Assignment 3: Data Exploration

Rosie Wu

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.

```
# this loads library packages tidyverse, lubricate, here
library(tidyverse);library(lubridate); library(here)
# now import the two csv files with read strings as factors in subcommands.
Neonics <- read.csv(here('Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv'), stringsAsFactors = TRUE)
Litter <- read.csv(here('Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv'), stringsAsFactors = TRUE)
# check work directory
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

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 are a class of widely used insecticides, and they can have detrimental impacts on the ecosystems, since it can be harmful for the pollinators as it travels through the food web. When it enters the nutrition pyramid of the food chain, it can also further harm the biodiversity and health of other living organisms. Therefore, we need to comprehend the harm of this toxin and form a comprensive framework regarding this insecticide in our agricultural regulations to ensure the environmental responsibility of this technique we implement so widely.

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: The woody debris and litter falling to the ground are essential sources of nutrition in the nutrient cycle for plant growth. The fallen debris can also help provide a great habitat for many microorganisms, insects, invertebrates etc. In addtion, sufficient level of woody debris and litter can help ensure good soil health and ensure the carbon sequestration process in nature while also prevents water overflow to a certain extent. Therefore, proper and suffcient wood debris falling to the ground in the forest is essential for the ecosystem and thus interests us.

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. The two main sampling design this research uses were Spatial Sampling Design and Temporal Sampling Design. 2. Spatial Sampling: Litter and fine woody debris sampling is practiced at NEON sites with woody vegetation >2m tall, and sampling for this product occurs only in tower plots that were randomized. Temporal Sampling Design: the research group set up ground traps sampled once a year, with a dimension of clip cells within a 20mx20m plot. 3. In the sampling process, the litterfall and fine woody debris sampling data yield mass data for plant functional groups from individual sampling sessions. Litter is collected from elevated traps, while fine woody debris is gathered from ground traps. All processed mass data are presented at the spatial level of a single trap and the temporal level of a single collection event.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
# Use dim to call out the dimensions of the Neonics dataset
dim(Neonics)
```

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

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...]

```
# The goal of this is to find the most common effect in the "Effect" column.
# first summary function on effect column, and sort
sort(summary(Neonics$Effect))
```

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

Answer: According to the sorted summary result, the Effect with highest values (top 3) are population, mortality, and behavior. Why might these be of interest? Because they are the most fundamental measures of the health of a species. If a certain species has lower population after the insectcide has been applied, that indicates the insecticides have negative impact on the reproduction of the species.When there is a high mortality rate, that means insectcides have been causing more deaths. Meanwhile, if the bahavior of a specie is changing in a noticeable pattern, then that means the effect may be significant as well. If the insecticide has siginificant effect on these aspects, then it means attention for this insecticide is needed.

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 function is still used.
# Since we need the 6 most commonly studied species (or the top 6 max)
# an argument maxsum= 6 is added inside the summary function.
summary(Neonics$Species.Common.Name, maxsum = 6)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee               (Other)
##                   152                   140                  3196
```

```
summary(Neonics$Species.Common.Name, maxsum = 7)
```

```
##             Honey Bee        Parasitic Wasp Buff Tailed Bumblebee
##                   667                   285                   183
##   Carniolan Honey Bee            Bumble Bee      Italian Honeybee
##                   152                   140                   113
##               (Other)
##                  3083
```

```
# I also used 7 to test out, since the most common is categorized as Other.
# But besides the "Other" all the other named 6 tops species are bees.
```

Answer: According to the output, all the top 6 species were bees. Compared to other species, bees tend to be classified as pollinators, especially honey and bumblebees. Pollinators are of interests over other insects, since they especially help with the reproductive process of plants and crops, which benefit the agriculture and ecosystems significantly.

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...]

```
# First manually view the dataframe and the specific column
# View(Neonics)
# use class function to classify the column
class(Neonics$Conc.1..Author.)
```

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

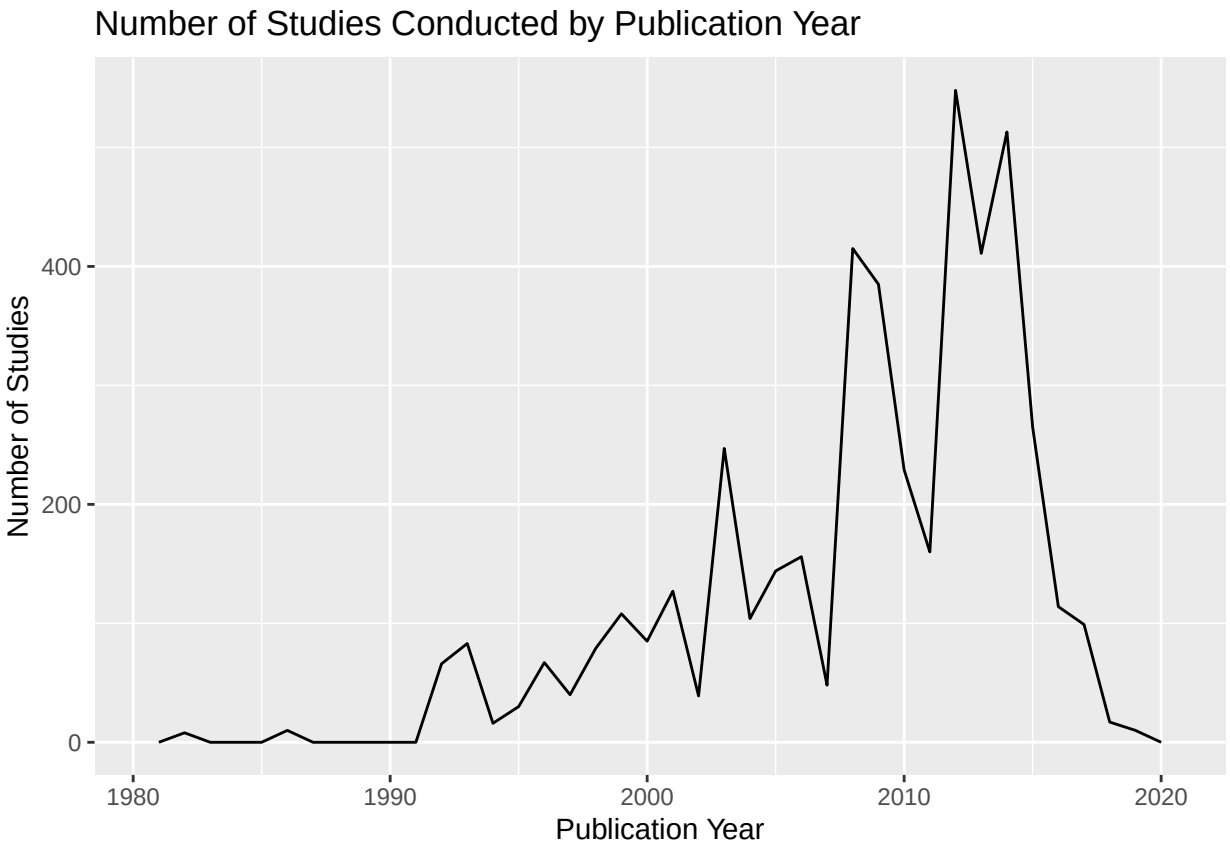
```
# The return for the class should be factor
```

Answer: This is not numeric but a "factor" instead. Because: after exploring the designated column in the dataframe, there appears to be "/" after some values in some cells, which make these not completely numeric values but characters.

Explore your data graphically (Neonics)

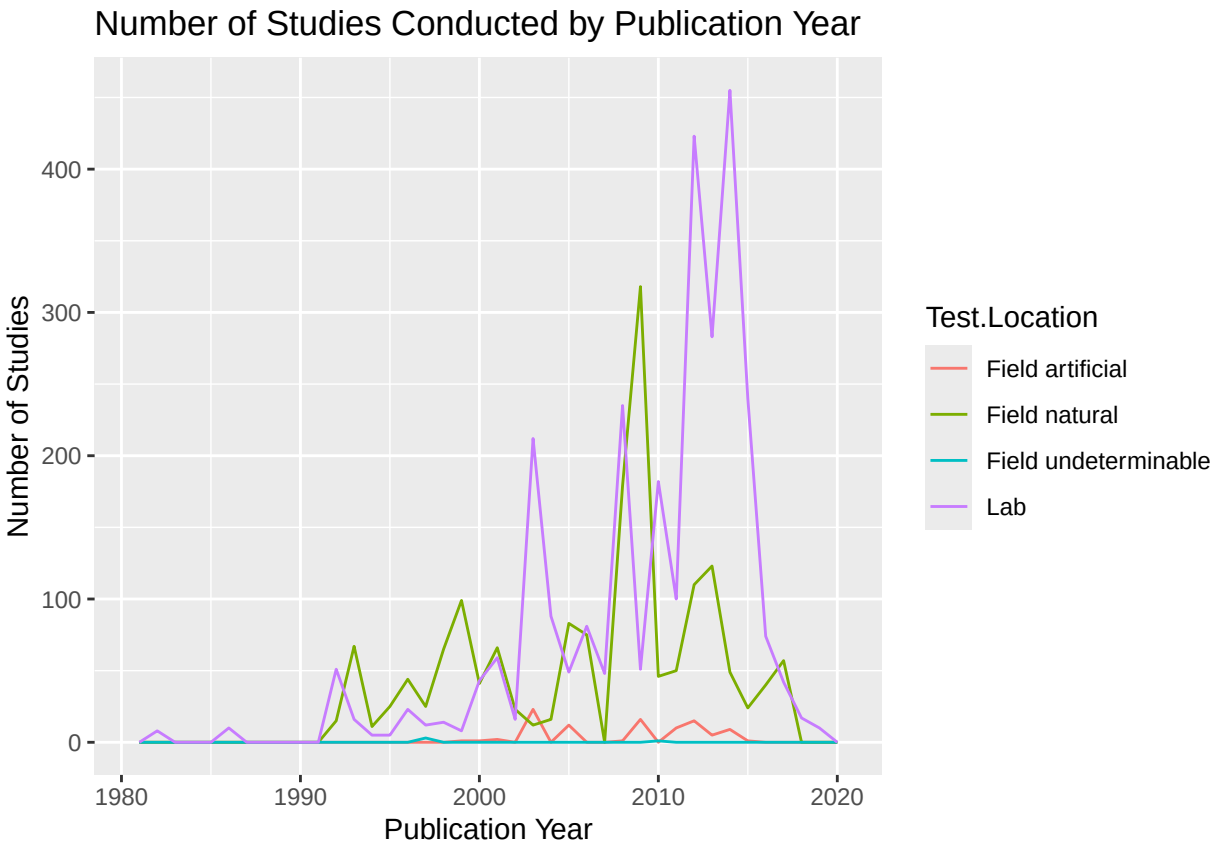
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
# first import the ggplot library
library(ggplot2)
# load the dataframe into the plot and set x axis as Publication year
# set bin width along with the chart tile and axis titles.
# y supposed to be the count of studies by publication year
ggplot(Neonics, aes(x = Publication.Year)) +
geom_freqpoly(binwidth = 1)+
  labs(title = "Number of Studies Conducted by Publication Year",
       x = "Publication Year",
       y = "Number of Studies")
```



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

```
# load the dataframe into the plot and set x axis as Publication year
# set bin width to 1 and set color by locations, along with the chart and axis titles.
# x and y kept the same
ggplot(Neonics, aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth = 1) +
  labs(title = "Number of Studies Conducted by Publication Year",
       x = "Publication Year",
       y = "Number of Studies")
```



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

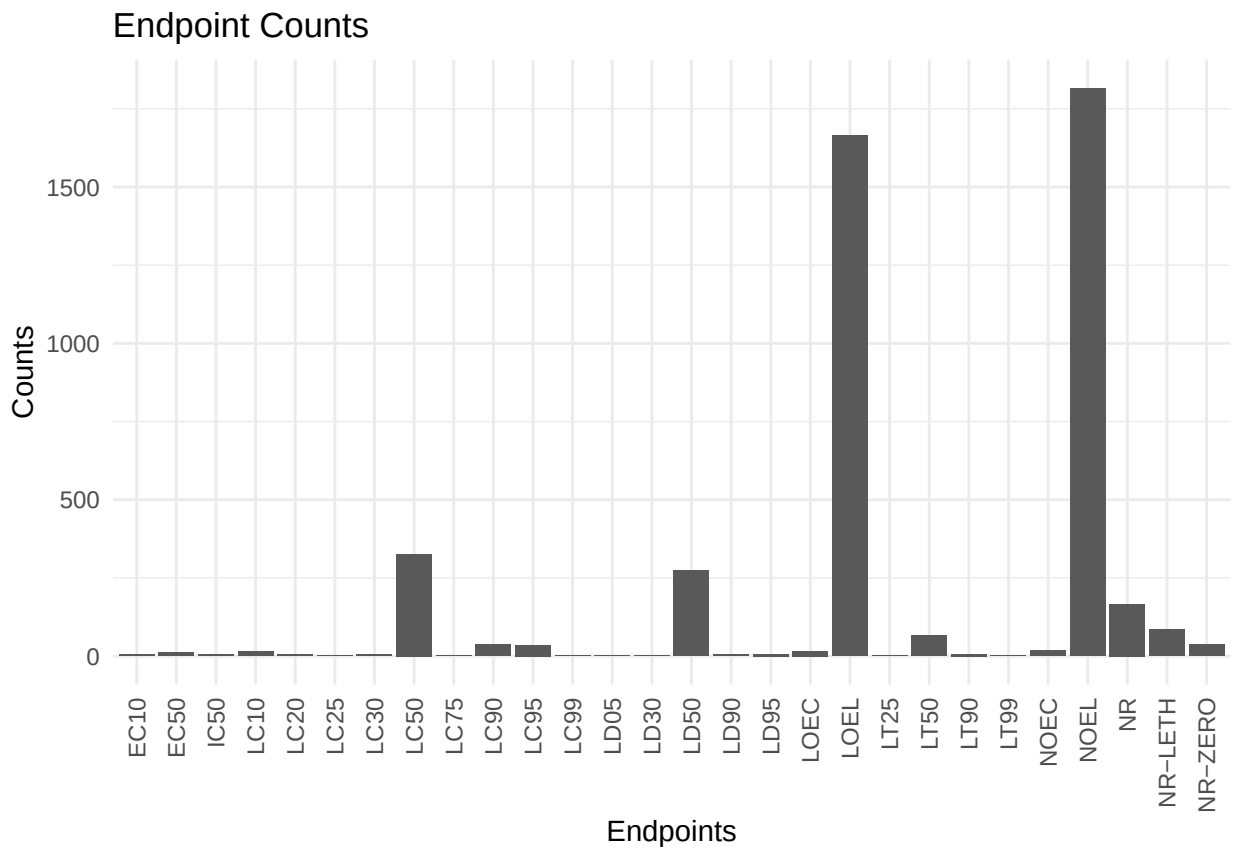
Answer: From the coloring of the graph, the plot with highest count is "Lab", and the count dramatically increased after around 2002, but it constantly drops or spikes dramatically overtime. The second highest is "Field Natural". Before 2010, the two locations rotate as the highest count, but Lab tends to be much higher after 2010, and it especially spiked around 2015.

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...]

```
# similar approach as previous one, use ggplot and then add geom_bar as bar gragh
# X axis being Endpoints
# y auto becomes Count.
# add the tile of chart and axis
# format the labels as the tip suggests.

ggplot(data = Neonics, aes(x = Endpoint )) +
  geom_bar() +
  labs(title = "Endpoint Counts", x = "Endpoints", y = "Counts") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Answer: The highest two counts fall under Enpoints LOEL and NOEL, which are Lowest-observable-effect-level– "lowest dose (concentration) producing effects that were significantly different (as reported by authors) from responses of controls (LOEAL/LOEC)" and No-observable-effect-level – "highest dose (concentration) producing effects not significantly different from responses of controls according to author's reported statistical test". (quoted from ECOTOX_CodeAppendix)

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.

```
# first see the class of collectDate as variable
class(Litter$collectDate)
```

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

```
# The output of the class is factor, so need to convert this to date
Litter$collectDate <- as.Date(Litter$collectDate, format = '%Y-%m-%d')
# check again
class(Litter$collectDate)
```

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

```r
# Use unique function to
unique(Litter$collectDate)
```

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

```r
# Answer for the dates that litter was sampled in August 2018 are 2018-08-02 and 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`?

```r
# using just unique function
unique(Litter$plotID)
```

```
##  [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
```

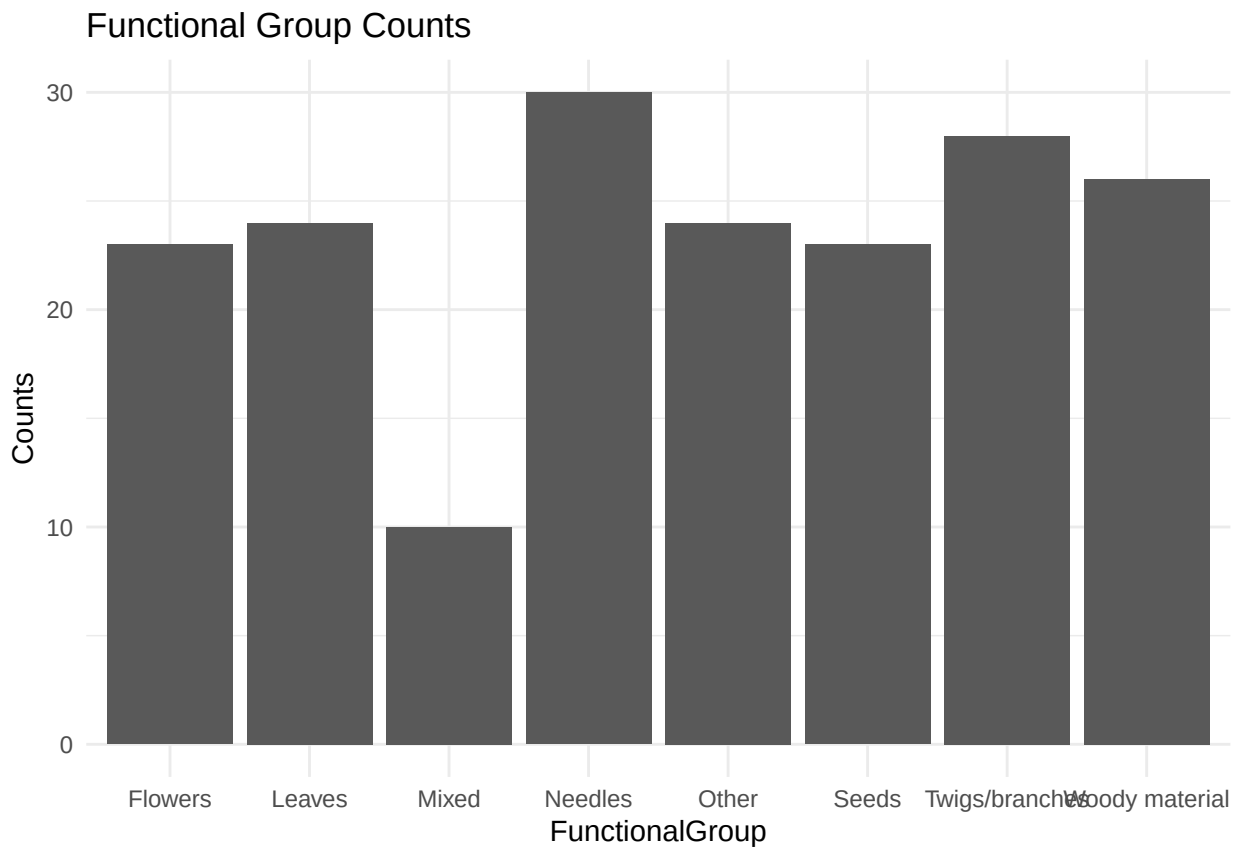
```r
# use summary for this column too
summary(Litter$plotID)
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: Unique function only listed out all PlotIDs and gave the list in its Descending order by count of the IDs, whereas the summary gave the Plot Ids by descending order along with its count.

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.

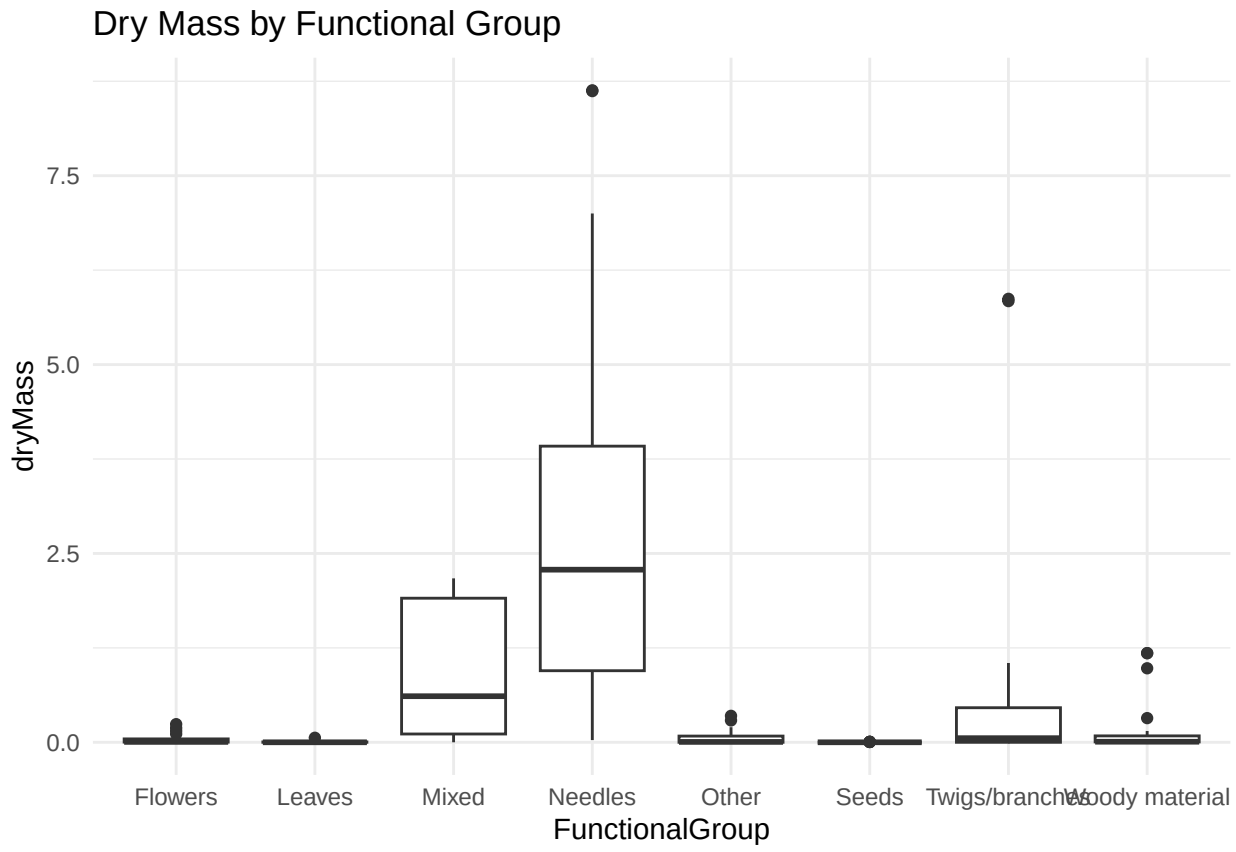
```r
# similar approach as #11, use ggplot and then add geom_bar as bar gragh
# X axis being functionalGroup
# y auto becomes Count.
# add the tile of chart and axis
ggplot(data = Litter, aes(x = functionalGroup)) +
  geom_bar() +
  labs(title = "Functional Group Counts", x = "FunctionalGroup", y = "Counts") +
  theme_minimal()
```



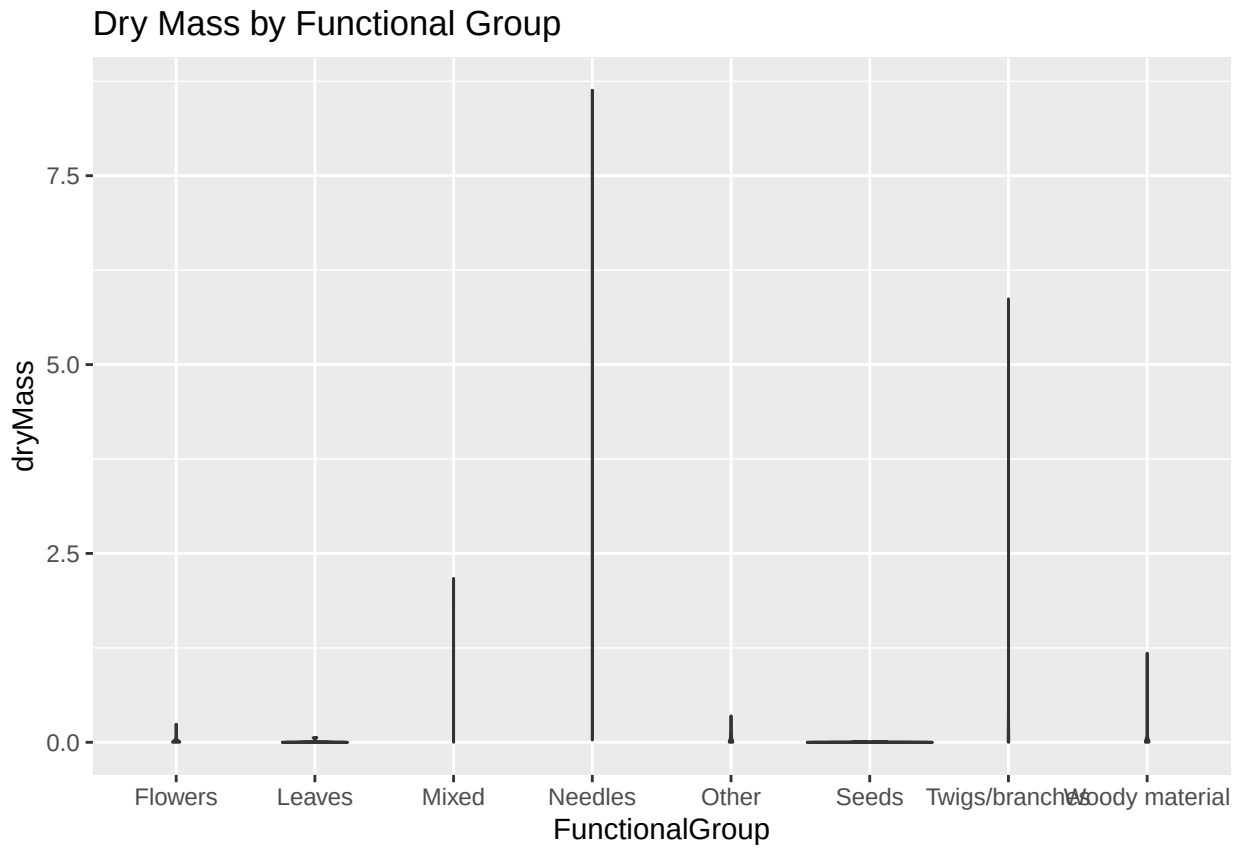

```
# This graph shows that the Litter types are fairly distributed across Niwot Ridge sites,
# with Mixed lower than everyone else.
```

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

```
# similar approach as #11, use ggplot and then add geom_bar as boxplot,
# X axis being functionalGroup
# y be dryMass, since it asks for DryMass by Functional Group
ggplot(Litter)+
  geom_boxplot(aes(x = functionalGroup, y = dryMass, group = cut_width(functionalGroup, 1))) +
  labs(title = "Dry Mass by Functional Group", x = "FunctionalGroup", y = "dryMass") +
  theme_minimal()
```




```
# keep other config the same and do the Violin plot
# also add quantiles as the previous lecture suggests
ggplot(Litter)+
  geom_violin(aes(x = functionalGroup, y = dryMass, group = cut_width(functionalGroup, 1)),
              draw_quantiles = c(0.25, 0.5, 0.75)) +
  labs(title = "Dry Mass by Functional Group", x = "FunctionalGroup", y = "dryMass")
```



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

> Answer: From the visualization perspective, the boxplot is easier to read its quantiles, ranges, and outliers, and it can be more easily understood as well.

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

> Answer: Needles tend to have highest biomass overall at these sites. This can be reflected from its range and average compared to others in the graph.