Assignment 3: Data Exploration

Nadia Barbo

Spring 2023

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 Sakai.

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. Check your working directory, load necessary packages (tidyverse, lubridate), 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.

```
getwd() #checking my working directory
```

```
## [1] "C:/Users/nadia/Documents/Duke_/EDA-Spring2023"
```

```
library(tidyverse); library(lubridate) #loading necessary packages
Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors
↪  = TRUE) #uploading ECOTOX dataset
Litter <- read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv",
↪  stringsAsFactors = TRUE) #uploading NEON dataset
```

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: Insecticides, such as neonicontinoids, are commonly used on agricultural crops to get rid of pests that eat/damage the crop. It is important to understand the ecotoxicology behind pesticides such as these in order to understand how they may potentially cause harm to non-target organisms since the spraying and wide distribution of pesticides results in their presence in multiple media that can lead to exposure. Neonicotinoids are a class of neurotoxic insecticides that are highly selective to insects and do not directly harm humans due to the difference in nicotine receptors between humans and insects. Neonicotinoids irreversibley bind with the postsynaptic nicotinic acetylcholine receptor in insects, causing constant activation and paralysis which eventually results in the insect's death. While neonicotinoids are termed "selective" because they primarily target insects, they target insects broadly, so "pest" and "non-pest" insects are at risk. This has led to the extermination of essential pollinators such as bees and butterflies. Neonicotinoids have gained a lot of critisicm for their killing of bee populations (colony collapse) which has a huge economic impact as well as societal impact since people love bees.

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 falls to the ground in forests creates important nursery sites for plants and shelter for animals. It also creates a micro-environment that attracts diverse fungi and insects. The litter and woody debris on forest floors is important to the future soil composition of the forest and can play a major role in determining the nutrients available to new plants/trees in the forest. Woody plants contain cellulose, minerals, and proteins that are available to other organisms when the litter decomposes. Additionally, litter and woody debris is important for water filtration and moisture retention at the forest floor. Finally, allowing the natural accumulation and decomposition of litter and woody debris on forest floors (instead of collecting and burning it) modulates the amount of carbon dioxide in the atmosphere.

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 and woody debris consisted of different plant functional groups, which were: leaves, needles, twigs/branches, woody material such as cones and bark, seeds, flowers and non-woody reproductive structures, and other. They labeled all unsorted material as mixed. Mass of each functional group from each collection event was measured to an accuracy of 0.01g. Functional groups with a mass less than 0.01g were noted in order to designate their presence but it is important to know that they were not within detectable masses. 2. Litter was collected from 0.5m^2 PVC elevated traps (80cm above the ground) and it had to have a butt end diameter less than 2cm and a length less than 50cm. Fine wood debris was collected from 3m x 0.5m rectangular ground traps and it had to have a butt end diameter less than 2cm and greater than 50cm. 3. They conducted both spatial and temporal sampling. For spatial resolution, sampling sites were in areas that had woody vegetation greater than 2m tall with traps strategically placed within

plots based on forest stature/cover. Traps were either randomly placed in plots or placed in a targeted fashion. This was determined based on percentage of aerial cover of woody vegetation. In terms of temporal sampling, ground traps were placed 1 time a year and elevated traps were placed at varying times per year based on the vegetation present at the site. They sampled 1x every 2 weeks at deciduous sites (during senescence) and 1x every 1-2 months at evergreen sites. During the winter, sampling at deciduous sites may have been suspended for up to 6 months.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics) #checking dimensions of neonics dataset
```

```
## [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?

```
sort((summary(Neonics$Effect)), decreasing=TRUE) #getting summary statistics of neonic
↪  dataset's effect column from most common effects studied to least common effects
↪  studied
```

```
##       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
```

Answer: The two most common effects of neonics studied are population (1803) and mortality (1493). These effects are studied to a much greater degree than others, with the next most studied effect being behavior (360). Population and mortality may specifically be of interest because neonics are known insecticides that have the goal of killing target insects, therefore decreasing the population, but there is concern over how neonics impact off-target species of insects. Since the molecular target of neonics is known for insects, it is likely that the chemical will show the increased mortality and decreased population in off-target insects as well and this is why it is especially studied.

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: The `sort()` command can sort the output of the summary command...]

```
sort(summary(Neonics$Species.Common.Name), decreasing=TRUE) #getting the summary
↪  statistics of neonic dataset's species (common name) column from most commonly
↪  studied to least commonly studied
```

```
##                          (Other)                             Honey Bee
##                              670                                   667
##                   Parasitic Wasp                  Buff Tailed Bumblebee
##                              285                                   183
##                Carniolan Honey Bee                           Bumble Bee
##                              152                                   140
##                  Italian Honeybee                       Japanese Beetle
##                              113                                    94
##                Asian Lady Beetle                        Euonymus Scale
##                               76                                    75
##                        Wireworm                     European Dark Bee
##                               69                                    66
##                Minute Pirate Bug                   Asian Citrus Psyllid
##                               62                                    60
##                     Parastic Wasp                 Colorado Potato Beetle
##                               58                                    57
##                  Parasitoid Wasp                   Erythrina Gall Wasp
##                               51                                    49
##                     Beetle Order       Snout Beetle Family, Weevil
##                               47                                    47
##            Sevenspotted Lady Beetle                     True Bug Order
##                               46                                    45
##               Buff-tailed Bumblebee                       Aphid Family
##                               39                                    38
##                    Cabbage Looper                 Sweetpotato Whitefly
##                               38                                    37
##                    Braconid Wasp                       Cotton Aphid
##                               33                                    33
##                    Predatory Mite                 Ladybird Beetle Family
##                               33                                    30
##                       Parasitoid                       Scarab Beetle
##                               30                                    29
##                    Spring Tiphia                        Thrip Order
##                               29                                    29
##              Ground Beetle Family                   Rove Beetle Family
##                               27                                    27
##                    Tobacco Aphid                       Chalcid Wasp
##                               27                                    25
##             Convergent Lady Beetle                      Stingless Bee
##                               25                                    25
##                 Spider/Mite Class                  Tobacco Flea Beetle
##                               24                                    24
##                  Citrus Leafminer                      Ladybird Beetle
##                               23                                    23
##                         Mason Bee                            Mosquito
##                               22                                    22
##                    Argentine Ant                               Beetle
##                               21                                    21
##           Flatheaded Appletree Borer                 Horned Oak Gall Wasp
##                               20                                    20
##                  Leaf Beetle Family                    Potato Leafhopper
##                               20                                    20
##           Tooth-necked Fungus Beetle                       Codling Moth
##                               20                                    19
```

```
##          Black-spotted Lady Beetle                       Calico Scale
##                                 18                                 18
##               Fairyfly Parasitoid                        Lady Beetle
##                                 18                                 18
##             Minute Parasitic Wasps                          Mirid Bug
##                                 18                                 18
##                   Mulberry Pyralid                           Silkworm
##                                 18                                 18
##                     Vedalia Beetle               Araneoid Spider Order
##                                 18                                 17
##                         Bee Order                     Egg Parasitoid
##                                 17                                 17
##                       Insect Class            Moth And Butterfly Order
##                                 17                                 17
##       Oystershell Scale Parasitoid Hemlock Woolly Adelgid Lady Beetle
##                                 17                                 16
##              Hemlock Wooly Adelgid                               Mite
##                                 16                                 16
##                       Onion Thrip               Western Flower Thrips
##                                 16                                 15
##                     Corn Earworm                  Green Peach Aphid
##                                 14                                 14
##                         House Fly                          Ox Beetle
##                                 14                                 14
##                Red Scale Parasite                 Spined Soldier Bug
##                                 14                                 14
##             Armoured Scale Family                   Diamondback Moth
##                                 13                                 13
##                     Eulophid Wasp                  Monarch Butterfly
##                                 13                                 13
##                     Predatory Bug               Yellow Fever Mosquito
##                                 13                                 13
##                Braconid Parasitoid                       Common Thrip
##                                 12                                 12
##       Eastern Subterranean Termite                             Jassid
##                                 12                                 12
##                        Mite Order                          Pea Aphid
##                                 12                                 12
##                  Pond Wolf Spider            Spotless Ladybird Beetle
##                                 12                                 11
##              Glasshouse Potato Wasp                           Lacewing
##                                 10                                 10
##             Southern House Mosquito            Two Spotted Lady Beetle
##                                 10                                 10
##                        Ant Family                       Apple Maggot
##                                  9                                  9
```

Answer: Aside from "Other," the six most commonly studied species reported in this dataset are the Honey Bee, the Parasitic Wasp, the Buff Tailed Bumblebee, the Carolina Honey Bee, the Bumble Bee, and the Italian Honeybee. These species are all beneficial to agriculture and crops. Bees are efficient pollinators, which is essential for plants to reproduce/produce fruits and seeds. The parasitic wasp is beneficial to agriculture because it kills other insects that are harmful to crops (and itself is not harmful to crops). This means that those working in the agricultural industry are aided, and in some cases rely on, these insects. Inadvertenly killing

these insects through the application of neonicotinoids to crops may be doing more harm than good. Additionally, these insects are attracted to agricultural lands because they are vast and filled with pollen and other insects for the bees and wasps respectively, so they are likely often impacted by neonicotinoids in their natural environment. Other insects may be less common, may be less beneficial to agriculture, and/or may be less attracted to land where neonicotinoids are applied.

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?

```
class(Neonics$Conc.1..Author.) #finding the class of the Conc.1..Author column from the
↪  neonics dataset
```

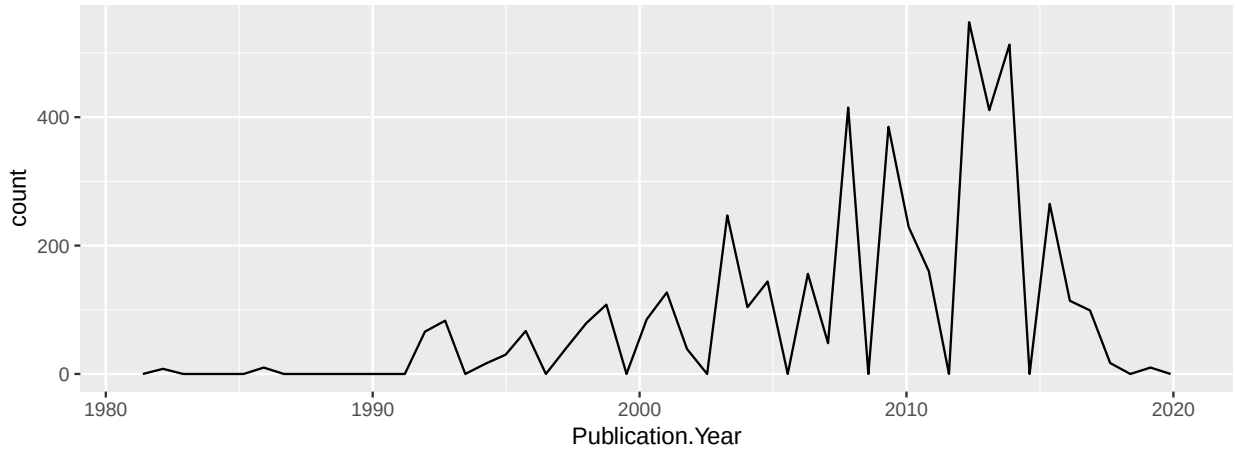
```
## [1] "factor"
```

Answer: The class of the Conc.1..Author column in the Neonics dataset is "factor." The class "factor" is used when there are discrete units that are not in a specific order. Numeric data can be ordered and mathematical calculations can be applied to it. This column's class is not numeric because the concentrations are reported in various different units so you cannot compare them numerically to one another.

Explore your data graphically (Neonics)

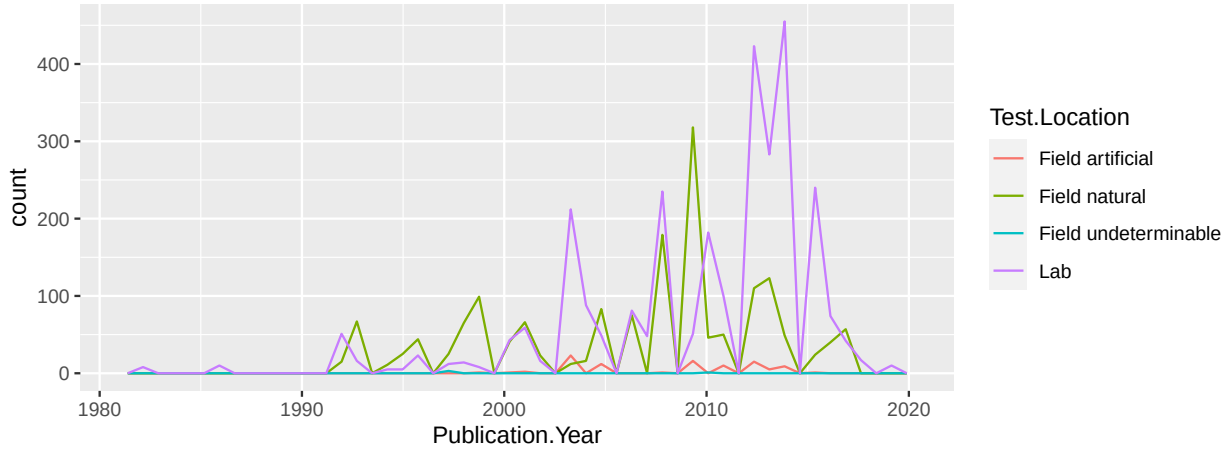
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50) #creating a frequency polygon graph
  ↪  showing the number of studies (y) published in a given year (x)
```



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

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color=Test.Location), bins = 50) #creating a
  ↪  frequency polygon graph showing the number of studies (y) published in a given year
  ↪  (x) with studies color coded based on test location
```



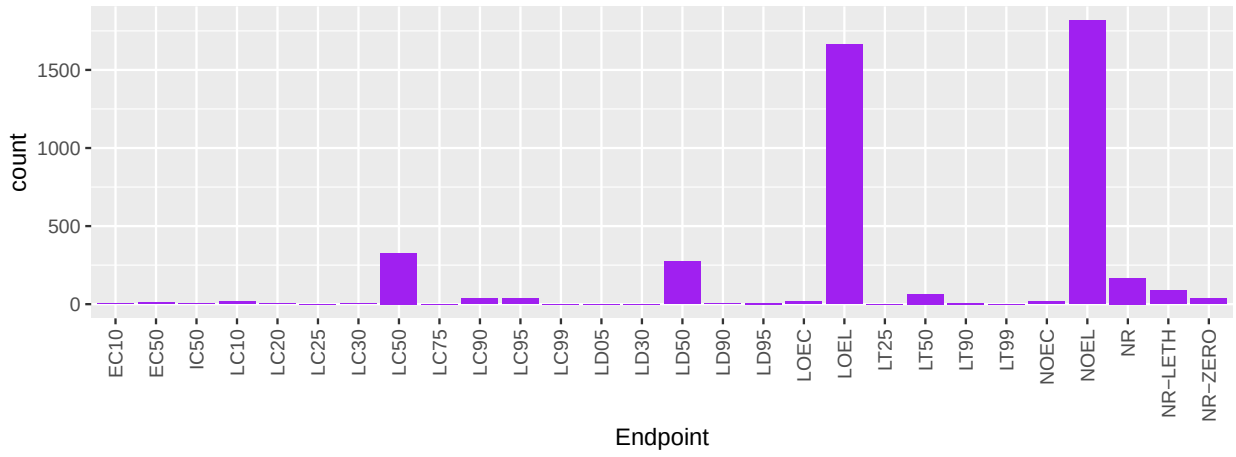
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are in the "lab" and in the "field natural." "Field natural" data collection did not begin until 1990, eight years after the earliest publication year, where "lab" data began from the start. Both lab based study and natural field based studies increased over time, but the total number of studies also increased over time. They both appear to always be the most common test locations, compared to "field artifial" and "field undeterminable." Field natural data collection peaked in 2007-2008, while lab based data peaked more recently in 2011-2013.

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...]

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar(fill="purple") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) #creating a bar
  ↪  graph of endpoint frequency (counts) in the neonics dataset and rotating the graph
```



Answer: The two most common endpoints from the Neonics dataset are No Observable Effect Level (NOEL) and Lowest Observable Effect Level (LOEL). A NOEL is the highest dose in which the species in question shows no significant effect from the exposure (compared to controls). A LOEL is the lowest dose in which the species in question shows a significant effect from the exposure (compared to controls).

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) #finding the class of "collectDate" from litter dataframe - it
↪  is "factor"
```

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

```
Litter$collectDate <- ymd(Litter$collectDate) #converting "collectDate" column into date
↪  format
class(Litter$collectDate) #confirming that the "collectDate" column was changed to "Date"
↪  class
```

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

```
unique(Litter$collectDate) #determining which dates litter was collected in August of
↪  2018
```

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

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

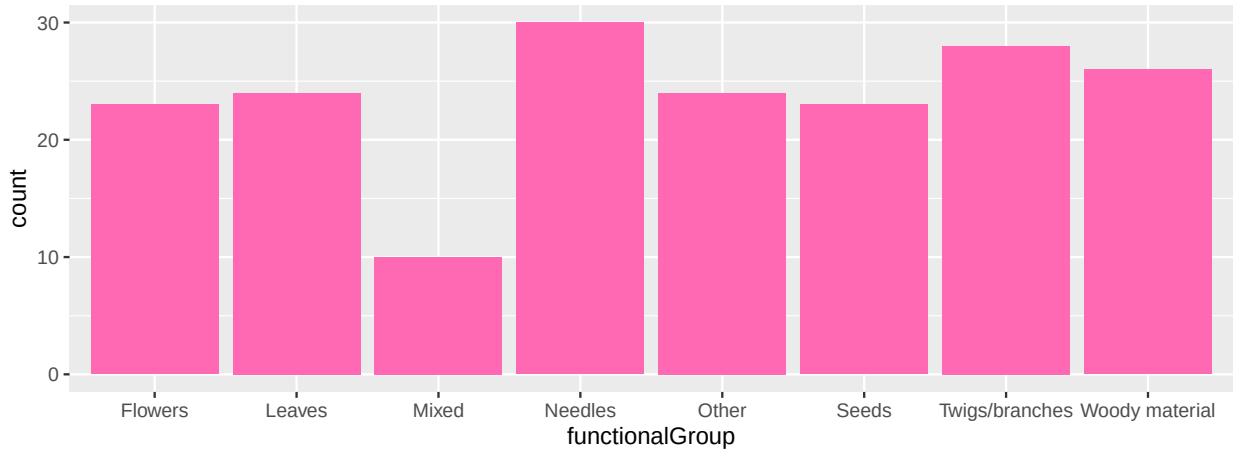
```
unique(Litter$plotID) #finding how many unique plots were sampled at Niwot Ridge (there
↪  are 12)
```

```
##  [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 unique function will output each unique value within the designated column/data one time and tell you how many unique values there are. The summary function will output each value within the designated column/data and along with the value it will tell you how many times that value occured within the desigdated data.

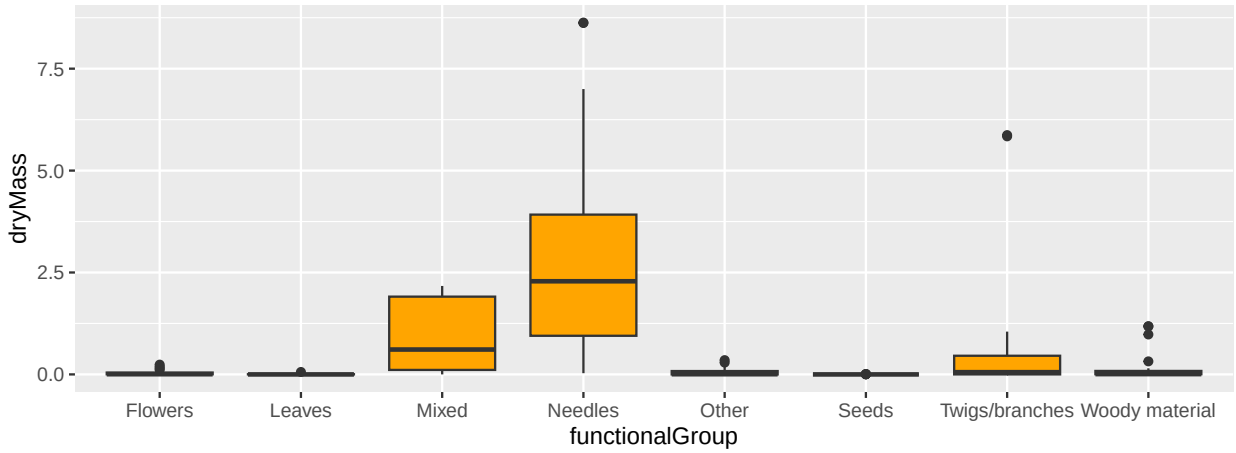
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar(fill="hot pink") #creating a bar graph of functional group counts in the
  ↪  litter dataset
```

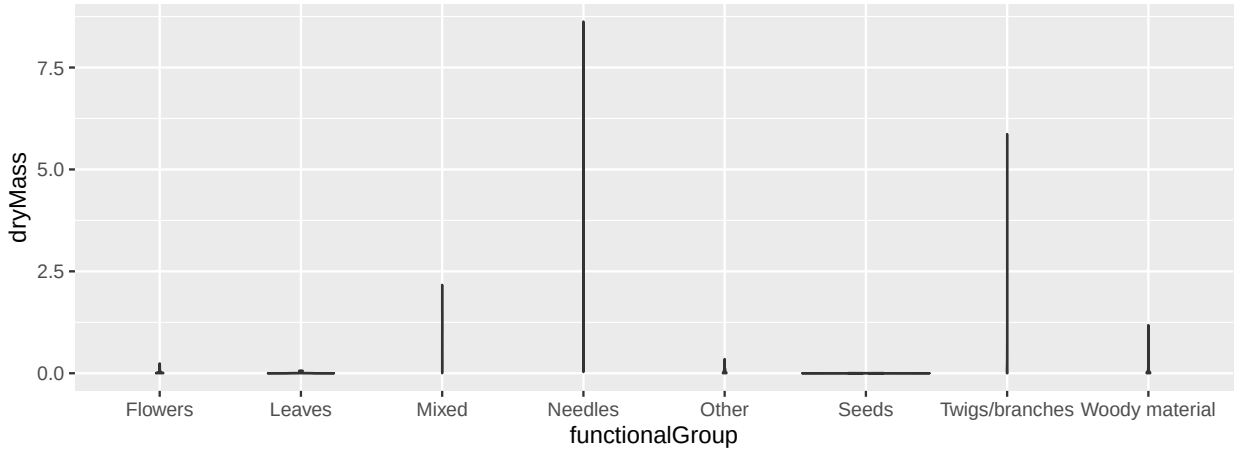


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x = dryMass, y = functionalGroup), fill="orange") +
  coord_flip() #making a boxplot of dryMass by functionalGroup from the litter dataset
  ↪  and flipping the coordinates (because I think it looks better)
```




```
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup)) +
  coord_flip() #making a violin plot of dryMass by functionalGroup from the litter
  ↪  dataset and flipping the coordinates (because I think it looks better)
```



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

> Answer: Boxplots are great for visualizing summary statistics of data, including the median and IQR, which can inform the general spread of data (but not the exact distribution). When points fall outside of 1.5 * IQR, they are plotted as outliers outside of the box of the boxplot. This allows for us to still be able to clearly visualize the boxplot's represented data. Violin plots show both the range of values and the distribution of values. While you can still visualize the median and IQR within a violin plot, any values outside of 1.5 * IQR are plotted like the rest of the distribution data, therefore spreading out the plot. In our data, the boxplot is a better visualization tool because the distribution of dryMass data for the needles, mixed, and twigs/branches functional groups is wide, therefore making their violin plots stretch out across the plot in a straight line. These three functional groups also have dryMasses that are mcuh greater than that for the other function groups, which lengthens the y axis, so we cannot see the smaller dryMass values for flowers, leaves, other, seeds, or woody materials well in the violin plot. It looks some of the functional groups with smaller dryMass values may have some distribution within them but we can't see it well because the large y axis. While this lengthening of the y axis is true in the boxplot as well, we are still able to see that there are defined boxes for all plots and we are able to see outliers more clearly.

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

> Answer: When looking at the medians, "Needles" and "Mixed" functional group categories tend to give the highest biomass at these sites.