Assignment 3: Data Exploration

Nancy Bao

OVERVIEW

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

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

Set up your R session

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

```
#Set/Check working directory
getwd()
```

```
## [1] "/Users/Nancy/Desktop/Semester 4/ENV 872L/Environmental_Data_Analytics_2021"
```

```
#my working directory was already set to the:
#"/Users/Nancy/Desktop/Semester 4/ENV 872L/Environmental_Data_Analytics_2021"

#Load packages
library(tidyverse)
library(dplyr)
library(ggplot2)

#Set relative file path for Neonics
Neonics<-read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv",stringsAsFactors=T)
#need to set stringsAsFactors=T to summarize character variables

#Set relative file path for Litter
Litter<-read.csv("./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv")
```

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: Ecotoxicology of neonicotinoids on insects is important for understanding the acute and chronic adverse effects that these pesticides may have on beneficial insects in our ecosystem like honeybees. Through my courses in ecotoxicology and environmental toxicology at Duke, these pesticides are detrimental to the growth and development of many insect species, which are consequential to higher order trophic levels that depend on honeybee pollination and consumption for survival.

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 critical nutritional components to the forest floor. Through my undergraduate research in soil science, plant nutrient, and dendrology,I learned that the organic matter, slowly released from the lignin, cellulose, and polyphenolic compounds are critical food sources and habitats for microbes, fungi, and other organisms in the aquatic and terrestrial ecosystems. Tracking the thickness and components that make up the litter and woody debris are ways to gauge forest and soil health in forest ecosystem management. My research was focused on N and C cycling, which is strongly driven by microbial breakdown and natural inputs of these nutrients from decaying plant matter.

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: The following pieces of information were obtained from the NEON_Litterfall_UserGuide.pdf:
   * Litter and woody debris were sampled from pairs of ground traps (sampled once a year;used for woody debris) and elevated PVC traps (used for litter; sampled bimonthly(once in 2 months)/monthly for evergreen forests and biweekly-once every 2 weeks for deciduous forests) at the terrestrial NEON sites. * Litter was defined as matter with the following dimensions for sampling: less than 50cm in length with a butt end diameter less than 2cm ;woody debris was defined as matter with the following dimensions for sampling: greater than 50 cm in length with a butt end diameter less than 2cm. *Litter and woody debris samples were sorted and quantified for dry biomass (in grams; ±0.01g) based on 8 functional group categories: leaves, twigs/branches, needles, mixed (not sorted), woody material, seeds, flowers and nonreproductive woody plant parts, and other (mosses and lichens); LOD for biomass was <0.01g.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
#Get dimensions
dim(Neonics)
```

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

```
#The dimensions of the dataset are 4,623 rows by 30 columns.
```

6. Using the `summary` function on the "Effects" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
#Most common effects studied
summary(Neonics$Effect)
```

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

```
# The top 5 most common effects studied are:
#Population, Mortality, Behavior, Feeding Behavior, and Reproduction
```

Answer: The top five most common effects studied based on number of observations include: Population (n=1803), Mortality (n=1493), Behavior (n=360), Feeding behavior (n=255), and Reproduction (n=197). When assessing ecotoxicity,it is important to consider the immediate health impacts that can be associated with or instigated by exposure to neonicotinoids. Effects such as mortality, reproduction, and behavior are critical effects that are detrimental to the growth and development of a species. They are good indicators for ecosystem health and are able to be observed for short-term consequences.

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.

```
#Determine 6 most commonly studied species using summary()
summary(Neonics$Species.Common.Name)
```

```
##                          Honey Bee                      Parasitic Wasp
##                                667                                285
##                Buff Tailed Bumblebee                  Carniolan Honey Bee
##                                183                                152
##                          Bumble Bee                     Italian Honeybee
##                                140                                113
##                     Japanese Beetle                   Asian Lady Beetle
##                                 94                                 76
##                     Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                   Minute Pirate Bug
##                                 66                                 62
##                Asian Citrus Psyllid                      Parastic Wasp
##                                 60                                 58
##              Colorado Potato Beetle                    Parasitoid Wasp
##                                 57                                 51
##                 Erythrina Gall Wasp                      Beetle Order
##                                 49                                 47
##         Snout Beetle Family, Weevil            Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order                Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                     Cabbage Looper
##                                 38                                 38
##                Sweetpotato Whitefly                      Braconid Wasp
##                                 37                                 33
##                       Cotton Aphid                     Predatory Mite
```

```
##                                33                                33
##              Ladybird Beetle Family                        Parasitoid
##                                30                                30
##                       Scarab Beetle                    Spring Tiphia
##                                29                                29
##                        Thrip Order               Ground Beetle Family
##                                29                                27
##                  Rove Beetle Family                    Tobacco Aphid
##                                27                                27
##                       Chalcid Wasp            Convergent Lady Beetle
##                                25                                25
##                     Stingless Bee                 Spider/Mite Class
##                                25                                24
##                 Tobacco Flea Beetle                 Citrus Leafminer
##                                24                                23
##                    Ladybird Beetle                         Mason Bee
##                                23                                22
##                            Mosquito                    Argentine Ant
##                                22                                21
##                              Beetle          Flatheaded Appletree Borer
##                                21                                20
##                 Horned Oak Gall Wasp                 Leaf Beetle Family
##                                20                                20
##                   Potato Leafhopper          Tooth-necked Fungus Beetle
##                                20                                20
##                       Codling Moth           Black-spotted Lady Beetle
##                                19                                18
##                       Calico Scale                 Fairyfly Parasitoid
##                                18                                18
##                        Lady Beetle              Minute Parasitic Wasps
##                                18                                18
##                          Mirid Bug                    Mulberry Pyralid
##                                18                                18
##                           Silkworm                     Vedalia Beetle
##                                18                                18
##                Araneoid Spider Order                         Bee Order
##                                17                                17
##                      Egg Parasitoid                     Insect Class
##                                17                                17
##             Moth And Butterfly Order        Oystershell Scale Parasitoid
##                                17                                17
## Hemlock Woolly Adelgid Lady Beetle               Hemlock Wooly Adelgid
##                                16                                16
##                              Mite                      Onion Thrip
##                                16                                16
##                Western Flower Thrips                     Corn Earworm
##                                15                                14
##                   Green Peach Aphid                        House Fly
##                                14                                14
##                           Ox Beetle                 Red Scale Parasite
##                                14                                14
##                  Spined Soldier Bug               Armoured Scale Family
##                                14                                13
##                    Diamondback Moth                     Eulophid Wasp
```

```
##                              13                              13
##                 Monarch Butterfly                 Predatory Bug
##                              13                              13
##             Yellow Fever Mosquito             Braconid Parasitoid
##                              13                              12
##                    Common Thrip    Eastern Subterranean Termite
##                              12                              12
##                           Jassid                      Mite Order
##                              12                              12
##                        Pea Aphid                Pond Wolf Spider
##                              12                              12
##          Spotless Ladybird Beetle          Glasshouse Potato Wasp
##                              11                              10
##                         Lacewing          Southern House Mosquito
##                              10                              10
##           Two Spotted Lady Beetle                     Ant Family
##                              10                               9
##                    Apple Maggot                         (Other)
##                               9                             670
```

Answer: The six most commonly studied species ranked from highest to lowest based on number of observations is the Honeybee (n=667), Parasitic wasp (n=285), Buff Tailed Bumblebee (n=183), Carniolan Honey Bee (n=152), Bumble Bee (n=140), and Italian honeybee (n=113). It is important to note that there is a category of Other (not specified species) which consists of n=670.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

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

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

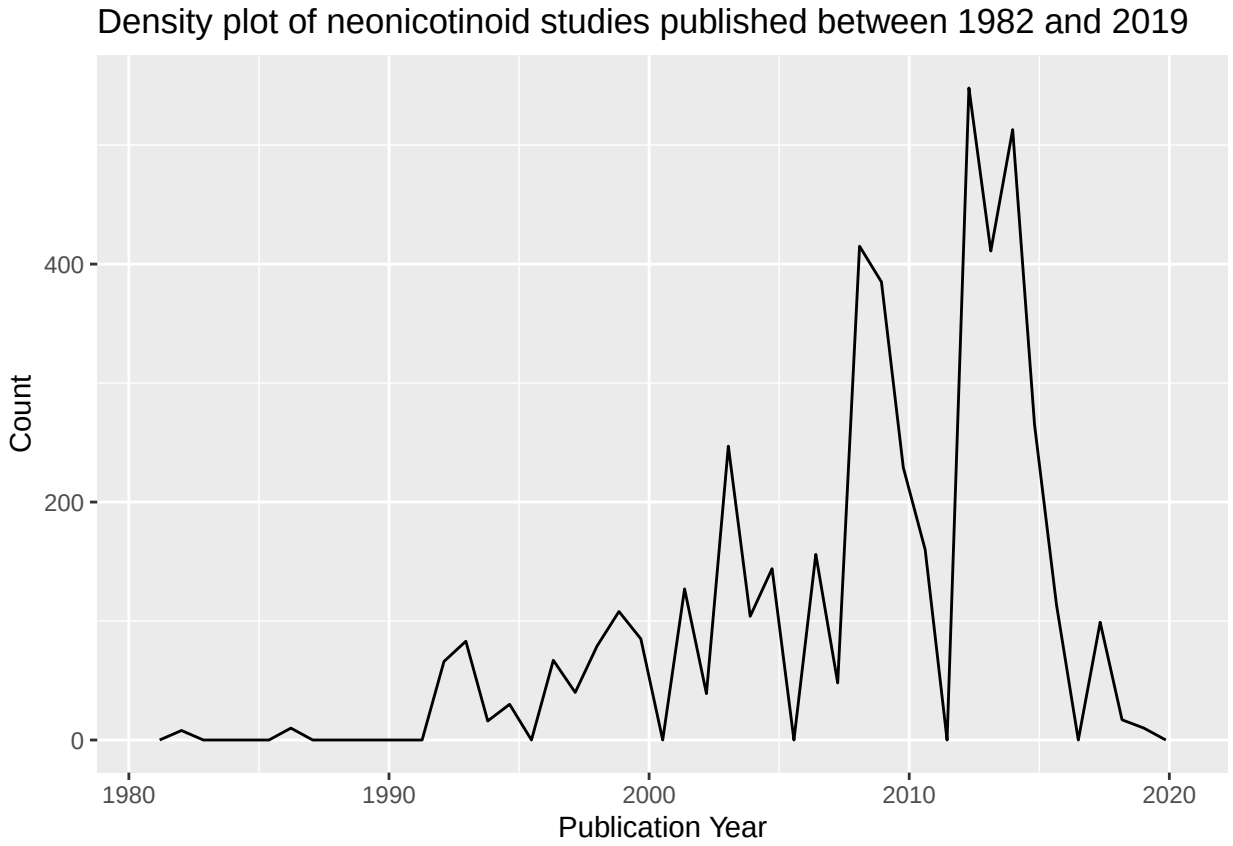
```
# class is factor
```

Answer: The class of Conc.1..Author is factor variable because it is categorical data. The Conc.1..Author is the effect concentration of neonicotinoids recorded from each study's exposure tests and the resultant concentrations recorded for some are below or above the threshold tested and so these have symbols like < or >, which makes these observations categories rather numeric values.

Explore your data graphically (Neonics)

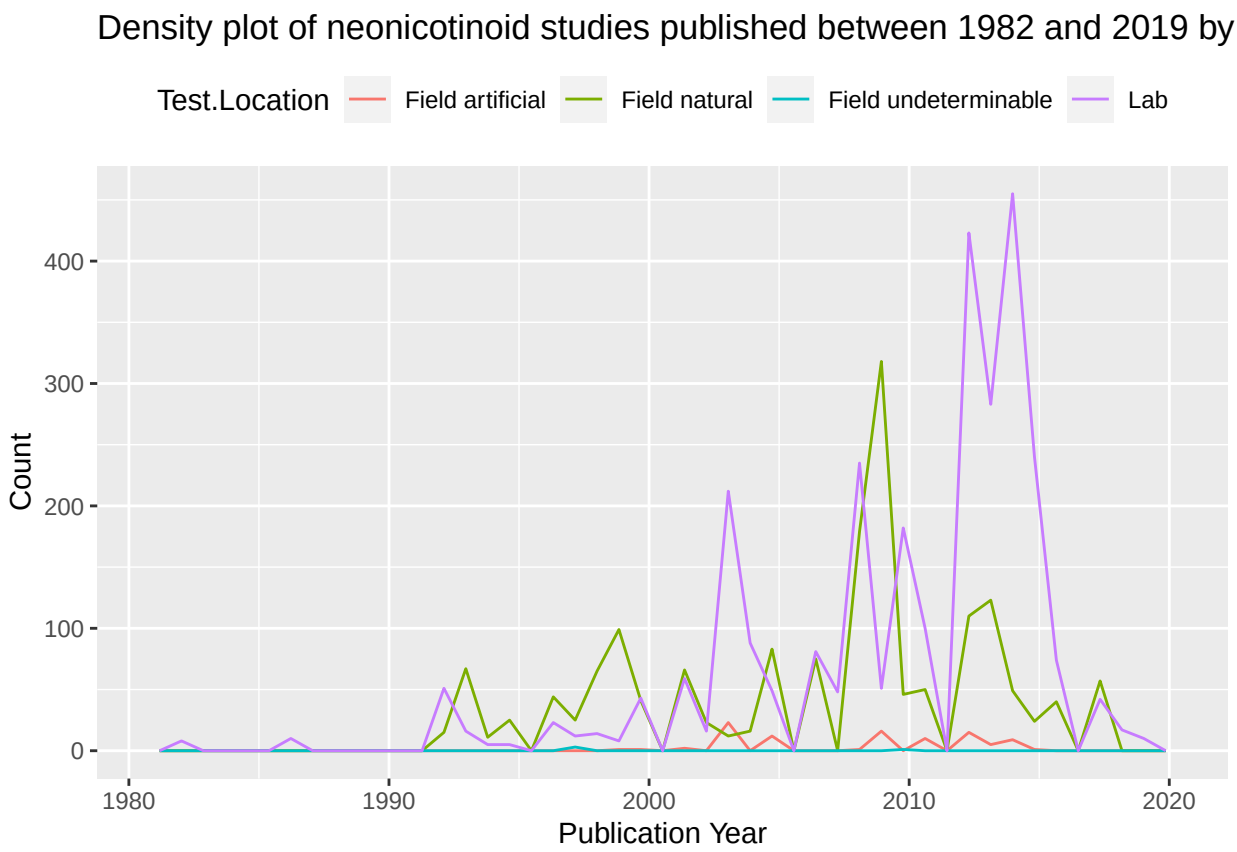
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
Neonic_studies_plot<-ggplot(Neonics)+geom_freqpoly(aes(x = Publication.Year), bins = 45)
Neonic_studies_plot_byYear<-Neonic_studies_plot+
  labs(title="Density plot of neonicotinoid studies published between 1982 and 2019",
  x ="Publication Year", y = "Count")
# renamed the x-axis and y-axis and plot title with the labs() function
# I chose 45 bins, because it provided an easy-to-read distribution.
Neonic_studies_plot_byYear #calling the plot
```



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

```
Neonic_density_plot<-ggplot(Neonics)+
  geom_freqpoly(aes(x = Publication.Year,color=Test.Location), bins = 45)+
  theme(legend.position = "top")+
  labs(title="Density plot of neonicotinoid studies published between 1982 and 2019 by location",
  x ="Publication Year", y = "Count")
#I changed the labels for the x-axis and y-axis using the labs()function
#I gave the density plot a title using the labs() function
#I moved the legend to the top with the theme(legend.position) function
# I chose 45 bins, because it provided an easy-to-read distribution.
Neonic_density_plot #calling the plot
```

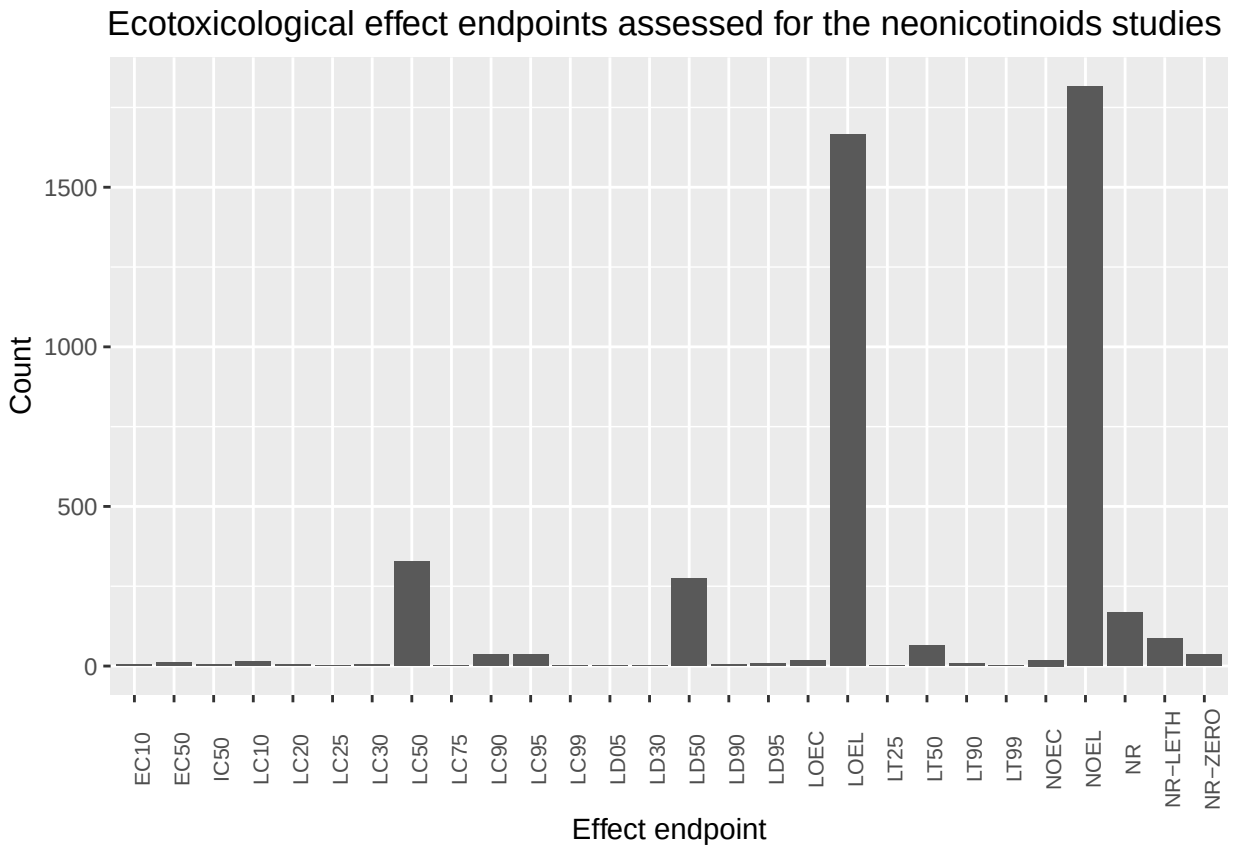


Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are the laboratory indoor setting and the natural field settings. As specified in the ECOTOXicology Database system metadata, the laboratory settings include greenhouses, indoor pots, and garden frames and the natural field settings include field surveys and agricultural sites. The trends do differ over time. From around 1992 to 2001 and from 2009 to 2010, the published studies were mainly from natural field locations. From around 2002 to 2004 and from 2011 to around 2020, the published studies were mainly from laboratory indoor locations. Published studies in 1982 to 1986 also had small peaks of laboratory indoor based studies.

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.

```
Endpoint_bar_graph<-ggplot(Neonics,aes(x=Endpoint))+geom_bar()+
  theme(axis.text.x = element_text(size = 8, angle = 90),
        plot.title = element_text(hjust = 0.5))+
  labs(title="Ecotoxicological effect endpoints assessed for the neonicotinoids studies ",
  x ="Effect endpoint", y = "Count")
#renamed axes and plot title and centered plot title
#I also changed the font and angled the x-axis categories because they were all bunched up.
# To do so, I used the theme(axis.text.x=element_text()) function
Endpoint_bar_graph # calling graph
```



Answer: The two most common endpoints are the LOEL (Lowest observable effect level) and the NOEL (no observable effect level) for the terrestrial ecosystem assessments. From the ECOTOX_CodeAppendix, the LOEL is the smallest dose of the neonicotinoid where you see a critical effect in the exposed terrestial species that is significantly different from the controls in the study. From the ECOTOX_CodeAppendix, the NOEL is the smallest dose of the neonicotinoid where no effects are observed that are significantly different from the controls in the exposure study.

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.

```
#Determine class of collectDate
class(Litter$collectDate)
```

```
## [1] "character"
```

```
#the class of collectDate is factor

#Change from factor to date
Litter$collectDate<-as.Date(Litter$collectDate)
#Check class of collectDate
class(Litter$collectDate)
```

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

```
#class is now Date
```

```
#Use unique function to to determine which dates litter was sampled
litter_sample_dates<-unique(Litter$collectDate)
#created object `litter_sample_dates` for unique(Litter$collectDate)
litter_sample_dates
```

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

```
#litter was sampled on 2018-08-02 and 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`?

```
#Class of plotID
class(Litter$plotID)
```

```
## [1] "character"
```

```
#class of plotID is character
#summary(Litter$plotID) provides the following output:
#   Length     Class      Mode
#      188   character  character

#Determine # of plots sampled at Niwot Ridge using unique()
plots_sampled_at_NR<-unique(Litter$plotID)
plots_sampled_at_NR
```

```
##  [1] "NIWO_061" "NIWO_064" "NIWO_067" "NIWO_040" "NIWO_041" "NIWO_063"
##  [7] "NIWO_047" "NIWO_051" "NIWO_058" "NIWO_046" "NIWO_062" "NIWO_057"
```

```
length(plots_sampled_at_NR) #length() tells me 12 plots were sampled
```

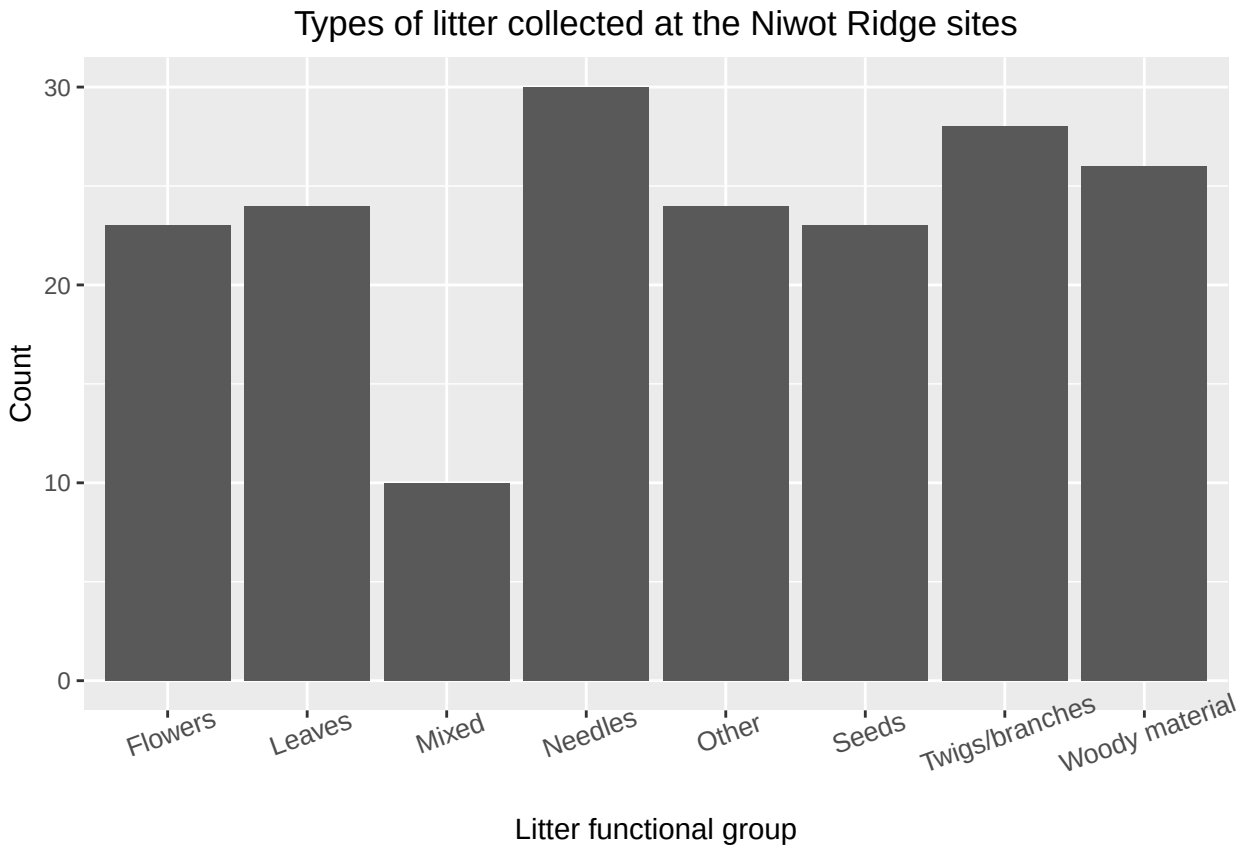
```
## [1] 12
```

```
#Unique returns:
#"NIWO_061" "NIWO_064" "NIWO_067" "NIWO_040"
#"NIWO_041" "NIWO_063" "NIWO_047" "NIWO_051"
#"NIWO_058" "NIWO_046" "NIWO_062" "NIWO_057"
```

Answer: The information from unique() is different from the information obtained from summary() because unique() shows you the exact plots that were sampled and removes any duplicate counts of the same plot name. I used the length() function on the object I created from the unique() function to tell me that there were 12 unique plots sampled in this dataset. plotID is a character variable, and when I use the summary() function it only summarizes the length of all the observations. So there are repeated counts of the plots and it does not return the unique fields like the unique() function. Not only do I know that 12 plots were sampled, I also know the exact plots which were:NIWO_061, NIWO_064, NIWO_067, NIWO_040, NIWO_041, NIWO_063, NIWO_047, NIWO_051, NIWO_058, NIWO_046, NIWO_062, NIWO_057.

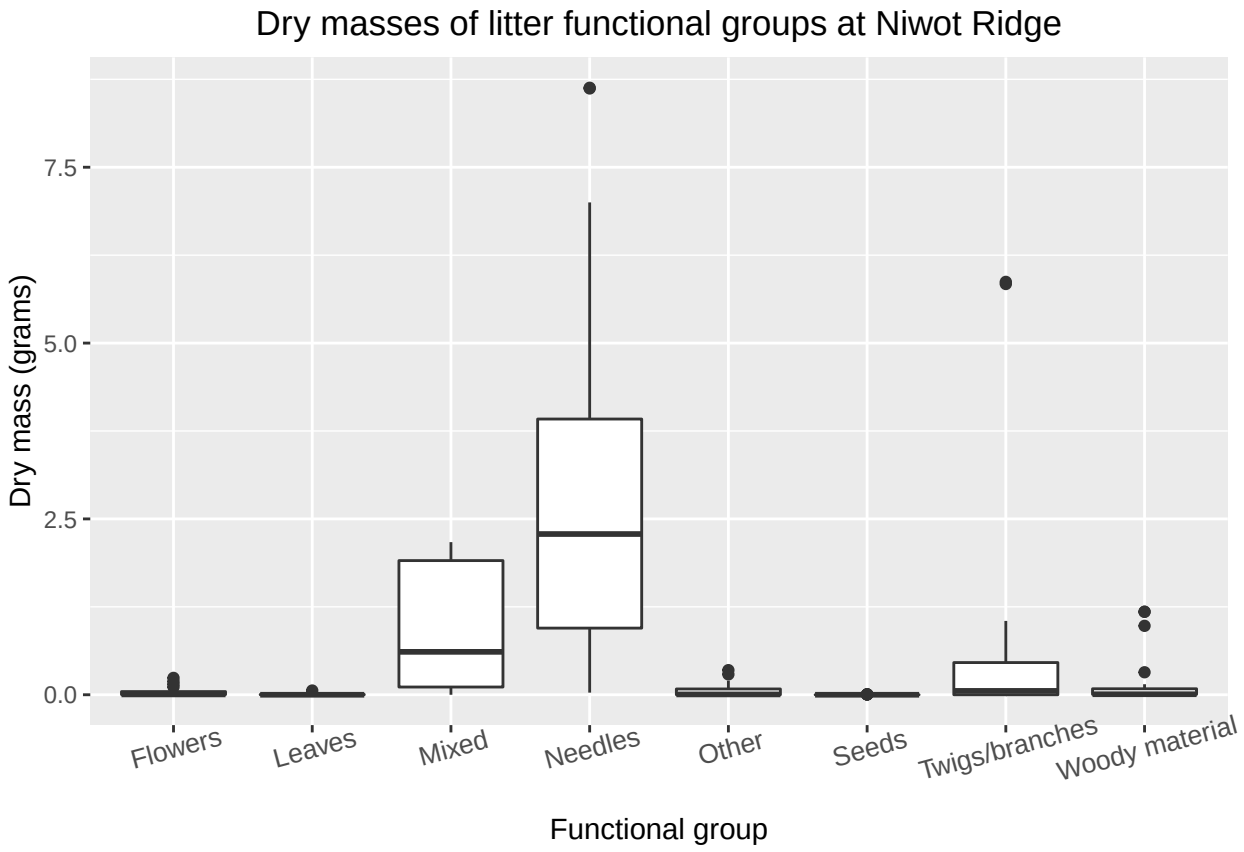
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```
Functional_group_bargraph<-ggplot(Litter,aes(x=functionalGroup))+geom_bar()+
  labs(title="Types of litter collected at the Niwot Ridge sites",
  x ="Litter functional group", y = "Count")+
  theme(axis.text.x = element_text(size = 10, angle = 20),plot.title = element_text(hjust = 0.5))
#renamed axes and plot title and center adjusted plot title
#titled text in x-axis b/c words were overlapping
Functional_group_bargraph #calling graph
```

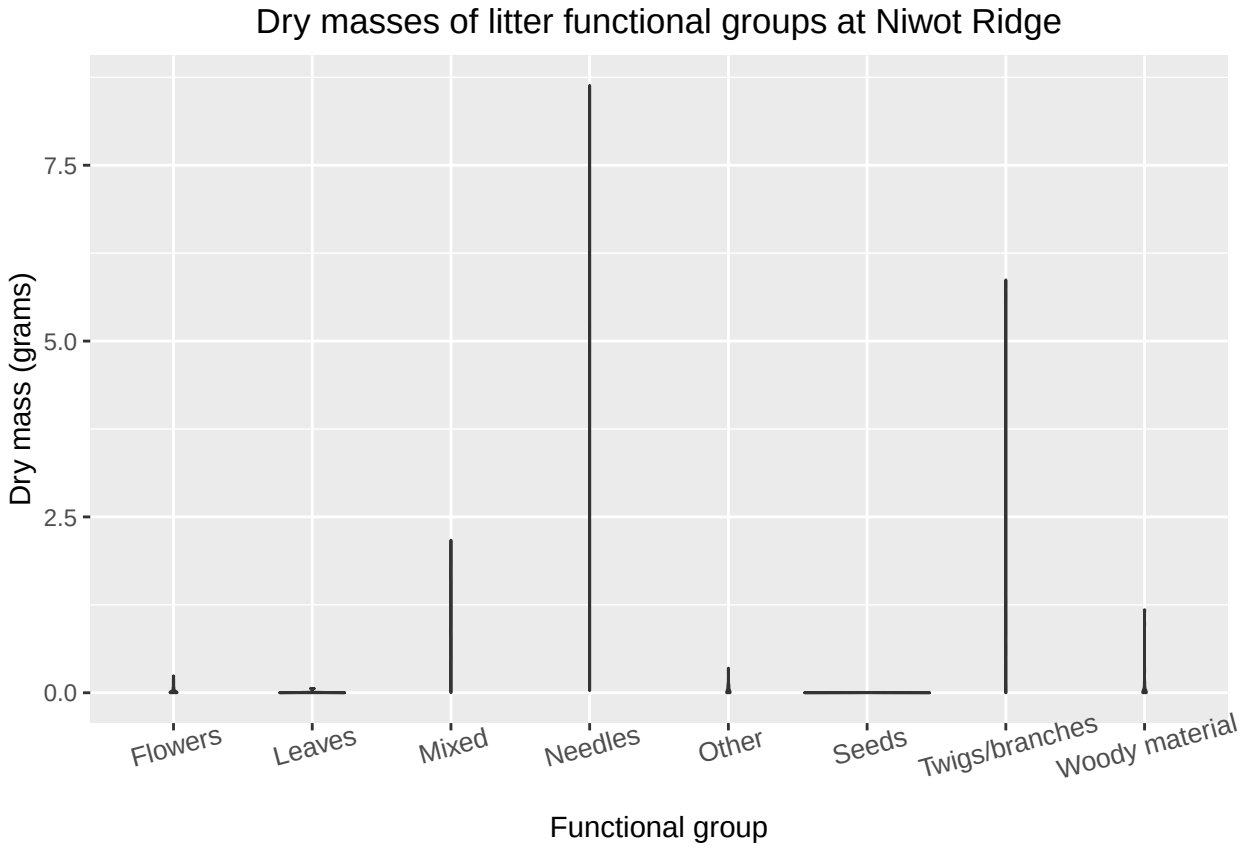


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#boxplot of dryMass by functionalGroup
dryMass_boxplot<-ggplot(Litter)+geom_boxplot(aes(x=functionalGroup, y=dryMass))+
  labs(title="Dry masses of litter functional groups at Niwot Ridge",
  x ="Functional group", y = "Dry mass (grams)")+
  theme(axis.text.x = element_text(size = 10, angle = 15),plot.title = element_text(hjust = 0.5))
dryMass_boxplot #calling the boxplot object
```




```
#violin plot of dryMass by functionalGroup
dryMass_violin<-ggplot(Litter)+
  geom_violin(aes(x=functionalGroup, y=dryMass))+
  labs(title="Dry masses of litter functional groups at Niwot Ridge",
  x ="Functional group", y = "Dry mass (grams)")+
  theme(axis.text.x = element_text(size = 10, angle = 15),plot.title = element_text(hjust = 0.5))
dryMass_violin #calling the violin plot object
```




```
#Changed functionalGroup from character to factor
#checked the sample size of the different functional groups
summary(as.factor(Litter$functionalGroup)) #sample sizes are relatively small
```

```
##        Flowers         Leaves          Mixed        Needles          Other
##             23             24             10             30             24
##          Seeds Twigs/branches Woody material
##             23             28             26
```

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

> Answer: The boxplot is a more effective visualizaiton option than the violin plot because I can visualize the distribution of the data for each functional group. I can see the outliers present in each functional group category. I can also better visualize the skewness of each distribution in the boxplot as opposed to the violin plot. The violin plot doesn't show the distribution variation well and we don't see the IQR, quartiles, and the median as we do with the boxplot. I also checked the sample size, which is relatively small as most are n<30 (with the exception of needles), and the violin plot shows a distorted version of the data, the distribution looks a lot smoother, which isn't the case with the boxplot. The outlier seen in the boxplot for twigs/branches is deceiving in the violin plot, which makes it look like twigs has the second highest biomass, but it is not.

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

> Answer: Based on the boxplots, needles have the highest dry biomass at the Niwot Ridge sites and Mixed (which the User Guide defines as unsorted material) has the second highest dry biomass at these sites.