024-10-19

```
# Read in the cleaned dataset and select records of California and Florida
df <- read.csv("strawberry_cleaned.csv")

# Find states that used chemicals in growing strawberry
states_with_chemical_info <- df %>%
  filter(!is.na(Chemical_Info)) %>%
  distinct(State) %>%
  pull(State)

print(states_with_chemical_info)
```

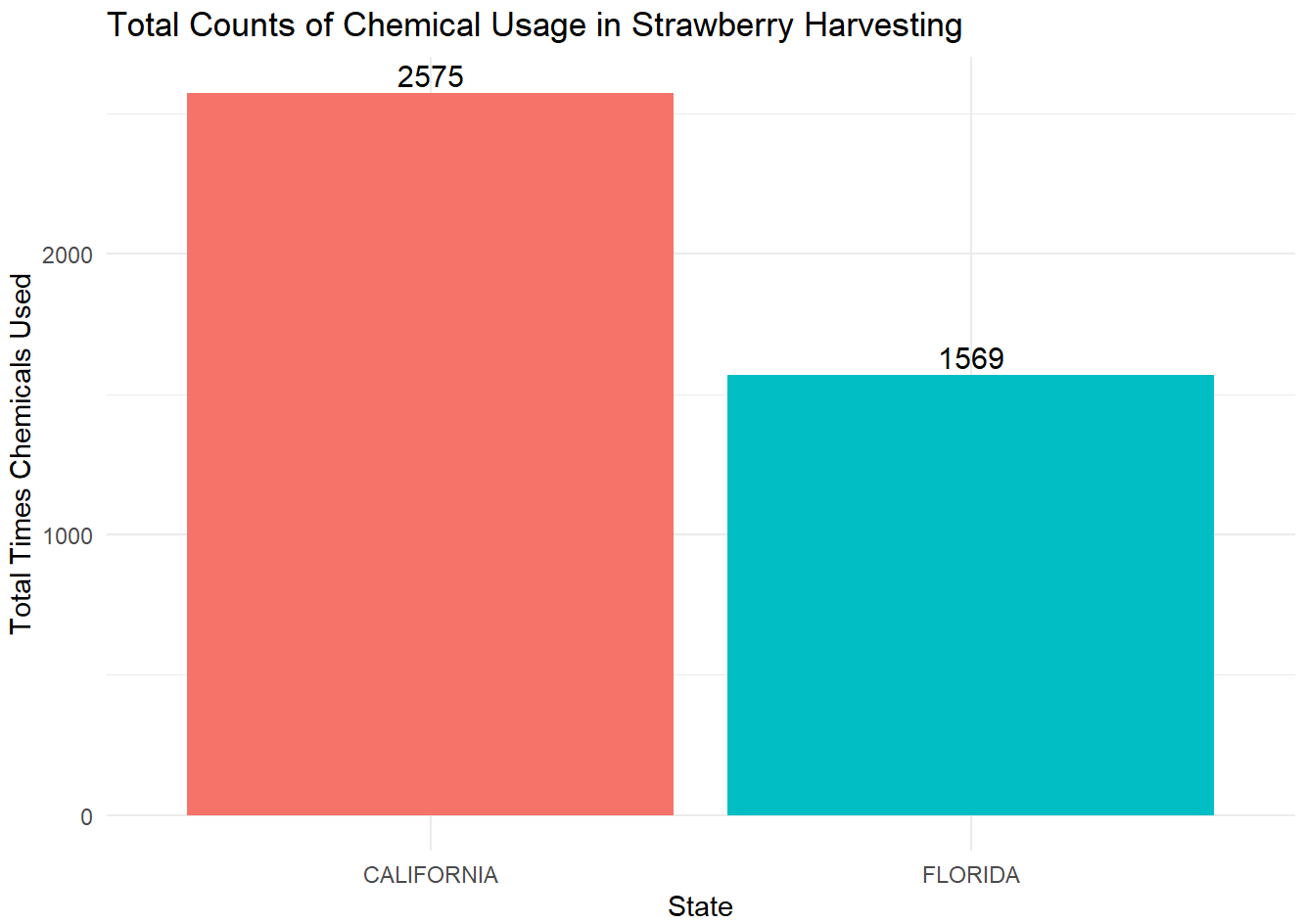
```
## [1] "CALIFORNIA" "FLORIDA"
```

So in the dataset, we only get chemical information related to California and Florida, we can filter out only those two states and compare the frequency of chemical used:

```
df1 <- df %>%
  select(-1) %>%
  filter(State == "CALIFORNIA" | State == "FLORIDA")

# Draw a bar plot comparing the frequency
chem_state <- df1 %>%
  group_by(State) %>%
  summarise(Total_obs=n()) %>%
  arrange(desc(Total_obs))
print(chem_state)
```

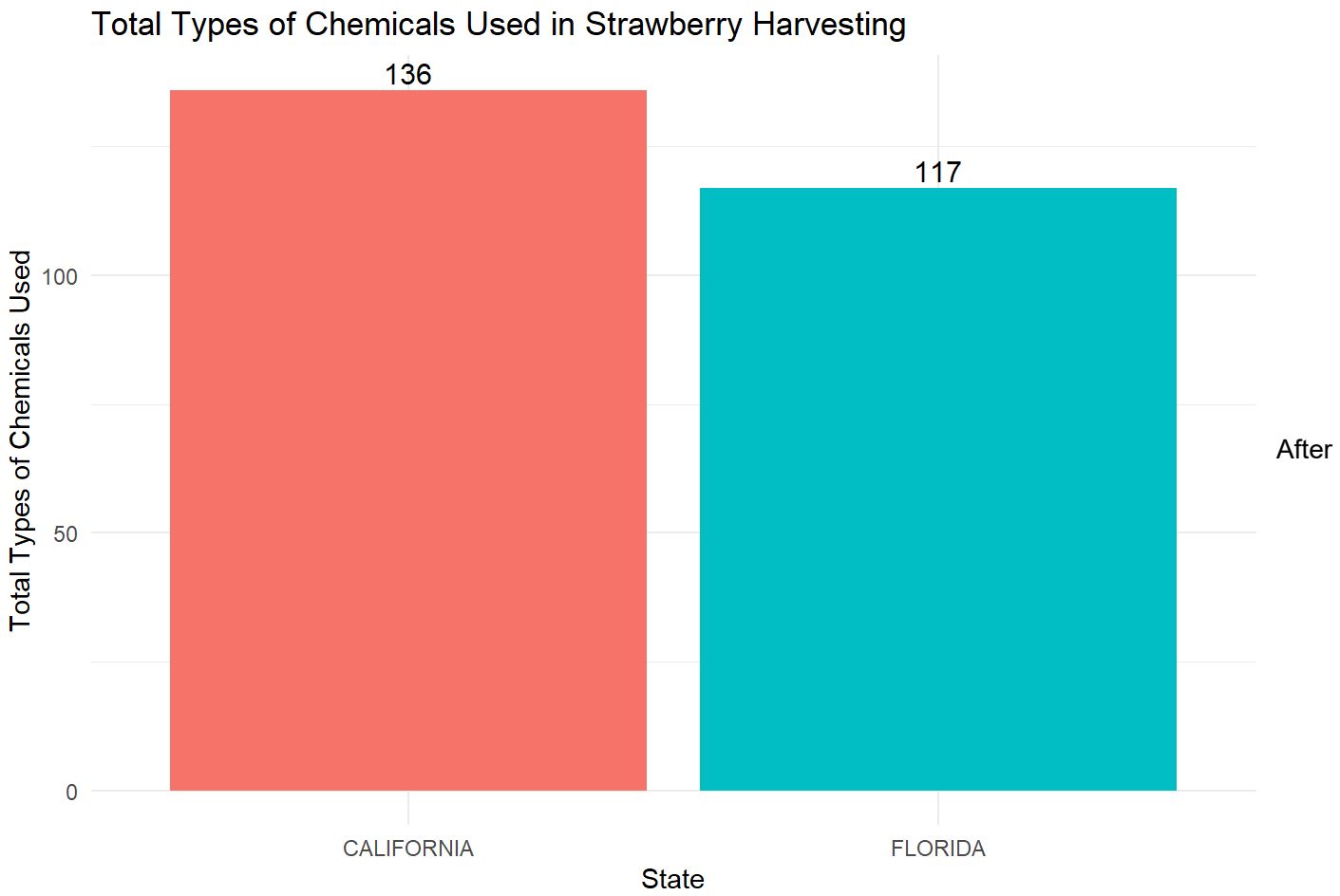
```
## # A tibble: 2 × 2
##   State      Total_obs
##   <chr>          <int>
## 1 CALIFORNIA      2575
## 2 FLORIDA         1569
```

```
ggplot(chem_state, aes(x = State, y = Total_obs, fill = State)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Total_obs), vjust = -0.3, size = 4) +
  theme_minimal() +
  labs(title = "Total Counts of Chemical Usage in Strawberry Harvesting",
       x = "State",
       y = "Total Times Chemicals Used") +
  theme(legend.position = "none")  # Remove legend
```



The frequency that California used chemicals is almost twice as the frequency of Florida.

```
# Split the California data and Florida data into two dataframes
df_cal <- df %>%
  filter(State == "CALIFORNIA")

df_flo <- df %>%
  filter(State == "FLORIDA")

# Find the distinct chemicals used in California and Florida
chem_comparison <- data.frame(
  State = c("CALIFORNIA", "FLORIDA"),
  Distinct_chemicals =
    c(length(unique(df_cal$Chemical_Info)),
      length(unique(df_flo$Chemical_Info)))
)


ggplot(chem_comparison, aes(x = State, y = Distinct_chemicals, fill = State)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = Distinct_chemicals), vjust = -0.3, size = 4) +
  theme_minimal() +
  labs(title = "Total Types of Chemicals Used in Strawberry Harvesting",
       x = "State",
       y = "Total Types of Chemicals Used") +
  theme(legend.position = "none")  # Remove Legend
```



finding the distinct chemicals in each state, California used 19 more types of chemicals than Florida, and there are 58 types of chemicals used in California but not in Florida:

```
# Find out chemicals used by California but not used by Florida
diff_chem1 <- setdiff(unique(df_cal$Chemical_Info), unique(df_flo$Chemical_Info))
print(diff_chem1)
```

```
##  [1] "CYCLANILIPROLE"
##  [2] "PERMETHRIN"
##  [3] "ISARIA FUMOSOROSEA STRAIN FE 9901"
##  [4] "BACILLUS AMYLOLIQUEFACIENS STRAIN D747"
##  [5] "BLAD"
##  [6] "BT SUBSP KURSTAKI EVB-113-19"
##  [7] "POLYOXIN D ZINC SALT"
##  [8] "QUINOLINE"
##  [9] "TRIFLOXYSTROBIN"
## [10] "PENDIMETHALIN"
## [11] "ACEQUINOCYL"
## [12] "AZADIRACHTIN"
## [13] "BEAUVERIA BASSIANA"
## [14] "BT KURSTAKI SA-11"
## [15] "CANOLA OIL"
## [16] "CHROMOBAC SUBTSUGAE PRAA4-1 CELLS AND SPENT MEDIA"
## [17] "ETOXAZOLE"
## [18] "FENBUTATIN-OXIDE"
## [19] "NEEM OIL"
## [20] "NEEM OIL, CLAR. HYD."
## [21] "PYRIDABEN"
## [22] "CAPSICUM OLEORESIN EXTRACT"
## [23] "GARLIC OIL"
## [24] "IRON PHOSPHATE"
## [25] "METALDEHYDE"
## [26] "METAM-SODIUM"
## [27] "BACILLUS AMYLOLIQUEFACIENS MBI 600"
## [28] "BACILLUS PUMILUS"
## [29] "COPPER OCTANOATE"
## [30] "POTASSIUM BICARBON."
## [31] "STREPTOMYCES LYDICUS"
## [32] "BT KURSTAKI EG7841"
## [33] "BT SUB AIZAWAI GC-91"
## [34] "BUPROFEZIN"
## [35] "BURKHOLDERIA A396 CELLS & MEDIA"
## [36] "HELICOVERPA ZEA NPV"
## [37] "PETROLEUM DISTILLATE"
## [38] "POTASSIUM SALTS"
## [39] "PYRIPROXYFEN"
## [40] "CAPRIC ACID"
## [41] "CAPRYLIC ACID"
## [42] "MINERAL OIL"
## [43] "PAECILOMYCES FUMOSOR"
## [44] "POTASSIUM SILICATE"
## [45] "BACILLUS SUBT. GB03"
## [46] "TRICHODERMA HARZ."
## [47] "GLUFOSINATE-AMMONIUM"
## [48] "SULFENTRAZONE"
## [49] "CHLORPYRIFOS"
## [50] "SOYBEAN OIL"
## [51] "ZETA-CYPERMETHRIN"
## [52] "AUREOBASIDIUM PULLULANS DSM 14940"
```

```
## [53] "AUREOBASIDIUM PULLULANS DSM 14941"
## [54] "BT KURSTAKI SA-12"
## [55] "GLIOCLADIUM VIRENS"
## [56] "TRICHODERMA VIRENS STRAIN G-41"
## [57] "EMAMECTIN BENZOATE"
## [58] "SPIROTETRAMAT"
```

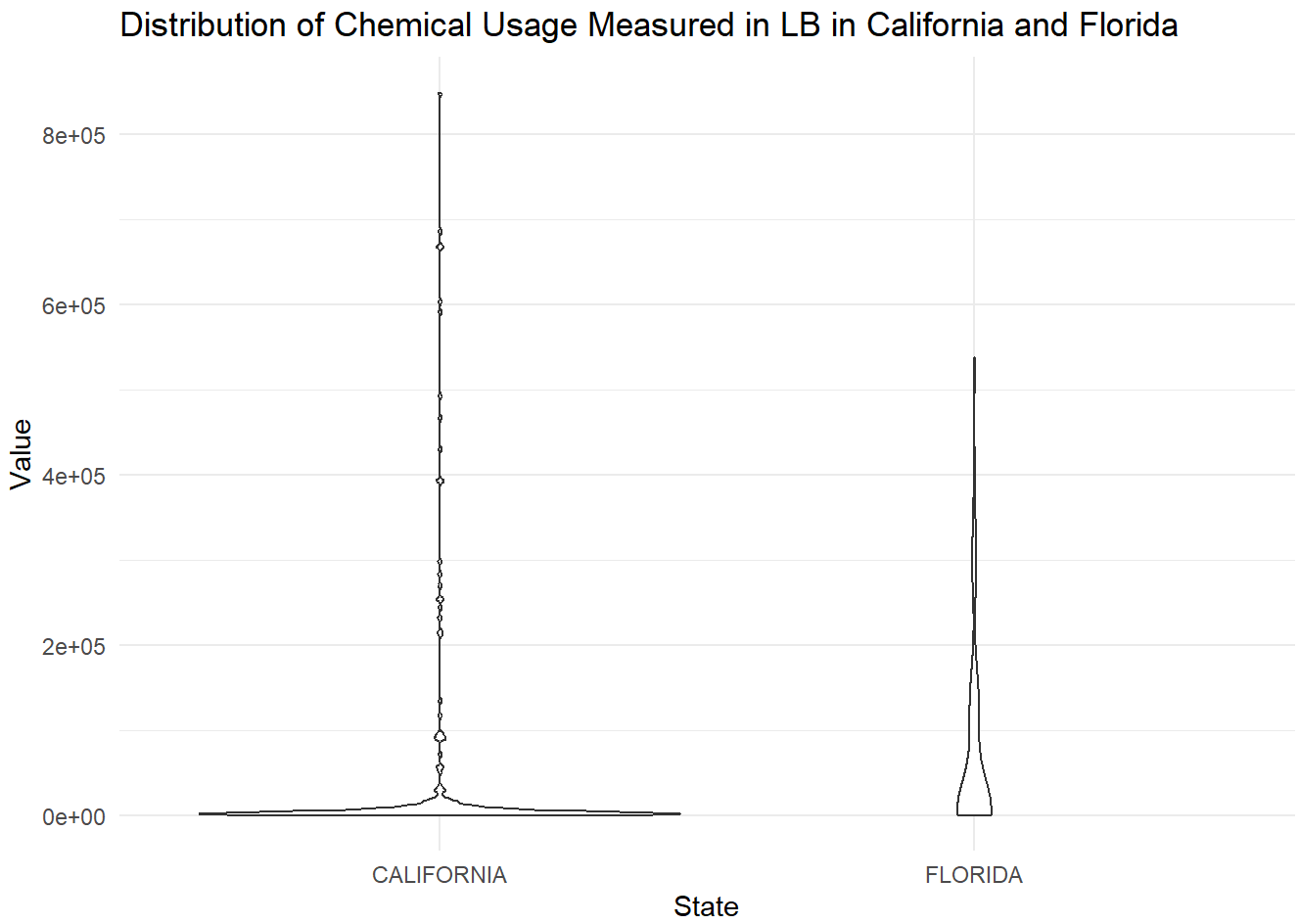
on the other hand, there are 39 types of chemicals used in Florida but not in California:

```
diff_chem2 <- setdiff(unique(df_flo$Chemical_Info), unique(df_cal$Chemical_Info))
print(diff_chem2)
```

```
##  [1] "PYRIOFENONE"              "ZOXAMIDE"
##  [3] "METSULFURON-METHYL"       "PENOXSULAM"
##  [5] "S-METOLACHLOR"            "BETA-CYFLUTHRIN"
##  [7] "ETHYL 2E;4Z-DECADIENOATE" "OXAMYL"
##  [9] "CUPRAMMONIUM ACETATE"     "DODECADIEN-1-OL"
## [11] "FLUENSULFONE"             "GIBBERELLIC ACID"
## [13] "CHLOROTHALONIL"           "COPPER CHLORIDE HYD."
## [15] "CYMOXANIL"                "FAMOXADONE"
## [17] "MANCOZEB"                 "2,4-D, DIMETH. SALT"
## [19] "CLETHODIM"                "METHOMYL"
## [21] "CYTOKININS"               "INDOLEBUTYRIC ACID"
## [23] "COPPER ETHANOLAMINE"      "DIMETHENAMID"
## [25] "FLUROXYPYR 1-MHE"         "HALOSULFURON-METHYL"
## [27] "KANTOR"                   "FENAZAQUIN"
## [29] "ETHEPHON"                 "DODINE"
## [31] "FLUTOLANIL"               "2,4-D, TRIISO. SALT"
## [33] "CYPERMETHRIN"             "ALKYL. DIM. BENZ. AM"
## [35] "DECYLDIMETHYLOCTYL"       "DIDECYL DIM. AMMON."
## [37] "DIMETHYLDIOCTYL"          "MUSTARD OIL"
## [39] "DIMETHYL DISULFIDE DMDS"
```

We can also show the distribution of the amount of chemical used in California and Florida through a violin plot:

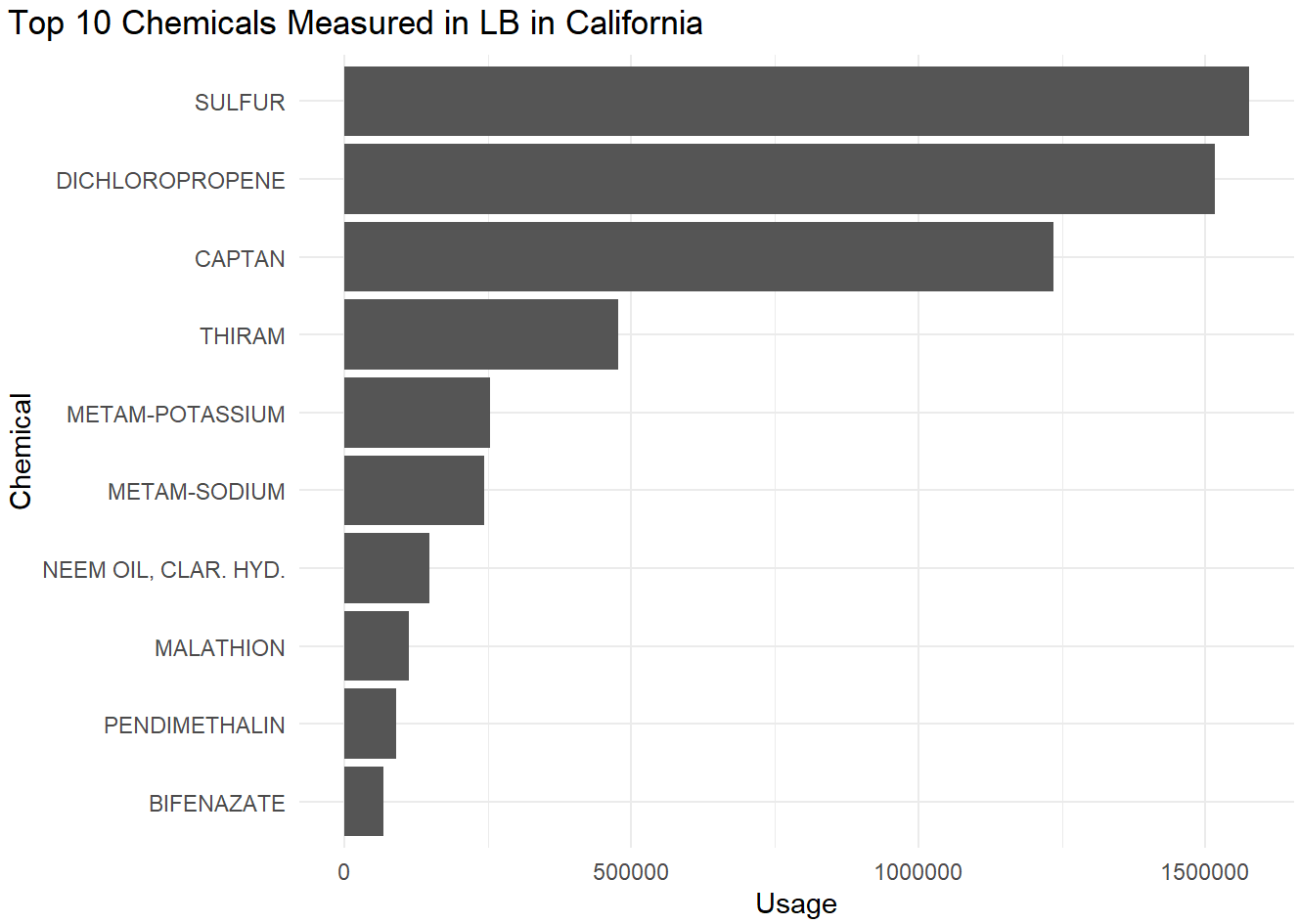
```
df1 <- df1 %>%
  filter(!is.na(Value)) %>%
  filter(Measurement == "MEASURED IN LB")

ggplot(df1, aes(x=State, y=Value)) +
  geom_violin() +
  theme_minimal() +
  labs(title="Distribution of Chemical Usage Measured in LB in California and Florida")
```



From this plot, when measurement of chemicals are in LB, California have more outliers than Florida, meaning that the distribution of chemical usage in California is more spread out, and Florida's distribution is more concentrated. Also, both California and Florida have concentrated values at bottom, meaning that most of chemicals usage are quite low.

We can first explore top chemicals used in California:

```
# Create a new column to record the frequency of each chemical used
chem_measure <- df_cal %>%
  group_by(Measurement) %>%
  filter(Measurement == "MEASURED IN LB") %>%
  filter(Chemical_Info != "TOTAL")

# Plot the top chemicals used in California
top_chem <- chem_measure %>%
  filter(!is.na(Value)) %>%
  filter(!is.na(Chemical_Info)) %>%
  group_by(Chemical_Info, Domain) %>%
  summarise(Value = sum(Value, na.rm = TRUE)) %>%
  arrange(desc(Value))

ggplot(top_chem[1:10,],
       aes(x=reorder(Chemical_Info, Value), y=Value))+
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  labs(x = "Chemical", y = "Usage",
       title = "Top 10 Chemicals Measured in LB in California")+
  theme(plot.title.position = "plot")
```



```
domain_chem <- chem_measure %>%
  filter(!is.na(Value)) %>%
  group_by(Domain) %>%
  summarise(count = n()) %>%
  arrange(desc(count))
print(domain_chem)
```

```
## # A tibble: 4 × 2
##   Domain                count
##   <chr>                 <int>
## 1 CHEMICAL, INSECTICIDE   100
## 2 CHEMICAL, FUNGICIDE      98
## 3 CHEMICAL, OTHER          14
## 4 CHEMICAL, HERBICIDE      12
```

From the table above, the insecticide group has the most types of chemicals.

Plot above shows the most used chemicals in LB in strawberry harvesting, we can further explore the properties of those chemicals:

```
# Find the index of GHS classification
GHS_searcher<-function(result_json_object){
  # check if the chemicals in the database first
  if (is.null(result_json_object) ||
      is.null(result_json_object[["result"]]) ||
      is.null(result_json_object[["result"]][["Hierarchies"]]) ||
      is.null(result_json_object[["result"]][["Hierarchies"]][["Hierarchy"]])){
    return("did not find the chemical in the database")
  }

  result<-result_json_object
  for (i in 1:length(result[["result"]][["Hierarchies"]][["Hierarchy"]])){
    if(result[["result"]][["Hierarchies"]][["Hierarchy"]][[i]][["SourceName"]]=="GHS Classificat
ion (UNECE)"){
      return(i)
    }

  }
}

# Use the output from GHS_searcher to access the hazard information
hazards_retriever<-function(index,result_json_object){

  # Check if GHS_searcher did not find the index
  if (is.character(index) && index == "did not find the chemical in the database") {
    return(index)
  }

  result<-result_json_object
  hierarchy<-result[["result"]][["Hierarchies"]][["Hierarchy"]][[index]]
  i<-1
  output_list<-rep(NA,length(hierarchy[["Node"]]))
  while(str_detect(hierarchy[["Node"]][[i]][["Information"]][["Name"]],"H") & i<length(hierarchy
[["Node"]])){
    output_list[i]<-hierarchy[["Node"]][[i]][["Information"]][["Name"]]
    i<-i+1
  }
  return(output_list[!is.na(output_list)])
}

# Find the chemical information for the top chemicals
access_hazard <- function(chemical){
  results <- list()
  for (chem in chemical){
      result <- get_pug_rest(identifier = chem,
                             namespace = "name",
                             domain = "compound",
                             operation="classification",
                             output = "JSON")
      ghs_index <- GHS_searcher(result)
      hazards <- hazards_retriever(ghs_index, result)
      results[[chem]] <- list(
```

```
        chemical_name = chem,
        chemical_hazards = ifelse(hazards == "did not find the chemical in the database", charac
ter(0), hazards)
      )

  }

  return(results)
}
```

```
chem_vec <- top_chem[1:10,]$Chemical_Info
hazard_info <- access_hazard(chem_vec)
hazards_df <- do.call(rbind, lapply(hazard_info, function(x) {
  data.frame(
    chemical_name = x$chemical_name,
    hazards = paste(x$chemical_hazards, collapse = ";"),
    num_hazards = length(x$chemical_hazards) - 1,
    stringsAsFactors = FALSE
  )
}))


hazards_df %>%
  flextable() %>%
  theme_vanilla() %>%
  fontsize(size = 10) %>%
  width(j = "chemical_name", width = 2.5) %>%  # Set width for first column
  width(j = "hazards", width = 5) %>%  # Set width for second column
  align(align = "left", part = "all") %>%
  set_table_properties(layout = "autofit")
```

chemical_name	hazards	num_hazards
SULFUR	H228: Flammable solid [Danger Flammable solids];H200: Physical Hazards;Hazard Statement Codes;H315: Causes skin irritation [Warning Skin corrosion/irritation];H300: Health Hazards;H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H373: May causes damage to organs through prolonged or repeated exposure [Warning Specific target organ toxicity, repeated exposure];H413: May cause long lasting harmful effects to aquatic life [Hazardous to the aquatic environment, long-term hazard];H400: Environmental Hazards	10
DICHLOROPROPENE	NA	0
CAPTAN	H303: May be harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H315: Causes skin irritation [Warning Skin corrosion/irritation];H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H318: Causes serious eye damage [Danger Serious eye damage/eye irritation];H330: Fatal if inhaled [Danger Acute toxicity, inhalation];H331: Toxic if inhaled [Danger Acute toxicity, inhalation];H340: May cause genetic defects [Danger Germ cell mutagenicity];H351: Suspected of causing cancer [Warning	15

chemical_name	hazards	num_hazards
	Carcinogenicity];H361: Suspected of damaging fertility or the unborn child [Warning Reproductive toxicity];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	
THIRAM	H302: Harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H302+H332: Harmful if swallowed or if inhaled [Warning Acute toxicity, oral; acute toxicity, inhalation];H315: Causes skin irritation [Warning Skin corrosion/irritation];H316: Causes mild skin irritation [Warning Skin corrosion/irritation];H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation];H320: Causes eye irritation [Warning Serious eye damage/eye irritation];H330: Fatal if inhaled [Danger Acute toxicity, inhalation];H332: Harmful if inhaled [Warning Acute toxicity, inhalation];H340: May cause genetic defects [Danger Germ cell mutagenicity];H341: Suspected of causing genetic defects [Warning Germ cell mutagenicity];H361: Suspected of damaging fertility or the unborn child [Warning Reproductive toxicity];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H373: May causes damage to organs through prolonged or repeated exposure [Warning Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	19
METAM-POTASSIUM	H302: Harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H312: Harmful in contact with skin [Warning Acute toxicity, dermal];H314: Causes severe skin burns and eye damage [Danger Skin corrosion/irritation];H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H332: Harmful if inhaled [Warning Acute toxicity, inhalation];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	9
METAM-SODIUM	H302: Harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H311: Toxic in contact with skin [Danger Acute toxicity, dermal];H312: Harmful in contact with skin [Warning Acute toxicity, dermal];H314: Causes severe skin burns and eye damage [Danger Skin corrosion/irritation];H315: Causes skin irritation [Warning Skin corrosion/irritation];H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H318: Causes serious eye damage [Danger Serious eye damage/eye irritation];H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation];H320: Causes eye irritation [Warning Serious eye damage/eye irritation];H332: Harmful if inhaled [Warning Acute toxicity, inhalation];H335: May cause respiratory irritation [Warning Specific target organ toxicity, single exposure; Respiratory tract irritation];H351: Suspected of causing cancer [Warning Carcinogenicity];H360: May damage fertility or the unborn child [Danger Reproductive toxicity];H361: Suspected of damaging fertility or the unborn child [Warning Reproductive toxicity];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H371: May cause	21

chemical_name	hazards	num_hazards
	damage to organs [Warning Specific target organ toxicity, single exposure];H373: May causes damage to organs through prolonged or repeated exposure [Warning Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	
NEEM OIL, CLAR. HYD.	NA	0
MALATHION	H302: Harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H320: Causes eye irritation [Warning Serious eye damage/eye irritation];H331: Toxic if inhaled [Danger Acute toxicity, inhalation];H341: Suspected of causing genetic defects [Warning Germ cell mutagenicity];H350: May cause cancer [Danger Carcinogenicity];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H373: May causes damage to organs through prolonged or repeated exposure [Warning Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	13
PENDIMETHALIN	H302: Harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H351: Suspected of causing cancer [Warning Carcinogenicity];H361: Suspected of damaging fertility or the unborn child [Warning Reproductive toxicity];H361d: Suspected of damaging the unborn child [Warning Reproductive toxicity];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	10
BIFENAZATE	H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H300: Health Hazards;Hazard Statement Codes;H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation];H320: Causes eye irritation [Warning Serious eye damage/eye irritation];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H373: May causes damage to organs through prolonged or repeated exposure [Warning Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard];Hazardous to the aquatic environment, acute hazard;Environmental Hazards;Hazard Classes;Hazardous to the aquatic environment, long-term hazard	13

Through the chemical information presented in the above table, eight out of the top ten chemicals have hazards information. Especially for Metam-sodium, which has 21 kinds of harzards information, and it can also cause organ damage with single exposure. Also, most of those chemicals are very toxic to aquatic life, which will harm the

sustainable development of the environment. Additionally, three of the top used chemicals belong to insecticide group, three of them belong to the fungicide group, another three belong to the other group, including metam-sodium, and one belongs to the herbicide group. This suggests that many of the chemicals used in California are distributed into pest control or fungi control. This phenomenon could be the result of large production of strawberry and warm and hot weather in California.

Now, we can further analyze the top chemicals used in strawberry growing in Florida:

```
# Select records with Florida only chemicals
df_measure_lb <- df_flo %>%
  filter(Measurement=="MEASURED IN LB") %>%
  filter(!is.na(Value)) %>%
  filter(Chemical_Info != "TOTAL")

df_measure_lb1 <- df_flo %>%
  filter(Measurement=="MEASURED IN LB") %>%
  filter(Chemical_Info != "TOTAL")

print(length(df_measure_lb$Value))
```

```
## [1] 35
```

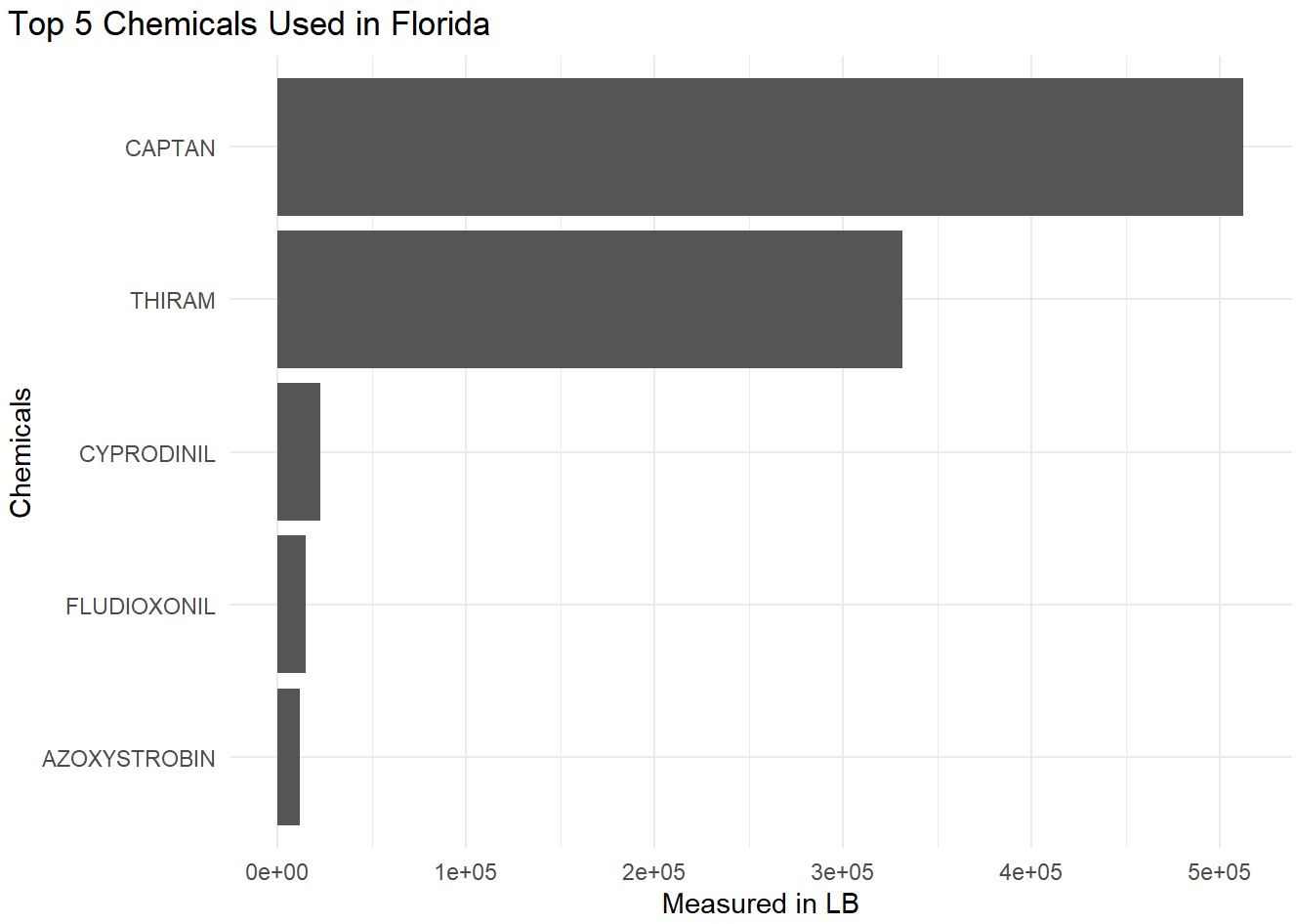
```
print(length(df_measure_lb1$Value))
```

```
## [1] 225
```

Originally, when chemical measurement is in LB, there are 225 observations, but when filtering out the records with empty “Value”, there are only 35 observations left. Since there are too many NAs in Value column, I try to build a linear regression based on the observations with non-NA Value column and use this model to predict the NAs in the Value column. However, the model failed to predict the values for two reasons: first of all, the proportion of NA is too large, which makes the result inaccurate. Secondly, in the model, Chemical_Info is an important variable, which is treated as factors with different levels, but in prediction process, there are new chemicals used in Florida, which creates new levels that the model haven’t seen before. Thus, the model cannot impute meaningful values into data, and I can only working with the non-NA values already existing in the dataset.

Here are the top five chemicals that are used most in Florida when measurement in LB:

```
# Create a new column to record the frequency of each chemical used
top_chem <- df_measure_lb %>%
  group_by(Chemical_Info, Domain) %>%
  summarise(Value = sum(Value, na.rm = TRUE)) %>%
  arrange(desc(Value))

# Plot the top 10 chemicals used in California
ggplot(top_chem[1:5,], aes(x=reorder(Chemical_Info, Value), y=Value))+
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  labs(x = "Chemicals", y = "Measured in LB",
       title = "Top 5 Chemicals Used in Florida")+
  theme(plot.title.position = "plot")
```



```
chem_vec_flo <- top_chem[1:5,]$Chemical_Info
hazard_info_flo <- access_hazard(chem_vec_flo)

hazards_df_flo <- do.call(rbind, lapply(hazard_info_flo, function(x) {
  data.frame(
    chemical_name = x$chemical_name,
    hazards = paste(x$chemical_hazards, collapse = ";"),
    num_hazards = length(x$chemical_hazards) - 1,
    stringsAsFactors = FALSE
  )
}))

hazards_df_flo %>%
  flextable() %>%
  theme_vanilla() %>%
  fontsize(size = 10) %>%
  width(j = "chemical_name", width = 2.5) %>%  # Set width for first column
  width(j = "hazards", width = 5) %>%  # Set width for second column
  align(align = "left", part = "all") %>%
  set_table_properties(layout = "autofit")
```

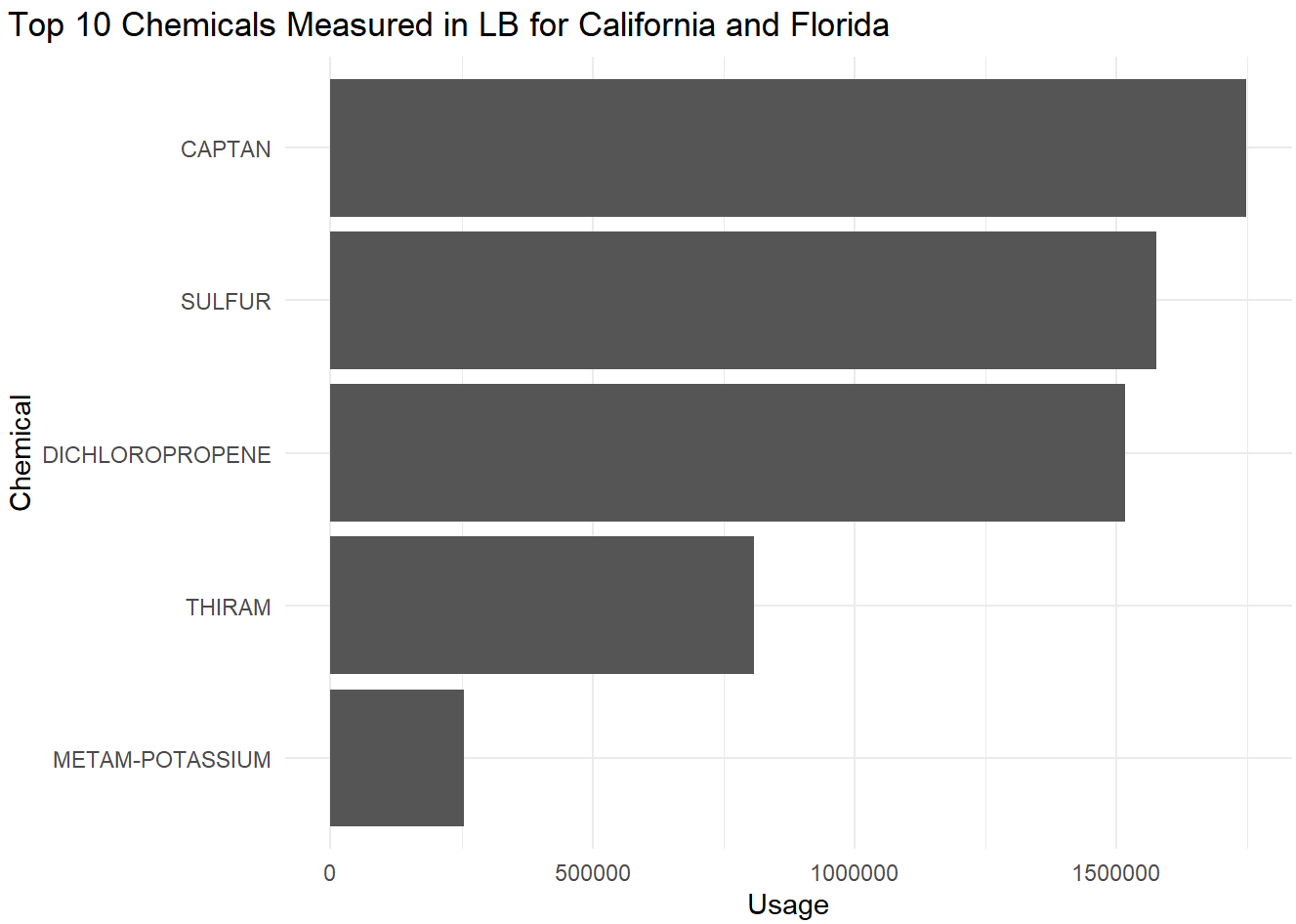
chemical_name	hazards	num_hazards
CAPTAN	H303: May be harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H315: Causes skin irritation [Warning Skin corrosion/irritation];H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H318: Causes serious eye damage [Danger Serious eye damage/eye irritation];H330: Fatal if inhaled [Danger Acute toxicity, inhalation];H331: Toxic if inhaled [Danger Acute toxicity, inhalation];H340: May cause genetic defects [Danger Germ cell mutagenicity];H351: Suspected of causing cancer [Warning Carcinogenicity];H361: Suspected of damaging fertility or the unborn child [Warning Reproductive toxicity];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	15
THIRAM	H302: Harmful if swallowed [Warning Acute toxicity, oral];H300: Health Hazards;Hazard Statement Codes;H302+H332: Harmful if swallowed or if inhaled [Warning Acute toxicity, oral; acute toxicity, inhalation];H315: Causes skin irritation [Warning Skin corrosion/irritation];H316: Causes mild skin irritation [Warning Skin corrosion/irritation];H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation];H320: Causes eye irritation [Warning Serious eye damage/eye irritation];H330: Fatal if inhaled [Danger Acute toxicity, inhalation];H332: Harmful if inhaled [Warning Acute toxicity, inhalation];H340: May cause genetic defects [Danger Germ cell mutagenicity];H341: Suspected of causing genetic defects [Warning Germ cell mutagenicity];H361: Suspected of damaging fertility or the unborn child [Warning Reproductive toxicity];H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H373: May causes damage to organs through	19

chemical_name	hazards	num_hazards
	prolonged or repeated exposure [Warning Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	
CYPRODINIL	H315: Causes skin irritation [Warning Skin corrosion/irritation];H300: Health Hazards;Hazard Statement Codes;H317: May cause an allergic skin reaction [Warning Sensitization, Skin];H319: Causes serious eye irritation [Warning Serious eye damage/eye irritation];H372: Causes damage to organs through prolonged or repeated exposure [Danger Specific target organ toxicity, repeated exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H401: Toxic to aquatic life [Hazardous to the aquatic environment, acute hazard];H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard];H411: Toxic to aquatic life with long lasting effects [Hazardous to the aquatic environment, long-term hazard];Hazardous to the aquatic environment, acute hazard;Environmental Hazards;Hazard Classes;Hazardous to the aquatic environment, long-term hazard	14
FLUDIOXONIL	H320: Causes eye irritation [Warning Serious eye damage/eye irritation];H300: Health Hazards;Hazard Statement Codes;H351: Suspected of causing cancer [Warning Carcinogenicity];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	6
AZOXYSTROBIN	H331: Toxic if inhaled [Danger Acute toxicity, inhalation];H300: Health Hazards;Hazard Statement Codes;H370: Causes damage to organs [Danger Specific target organ toxicity, single exposure];H400: Very toxic to aquatic life [Warning Hazardous to the aquatic environment, acute hazard];H400: Environmental Hazards;H410: Very toxic to aquatic life with long lasting effects [Warning Hazardous to the aquatic environment, long-term hazard]	6

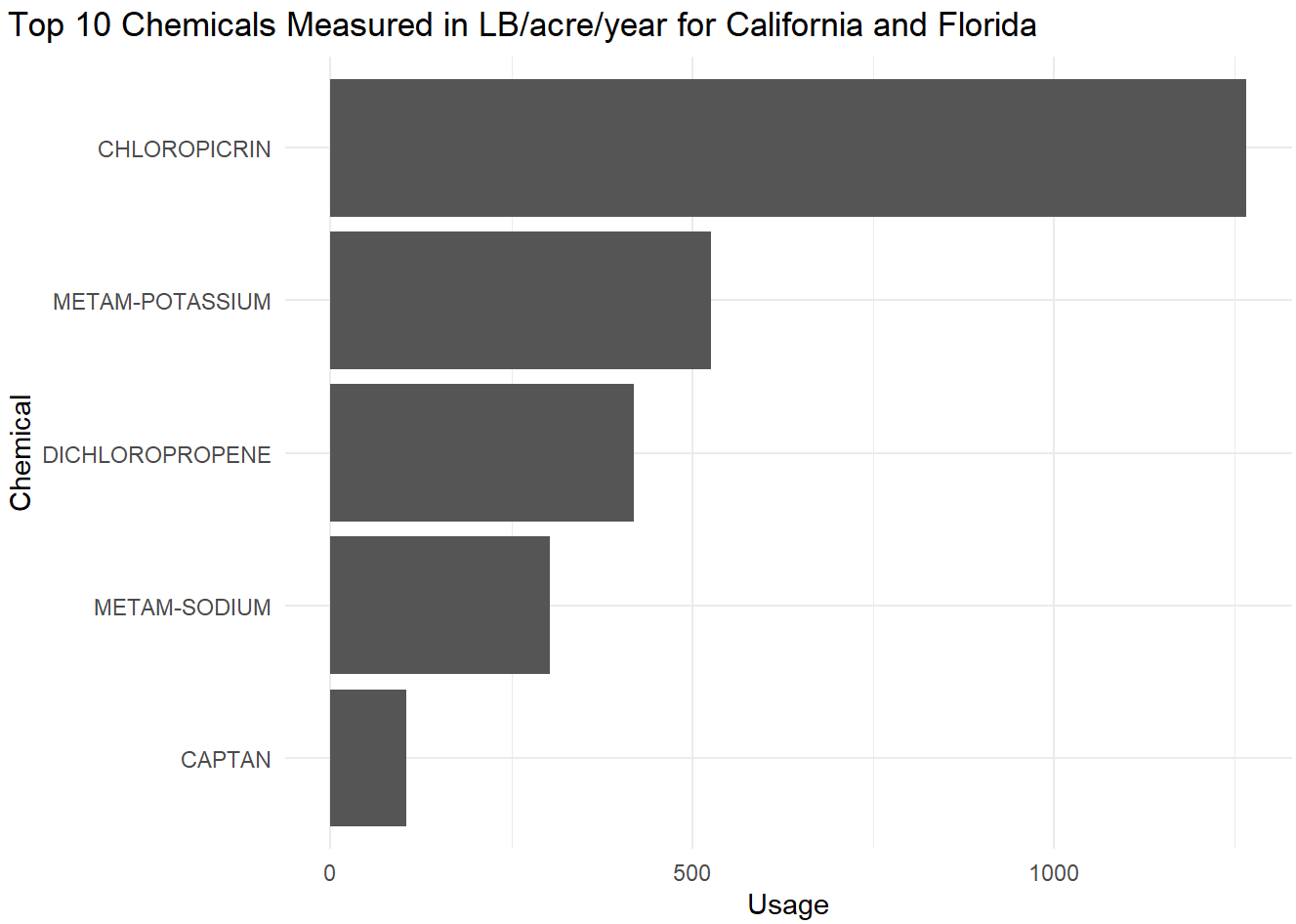
Through the table shown above, all the chemicals used in Florida have hazard information. Especially for thiram, which has 19 kinds of hazard information and can cause irritation and allergic reaction to skin, and it also causes damage to organs with single exposure. Additionally, it has toxic effects to aquatic lives, which is bad for the environment. Also, all of the top 5 chemicals used are fungicide, this may relate to the location of Florida, which is already in the tropical area. The warmer weather brings more fungi, so the farmers there need to use more fungicide for fungi or spores control.

After analyzing the chemical measurement in LB, we can compare it with another type of measurement “measurement in LB/acres/year” to see the difference between top chemicals under this type of measurement:

```
df1 <- df %>%
  select(-1) %>%
  filter(State == "CALIFORNIA" | State == "FLORIDA") %>%
  filter((Measurement == "MEASURED IN LB") |
  (Measurement == "MEASURED IN LB / ACRE / YEAR")) %>%
  filter(!is.na(Value))

chem_lb <- df1 %>%
  group_by(Measurement) %>%
  filter(Measurement == "MEASURED IN LB") %>%
  filter(Chemical_Info != "TOTAL")

chem_lb_acre <- df1 %>%
  group_by(Measurement) %>%
  filter(Measurement == "MEASURED IN LB / ACRE / YEAR") %>%
  filter(Chemical_Info != "TOTAL")

# Plot the top chemicals used in California
top_chem_lb <- chem_lb %>%
  filter(!is.na(Chemical_Info)) %>%
  group_by(Chemical_Info, Domain) %>%
  summarise(Value = sum(Value, na.rm = TRUE)) %>%
  arrange(desc(Value))

top_chem_lb_acre <- chem_lb_acre %>%
  filter(!is.na(Chemical_Info)) %>%
  group_by(Chemical_Info, Domain) %>%
  summarise(Value = sum(Value, na.rm = TRUE)) %>%
  arrange(desc(Value))

ggplot(top_chem_lb[1:5,],
       aes(x=reorder(Chemical_Info, Value), y=Value))+
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  labs(x = "Chemical", y = "Usage",
       title = "Top 10 Chemicals Measured in LB for California and Florida")+
  theme(plot.title.position = "plot")
```




```
ggplot(top_chem_lb_acre[1:5,],
       aes(x=reorder(Chemical_Info, Value), y=Value))+
  geom_bar(stat = "identity")+
  theme_minimal()+
  coord_flip()+
  labs(x = "Chemical", y = "Usage",
       title = "Top 10 Chemicals Measured in LB/acre/year for California and Florida")+
  theme(plot.title.position = "plot")
```



From the comparison of those two plots, the usage of dichloropropene are consistent under those two measurements, which means that the total usage in LB and usage per acre for dichloropropene are all very high. However, it is a different story for captan. Captan is the top chemical in total usage in LB but only the fifth highest in usage per acre, which means that the total area of using captan may be large, but the usage per acre may be lower. Also, the top chemical in usage per acre, chloropicrin, does not appear in the top 5 chemicals in total usage in LB, and this could be related to that the area of using chloropicrin is small.

In summary, on one hand, by comparing the chemicals used in California and Florida, California used more types of chemicals than Florida, also, the counts of chemicals used in growing strawberry were also higher for California, which is almost two times of Florida's value. Additionally, for chemical measurement in LB, California has more outliers and more spread out than Florida, and the chemicals used in California are distributed in different groups, but Florida seems to favor chemicals in fungicide group. On the other hand, California and Florida both used chemicals that are toxic to the aquatic lives, environment and human's health, and it is important to suggest to the goverment to restrict the amount of harmful chemicals used in insecticides and fungicides, and they should be changed to more environmental friendly substitutes.