```
---
title: "chem_"
author: "Guangyu Ding"
date: "2024-10-31"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Consider the survey_d_chem file.

```{r}
library(dplyr)
library(readr)
library(tidyr)
library(ggplot2)

chem_data <- read_csv("survey_d_chem.csv")

## Filter data for California, because California is one of the major strawberry
## producing regions in the United States. Observing the data here is more
## convincing.
chem_data_ca <- chem_data %>%
  filter(State == "CALIFORNIA")

# Clean the data by removing non-numeric values and converting to numeric
chem_data_ca <- chem_data_ca %>%
  filter(!Value %in% c("(D)", "(NA)")) %>%
  mutate(Value = as.numeric(Value))

# Summarize total applications per year for each chemical category
chem_summary <- chem_data_ca %>%
  group_by(Year, col2) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE)) %>%
  spread(key = col2, value = Total_Value)

# Print the summary
print(chem_summary)

## Visualize the data by making line chart
chem_data <- chem_data %>%
  filter(!Value %in% c("(D)", "(NA)")) %>%
  mutate(Value = as.numeric(Value))

# Summarize total applications per year across all chemical types
```
```

```
chem_summary <- chem_data %>%
  group_by(Year) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE))

ggplot(chem_summary, aes(x = Year, y = Total_Value)) +
  geom_line(color = "darkgreen", size = 1) +
  geom_point(color = "darkgreen", size = 2) +
  labs(title = "Total Chemical Application Trends",
       x = "Year",
       y = "Total Applications (in pounds)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Now separate the values by the categories listed in col2 column. Through
this we can see which chemicals are being used more every year.

```{r}
# Summarize total applications per year for each chemical category (col2)
chem_summary1 <- chem_data %>%
  group_by(Year, col2) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE))

# Spread the data to have categories as columns (if needed)
chem_summary_wide <- chem_summary1 %>%
  spread(key = col2, value = Total_Value)

# Print the summary data for each chemical category
print(chem_summary1)
## Obviously, Insecticides are the most commonly used chemical substances for
## each year.

## We can see the trend of usage of each types of chemical substances by making
## a line chart to show all types of chemical substances used from 2018 to 2023.
chem_data_ca <- chem_data_ca %>%
  filter(!Value %in% c("(D)", "(NA)")) %>%
  mutate(Value = as.numeric(Value))

chem_summary <- chem_data_ca %>%
  group_by(Year, col2) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE)) %>%
  spread(key = col2, value = Total_Value)

chem_long <- chem_summary %>%
  gather(key = "Chemical_Type", value = "Total_Value", -Year)

ggplot(chem_long, aes(x = Year, y = Total_Value, color = Chemical_Type, group =
```

```
Chemical_Type)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "Chemical Application Trends in California",
       x = "Year",
       y = "Total Applications (in pounds)",
       color = "Chemical Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

## Based on the plot, the usage of insecticide decreases from 2018 to 2023.
## The fungicide had an increase in 2019, but decreased since then.
## The herbicide has an overall trend of decreasing, but experienced a rebound in 2021.
## All other chemical substances has a similar trend with fungicide, but decreased more
from 2019 to 2021, less from 2021 to 2023.


## We can make more specific observation on each types of chemical substances. i.e.
Fungicide
# Filter data for California and fungicides
fungicide_data_ca <- chem_data %>%
  filter(State == "CALIFORNIA", col2 == "FUNGICIDE")

# Clean the data by removing non-numeric values and converting to numeric
fungicide_data_ca <- fungicide_data_ca %>%
  filter(!Value %in% c("(D)", "(NA)")) %>%
  mutate(Value = as.numeric(Value))

# Summarize total applications per year for fungicides
fungicide_summary <- fungicide_data_ca %>%
  group_by(Year) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE))

# Plot the trends using ggplot2
ggplot(fungicide_summary, aes(x = Year, y = Total_Value)) +
  geom_line(color = "blue", size = 1) +
  geom_point(color = "blue", size = 2) +
  labs(title = "Fungicide Application Patterns in California",
       x = "Year",
       y = "Total Applications (in pounds)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


Now consider the suevey_d_fert file.By observing this file, we can get the usage
of the fertilizer

```{r}
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)

fert_data <- read_csv("survey_d_fert.csv")

fert_data <- fert_data %>%
  filter(!Value %in% c("(D)", "(NA)")) %>%
  mutate(Value = as.numeric(Value))

# Summarize total applications per year for each fertilizer type
fert_summary <- fert_data %>%
  group_by(Year, chem_name) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE))

# Convert data to long format for easier plotting
fert_long <- fert_summary %>%
  spread(key = chem_name, value = Total_Value)

# Plot the trends
ggplot(fert_summary, aes(x = Year, y = Total_Value, color = chem_name, group =
chem_name)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "Fertilizer Application Trends",
       x = "Year",
       y = "Total Applications (in pounds)",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

By observation, from March 2018 to June 2020, nitrogen was the one being used
most. For the rest of time period from 2018 to 2023, Potash was at the first place.

Now visualize the usage of fertilizer in California. Compare the overall data
in the United States with the data in California.

```{r}
library(dplyr)
library(readr)
library(ggplot2)
library(tidyr)

fert_data_ca <- read_csv("fert_data_ca.csv")
```

```
summary(fert_data_ca)
str(fert_data_ca)
head(fert_data_ca)
fert_data_ca <- fert_data_ca %>%
  filter(!Value %in% c("(D)", "(NA)")) %>%
  mutate(Value = as.numeric(Value))
fert_summary <- fert_data_ca %>%
  group_by(Year, chem_name) %>%
  summarise(Total_Value = sum(Value, na.rm = TRUE))

print(fert_summary)

# Visualize trends in fertilizer applications over the years
ggplot(fert_summary, aes(x = Year, y = Total_Value, color = chem_name, group =
chem_name)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "Fertilizer Application Trends in California",
       x = "Year",
       y = "Total Applications (in pounds)",
       color = "Fertilizer Type") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

We find that the usage of Nitrogen is ahead in California. Maybe other places
in United State can change from using potash to nitrogen.

Now consider the files fung, fung_ca_only and fung_fl_only to observe which chem
is used most.

```{r}
library(dplyr)
library(readr)
library(ggplot2)

fung_fl <- read_csv("fung_fl_only.csv")
fung_data <- read_csv("fung.csv")
fung_ca <- read_csv("fung_ca_only.csv")

summary(fung_fl)
summary(fung_data)
summary(fung_ca)

# Combine the data into one DataFrame for analysis
combined_fung <- bind_rows(fung_fl, fung_data, fung_ca)
```

```
# Summarize or clean the data if necessary (e.g., handle missing values)
combined_fung <- combined_fung %>%
  filter(!is.na(chem_index), !is.na(chem_name))

# Create a bar plot to visualize chem_name versus chem_index
ggplot(combined_fung, aes(x = chem_name, y = chem_index, fill = chem_name)) +
  geom_bar(stat = "identity") +
  labs(title = "Chem Index Values by Chem Name",
       x = "Chem Name",
       y = "Chem Index") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = "Set3")
```

The plot produced seems not that useful. Here I think what we need to do is
to sort the data in the three fung tables in descending order to obtain the
usage quantities of each chemical substance nationwide, in California, and
in Florida.

To prove the series of results above, we can directly observe the relevant
data of the harvest quantity.The files sur_CA_19_23 and sur_fl_19_23 are used.
```{r}
# Load necessary libraries
library(ggplot2)
library(dplyr)
library(readr)

# Load the data from the CSV files
data_ca <- read_csv("sur_CA_19_23.csv")
data_fl <- read_csv("sur_fl_19_23.csv")

# Extract the years for plotting
years <- c("value_19", "value_20", "value_21", "value_22", "value_23")

# Convert the selected columns to numeric after removing commas
convert_to_numeric <- function(x) as.numeric(gsub(",", "", x))

# Visualize PRICE RECEIVED trends for California and Florida
price_ca <- data_ca %>% filter(product_price == "PRICE RECEIVED") %>%
select(all_of(years)) %>% unlist() %>% as.numeric()
price_fl <- data_fl %>% filter(product_price == "PRICE RECEIVED") %>%
select(all_of(years)) %>% unlist() %>% as.numeric()

# Plot PRICE RECEIVED trends
plot_data <- data.frame(
```

```
  Year = rep(years, 2),
  Price = c(price_ca, price_fl),
  State = rep(c("California", "Florida"), each = length(years))
)

ggplot(plot_data, aes(x = Year, y = Price, group = State, color = State)) +
  geom_line() + geom_point() +
  ggtitle("PRICE RECEIVED Trend (2019-2023)") +
  xlab("Year") + ylab("Price ($/CWT)") +
  theme_minimal()

# Visualize HARVESTED acres trends for California and Florida
harvested_ca <- data_ca %>% filter(product_price == "HARVESTED") %>%
select(all_of(years)) %>% unlist() %>% convert_to_numeric()
harvested_fl <- data_fl %>% filter(product_price == "HARVESTED") %>%
select(all_of(years)) %>% unlist() %>% convert_to_numeric()

# Plot HARVESTED acres trends
harvested_data <- data.frame(
  Year = rep(years, 2),
  Acres = c(harvested_ca, harvested_fl),
  State = rep(c("California", "Florida"), each = length(years))
)

ggplot(harvested_data, aes(x = Year, y = Acres, group = State, color = State)) +
  geom_line() + geom_point() +
  ggtitle("HARVESTED Acres Trend (2019-2023)") +
  xlab("Year") + ylab("Acres") +
  theme_minimal()

# Calculate average HARVESTED acres for each state
avg_harvested_ca <- mean(harvested_ca)
avg_harvested_fl <- mean(harvested_fl)

# Print the averages
avg_harvested_ca
avg_harvested_fl
```

Maybe one conclustion we can get is for strawberriews cultivation, we may use nitrogen as the fertilizer.

Since the assignment is related to the chemical usage of cultivating strawberries, several data related to
the price of the strawberries may not be consider in my report.