Assignment 3: Data Exploration

Julia Kagiliery

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.

```
library(tidyverse)
library(lubridate)
library(here)
library(gridExtra)

getwd()
```

```
## [1] "/Users/juliakagiliery/Library/Mobile Documents/com~apple~CloudDocs/GitHub Links/EDAClas2025"
```

```
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)
```

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: I might be interested in the ecotoxicology of these pesticides on insects because I want to understand the impact of using pesticides on agricultural feilds for surrounding areas. The pesticides may become runoff from the farms and then start impacting non-target insects (like bees or lady bugs which provide benefits to the envionment like polination and pest control). Understanding this ecotoxicology of different pesticides might help me select a pesticide that minimzies the adverse effects for non-target species.

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: Litter and woody debris that fall to the ground in forests can become fuel for forest fires. Having an excess amount of this debris on the ground might be correlated to the severity of a possible fire and hence understanding the litter and debris might inform forest managment practices (such as controlled burns) to protect against forest fire.

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. Litter is defined as material that drops from the forest canopy with a butt end diameter less than 2 cm and a length less than 50 cm which is collected in an elevanted (at 80 cm off the ground) 0.5m^2 PVC trap. Fine wood debries is definied material that is dropped from the forest canpoy but exceeds the 2 cm diameter and the 50 cm length which is collected in ground traps which have an area of 3 m x 0.5 m. 2. The sampling is executed at sites that contain vegetation that are taller than 2 meters. 3. Ground traps are sampled once per year, frequency of elevated trap sampling depends on the vegetation present at the site which rannges from once every other week to once every one are two months.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
print(dim(Neonics))
```

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

Answer: The dataset has the dimensions 4623 x 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...]

```
NeonicsSummaryEffects <- summary(Neonics$Effect)
NeonicsSummaryEffects <- sort(NeonicsSummaryEffects, decreasing = TRUE)
head(NeonicsSummaryEffects)
```

```
##       Population        Mortality         Behavior Feeding behavior
##             1803             1493              360              255
##     Reproduction      Development
##              197              136
```

Answer: The three most common effects studied include population, mortality, and behavior. Population and mortality are by far the two most studied with 1803 and 1493 studies. These two effects are likely the most studied because population ammount (which is related to mortality) can impact the ecosystem as a whole. If the insect (as most insects are) is a part of the food chain or provide ecosystem functions (like controlling pests or polinating), then the ammount of insects around to serve these roles has implications for the entire system (whereas maybe development or behavior does not has as wide of an ecosystem impact). Furthermore, population and mortality are easy to measure (as it involves counting the number of organisms and or figuring out if they are alive) and therefore relatively easy to measure whereas other effects on development or behavior might be more complicated to study.

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...]

```
NeonicsSummaryInsect <- summary(Neonics$Species.Common.Name, maxsum = 7)
NeonicsSummaryInsect <- sort(NeonicsSummaryInsect, decreasing = TRUE)
print(NeonicsSummaryInsect)
```

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

Answer: I printed the first 7 species (the first listed species is "other" which is not a species, it is likely just a collection of odd one off species that were not coded for by ecotox). The six SPECIES most studied are the honey bee, the parasitic wasp, the buff tailed bumblebee, the carniolan honey bee, the bumble bee, and the italian honeybee. These species are all polinators which means they provide ecosystem services and are of economic importance as they contribute to agriculture yeild.

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...]

```
#view(Neonics)
print(class(Neonics$Conc.1..Author.))
```

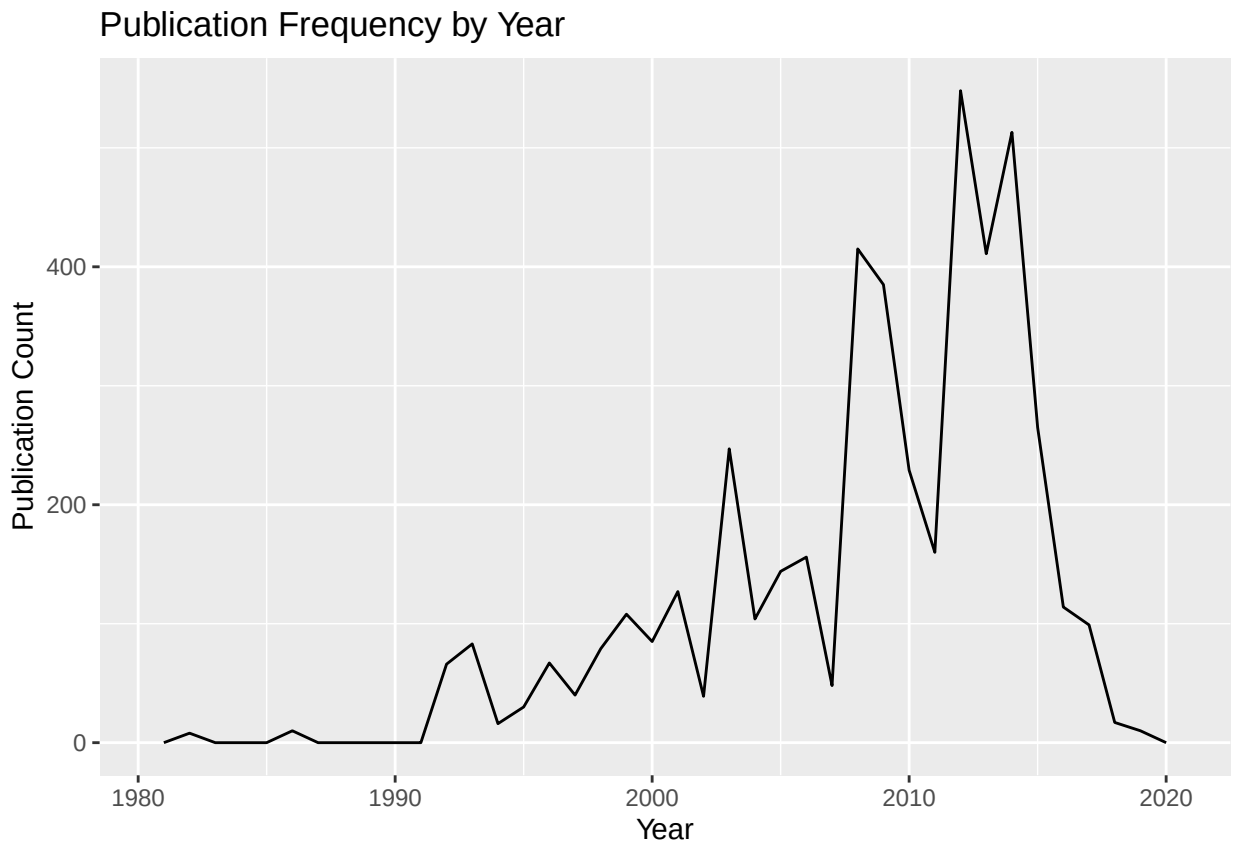
```
## [1] "factor"
```

Answer: This column is of the class factor because it includes opperators like ~, reports "NR" in some places, and it has slashes following some numbers.

Explore your data graphically (Neonics)

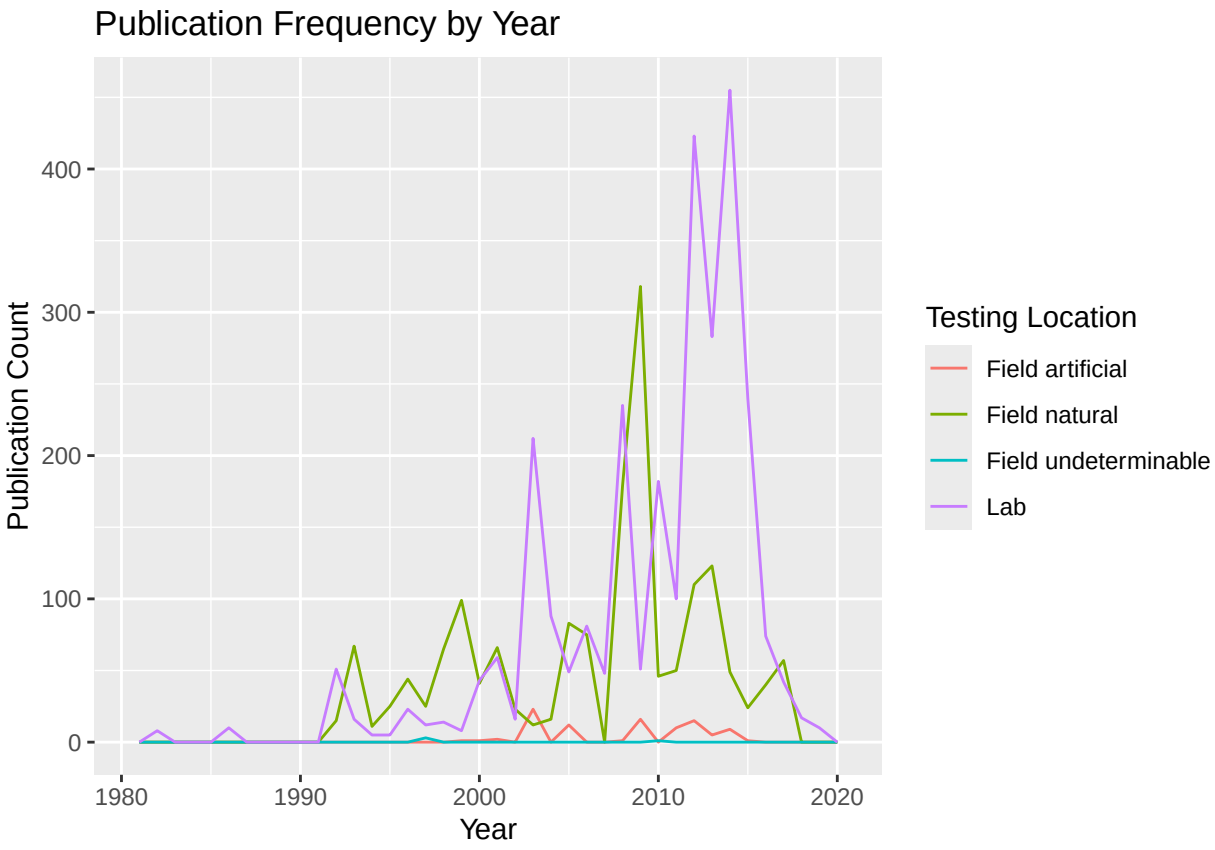
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
Neonics |>
  ggplot(aes(x = Publication.Year)) +
  geom_freqpoly(binwidth = 1) +  # show every year of Data
  labs(title = "Publication Frequency by Year", x = "Year", y = "Publication Count")
```



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

```r
Neonics |>
  ggplot(aes(x = Publication.Year, color = Test.Location)) +
  geom_freqpoly(binwidth = 1) +  # show every year of Data
  labs(
    title = "Publication Frequency by Year",
    x = "Year",
    y = "Publication Count",
    color = "Testing Location"
  )
```



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

> Answer: The most common testing lactions are the lab and natural field sites. The lab seems to be the dominat site from the early 1980s to the early 1990s when field sites over take that until the year 2000. The lab becomes more popular again in the mid 200s. Just before 2010, natural field research peaks, but since then, it appears that lab research has largely been the most published.

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...]

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

Answer:The two most common endpoints are NOEL and LOEL which stand for No Observed Effect Limit and Lowest Observed Effect Limit

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) # this variable is currently a factor
```

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

```
Litter$collectDate <- as.Date(Litter$collectDate)
class(Litter$collectDate) # now it is a date
```

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

```r
print(unique(Litter$collectDate)) #here are the two unique days sampled on
```

```
## [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`?

```r
print(length(unique(Litter$plotID))) # I want to know the length of the vector which includes the uniqu
```

```
## [1] 12
```

```r
LitterplotIDSummary <- summary(Litter$plotID)
print(LitterplotIDSummary)
```

```
## 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
```

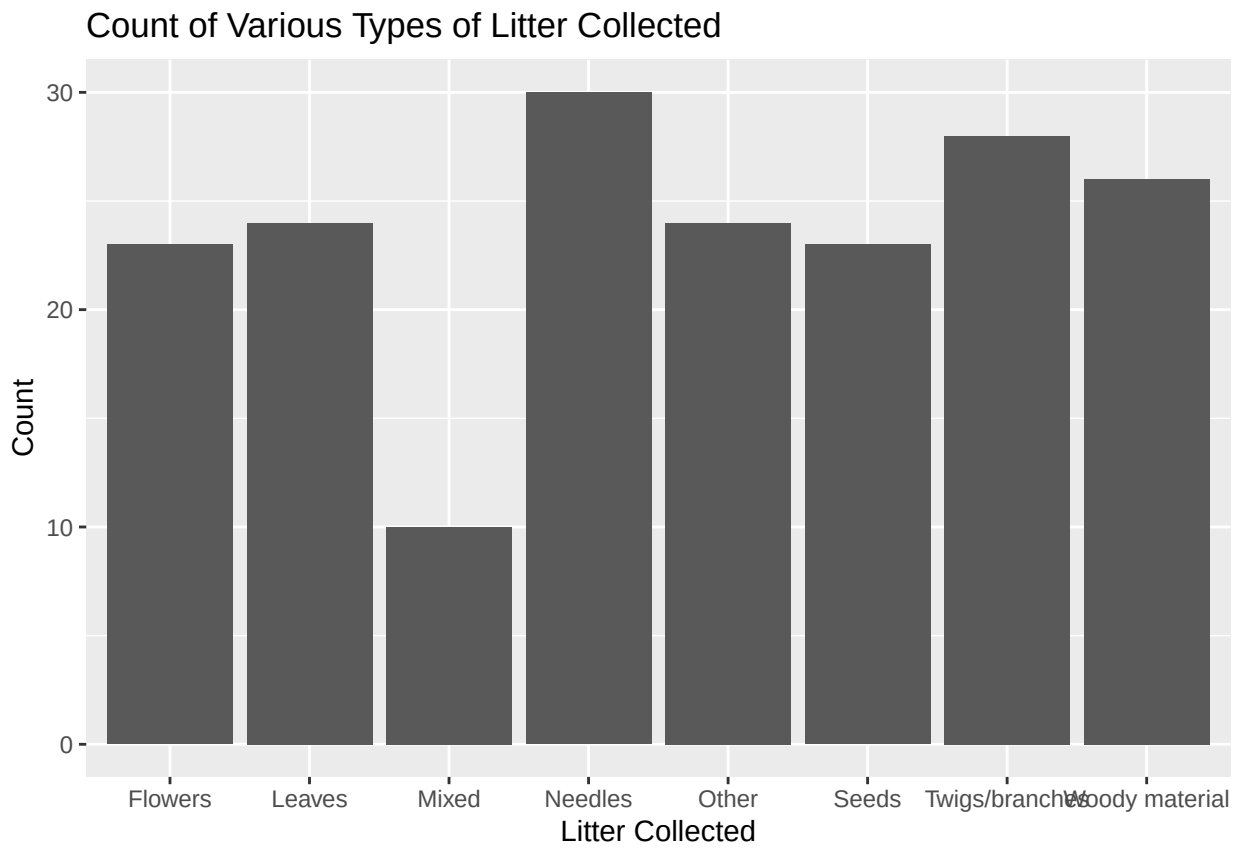
```r
print(length(LitterplotIDSummary))
```

```
## [1] 12
```

Answer: unique() as a function tells me what the unique plot IDs are. It reports every value given in the plot ID column but does not tell me about the relative occurence of those IDs, just that the exist. Summary gives me both the unique ID and the number of times that ID appears in the data set. I can figure out how many unique values there are from either method, but the summary function may be more useful given that it tells me about relative occurence of each ID type.

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
Litter |>
  ggplot(aes(x = functionalGroup)) +
  geom_bar() +
  labs(title = "Count of Various Types of Litter Collected", x = "Litter Collected", y = "Count")
```

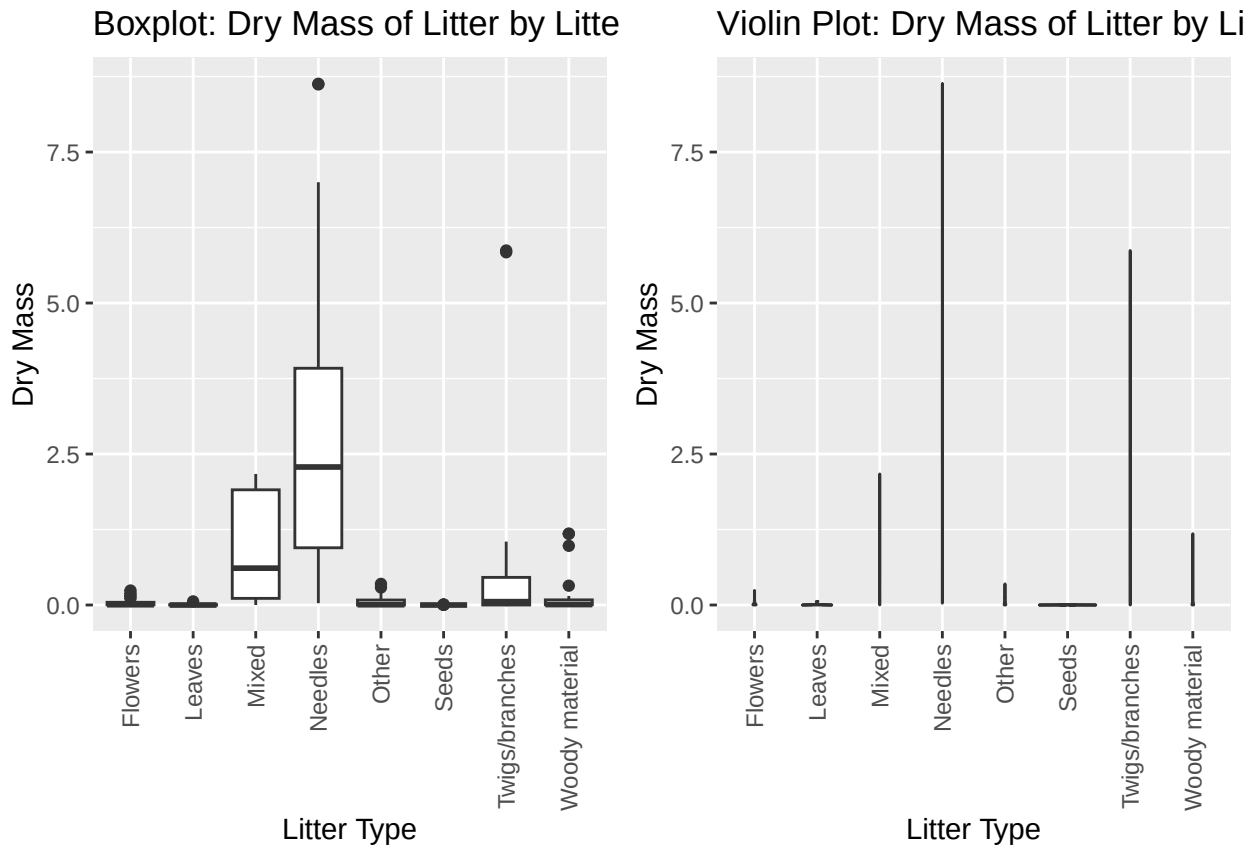


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
BoxplotLitter <- Litter |>
  ggplot(aes(y = dryMass, x = functionalGroup)) +
  geom_boxplot() +
  labs(title = "Boxplot: Dry Mass of Litter by Litter Type", x = "Litter Type", y = "Dry Mass") +
  theme(axis.text.x = element_text(
    angle = 90,
    vjust = 0.5,
    hjust = 1
  ))

Violinplotlitter <- Litter |>
  ggplot(aes(y = dryMass, x = functionalGroup)) +
  geom_violin() +
  labs(title = "Violin Plot: Dry Mass of Litter by Litter Type", x = "Litter Type", y = "Dry Mass") +
  theme(axis.text.x = element_text(
    angle = 90,
    vjust = 0.5,
    hjust = 1
  ))

grid.arrange(BoxplotLitter, Violinplotlitter, ncol = 2)
```



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

Answer: Violin plots can be great because they are helpful in visualizing the shapes of distibutions. Bimodal and unimodal data look the same on a boxplot (the standard deviation and average of these datasets can be the same) but it will show up differently on a violin plot. Boxplots can be better in that they make it easier to observe outliers (like the one twigs and branches point that is definitely more than 1.5 * IQR greater than any other dray mass data point in the same litter type). The box plot also shows us where the IQR (interquartile range), average, and extreme values are which is not true for this violin plot. There is likely just not a large enough dataset to make a violin plot readable (this is like trying to picture some sort of distribution with only a small number of points. Without enough data points, I cannot reasonable tell you the difference between unimodial, biomodal, or multimodal distributions).

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

Answer: It appears that needles and mixed litter respectively are the two highest biomasses. However, the varaible spread of these dry masses (in which their quartiles overlap with eachother and even other litter types) make it hard to draw a definitive, statistically significant conclusion. If we really wanted to know, we could compute a test statistic of some kind (maybe a t-test) to determine if the avearge of these two litter types are statistically different from each other and the other classes.