Assignment 3: Data Exploration

Mazhar Bhuyan

Fall 2024

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

Directions

1. Rename this file `<FirstLast>_A03_DataExploration.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Assign a useful **name to each code chunk** and include ample **comments** with your code.
5. Be sure to **answer the questions** in this assignment document.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file.
7. After Knitting, submit the completed exercise (PDF file) to the dropbox in Canvas.

TIP: If your code extends past the page when knit, tidy your code by manually inserting line breaks.

TIP: If your code fails to knit, check that no `install.packages()` or `View()` commands exist in your code.

Set up your R session

1. Load necessary packages (tidyverse, lubridate, here), check your current working directory and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively. Be sure to include the subcommand to read strings in as factors.

```
install.packages("tidyverse")
install.packages("lubridate")
install.packages("here")
library(tidyverse)
library(conflicted)
conflicted::conflict_prefer("filter", "dplyr")
library(lubridate)
library(here)
library(forecast)
#needed to install conflicted package because "filter".  functions conflicts in "dplyr" package.
```

```
Neonics <-read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",
                   stringsAsFactors = TRUE,
                   colClasses = c("CAS.Number"= "factor",
                                  "Reference.Number" = "factor",
                                  "Publication.Year" = "factor")
)
write.csv(Neonics,
          file = here("./Data/Processed/Neonics.csv"),
          row.names = FALSE)

Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
                   stringsAsFactors = TRUE)
  #Not sure about dryMass and remarks columns. These two columns are number and logical data.     Shoul

write.csv(Litter,
           file = here("./Data/Processed/Litter.csv"),
           row.names = FALSE)

# stringASFactors Converts all character columns into factors, Useful for \ categorical data e.g., spec
```

Learn about your system

2. The neonicotinoid dataset was collected from the Environmental Protection Agency's ECOTOX Knowledgebase, a database for ecotoxicology research. Neonicotinoids are a class of insecticides used widely in agriculture. The dataset that has been pulled includes all studies published on insects. Why might we be interested in the ecotoxicology of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Neonicotinoids, although, widely used for protecting crops from insecticides, it has also usherd other kind of hazard. Since the use is not targeted towards specific kinds. of insects it is creating detrimental impact on bees and pollinators insects. The

3. The Niwot Ridge litter and woody debris dataset was collected from the National Ecological Observatory Network, which collectively includes 81 aquatic and terrestrial sites across 20 ecoclimatic domains. 32 of these sites sample forest litter and woody debris, and we will focus on the Niwot Ridge long-term ecological research (LTER) station in Colorado. Why might we be interested in studying litter and woody debris that falls to the ground in forests? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Study of litter and woody debris deposit in the forest is important to understant the forest health. The amount and type of deposit gives significant insight about nutrient cycling. These nutrients help growing new plants. It also retains water for plants and the forest ecosystem.Wody debris stores significant amount of carbons. These all data can be used for forecasting and carbon budget, efficiency and project the future.

4. How is litter and woody debris sampled as part of the NEON network? Read the NEON_Litterfall_UserGuide.pdf document to learn more. List three pieces of salient information about the sampling methods here:

   Answer: 1. Spatial Sampling: Sampling occurs in 20 m $\times$ 20 m or 40 m $\times$ 40 m plots at forested sites. 2. Spatial Sampling : Elevated traps are 0.5 m^2 PVC baskets placed ~80 cm above the ground. 3. Temporal Sampling : Ground traps are sampled once per year. Elevated traps are sampled every 2 weeks in deciduous forests during senescence and every 1–2 months in evergreen forests

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

```
## [1] 4623   30
```

```
dim(Litter)
```

```
## [1] 188  19
```

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest? [Tip: The `sort()` command is useful for listing the values in order of magnitude...]

```
sort(summary(Neonics$Effect), decreasing = TRUE)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development        Avoidance         Genetics
##              197              136              102               82
##        Enzyme(s)           Growth       Morphology    Immunological
##               62               38               22               16
##     Accumulation     Intoxication     Biochemistry          Cell(s)
##               12               12               11                9
##       Physiology        Histology       Hormone(s)
##                7                5                1
```

```
  # Summary function lists the effect of neonicotinoids as stated as observed mortality, growth   etc.

#sort(table(Neonics$Effect), decreasing = TRUE)
# we can also sort the summary of the column using 'table' function
```

Answer: Researchers most commonly study the population of different species, as it serves as a key indicator of the detrimental effects of neonicotinoids. Mortality rates also provide crucial insight into the impact of this insecticide. In addition to population and mortality, scientists examine insect behavior and feeding patterns to assess sublethal effects. Tracking these impacts allows researchers to understand and quantify the overall effects of neonicotinoids.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name). What do these species have in common, and why might they be of interest over other insects? Feel free to do a brief internet search for more information if needed.[TIP: Explore the help on the `summary()` function, in particular the `maxsum` argument...]

```
summary(Neonics$Species.Common.Name, maxsum = 15)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
```

```
##                  152                  140                  113
##       Japanese Beetle     Asian Lady Beetle        Euonymus Scale
##                   94                   76                   75
##             Wireworm     European Dark Bee     Minute Pirate Bug
##                   69                   66                   62
##  Asian Citrus Psyllid         Parastic Wasp              (Other)
##                   60                   58                 2523
```

Answer: The species categories are sorted based on the most commonly studied species. Among the top 15 species, most are honey bees. As major pollinators, honey bees play a crucial role in agriculture. When exposed to non-targeted pesticides, their colonies face the risk of extinction, which could have devastating effects on crop production and human survival. This threat makes honey bees a priority for research and conservation efforts over other species.

8. Concentrations are always a numeric value. What is the class of `Conc.1..Author.` column in the dataset, and why is it not numeric? [Tip: Viewing the dataframe may be helpful...]

```
class(Neonics$Conc.1..Author.)
```

```
## [1] "factor"
```

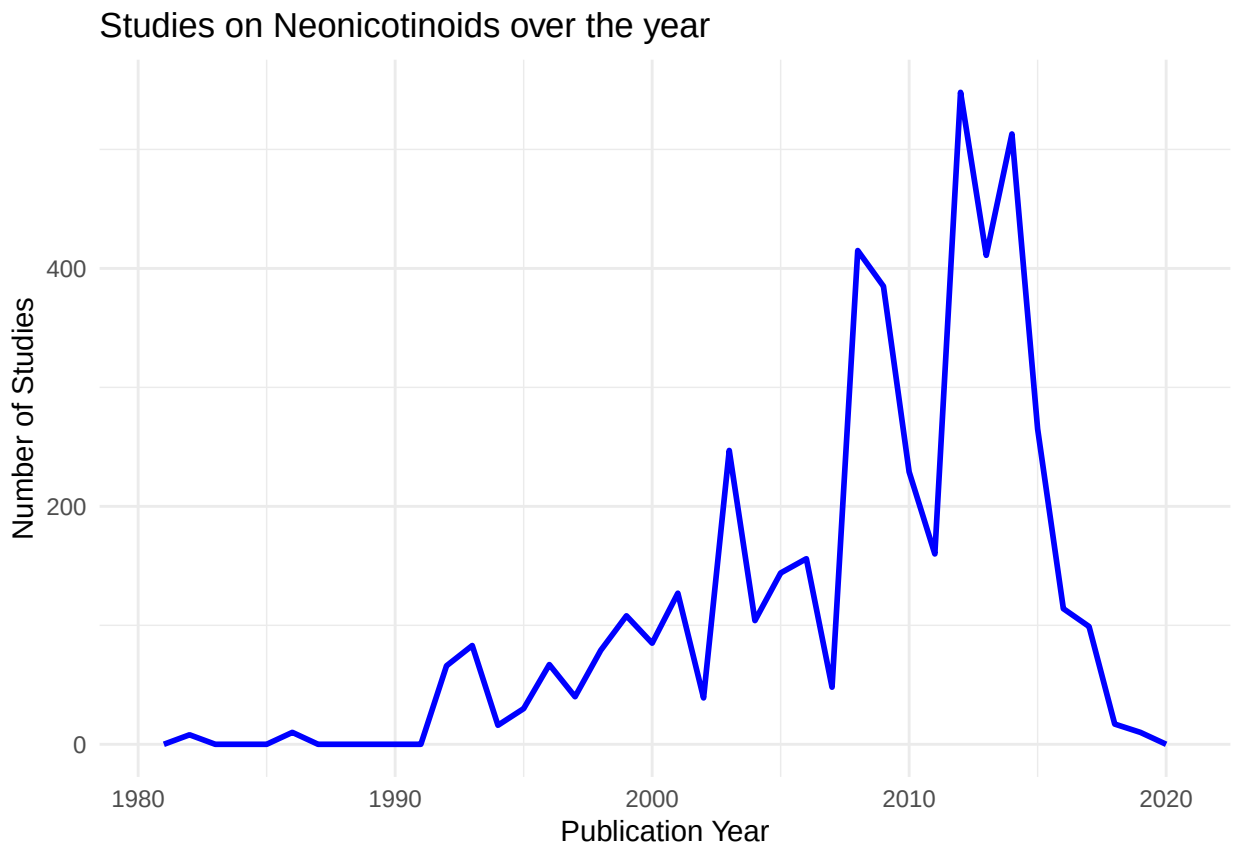
Answer: The Conc.1..Author column is a factor because it contains measurements of active ingredient concentrations expressed in different units, as specified in the Conc.1.Units..Author column. If all measurements used the same unit, the column could be numeric instead of a factor.

Explore your data graphically (Neonics)

9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
Neonics_Pub_Year <- Neonics %>%
      mutate(Publication_Year = as.numeric(as.character(Publication.Year)))

# Before plotting the column we need to convert the Publication.Year from factor to number. "as.numeric

ggplot(Neonics_Pub_Year, aes(x= Publication_Year)) +
  geom_freqpoly(binwidth = 1, color = "blue", size = 1) +
  ggtitle("Studies on Neonicotinoids over the year") +
  xlab("Publication Year") +
  ylab("Number of Studies") +
  theme_minimal()
```

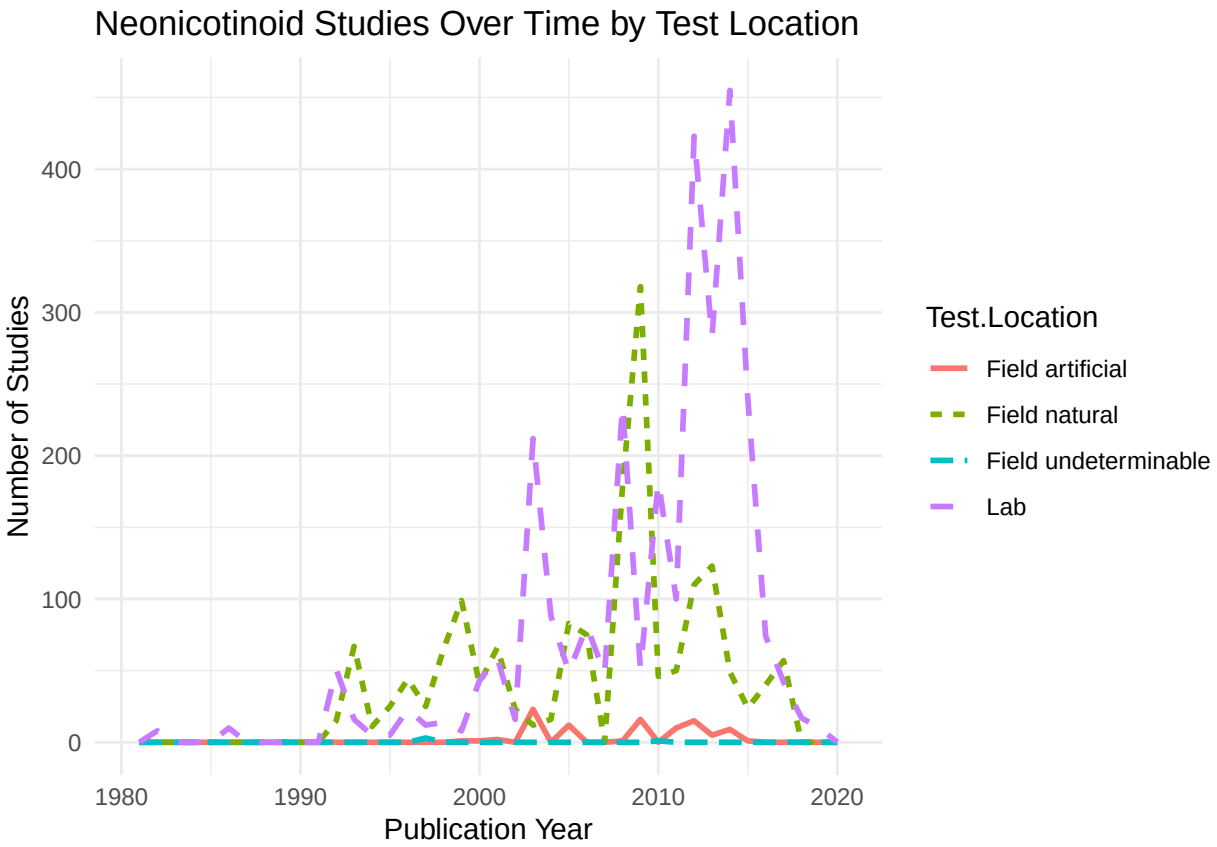
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```




```
#geom_freqpoly creates histogram.
#binwidth= yearly data. 2 means that the data is plotted 2 years apart
#size indicate the line width of the plot
#theme_minimal clears gridline and background elements
```

10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(Neonics_Pub_Year, aes(x = Publication_Year,
                             color = Test.Location,
                             linetype = Test.Location)) +
      geom_freqpoly(binwidth = 1, size = 1) +
      ggtitle("Neonicotinoid Studies Over Time by Test Location") +
      xlab("Publication Year") +
      ylab("Number of Studies") +
    theme_minimal()
```



Interpret this graph. What are the most common test locations, and do they differ over time?

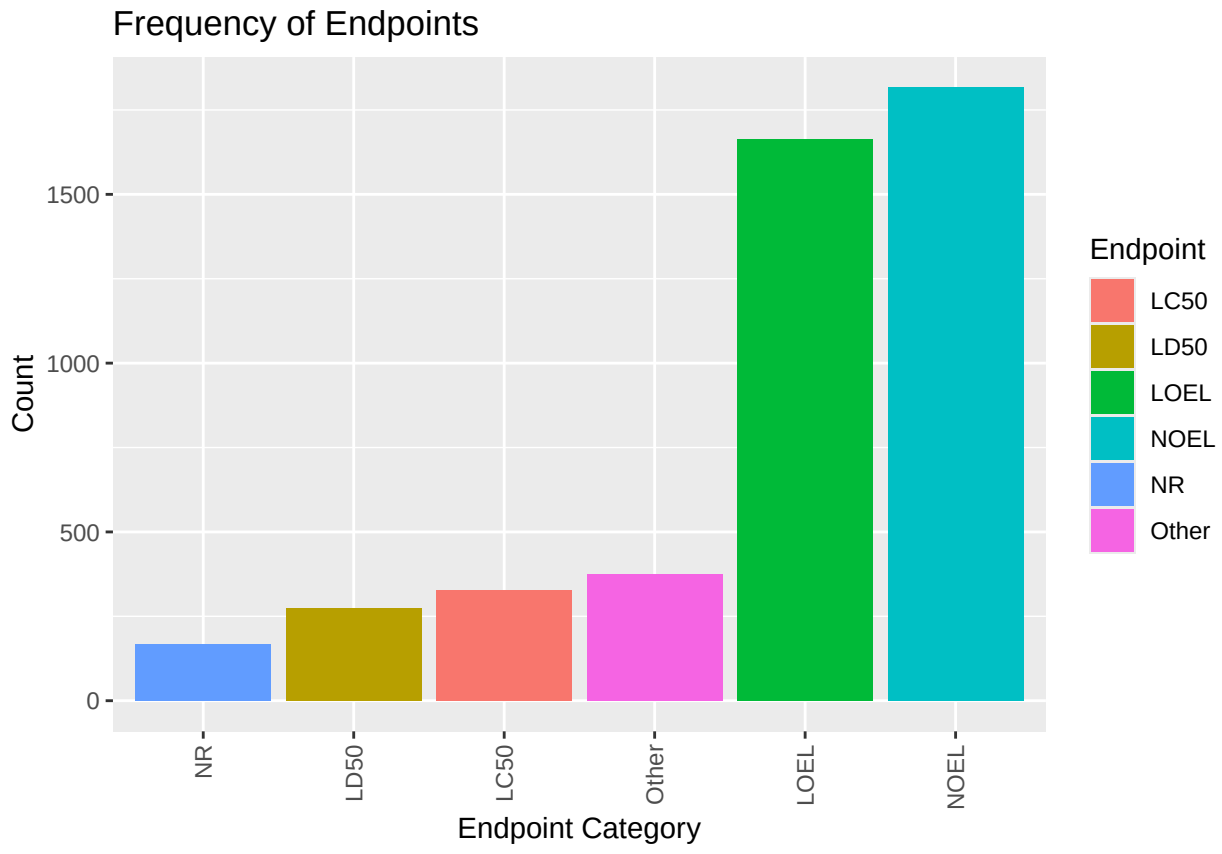
Answer: The most common test location is lab. Testing locations differ over time. From 1990 to 2010 period the most common test location was Field Artificial which has changed to Lab after 2010. At the same time Field Artificial test location has plummeted.

11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

[**TIP**: Add `theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))` to the end of your plot command to rotate and align the X-axis labels...]

```
#Endpoints counts are mostly concentrated with top 5 types. A general bar chart makes no sense. I want
#fct_lump can do the grouping.Easier version I guess.

# sorting in decreasing order.

Endpoint_Count <- Neonics %>%
  count(Endpoint, sort = TRUE)

# Groups the data by Endpoint.
# Counts how many times each unique Endpoint appears.
# Creates a new column named n, which holds the count of each unique Endpoint.

# Separating top 5
```

```r
top_5 <- head(Endpoint_Count, 5)

Remaining_Endpoint <- sum(Endpoint_Count$n[-(1:5)])
# Counting the Endpoints those are not in top 5 and group them in other category

Other_count <- data.frame(Endpoint = "Other", n = Remaining_Endpoint)

Endpoint_Count_Grouped <- bind_rows(top_5, Other_count)

ggplot(Endpoint_Count_Grouped, aes(x = reorder(Endpoint, n),
                                   y= n,
                                   fill = Endpoint)) +
  geom_bar(stat = "identity") +
  ggtitle("Frequency of Endpoints") +
  xlab("Endpoint Category") +
  ylab("Count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```r
# reorder function sorted the Endpoints based on frequency
# stats= identify supply the precomputed counts.
```

Answer: Two most common endpoints are:

NOEL: No-observable-effect-level: highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test.

LOEL: Lowest-observable-effect-level: lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)

Explore your data (Litter)

12. Determine the class of collectDate. Is it a date? If not, change to a date and confirm the new class of the variable. Using the `unique` function, determine which dates litter was sampled in August 2018.

```
class(Litter$collectDate)
```

```
## [1] "factor"
```

```
Litter$collectDate <- ymd(Litter$collectDate)
class(Litter$collectDate)
```

```
## [1] "Date"
```

```
unique_dates_august_2018 <- unique(Litter$collectDate[month(Litter$collectDate) == 8 & year(Litter$colle
print(unique_dates_august_2018)
```

```
## [1] "2018-08-02" "2018-08-30"
```

13. Using the `unique` function, determine how many different plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```
summary(Litter)
```

```
##                                     uid                     namedLocation
##  028eea3d-5c20-4afc-bb7e-a05bab305152:   1   NIWO_040.basePlot.ltr:20
##  06789d7b-b742-41d9-8556-79d23c193dc0:   1   NIWO_041.basePlot.ltr:19
##  07780a1e-8af9-4b8a-bb9b-be8add15a1e0:   1   NIWO_046.basePlot.ltr:18
##  0a6cae78-ea42-4e68-98c6-9d929068a38a:   1   NIWO_061.basePlot.ltr:17
##  0ae1c621-387e-42a9-bcf3-7ad1c9b97ab4:   1   NIWO_067.basePlot.ltr:17
##  0b274782-8e52-4f6a-bb17-36daa821f929:   1   NIWO_058.basePlot.ltr:16
##  (Other)                             :182   (Other)              :81
##  domainID   siteID         plotID            trapID        weighDate
##  D13:188   NIWO:188   NIWO_040:20   NIWO_040_205:20   2018-08-06:91
##                       NIWO_041:19   NIWO_041_059:19   2018-09-05:97
##                       NIWO_046:18   NIWO_046_155:18
##                       NIWO_061:17   NIWO_061_169:17
##                       NIWO_067:17   NIWO_067_017:17
##                       NIWO_058:16   NIWO_058_101:16
##                       (Other) :81   (Other)     :81
##        setDate    collectDate                  ovenStartDate
##  2018-07-05:91   Min.   :2018-08-02   2018-08-02T21:00Z:91
##  2018-08-02:97   1st Qu.:2018-08-02   2018-08-30T22:30Z:97
##                  Median :2018-08-30
##                  Mean   :2018-08-16
##                  3rd Qu.:2018-08-30
##                  Max.   :2018-08-30
```

```
##
##             ovenEndDate                       fieldSampleID
##  2018-08-06T18:02Z:91   NEON.LTR.NIWO041059.20180830: 11
##  2018-09-05T19:30Z:97   NEON.LTR.NIWO040205.20180802: 10
##                         NEON.LTR.NIWO040205.20180830: 10
##                         NEON.LTR.NIWO046155.20180802: 10
##                         NEON.LTR.NIWO058101.20180802:  9
##                         NEON.LTR.NIWO061169.20180802:  9
##                         (Other)                     :129
##                            massSampleID      samplingProtocolVersion
##  NEON.LTR.NIWO040205.20180802.MXT:  2   NEON.DOC.001710vE:188
##  NEON.LTR.NIWO040205.20180802.NDL:  2
##  NEON.LTR.NIWO040205.20180830.MXT:  2
##  NEON.LTR.NIWO040205.20180830.NDL:  2
##  NEON.LTR.NIWO041059.20180830.MXT:  2
##  NEON.LTR.NIWO041059.20180830.NDL:  2
##  (Other)                          :176
##        functionalGroup    dryMass       qaDryMass remarks
##  Needles       :30     Min.   :0.0000   N:168     Mode:logical
##  Twigs/branches:28     1st Qu.:0.0000   Y: 20     NA's:188
##  Woody material:26     Median :0.0050
##  Leaves        :24     Mean   :0.6115
##  Other         :24     3rd Qu.:0.3200
##  Flowers       :23     Max.   :8.6300
##  (Other)       :33
##                        measuredBy
##  kstyers@battelleecology.org:91
##  szrillo@battelleecology.org:97
##
##
##
##
##
```

```
unique_plot <- unique(Litter$plotID)
print(unique_plot)
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

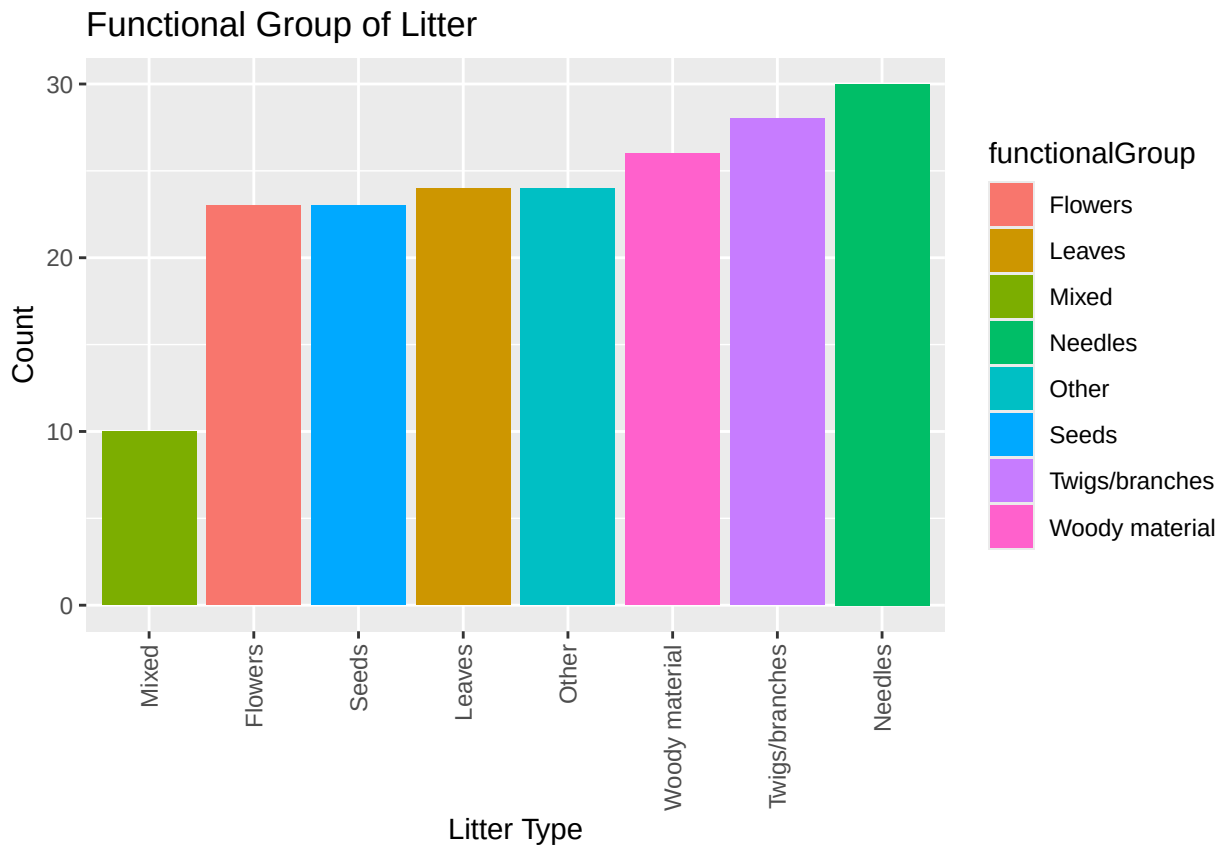
Answer: The summary function provides an overview of the data, including count, frequency, mean, and median for numeric columns. In contrast, unique focuses on a specific variable, showing only its distinct values.

14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
class(Litter$functionalGroup)
```

```
## [1] "factor"
```

```
Litter_type_count <- Litter %>%
    count(functionalGroup, sort = TRUE)


ggplot(Litter_type_count, aes(x = reorder(functionalGroup, n),
                              y = n,
                              fill = functionalGroup)) +
  geom_bar(stat = "identity") +
  ggtitle("Functional Group of Litter") +
  xlab("Litter Type") +
  ylab("Count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```

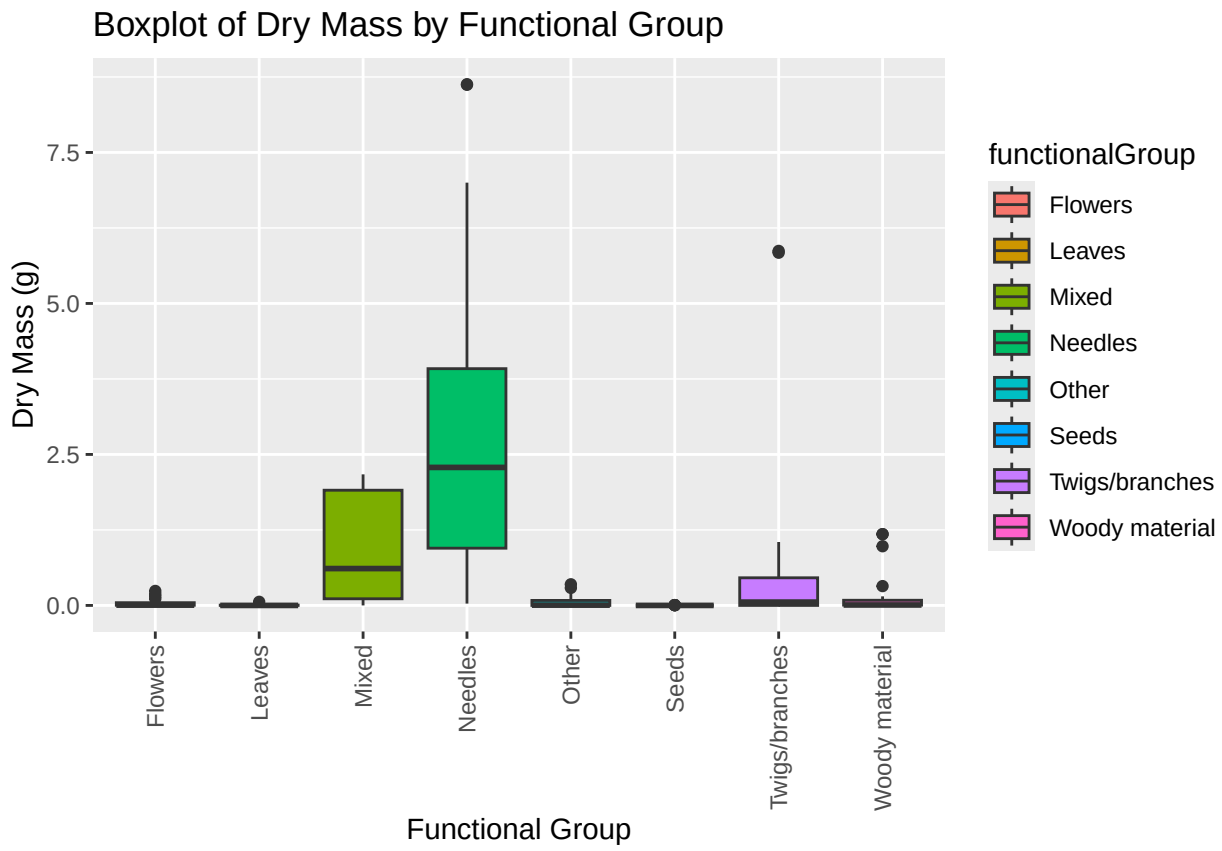


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

###Boxplot

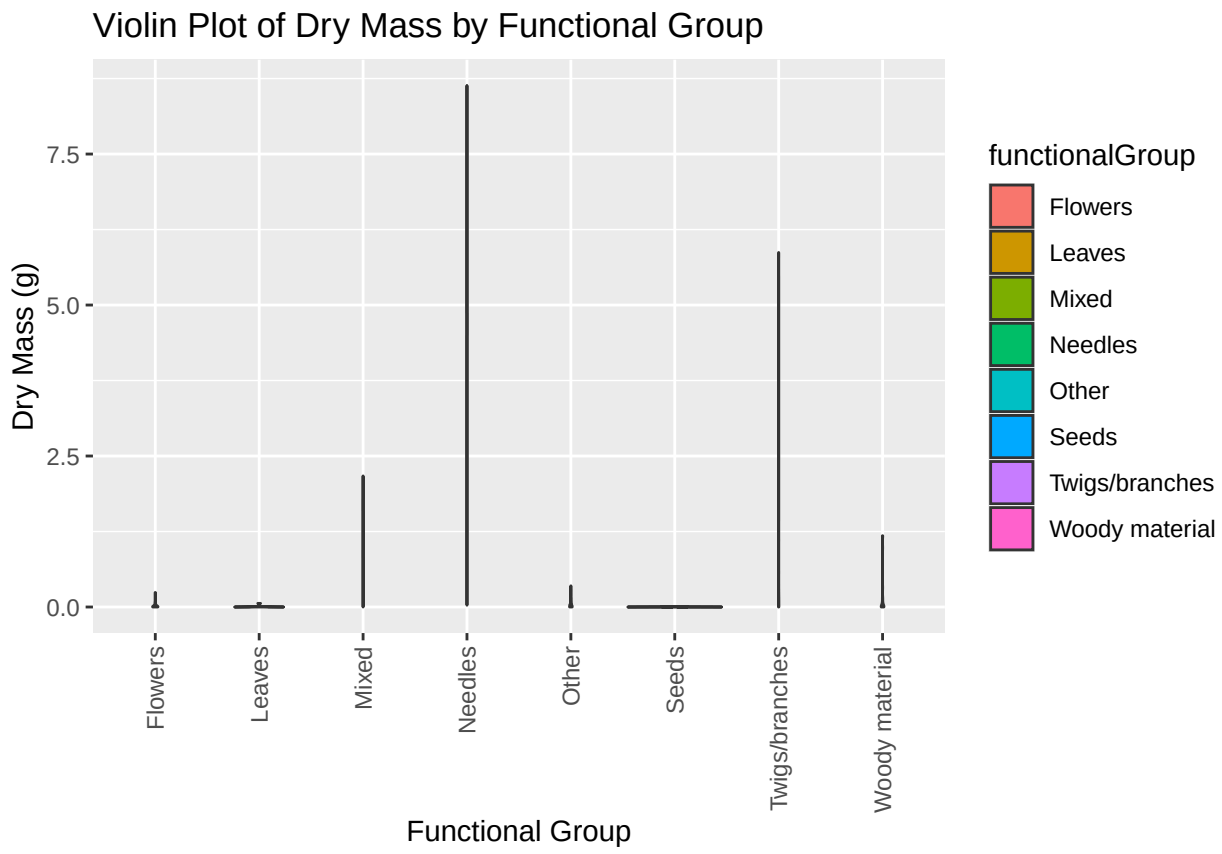
```
Litter_clean <- Litter %>%
  filter(!is.na(dryMass))

ggplot(Litter_clean, aes(x = functionalGroup, y = dryMass,
                         fill = functionalGroup)) +
  geom_boxplot() +
```

```
  xlab("Functional Group") +
  ylab("Dry Mass (g)") +
  ggtitle("Boxplot of Dry Mass by Functional Group") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



###Violin Plot

```
ggplot(Litter_clean, aes(x = functionalGroup, y = dryMass,
                         fill = functionalGroup)) +
  geom_violin() +
  xlab("Functional Group") +
  ylab("Dry Mass (g)") +
  ggtitle("Violin Plot of Dry Mass by Functional Group") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```




```
#theme_minimal()
```

Why is the boxplot a more effective visualization option than the violin plot in this case?

> Answer: Boxplots is more effective visualization option in this case beacuse the dataset is considerably large. Boxplot offers clear information about quartiles and outliers, making it easier to interpret the spread and central tendency of dry mass values across different functional groups. Since many functional groups have little to no dry mass, the violin plot appears as thin, barely visible shapes, making comparisons harder. The boxplot, on the other hand, clearly shows the differences between groups with distinct medians and quartiles, making it easier to interpret.

What type(s) of litter tend to have the highest biomass at these sites?

> Answer: Needles