loading/cleaning/wrangling/exploring

Edward Cruz

2024-07-29

The structure below is one possible setup for a data analysis project (including the course project). For a manuscript, adjust as needed. You don’t need to have exactly these sections, but the content covering those sections should be addressed.

This uses MS Word as output format. See here for more information. You can switch to other formats, like html or pdf. See the Quarto documentation for other formats.

```
Warning: package 'ggplot2' was built under R version 4.3.3

Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'scales' was built under R version 4.3.3

Warning: package 'showtext' was built under R version 4.3.3

Warning: package 'sysfonts' was built under R version 4.3.3

Warning: package 'ggimage' was built under R version 4.3.3
```

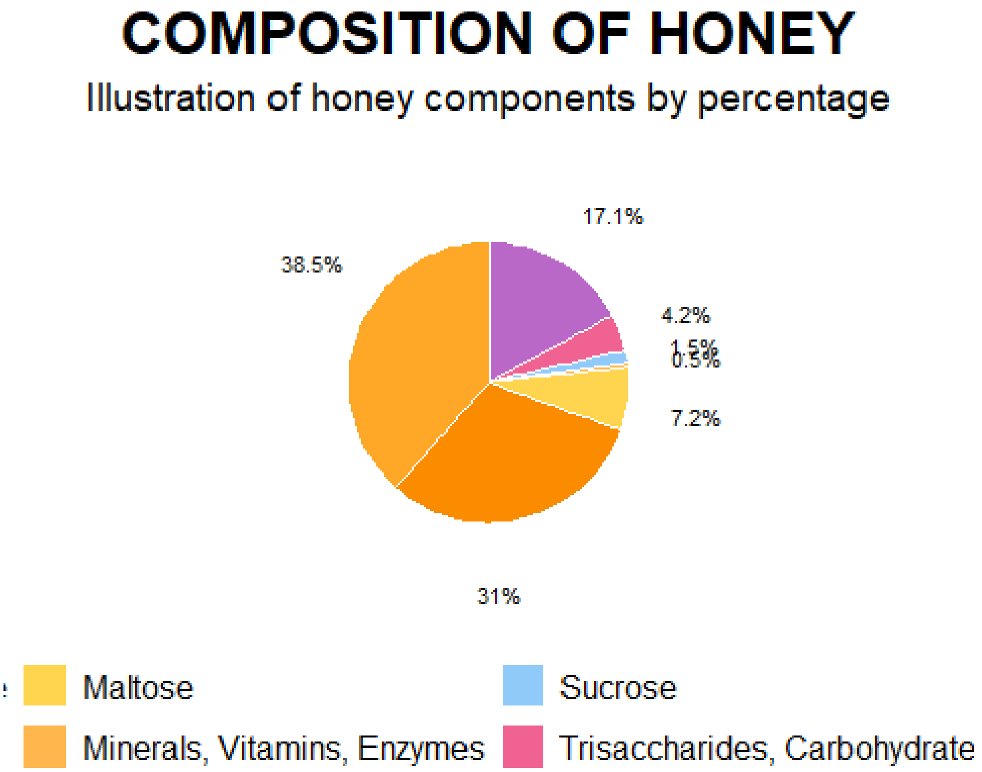
Summary/Abstract

Group Ten is pursuing the historical data collected by various public agencies to determine if the impact of Varroa, American Foulbrood, and global warming has demonstrated an increase in hive losses across the United States and Texas. The research will be accomplished using data sets derived from the agencies National Agricultural Statistics Service, Agricultural Statistics Board, and United States Department of Agriculture (USDA). Data sets containing several years of hive losses based on varroa and bacterium losses including recent possible environmental thermal global warming. A visualization of outcomes using R demonstrating and validating possible detrimental effects on honeybee colonies in the United States and Texas wrought by the negative impact of mites, bacterium, and global warming that will affect honey production and inevitably impact food production.

Introduction

```
# Create a data frame with the composition of honey
honey_data <- data.frame(
  component = c("Fructose", "Glucose", "Water", "Maltose", "Trisaccharides, C
arbohydrates", "Sucrose", "Minerals, Vitamins, Enzymes"),
  percentage = c(38.5, 31.0, 17.1, 7.2, 4.2, 1.5, 0.5)
)

# Define colors for the segments
honey_data$color <- c("#FFA726", "#FB8C00", "#FFD54F", "#FFB74D", "#90CAF9",
"#F06292", "#BA68C8")

# Plot with ggplot2
ggplot(honey_data, aes(x = "", y = percentage, fill = component)) +
  geom_bar(width = 0.8, stat = "identity", color = "white") +
  coord_polar("y", start = 0) +
  scale_fill_manual(values = honey_data$color) +
  geom_text(aes(x= 1.8,label = paste0(percentage, "%")), position = position_
stack(vjust = 0.6), size = 3, color = "black") +
  labs(
    title = "COMPOSITION OF HONEY",
    subtitle = "Illustration of honey components by percentage",
    fill = NULL
  ) +
  theme_void() +
  theme(
    plot.title = element_text(size = 22, face = "bold", hjust = 0.5, family =
"lobster"),
    plot.subtitle = element_text(size = 14, hjust = 0.5),
    legend.position = "bottom",
    legend.text = element_text(size = 12)
  )
```

```
Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font family
not found in Windows font database

Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : fon
t
family not found in Windows font database

Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
font family not found in Windows font database
```



Honeybees have been nature’s pollinators for centuries, with documented relationships with humans dating back to ancient Egyptians and Hindus. Historically, humans have maintained beehives, using honey as medicine in cultures such as the Egyptians, Assyrians, Chinese, Greeks, and Romans. The natural antibacterial properties of honey made it a valuable treatment for wounds, preventing infection, a practice used by Romans and Russians during World War I. Honeybees and other pollinators are critical for food production and nutritional security, yet bees face a variety of survival challenges. Currently, Varroa mites impact bee colonies and this mite infestation, a tiny red-brown parasite that can live on adult honeybees and reproduce on larvae and pupae in the developing brood. Another major threat is American Foulbrood Disease (AFB), caused by the bacterium Paenibacillus larvae. A disease that is fatal to honeybee larvae and found worldwide. The only effective control measure is to incinerate and destroy infected hives and live bees mitigating the infectious spread to other colonies. In addition, the exploration of climate change impacting honeybee colony losses has only recently been researched. While there are correlations between higher winter temperatures and greater colony losses, the effects of warmer autumn and winter temperatures on colony population dynamics and age structure as potential causes of reduced colony survival have not yet been fully investigated.Index Catalog // USDA Economics, Statistics and Market Information System. (n.d.-b). Index Catalog // USDA Economics, Statistics and Market Information System. (n.d.-a). , USDA - National Agricultural Statistics Service - Surveys - honey bee surveys and reports. (n.d.). https://usda.library.cornell.edu/catalog?f%5Bkeywords_sim%5D%5B%5D=honey+bees&locale=en https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Bee_and_Honey/

You can add options to executable code like this

```
library(ggplot2)
library(sf)
```

```
Warning: package 'sf' was built under R version 4.3.3
```

```
Linking to GEOS 3.11.2, GDAL 3.8.2, PROJ 9.3.1; sf_use_s2() is TRUE
```

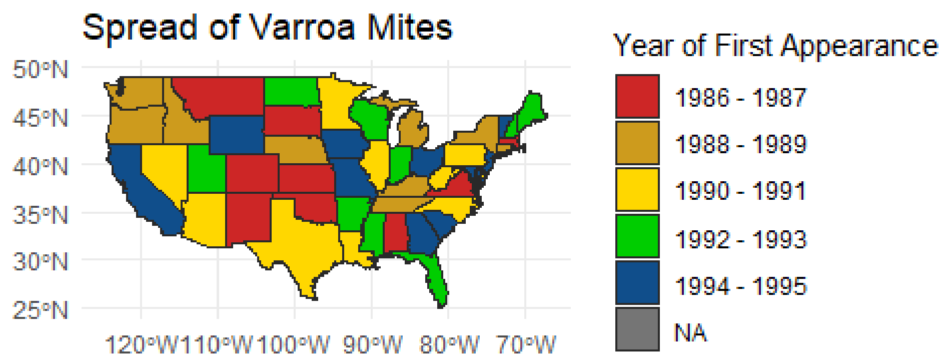
```
library(maps)
```

```
Warning: package 'maps' was built under R version 4.3.3
```

```
states <- st_as_sf(map("state", plot = FALSE, fill = TRUE))
```

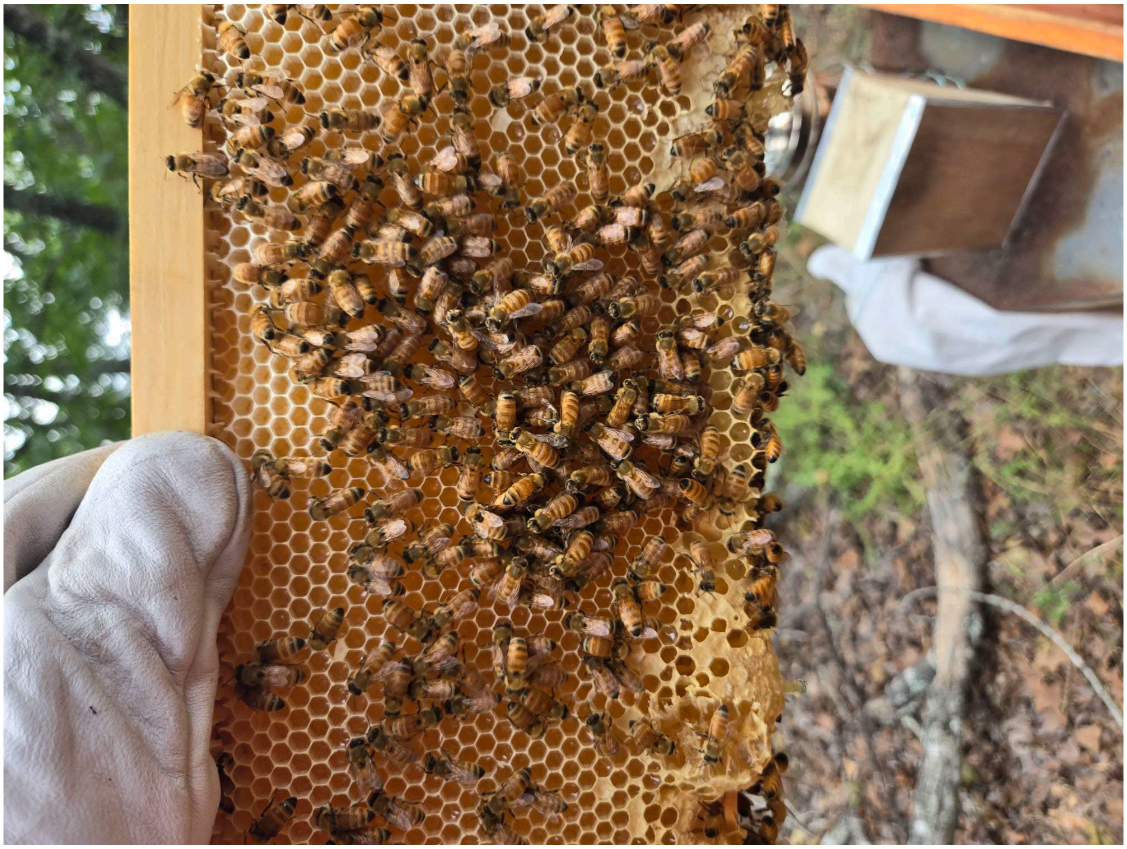
```
varroa_data <- data.frame(
  state = c("alabama", "alaska", "arizona", "arkansas", "california", "colora
do",
            "connecticut", "delaware", "florida", "georgia", "hawaii", "idaho
",
            "illinois", "indiana", "iowa", "kansas", "kentucky", "louisiana",
            "maine", "maryland", "massachusetts", "michigan", "minnesota",
            "mississippi", "missouri", "montana", "nebraska", "nevada", "new
hampshire",
            "new jersey", "new mexico", "new york", "north carolina", "north
dakota",
            "ohio", "oklahoma", "oregon", "pennsylvania", "rhode island",
            "south carolina", "south dakota", "tennessee", "texas", "utah",
            "vermont", "virginia", "washington", "west virginia", "wisconsin"
, "wyoming"),
  year = c("1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995",
           "1986 - 1987", "1988 - 1989", "1990 - 1991", "1992 - 1993", "1994
- 1995")
)
```

```
states <- merge(states, varroa_data, by.x = "ID", by.y = "state", all.x = TRU
E)

ggplot(data = states) +
  geom_sf(aes(fill = year), color = "#292929", size = 0.2) +
  scale_fill_manual(
    values = c(
      "1986 - 1987" = "#CD2626",
      "1988 - 1989" = "#CD9B1D",
      "1990 - 1991" = "#FFD700",
      "1992 - 1993" = "#00CD00",
      "1994 - 1995" = "#104E8B"
    ),
    na.value = "#757575"
  ) +
  theme_minimal() +
  labs(
    title = "Spread of Varroa Mites",
    fill = "Year of First Appearance"
  )
```



Written By:sdns6mchl4. (2016, February 24). Varroa mite spread in the United States. Beesource Beekeeping Forums. https://www.beesource.com/threads/varroa-mite-spread-in-the-united-states.365462/ ## General Background Information

Uncapped Honey Floresville,Texas Hive



Capped Honey one Month Later same hive frame- Italian bees Floresville, Texas

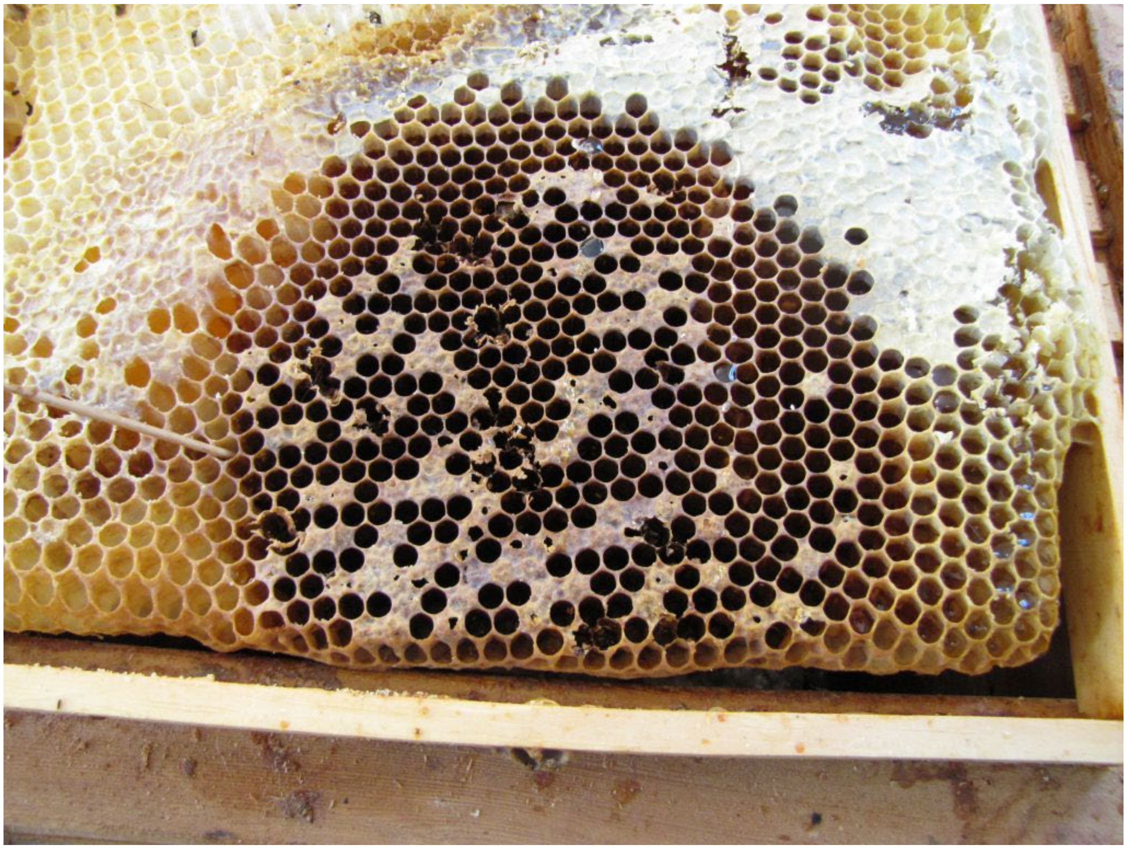
Description of data and data source

Bee colonies maintained by beekeepers are considered livestock by the USDA due to their ability to produce honey, a consumable food item, and their essential role in assisting farmers with pollination crop seasons. Given the importance of bee colonies in agriculture, it was logical to source data from the following two authoritative websites: 1. USDA National Agricultural Statistics Service (NASS): This site provides comprehensive agricultural data, including statistics on honey production and colony health. 2. Bee Informed Partnership: This site offers detailed insights and research on bee colony management and health, contributing valuable information on the status and trends of bee populations. Index Catalog // USDA Economics, Statistics and Market Information System. (n.d.-a).
https://usda.library.cornell.edu/catalog?f%5Bkeywords_sim%5D%5B%5D=honey+bees&locale=en
USDA - National Agricultural Statistics Service - Surveys - honey bee surveys and reports. (n.d.). https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Bee_and_Honey/

Questions/Hypotheses to be addressed

Hypothesis: “The negative impacts of mites, bacterium, and global warming have detrimental effects on honeybee colonies in the United States and Texas, which in turn will lead to a decline in honey production and negatively impact food production.” This hypothesis can be tested and validated through a visualization of outcomes using R, demonstrating the relationship between these factors and their effects on honeybee colonies.

Bacterium Infection Foul Brood.



Dead bees resulting from extreme heat found in hive.

To cite other work (important everywhere, but likely happens first in introduction), make sure your references are in the bibtex file specified in the YAML header above and have the right bibtex key. Then you can include like this:

Examples of reproducible research projects can for instance be found in [@mckay2020; @mckay2020a].

Methods

Describe your methods. That should describe the data, the cleaning processes, and the analysis approaches. You might want to provide a shorter description here and all the details in the supplement.

Schematic of workflow

Sometimes you might want to show a schematic diagram/figure that was not created with code (if you can do it with code, do it). **?@fig-schematic** is an example of some - completely random/unrelated - schematic that was generated with Biorender. We store those figures in the `assets` folder.

Data aquisition

We got our data from the United States Department of Agriculture (USDA).

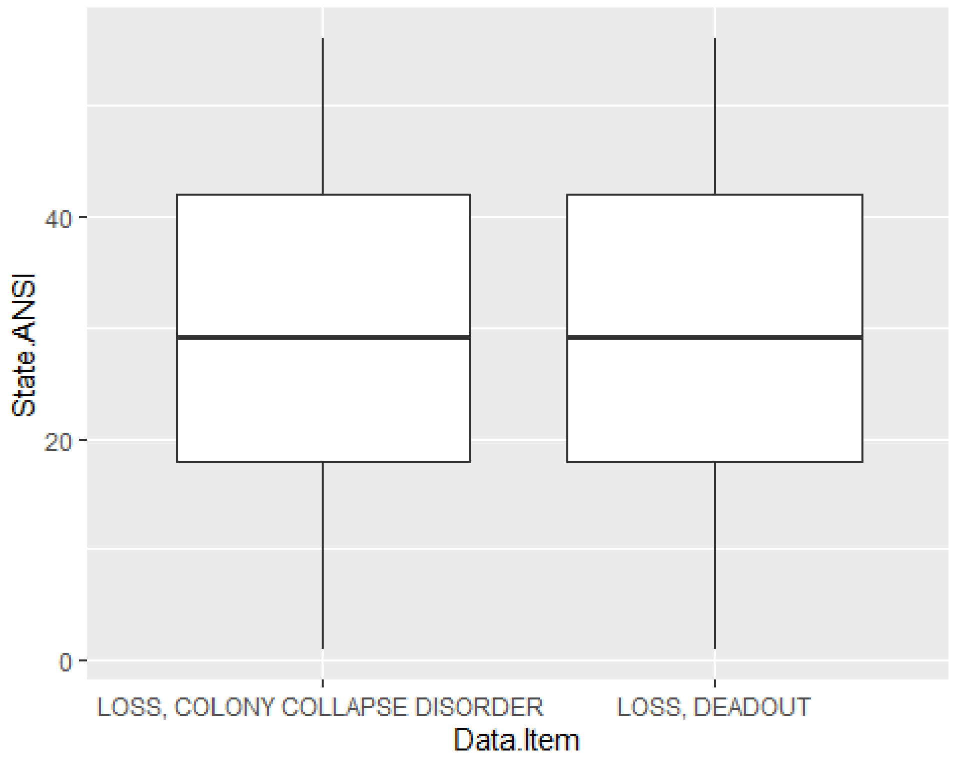
Data import and cleaning

We decided to clean out our data from a few different datasets. We had to remove blank spaces and columns that were not pertinent to our analysis. We then filtered out other observations that did not directly deal with the data we are exploring. We are looking for cause of death to bee colonies and how they are affected by mites and climate change so we wanted to single out data that represented the losses so we can explore the different states by year and determine how the colonies were affected.

```
library(readxl)
library(tidyverse)

Warning: package 'tidyverse' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.3

Warning: package 'readr' was built under R version 4.3.3

Warning: package 'stringr' was built under R version 4.3.3

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0
──
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ lubridate 1.9.3     ✓ tibble    3.2.1
✓ purrr     1.0.2     ✓ tidyr     1.3.1
✓ readr     2.1.5
── Conflicts ────────────────────────────────────────── tidyverse_conflicts()
──
✕ readr::col_factor() masks scales::col_factor()
✕ purrr::discard()    masks scales::discard()
✕ dplyr::filter()     masks stats::filter()
✕ dplyr::lag()        masks stats::lag()
```

```
✖ purrr::map()        masks maps::map()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all co
nflicts to become errors
```

```r
library(ggplot2)
library(knitr)

data <- read.csv("C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science Progr
am/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-EdwardC
ruz/DatabyState.csv")

View(data)

# Select all columns except 3, 6, and 9
Data_Clean <- dplyr::select(data, -c(3, 6, 9))

# Output cleaned data file to a csv file.
write.csv(Data_Clean, "C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science
Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-Ed
wardCruz/Databystate_Clean.csv")

view(Data_Clean)

data <- read.csv("C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science Progr
am/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-EdwardC
ruz/DatabyState.csv")

# Select all columns except 3, 6, and 9
Data_Clean <- dplyr::select(data, -c(3, 6, 9))

# Output cleaned data file to a csv file.
write.csv(Data_Clean, "C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science
Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-Ed
wardCruz/Databystate_Clean.csv")

# Assuming your data frame is named "data"
# Filter for rows where "Loss" or "Deadout" is present in any column (case-in
sensitive)
library(stringr)  # Load stringr package for regular expressions
data_filtered <- Data_Clean[rowSums(sapply(data, grepl, pattern = c("Loss"),
ignore.case = TRUE)) > 0, ]

# Output cleaned data file to a csv file.
write.csv(Data_Clean, "C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science
Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-Ed
wardCruz/Databystate_Filtered.csv")

view(Data_Clean)

ggplot(data_filtered, aes( Data.Item, State.ANSI)) + geom_boxplot()
```

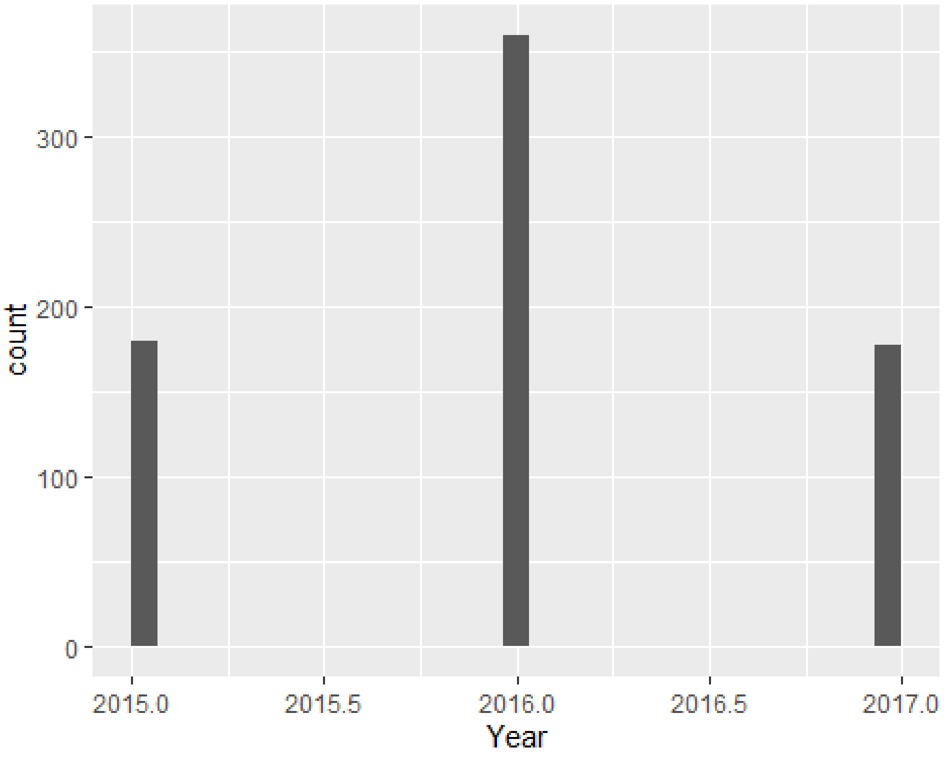


STATE ANSI IS THE CODE FOR STATE BY PERIOD AND YEAR. VALUE IS COUNT.

The above data set is organized by year, state, period of the year, type of loss, and count. There are only three observed years, with observations from 2015 to 2017. Loss by collapse disorder describes a colony that losses a majority of its worker bees. Loss by deadout describes a loss of the entire colony: workers, drones, larvae, pupa, and queen. The bar chart shown below depicts loss count by type and period. The periods roughly follow North American seasons; April through June is spring, January through March is winter, July through September is summer, and October through December is fall. We can observe a pattern in the losses increasing in the first half of the year (January through June), while losses decrease the second half. This could suggest that as the weather warms, colonies are affected by the increasing heat.

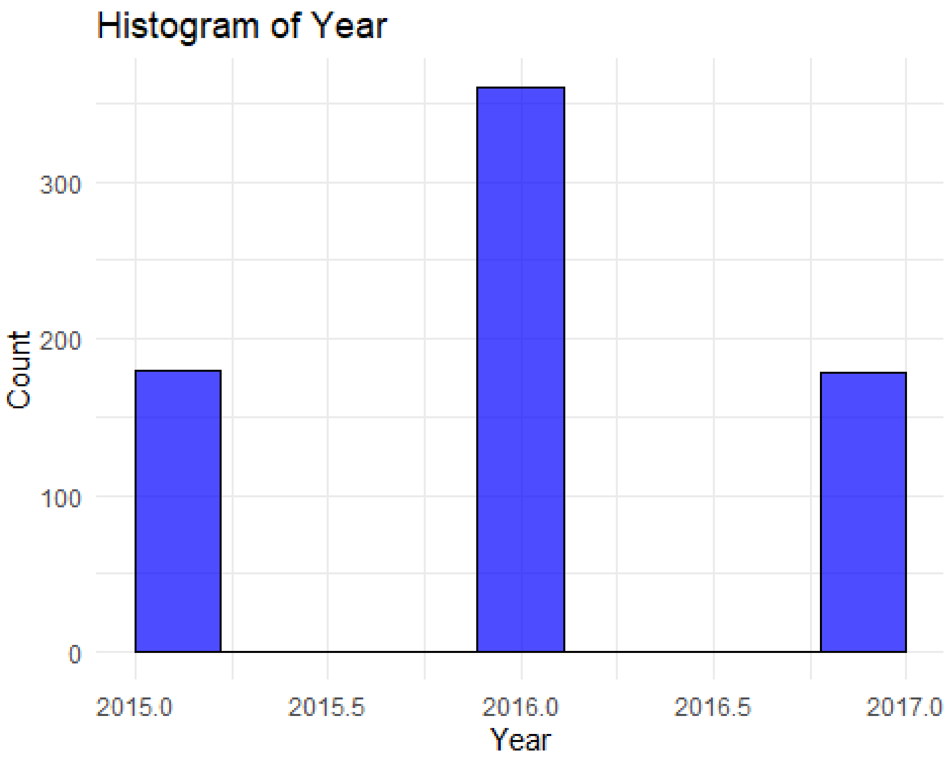
```
ggplot(data_filtered, aes(Year)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The above data set is organized by year, state, period of the year, type of loss, and count. There are only three observed years, with observations from 2015 to 2017. Loss by collapse disorder describes a colony that losses a majority of its worker bees. Loss by deadout describes a loss of the entire colony: workers, drones, larvae, pupa, and queen. The bar chart shown below depicts loss count by type and period. The periods roughly follow North American seasons; April through June is spring, January through March is winter, July through September is summer, and October through December is fall. We can observe a pattern in the losses increasing in the first half of the year (January through June), while losses decrease the second half. This could suggest that as the weather warms, colonies are affected by the increasing heat

```
ggplot(data_filtered, aes(x = Year)) +
  geom_histogram(bins = 10, fill = "blue", color = "black", alpha = 0.7) +  # Adding fill color, border color, and transparency
  labs(title = "Histogram of Year", x = "Year", y = "Count") +  # Adding labels
  theme_minimal()
```



```
library(readxl)
library(tidyverse)
library(dplyr)
library(mgcv)

Warning: package 'mgcv' was built under R version 4.3.3

Loading required package: nlme

Warning: package 'nlme' was built under R version 4.3.3


Attaching package: 'nlme'

The following object is masked from 'package:dplyr':

    collapse

This is mgcv 1.9-1. For overview type 'help("mgcv-package")'.

# Output cleaned data file to a csv file.
hcny_data <- read.csv("C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science
Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-Ed
wardCruz/hcny_CleanDraft.csv", header = FALSE, stringsAsFactors = FALSE)

head(hcny_data)
```

```
             V1           V2          V3      V4         V5    V6      V7
1         state varroa_mites other_pests disease pesticides other unknown
2  Pennsylvania         13.8         7.2     2.7        6.9   1.1     1.1
3         Texas         46.8        42.3    10.9       12.1  30.1    10.2
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6       6
```

This gives us the properties of out data set

```
str(hcny_data)

'data.frame':   5 obs. of  7 variables:
 $ V1: chr  "state" "Pennsylvania" "Texas" "Wisconsin" ...
 $ V2: chr  "varroa_mites" "13.8" "46.8" "67.2" ...
 $ V3: chr  "other_pests" "7.2" "42.3" "9.8" ...
 $ V4: chr  "disease" "2.7" "10.9" "47.8" ...
 $ V5: chr  "pesticides" "6.9" "12.1" "49.2" ...
 $ V6: chr  "other" "1.1" "30.1" "48.1" ...
 $ V7: chr  "unknown" "1.1" "10.2" "46.8" ...

summary(hcny_data)

      V1                 V2                 V3                 V4
 Length:5           Length:5           Length:5           Length:5
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
      V5                 V6                 V7
 Length:5           Length:5           Length:5
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character

view(hcny_data)

# Merge the first two rows to create a proper header
header <- hcny_data[1:4, ]
header <- sapply(header, function(x) paste(na.omit(x), collapse = " "))
colnames(hcny_data) <- header
head(header)

                                  V1                                   V2
"state Pennsylvania Texas Wisconsin"        "varroa_mites 13.8 46.8 67.2"
                                  V3                                   V4
          "other_pests 7.2 42.3 9.8"              "disease 2.7 10.9 47.8"
                                  V5                                   V6
          "pesticides 6.9 12.1 49.2"                "other 1.1 30.1 48.1"

# Remove the first two rows (header rows) and any completely blank rows
hcnydata_cleaned <- hcny_data[-c(1,2), ]
head(hcnydata_cleaned)

  state Pennsylvania Texas Wisconsin varroa_mites 13.8 46.8 67.2
3                              Texas                        46.8
```

```
4                          Wisconsin                        67.2
5                      United States                        50.9
  other_pests 7.2 42.3 9.8 disease 2.7 10.9 47.8 pesticides 6.9 12.1 49.2
3                     42.3                  10.9                     12.1
4                      9.8                  47.8                     49.2
5                     13.9                   6.5                     10.5
  other 1.1 30.1 48.1 unknown 1.1 10.2 46.8
3                30.1                  10.2
4                48.1                  46.8
5                13.6                     6

# Assuming hcny_data_cleaned is your cleaned data frame
colnames(hcnydata_cleaned) <- c("State", "Varroa_Mites", "Other_Parasites_1/"
, "Disease_2/", "Pesticides", "Other_3/", "Unknown")

head(hcnydata_cleaned)

          State Varroa_Mites Other_Parasites_1/ Disease_2/ Pesticides Other_3
/
3         Texas         46.8               42.3       10.9       12.1     30.
1
4     Wisconsin         67.2                9.8       47.8       49.2     48.
1
5 United States         50.9               13.9        6.5       10.5     13.
6
  Unknown
3    10.2
4    46.8
5       6

# Convert specified columns to numeric
hcnydata_cleaned <- hcnydata_cleaned %>%
  mutate(across(c("Varroa_Mites", "Other_Parasites_1/", "Disease_2/", "Pestic
ides", "Other_3/", "Unknown"), as.numeric))

# View the updated data frame
hcnydata_cleaned<-hcnydata_cleaned[-1,]

head(hcnydata_cleaned)

          State Varroa_Mites Other_Parasites_1/ Disease_2/ Pesticides Other_3
/
4     Wisconsin         67.2                9.8       47.8       49.2     48.
1
5 United States         50.9               13.9        6.5       10.5     13.
6
  Unknown
4    46.8
5     6.0

# Verify the changes
str(hcnydata_cleaned)
```

```
'data.frame':   2 obs. of  7 variables:
 $ State              : chr  "Wisconsin" "United States"
 $ Varroa_Mites       : num  67.2 50.9
 $ Other_Parasites_1/: num  9.8 13.9
 $ Disease_2/         : num  47.8 6.5
 $ Pesticides         : num  49.2 10.5
 $ Other_3/           : num  48.1 13.6
 $ Unknown            : num  46.8 6
```

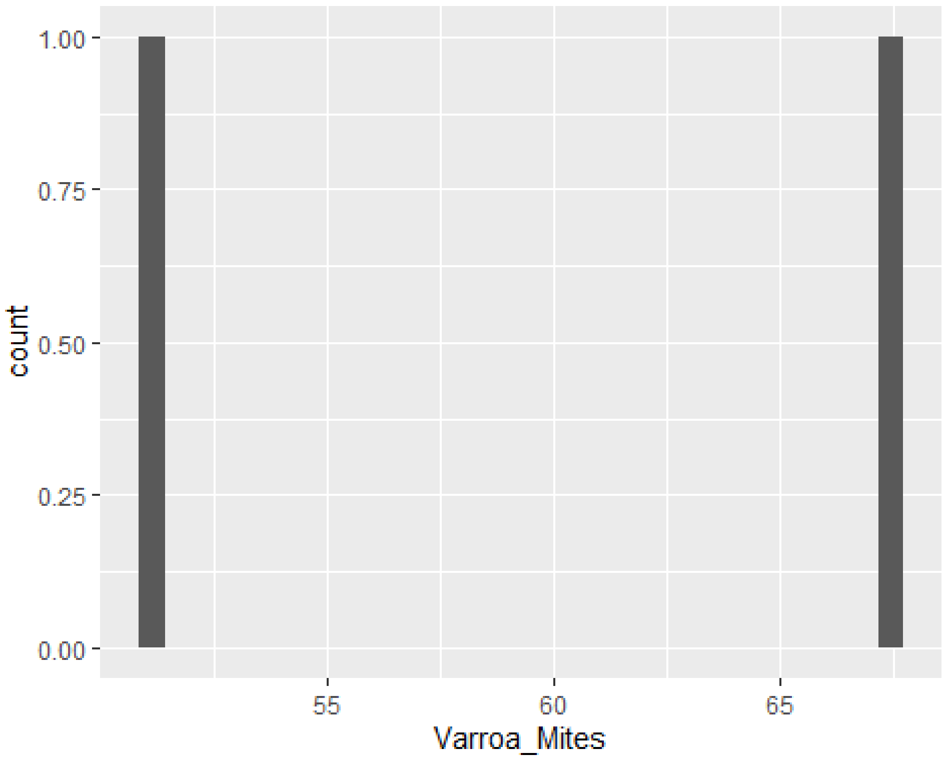
```
head(hcnydata_cleaned)
```

```
          State Varroa_Mites Other_Parasites_1/ Disease_2/ Pesticides Other_3
/
4     Wisconsin         67.2                9.8       47.8       49.2     48.
1
5 United States         50.9               13.9        6.5       10.5     13.
6
  Unknown
4    46.8
5     6.0
```

```
view(hcnydata_cleaned)
```

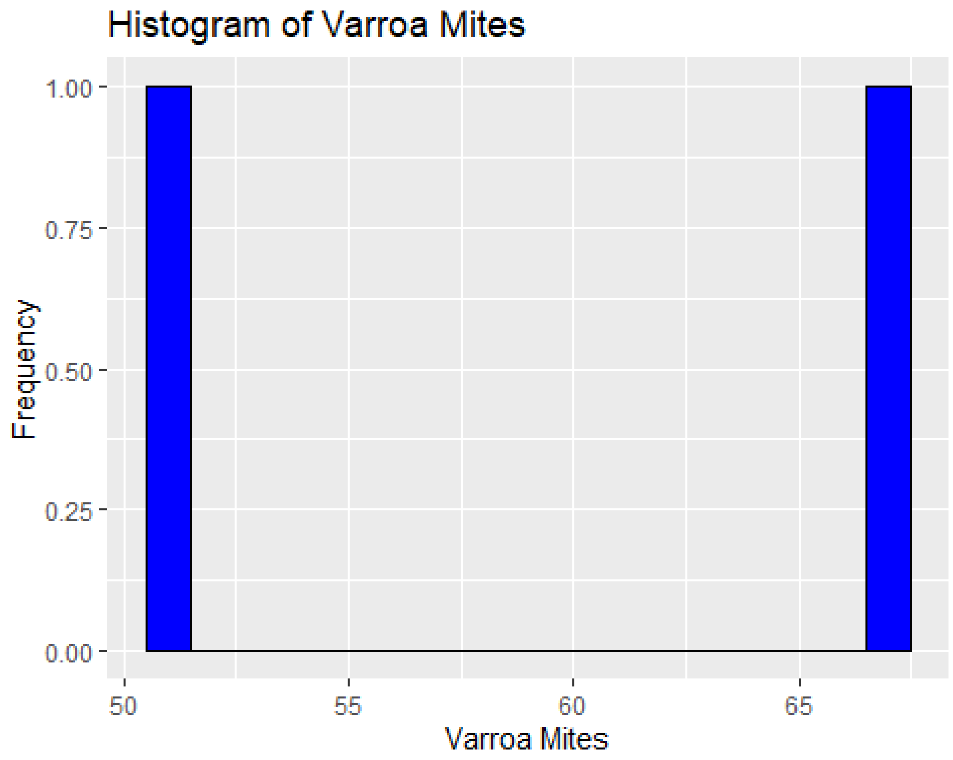
```
ggplot(hcnydata_cleaned, aes(Varroa_Mites)) + geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

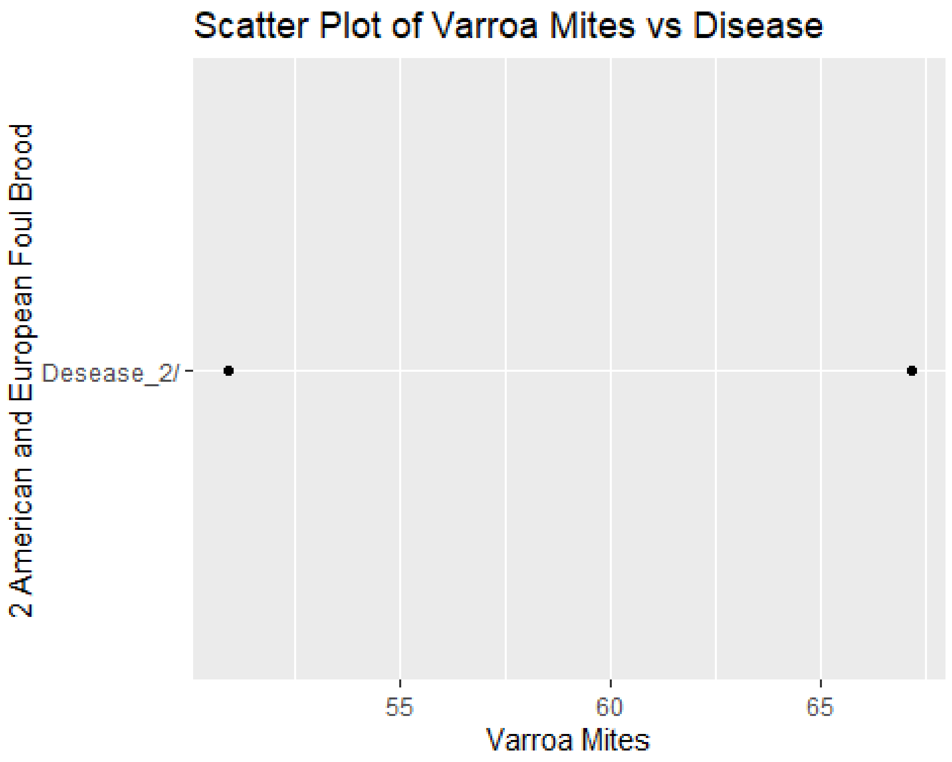


The above indicated the count of Varroa Mites in our data From our project, we found that varroa mites are the top cause for colony loss. Four states had the highest levels of varroa cases and were chosen for comparison, in the United States: Wisconsin, Texas, Ohio, and Pennsylvania. The box plot below shows one outlier in this top 5 areas, with the average being just above 50.

```
ggplot(hcnydata_cleaned, aes(x = Varroa_Mites)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Histogram of Varroa Mites",
       x = "Varroa Mites",
       y = "Frequency")
```



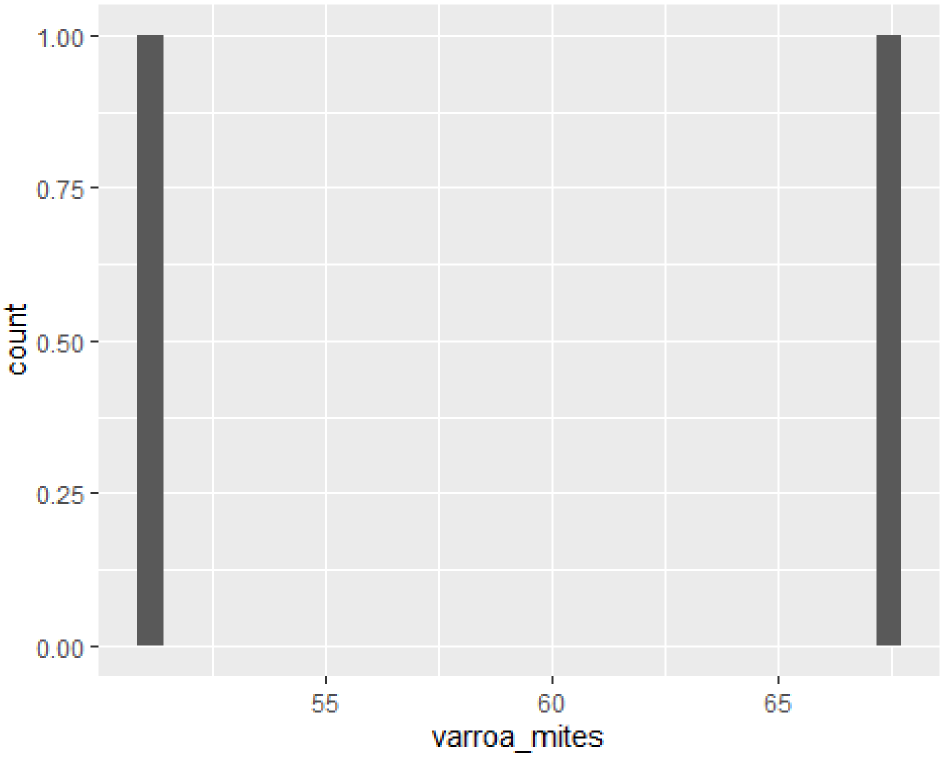
```
# Example scatter plot of two variables
ggplot(hcnydata_cleaned, aes(x = Varroa_Mites, y = "Desease_2/")) +
  geom_point() +
  labs(title = "Scatter Plot of Varroa Mites vs Disease",
       x = "Varroa Mites",
       y = "2 American and European Foul Brood")
```



The above plot illustrates the comparison between American and European Foul Brood and Varroa Mites

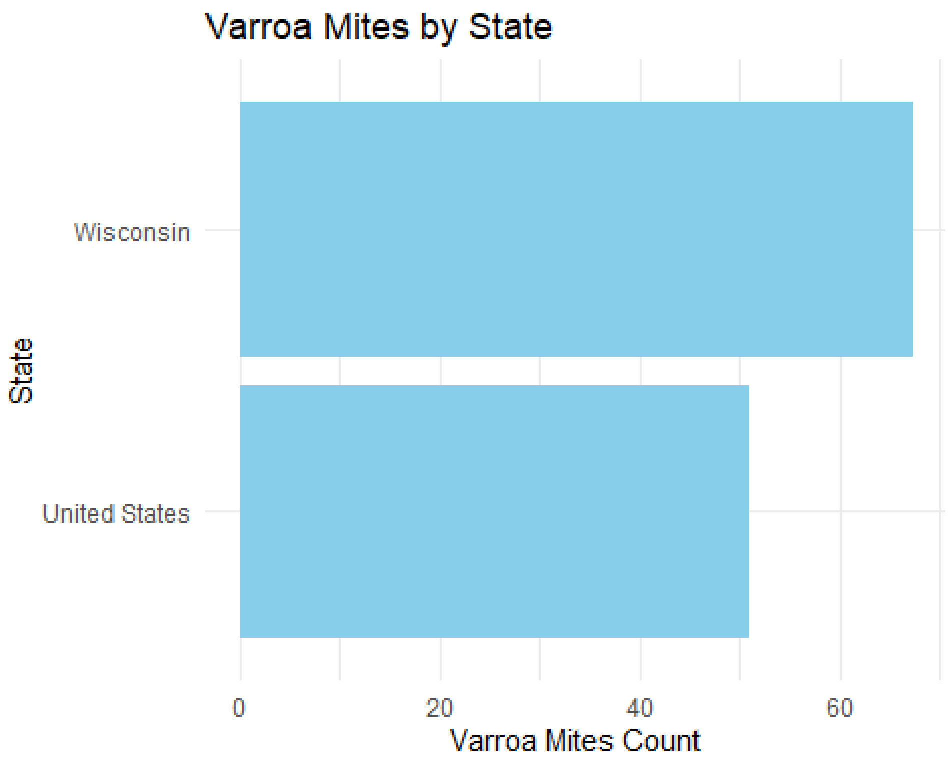
```
view(hcnydata_cleaned)

# Write the cleaned data to a new CSV file
write.csv(hcnydata_cleaned, "C:/Users/ecruz/OneDrive/Documents/UTSA - Data Sc
ience Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfo
lio-EdwardCruz/hcnydata_cleaned.csv", row.names = FALSE)

# Output cleaned data file to a csv file.
hcny_data <- read.csv("C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science
Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-Ed
wardCruz/hcny_CleanDraft.csv", header = FALSE, stringsAsFactors = FALSE)

head(hcny_data)

             V1           V2          V3      V4         V5    V6      V7
1         state varroa_mites other_pests disease pesticides other unknown
2  Pennsylvania         13.8         7.2     2.7        6.9   1.1     1.1
3         Texas         46.8        42.3    10.9       12.1  30.1    10.2
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6       6

str(hcny_data)

'data.frame':	5 obs. of  7 variables:
 $ V1: chr  "state" "Pennsylvania" "Texas" "Wisconsin" ...
 $ V2: chr  "varroa_mites" "13.8" "46.8" "67.2" ...
 $ V3: chr  "other_pests" "7.2" "42.3" "9.8" ...
 $ V4: chr  "disease" "2.7" "10.9" "47.8" ...
 $ V5: chr  "pesticides" "6.9" "12.1" "49.2" ...
 $ V6: chr  "other" "1.1" "30.1" "48.1" ...
 $ V7: chr  "unknown" "1.1" "10.2" "46.8" ...

summary(hcny_data)

      V1                 V2                 V3                 V4
 Length:5           Length:5           Length:5           Length:5
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
      V5                 V6                 V7
 Length:5           Length:5           Length:5
 Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character

# Step 2: Merge the first two rows to create a new header
new_header <- paste(hcny_data[1, ], hcny_data[2, ], sep = " ")

# Step 3: Set the new header
colnames(hcny_data) <- new_header

# Step 4: Remove the first three rows
hcny_data <- hcny_data[-c(1:3), ]


# Step 5: Remove completely blank rows
hcny_data <- hcny_data[rowSums(hcny_data != "") > 0, ]


# Display the resulting data frame
head(hcny_data)

  state Pennsylvania varroa_mites 13.8 other_pests 7.2 disease 2.7
4          Wisconsin              67.2             9.8        47.8
5      United States              50.9            13.9         6.5
  pesticides 6.9 other 1.1 unknown 1.1
4           49.2      48.1        46.8
5           10.5      13.6           6

# Assuming hcny_data is your cleaned data frame
colnames(hcny_data) <- c("state", "varroa_mites", "other_pests", "disease", "
pesticides", "other", "unknown")

head(hcny_data)

          state varroa_mites other_pests disease pesticides other unknown
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6       6
```

```
# Convert specified columns to numeric with warning suppression
hcnydata_cleaned <- hcny_data %>%
  mutate(across(c("varroa_mites", "other_pests", "disease", "pesticides", "ot
her", "unknown"), ~suppressWarnings(as.numeric(.))))

# Display the resulting data frame
head(hcnydata_cleaned)

          state varroa_mites other_pests disease pesticides other unknown
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6     6.0

head(hcnydata_cleaned)

          state varroa_mites other_pests disease pesticides other unknown
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6     6.0

# Verify the changes
str(hcnydata_cleaned)

'data.frame':   2 obs. of  7 variables:
 $ state       : chr  "Wisconsin" "United States"
 $ varroa_mites: num  67.2 50.9
 $ other_pests : num  9.8 13.9
 $ disease     : num  47.8 6.5
 $ pesticides  : num  49.2 10.5
 $ other       : num  48.1 13.6
 $ unknown     : num  46.8 6

head(hcnydata_cleaned)

          state varroa_mites other_pests disease pesticides other unknown
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6     6.0

ggplot(hcnydata_cleaned, aes(varroa_mites)) + geom_histogram()
```



```
view((hcnydata_cleaned))

# Write the cleaned data to a new CSV file
write.csv(hcnydata_cleaned, "C:/Users/ecruz/OneDrive/Documents/UTSA - Data Sc
ience Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfo
lio-EdwardCruz/hcny_CleanDraft.csv", row.names = FALSE)

# Assuming hcnydata_cleaned is your data frame
hcnydata_cleaned <- hcnydata_cleaned[apply(hcnydata_cleaned, 1, function(x) !
all(is.na(x))), ]

# Print the modified data frame to verify the changes
head(hcnydata_cleaned)

          state varroa_mites other_pests disease pesticides other unknown
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6     6.0

view(hcnydata_cleaned)

str(hcnydata_cleaned)

'data.frame':   2 obs. of  7 variables:
 $ state       : chr  "Wisconsin" "United States"
 $ varroa_mites: num  67.2 50.9
 $ other_pests : num  9.8 13.9
 $ disease     : num  47.8 6.5
 $ pesticides  : num  49.2 10.5
 $ other       : num  48.1 13.6
 $ unknown     : num  46.8 6

#update.packages('mgcv')
library(mgcv)

# Convert the 'state' column to a factor
hcnydata_cleaned$state <- as.factor(hcnydata_cleaned$state)

# Display the resulting data frame
head(hcnydata_cleaned)

          state varroa_mites other_pests disease pesticides other unknown
4     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
5 United States         50.9        13.9     6.5       10.5  13.6     6.0

# Step 3: Reindex the rows starting from 1
rownames(hcnydata_cleaned) <- NULL

# Identify the rows to exclude
rows_to_exclude <- c(46)

# Remove rows 45 to 52 while excluding row 46
data_hcny <- hcnydata_cleaned[-c(45:52)[-which(c(45:52) %in% rows_to_exclude)
], ]

library(ggplot2)
```

```r
# Create a bar graph to show state and varroa_mites
ggplot(data = data_hcny, aes(x = state, y = varroa_mites)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Varroa Mites by State",
       x = "State",
       y = "Varroa Mites Count") +
  theme_minimal() +
  coord_flip()
```



T Varroa Mite

The above image shows the count of Varroa Mites by States.Colorado and North Dakota comes highest in the plot with Oklahoma and New Mexico the lowest in count. This indicated that we have more of Varroa Mites in North Dakota and Colorado compared to Oklahoma and New Mexico ## Statistical analysis

Explain anything related to your statistical analyses.

Results

Texas Bee gathering pollen from natural Texas foliage

Exploratory/Descriptive analysis

Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.

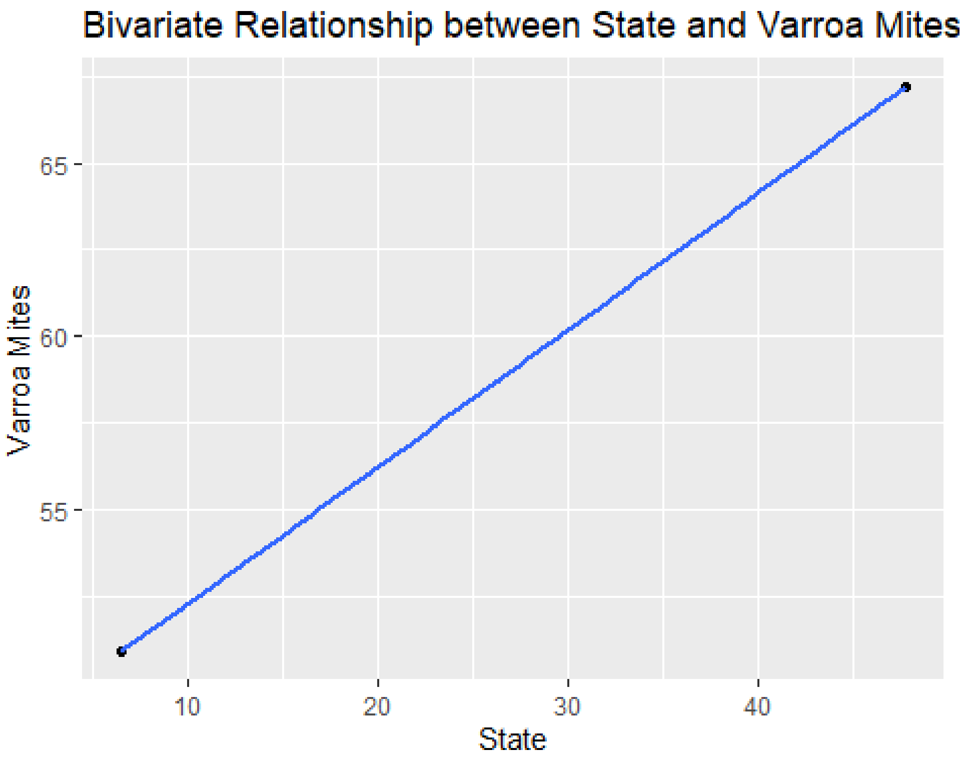
Remove all rows with any NA values

```
library(tidyr)
library(dplyr)


my_data <- data_hcny %>% drop_na()

# Verify that NAs have been removed
glimpse(my_data)
```

```
Rows: 2
Columns: 7
$ state        <fct> Wisconsin, United States
$ varroa_mites <dbl> 67.2, 50.9
$ other_pests  <dbl> 9.8, 13.9
$ disease      <dbl> 47.8, 6.5
$ pesticides   <dbl> 49.2, 10.5
$ other        <dbl> 48.1, 13.6
$ unknown      <dbl> 46.8, 6.0
```

Bivariate Analysis

```
## Example of a bivariate plot between 'state' and 'varroa_mites'
ggplot(my_data, aes(x = disease, y = varroa_mites)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Bivariate Relationship between State and Varroa Mites",
       x = "State",
       y = "Varroa Mites")
```



From the above graph we see that Varroa Mites concentrated more in certain states than others # Multivariable GLM

```
#install.packages("tidymodels")
library(tidymodels)

Warning: package 'tidymodels' was built under R version 4.3.3

── Attaching packages ────────────────────────────────────── tidymodels 1.2.0
──
```

```
✓ broom        1.0.6      ✓ rsample      1.2.1
✓ dials        1.2.1      ✓ tune         1.2.1
✓ infer        1.0.7      ✓ workflows    1.1.4
✓ modeldata    1.4.0      ✓ workflowsets 1.1.0
✓ parsnip      1.2.1      ✓ yardstick    1.3.1
✓ recipes      1.0.10
```

```
Warning: package 'broom' was built under R version 4.3.3

Warning: package 'dials' was built under R version 4.3.3

Warning: package 'infer' was built under R version 4.3.3

Warning: package 'modeldata' was built under R version 4.3.3

Warning: package 'parsnip' was built under R version 4.3.3

Warning: package 'recipes' was built under R version 4.3.3

Warning: package 'rsample' was built under R version 4.3.3

Warning: package 'tune' was built under R version 4.3.3

Warning: package 'workflows' was built under R version 4.3.3

Warning: package 'workflowsets' was built under R version 4.3.3

Warning: package 'yardstick' was built under R version 4.3.3
```

```
── Conflicts ────────────────────────────────────────── tidymodels_conflicts()
──
✖ nlme::collapse()  masks dplyr::collapse()
✖ purrr::discard()  masks scales::discard()
✖ dplyr::filter()   masks stats::filter()
✖ recipes::fixed()  masks stringr::fixed()
✖ dplyr::lag()      masks stats::lag()
✖ purrr::map()      masks maps::map()
✖ yardstick::spec() masks readr::spec()
✖ recipes::step()   masks stats::step()
• Use tidymodels_prefer() to resolve common conflicts.
```

```r
library(broom)
library(stats)
library(MASS)
```

```
Warning: package 'MASS' was built under R version 4.3.3


Attaching package: 'MASS'
```

```
The following object is masked from 'package:dplyr':

    select
```

```
# Specify the GLM model
glm_spec <- linear_reg() %>%
  set_engine("glm")

# Create a recipe for preprocessing the data
glm_recipe <- recipe(varroa_mites ~ other_pests + disease + pesticides + othe
r + unknown, data = my_data) %>%
  step_normalize(all_predictors())

# Create a workflow
glm_workflow <- workflow() %>%
  add_model(glm_spec) %>%
  add_recipe(glm_recipe)

# Fit the model
glm_fit <- fit(glm_workflow, data = my_data)

# Print the model summary using tidy()
model_summary <- tidy(glm_fit)
print(model_summary)
```

```
# A tibble: 6 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)     59.0       NaN       NaN     NaN
2 other_pests    -11.5       NaN       NaN     NaN
3 disease         NA          NA        NA      NA
4 pesticides      NA          NA        NA      NA
5 other           NA          NA        NA      NA
6 unknown         NA          NA        NA      NA
```

?@tbl-summarytable shows a summary of the data.

Note the loading of the data providing a **relative** path using the `../../` notation. (Two dots means a folder up). You never want to specify an **absolute** path like `C:\ahandel\myproject\results\` because if you share this with someone, it won't work for them since they don't have that path. You can also use the `here` R package to create paths. See examples of that below. I generally recommend the `here` package.

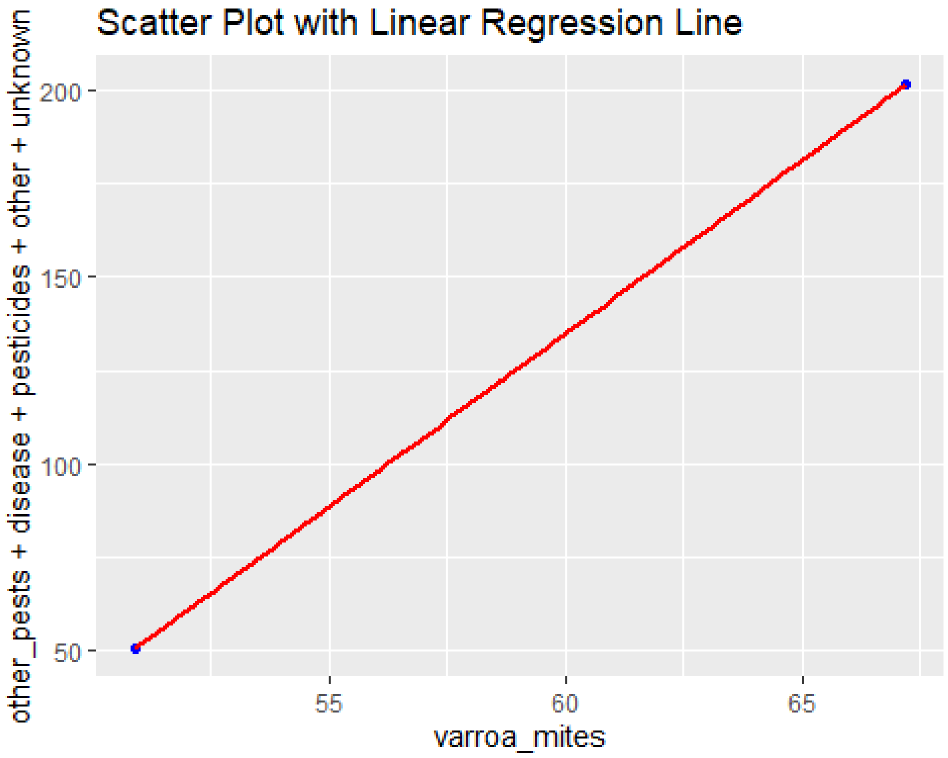
Basic statistical analysis

To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any "p<0.05 means statistical significance" interpretation is not valid.

?@fig-result shows a scatterplot figure produced by one of the R scripts.

```
plot <- ggplot(my_data, aes(x = varroa_mites , y =  other_pests + disease + p
esticides + other + unknown )) +
  geom_point(color = "blue") +              # Scatter plot
  geom_smooth(method = "lm", color = "red") + # Regression line
  ggtitle("Scatter Plot with Linear Regression Line") +
  xlab("varroa_mites") +
  ylab("other_pests + disease + pesticides + other + unknown")
print(plot)
```

```
`geom_smooth()` using formula = 'y ~ x'

Warning in qt((1 - level)/2, df): NaNs produced

Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



The above diagram indicates that varroa mites are proportionally related to other psests, diseases, pesticides and other unknown parasite

Full analysis

Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.

Example **?@tbl-resulttable2** shows a summary of a linear model fit.

```
saveRDS(my_data, file = "my_data.rds")
```

```
file.exists("my_data.rds")
```

```
[1] TRUE
```

```
getwd()
```

```
[1] "C:/Users/ecruz/OneDrive/Documents/UTSA - Data Science Program/Semester C
lasses/Practicum II Repository/P2-Practicum-II-Portfolio-EdwardCruz"
```

```
loaded_data <- readRDS("my_data.rds")
print(loaded_data)
```

```
          state varroa_mites other_pests disease pesticides other unknown
1     Wisconsin         67.2         9.8    47.8       49.2  48.1    46.8
2 United States         50.9        13.9     6.5       10.5  13.6     6.0
```

```
my_data <- readRDS("my_data.rds")
```

```
model <- lm(varroa_mites ~ other_pests + disease + pesticides + other + unkno
wn, data = my_data)
```

```
summary(model)
```

```


Call:
lm(formula = varroa_mites ~ other_pests + disease + pesticides + 
    other + unknown, data = my_data)

Residuals:
ALL 2 residuals are 0: no residual degrees of freedom!

Coefficients: (4 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  106.161        NaN     NaN      NaN
other_pests   -3.976        NaN     NaN      NaN
disease           NA         NA      NA       NA
pesticides        NA         NA      NA       NA
other             NA         NA      NA       NA
unknown           NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:    NaN 
F-statistic:   NaN on 1 and 0 DF,  p-value: NA
```

From the linear regression we see that only 49% of variation in the data can be explained by varroa mites, The p value is 0.01 which is statistically significant

```
library(broom)
library(knitr)

model_tidy <- tidy(model)

kable(model_tidy, caption = "Linear Model Fit Table")
```

Linear Model Fit Table

term	estimate	std.error	statistic	p.value
(Intercept)	106.16098	NaN	NaN	NaN
other_pests	-3.97561	NaN	NaN	NaN
disease	NA	NA	NA	NA
pesticides	NA	NA	NA	NA
other	NA	NA	NA	NA
unknown	NA	NA	NA	NA

The p value for varroa Mites is statistically significant at 0.009 when compared with other disease or pesticides.

Discussion

Summary and Interpretation

Summarize what you did, what you found and what it means.

Strengths and Limitations

Discuss what you perceive as strengths and limitations of your analysis.

Conclusions

What are the main take-home messages?

Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end

This paper [@leek2015] discusses types of analyses.

These papers [@mckay2020; @mckay2020a] are good examples of papers published using a fully reproducible setup similar to the one shown in this template.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal are available. You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

References