Assignment 3: Data Exploration

Hannah Nelson

Fall 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.

```
#checking that working directory is correct
getwd()
```

```
## [1] "/Users/hannahnelson/Desktop/env872/EDA-Fall2023"
```

```
#loading in packages
library(tidyverse)
library(lubridate)

#reading in and viewing each data set
```

```
Neonics <- read.csv(".//Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors = TRUE)
#View(Neonics)
Litter <- read.csv(".//Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv", stringsAsFactors = TRUE)
#View(Litter)
```

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: Studying the ecotoxicology of neonicotinoids on insects is important because understanding how a widely used class of insecticides impacts different species of insects is necessary knowledge for informed decsions to be made about know to use neonicotinods responsibly. Looking at data on this relationship can show which insect species are being sucessfully targetted, and if any species are developing resistance. This is helpful becuase it allows people using neonicotinoids to know which species they will be able to target, and if species they were are intending to target could react lethally to neonicotinoids.

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 are important to study becuase they play an important role in eocsystems. They play a role in nutrient cycling and carbon sequestration, act as a barrier to erosion, and provide habitat building materials for species. Looking at litter and woody debris falling to the forest floor over time can provide insight into the health of forest ecosystems, and the carbon sequestration potential of forests.

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 fine woody debris sampling is executed at terrestrial NEON sites that contain woody vegetation over 2 meters tall. 2. An elevated and ground litter trap are deployed for every 400 square meters of of plot area. 3. Target sampling frequency for elevated traps is determined by what vegetation is present at the site. On average, decidious forest sites are sampled every two weeks and evergreen forest sites are sampled once every one to two months.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#the data set has 4623 rows and 30 columns
```

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?

```
summary(Neonics$Effect) %>% sort(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
```

Answer: The most common effects studied are population & mortality. These effects are specifically of interest because the primary purpose of the data is to understand how effective neonicotinoids are on insect species, which can be best understood by looking at population and mortality data.

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...]

```
summary(Neonics$Species.Common.Name) %>% sort(decreasing=TRUE)
```

```
##                          (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: Six most commonly studied species: 1. Honey Bee 2. Parasitic Wasp 3. Buff Tailed Bumblebee 4. Carniolan Honey Bee 5. Bumble Bee 6. Italian Honeybee
These insects are likely of interest over other insects because many bee species are vulnerable or endangered. Understanding if the use of neonicotinoids is harming species of bees is important in the effort to save bee species from extinction and preserve biodiversity. While parasitic wasps are not endangered, they are also a species of insect that is beneficial to the environment and could be harmed by the use of neonicotinoids.

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.)
```

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

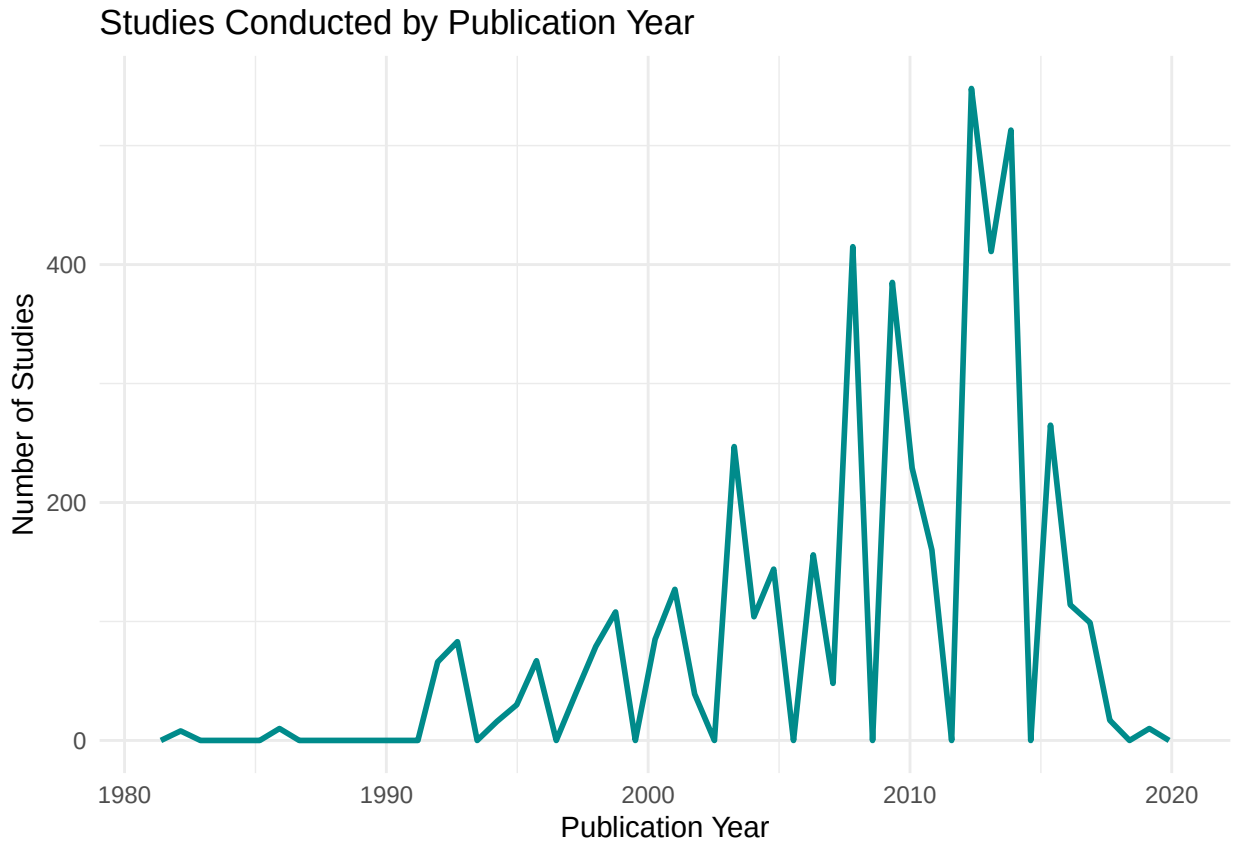
Answer: The class of "Conc.1..Author.." is factor. It is not numeric becuase the data is discrete, not continuous.

Explore your data graphically (Neonics)

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

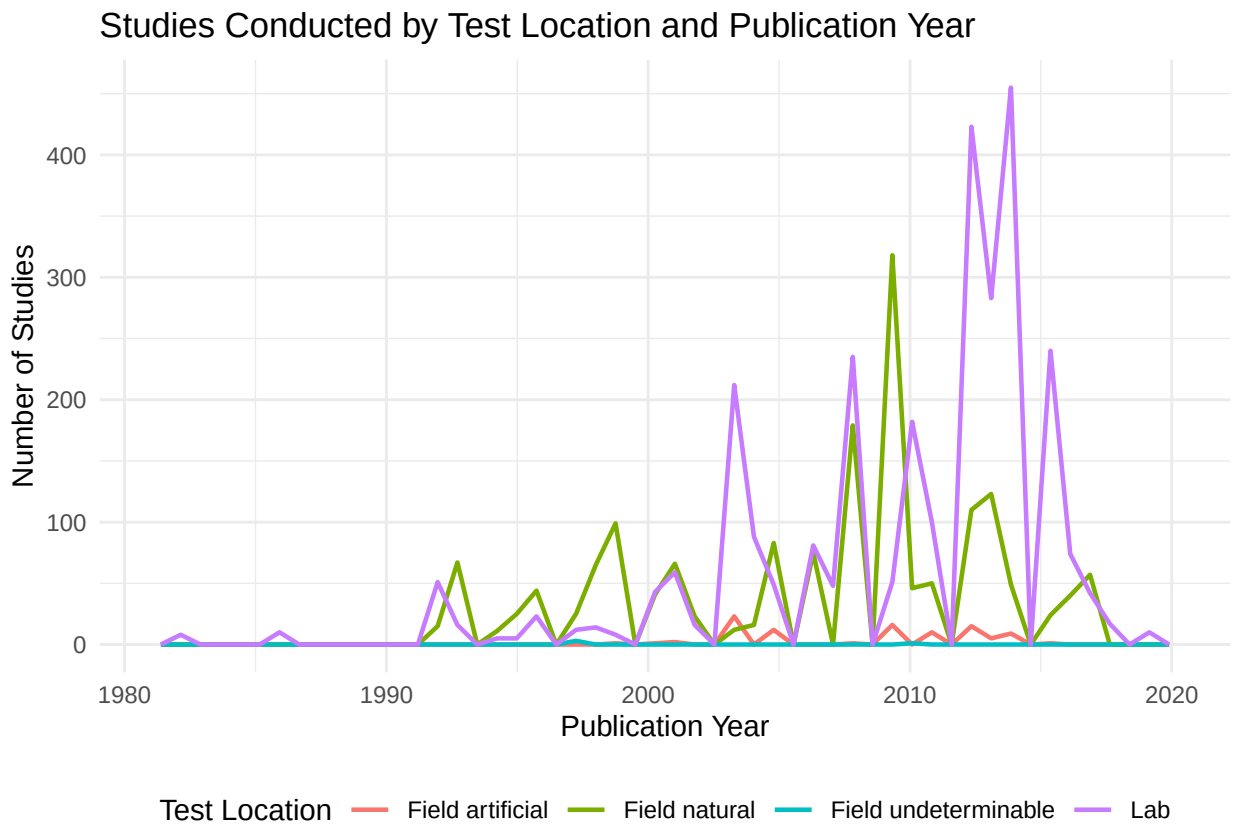
```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins=50, color="cyan4", size=1)+
  theme_minimal()+
  labs(title="Studies Conducted by Publication Year", x="Publication Year", y="Number of Studies")
```

```
## 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.
```



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, size=0.8)+
  theme_minimal()+
  theme(legend.position = "bottom")+
  scale_color_discrete(name="Test Location")+
  labs(title="Studies Conducted by Test Location and Publication Year", x="Publication Year", y="Number
```



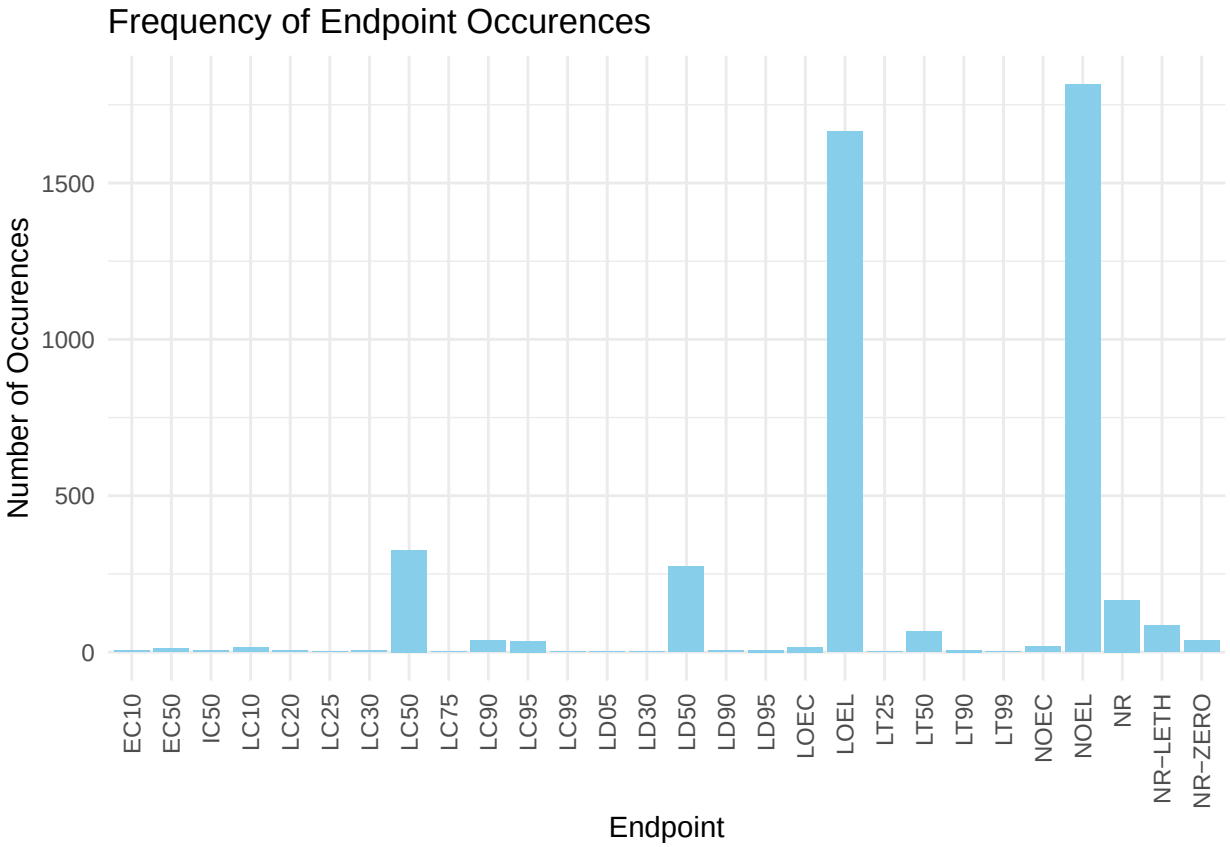
Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are lab & field natural. In the 1990s, field natural was the most common test location. Lab becuase the most common test location in the early 2000s, and has remained in that position except for a brief period around 2008-2009 when field natural overtook lab as the most common. The use of both lab and field natural locations for tests have increased over time with similar patterns.

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="skyblue") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(title="Frequency of Endpoint Occurences", x="Endpoint", y="Number of Occurences")
```



Answer: The two most common endpoints are NOEL & LOEL 1. NOEL (no observable effect level): The highest dose produced effects that were not significantly different from responses of controls 2. LOEL (lowest observable effect level): The lowest dose produced effects that were significantly different from the responses of 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)
```

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

```
#the class of "collectDate" is factor before transforming data

#transforming "collectDate" to date class
Litter$collectDate <- as.Date(Litter$collectDate)

class(Litter$collectDate)
```

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

```r
#after the transformation the class of "collectDate" is date

unique(Litter$collectDate)
```

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

```r
#litter was collected on the 2nd and 30th of August 2018
```

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`?

```r
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
#returns list of each unique plot ID


unique(Litter$plotID) %>% length()
```

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

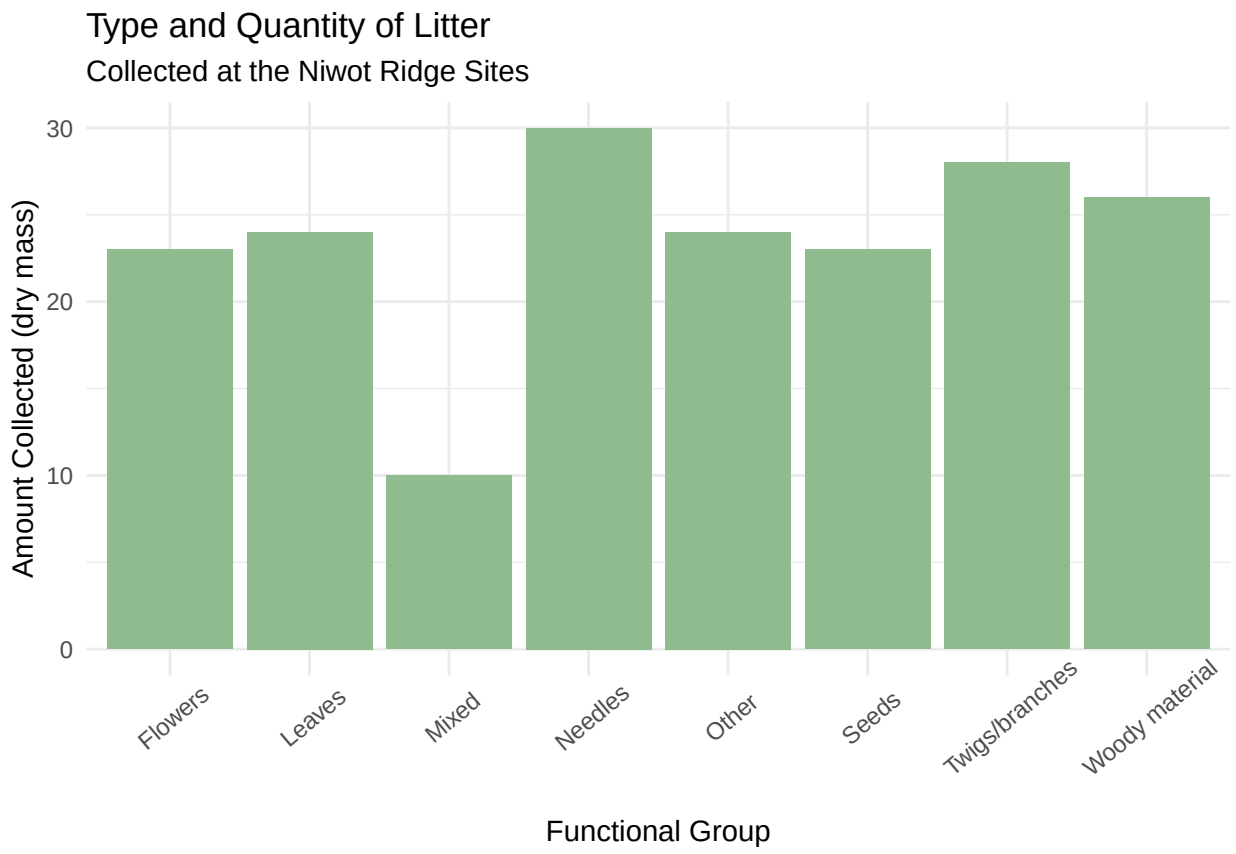
```r
#returns the length of the list


#There were 12 plots sampled at Niwot Ridge: NIWO_061, NIWO_064, NIWO_067, NIWO_040, NIWO_041, NIWO_063
```

Answer: Information obtained using unique is different than information obtained using summary because unique removes duplicate values to only give one of each, while summary would include duplicate values. This was important because plot ID and date values repeat multiple times in the data set. Using unique allowed those duplicate values to each only be seen once, which provided clarity in data exploration.

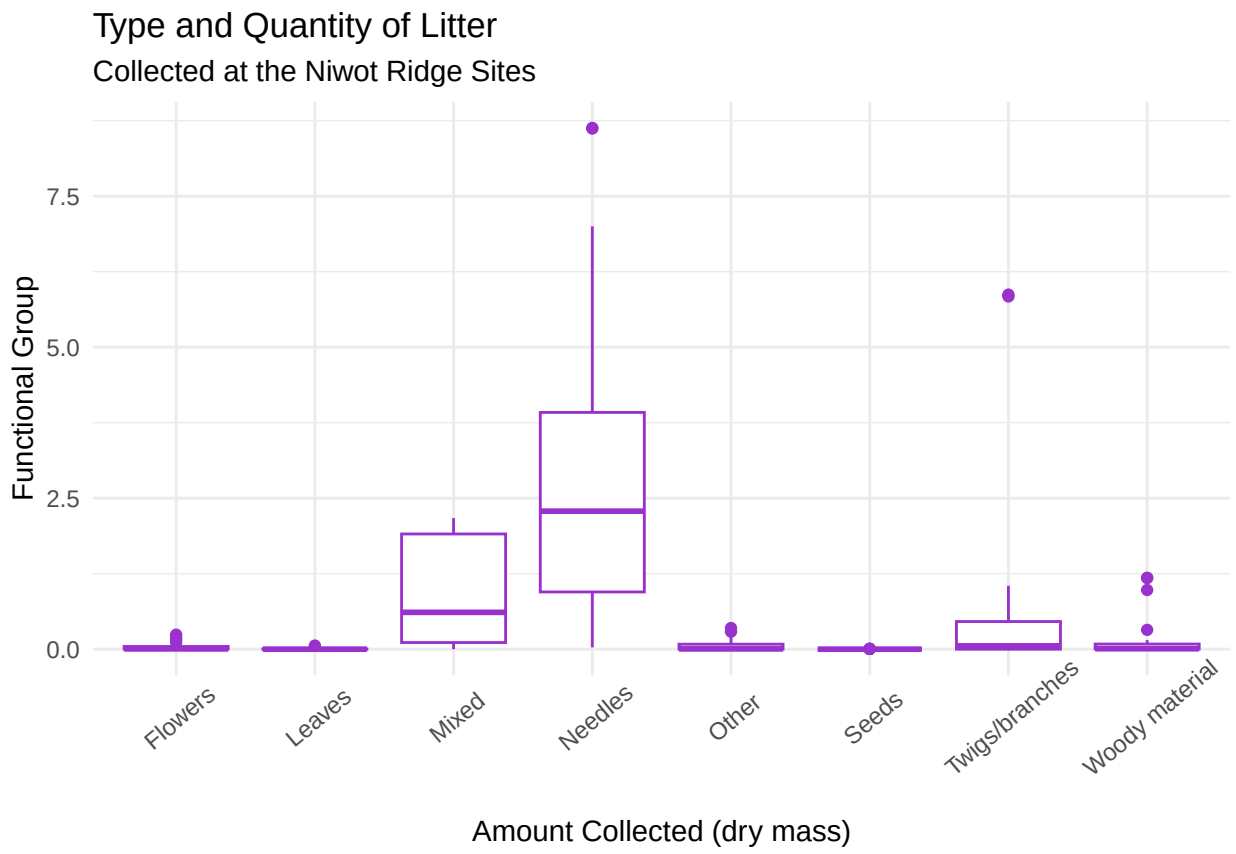
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x=functionalGroup))+
  geom_bar(fill="darkseagreen") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 40, vjust = .8))+
  labs(title="Type and Quantity of Litter", subtitle="Collected at the Niwot Ridge Sites", x= "Functiona
```

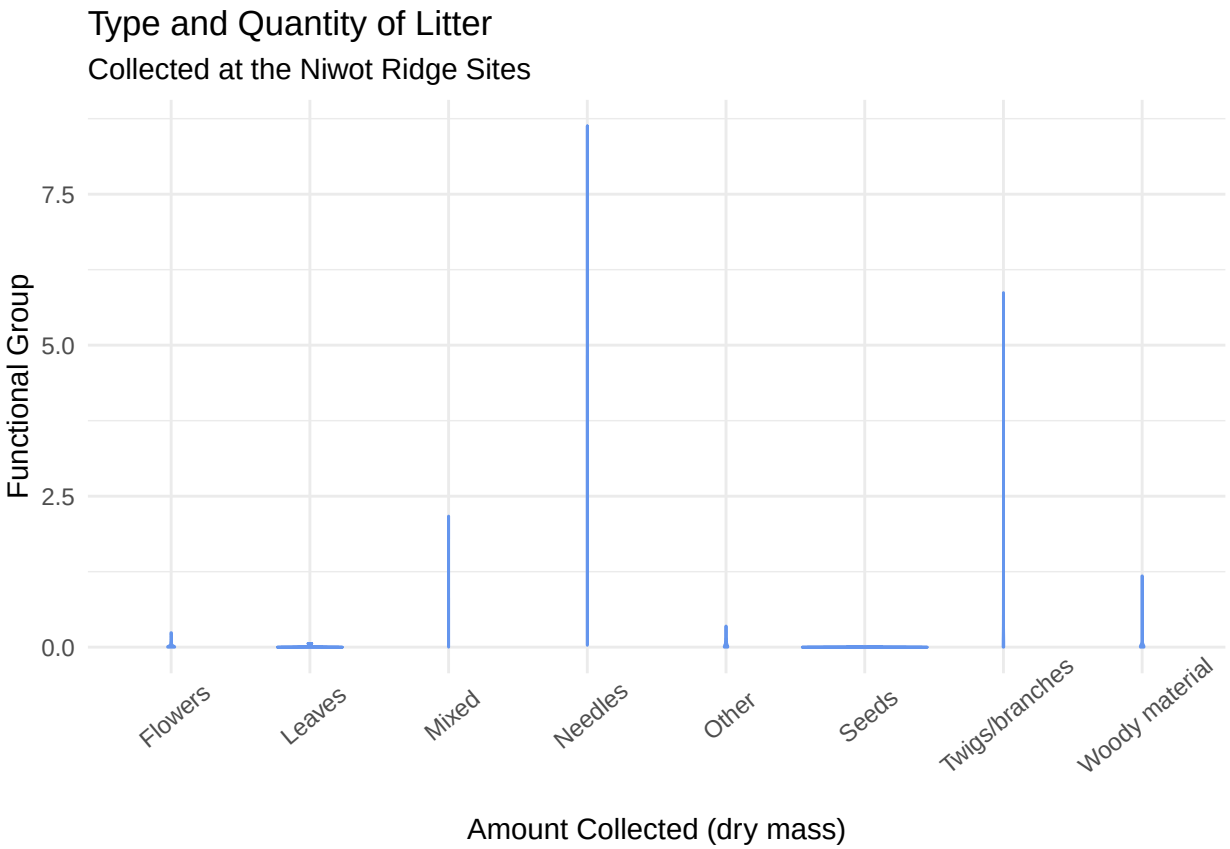


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter)+
  geom_boxplot(aes(x=functionalGroup, y=dryMass), color="darkorchid")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 40, vjust = .8))+
  labs(title="Type and Quantity of Litter", subtitle="Collected at the Niwot Ridge Sites", x="Amount Co
```




```
ggplot(Litter)+
  geom_violin(aes(x=functionalGroup, y=dryMass),
    draw_quantiles = c(0.25, 0.5, 0.75), color="cornflowerblue")+
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 40, vjust = .8))+
  labs(title="Type and Quantity of Litter", subtitle="Collected at the Niwot Ridge Sites", x="Amount Co
```



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

> Answer: A boxplot is more effective than a violin plot in this case becuase there is not enough distribution of data points within the interquartile ranges of each litter type for a violin plot to produce a useful visualization.

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

> Answer: Needles & mixed litter types tend to have the highest biomass at the Niwot Ridge sites.