Chemical

Guangyu Ding

2024-10-22

Consider the survey_d_chem file.

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

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. Thr
ough
## this we can see which chemicals are being used more every year.

```{r}
# Summarize total applications per year for each chemical category (col
2)
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 substanc
es for
## each year.

## We can see the trend of usage of each types of chemical substances b
y 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 2
023.
## 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, bu
t decreased more from 2019 to 2021, less from 2021 to 2023.



## We can make more specific observation on each types of chemical subs
tances. 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 numer
ic
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.