loading/cleaning/wrangling/exploring

Mojisola Popoola, Leonel Salazar, ED Cruz, Amelia

library(readxl)  
library(tidyverse)  
library(dplyr)  
library(readr)  
library(tidymodels)

# 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-EdwardCruz/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 Florida 49.5 21.8 7.1 18.2 15.1 7.7  
3 Georgia 30.2 12.2 2.4 3.9 9.8 7.1  
4 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3

str(hcny\_data)

'data.frame': 47 obs. of 7 variables:  
 $ V1: chr "state" "Florida" "Georgia" "Idaho" ...  
 $ V2: chr "varroa\_mites" "49.5" "30.2" "40.9" ...  
 $ V3: chr "other\_pests" "21.8" "12.2" "6.7" ...  
 $ V4: chr "disease" "7.1" "2.4" "5.9" ...  
 $ V5: chr "pesticides" "18.2" "3.9" "5.7" ...  
 $ V6: chr "other" "15.1" "9.8" "1.3" ...  
 $ V7: chr "unknown" "7.7" "7.1" "1.4" ...

summary(hcny\_data)

V1 V2 V3 V4   
 Length:47 Length:47 Length:47 Length:47   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
 V5 V6 V7   
 Length:47 Length:47 Length:47   
 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 Florida varroa\_mites 49.5 other\_pests 21.8 disease 7.1 pesticides 18.2  
4 Idaho 40.9 6.7 5.9 5.7  
5 Illinois 25.6 13.3 3.2 6.8  
6 Indiana 23.8 5.6 1.6 7.4  
NA <NA> <NA> <NA> <NA> <NA>  
8 Kansas 35.5 2 0.1 21.7  
9 Kentucky 8 2.9 1 1.3  
 other 15.1 unknown 7.7  
4 1.3 1.4  
5 6.7 3.7  
6 18.3 3  
NA <NA> <NA>  
8 3.4 2  
9 0.5 5.6

# 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 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3  
NA <NA> <NA> <NA> <NA> <NA> <NA> <NA>  
8 Kansas 35.5 2 0.1 21.7 3.4 2  
9 Kentucky 8 2.9 1 1.3 0.5 5.6

# Convert specified columns to numeric with warning suppression  
hcnydata\_cleaned <- hcny\_data %>%  
 mutate(across(c("varroa\_mites", "other\_pests", "disease", "pesticides", "other", "unknown"), ~suppressWarnings(as.numeric(.))))  
  
# Display the resulting data frame  
head(hcnydata\_cleaned)

state varroa\_mites other\_pests disease pesticides other unknown  
4 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3.0  
NA <NA> NA NA NA NA NA NA  
8 Kansas 35.5 2.0 0.1 21.7 3.4 2.0  
9 Kentucky 8.0 2.9 1.0 1.3 0.5 5.6

head(hcnydata\_cleaned)

state varroa\_mites other\_pests disease pesticides other unknown  
4 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3.0  
NA <NA> NA NA NA NA NA NA  
8 Kansas 35.5 2.0 0.1 21.7 3.4 2.0  
9 Kentucky 8.0 2.9 1.0 1.3 0.5 5.6

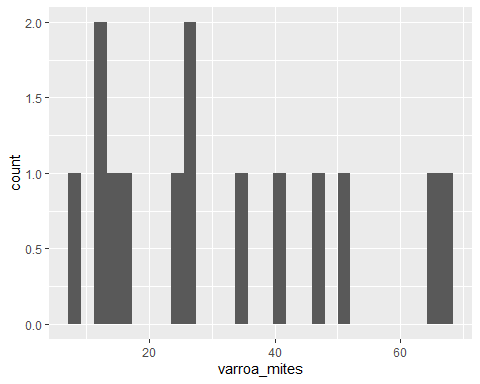
# Verify the changes  
str(hcnydata\_cleaned)

'data.frame': 44 obs. of 7 variables:  
 $ state : chr "Idaho" "Illinois" "Indiana" NA ...  
 $ varroa\_mites: num 40.9 25.6 23.8 NA 35.5 8 NA NA NA NA ...  
 $ other\_pests : num 6.7 13.3 5.6 NA 2 2.9 NA NA NA NA ...  
 $ disease : num 5.9 3.2 1.6 NA 0.1 1 NA NA NA NA ...  
 $ pesticides : num 5.7 6.8 7.4 NA 21.7 1.3 NA NA NA NA ...  
 $ other : num 1.3 6.7 18.3 NA 3.4 0.5 NA NA NA NA ...  
 $ unknown : num 1.4 3.7 3 NA 2 5.6 NA NA NA NA ...

head(hcnydata\_cleaned)

state varroa\_mites other\_pests disease pesticides other unknown  
4 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3.0  
NA <NA> NA NA NA NA NA NA  
8 Kansas 35.5 2.0 0.1 21.7 3.4 2.0  
9 Kentucky 8.0 2.9 1.0 1.3 0.5 5.6

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 Science Program/Semester Classes/Practicum II Repository/P2-Practicum-II-Portfolio-EdwardCruz/hcny\_CleanDraft.csv", row.names = FALSE)

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Conclusion- all code below is test code right now\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# 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 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3.0  
8 Kansas 35.5 2.0 0.1 21.7 3.4 2.0  
9 Kentucky 8.0 2.9 1.0 1.3 0.5 5.6  
14 Michigan 16.5 1.7 1.8 2.1 3.3 10.3

view(hcnydata\_cleaned)

str(hcnydata\_cleaned)

'data.frame': 14 obs. of 7 variables:  
 $ state : chr "Idaho" "Illinois" "Indiana" "Kansas" ...  
 $ varroa\_mites: num 40.9 25.6 23.8 35.5 8 16.5 12.6 13.1 26.8 65.3 ...  
 $ other\_pests : num 6.7 13.3 5.6 2 2.9 1.7 3.5 6.4 3.9 33.5 ...  
 $ disease : num 5.9 3.2 1.6 0.1 1 1.8 1 0.8 0.6 0.7 ...  
 $ pesticides : num 5.7 6.8 7.4 21.7 1.3 2.1 3.2 0.5 0.5 5.7 ...  
 $ other : num 1.3 6.7 18.3 3.4 0.5 3.3 9.1 4.8 0.6 1 ...  
 $ unknown : num 1.4 3.7 3 2 5.6 10.3 2.4 3.4 3.7 4.4 ...

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

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

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

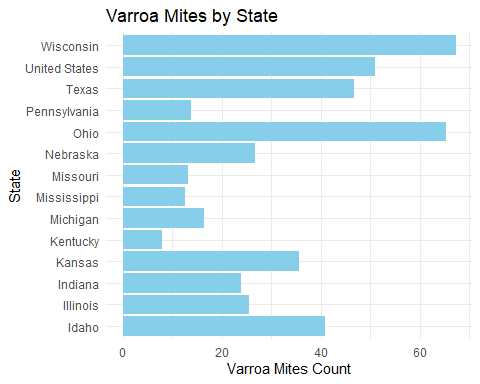
# Convert the 'state' column to a factor  
hcnydata\_cleaned$state <- as.factor(hcnydata\_cleaned$state)  
  
# Display the resulting data frame  
head(hcnydata\_cleaned)

state varroa\_mites other\_pests disease pesticides other unknown  
4 Idaho 40.9 6.7 5.9 5.7 1.3 1.4  
5 Illinois 25.6 13.3 3.2 6.8 6.7 3.7  
6 Indiana 23.8 5.6 1.6 7.4 18.3 3.0  
8 Kansas 35.5 2.0 0.1 21.7 3.4 2.0  
9 Kentucky 8.0 2.9 1.0 1.3 0.5 5.6  
14 Michigan 16.5 1.7 1.8 2.1 3.3 10.3

# 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)  
  
# 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()



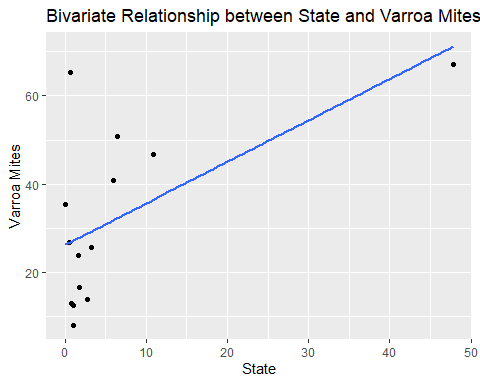
# Remove all rows with any NA values

library(tidyr)  
  
  
my\_data <- data\_hcny %>% drop\_na()  
  
# Verify that NAs have been removed  
glimpse(my\_data)

Rows: 14  
Columns: 7  
$ state <fct> Idaho, Illinois, Indiana, Kansas, Kentucky, Michigan, Mis…  
$ varroa\_mites <dbl> 40.9, 25.6, 23.8, 35.5, 8.0, 16.5, 12.6, 13.1, 26.8, 65.3…  
$ other\_pests <dbl> 6.7, 13.3, 5.6, 2.0, 2.9, 1.7, 3.5, 6.4, 3.9, 33.5, 7.2, …  
$ disease <dbl> 5.9, 3.2, 1.6, 0.1, 1.0, 1.8, 1.0, 0.8, 0.6, 0.7, 2.7, 10…  
$ pesticides <dbl> 5.7, 6.8, 7.4, 21.7, 1.3, 2.1, 3.2, 0.5, 0.5, 5.7, 6.9, 1…  
$ other <dbl> 1.3, 6.7, 18.3, 3.4, 0.5, 3.3, 9.1, 4.8, 0.6, 1.0, 1.1, 3…  
$ unknown <dbl> 1.4, 3.7, 3.0, 2.0, 5.6, 10.3, 2.4, 3.4, 3.7, 4.4, 1.1, 1…

# 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")



# Multivariable GLM

install.packages("tidymodels")

Warning: package 'tidymodels' is in use and will not be installed

library(tidymodels)  
library(broom)  
  
# 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 + other + 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) 31.9 3.25 9.82 0.00000972  
2 other\_pests 13.1 3.87 3.39 0.00950   
3 disease 9.97 16.7 0.597 0.567   
4 pesticides 13.9 7.89 1.76 0.116   
5 other -10.6 7.75 -1.37 0.207   
6 unknown -3.33 13.6 -0.245 0.812