Assignment 3: Data Exploration

Elise Harrigan

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.

```
library(tidyverse)

Neonics <- read.csv("~/Desktop/Duke/Spring 2021/EnvDataAnalytics_872/Environmental_Data_Analytics_2021/

Litter <- read.csv("~/Desktop/Duke/Spring 2021/EnvDataAnalytics_872/Environmental_Data_Analytics_2021/Da
```

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 ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

   Answer:Neonicotinoids are used to keep crops safe and healthy from insects that may interfere with the productivity or viability of the crop. It is an insecticide that directly targets certain species of insects that tend to feed or growth in field crops. With the use of insecticides, there can be unforeseen consequences as pollinating insects, which are necessary, might also be negatively affected by the usage of the product.

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: Forest litter and woody debris is critical to determine forest health. In the case of Colorado, a dry and arid climate, too much forest litter and debris can lead to increase in wildfires. If there is a lighting storm or manmade fires started, a build up of litter can dry out and create mass fire outbreaks.

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: *Sampling of woody debris and litter occurs in sites over 2m tall in woody vegetation* Sites with forested tower airsheds the litter sampling plots are 20 40x40m tower plots and 26 20x20m plots.There is one elevated trap and one ground trap every 400 m^2 plot area which results in 1-4 trap pairs per plot. * In plots with greater than 50% of forest cover, traps are randomized and utilized on a grid cell location. For plots less than 50% coverage or patchy vegetation, traps are placed in targeted areas based on the vegetaion.

Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

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

```
#Neonics has 4623 rows and 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?

```
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
```

Answer: the most common effects that are studied are Population, Mortality and Feeding behavior. These are most common because it is important to know how many insects you are dealing iwth, the mortality rate and how they are able to feed and continue to grow in their life cycle.

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.

```
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 most common thread among the top species are they they are pollinators. Insecticides are dangerous to pollinator communities such as honey bee, bumble bee and wasp. Although insecticides are designed to rid the crops of infestation of pests, it has a negative effect on other important insects such as bees. Honey Bee Parasitic Wasp 667 285 Buff Tailed Bumblebee Carniolan Honey Bee 183 152 Bumble Bee Italian Honeybee 140 113

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(Neonics$Conc.1..Author.)
```

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

```
head(Neonics$Conc.1..Author.)
```

```
## [1] 27.2 19.7 47   25   13   268
```

```
## 1006 Levels: <0.0004 <0.025 <0.088 <0.5 <1.5 <10/ <2.5/ <4.00 <5.00 ... NR/
```

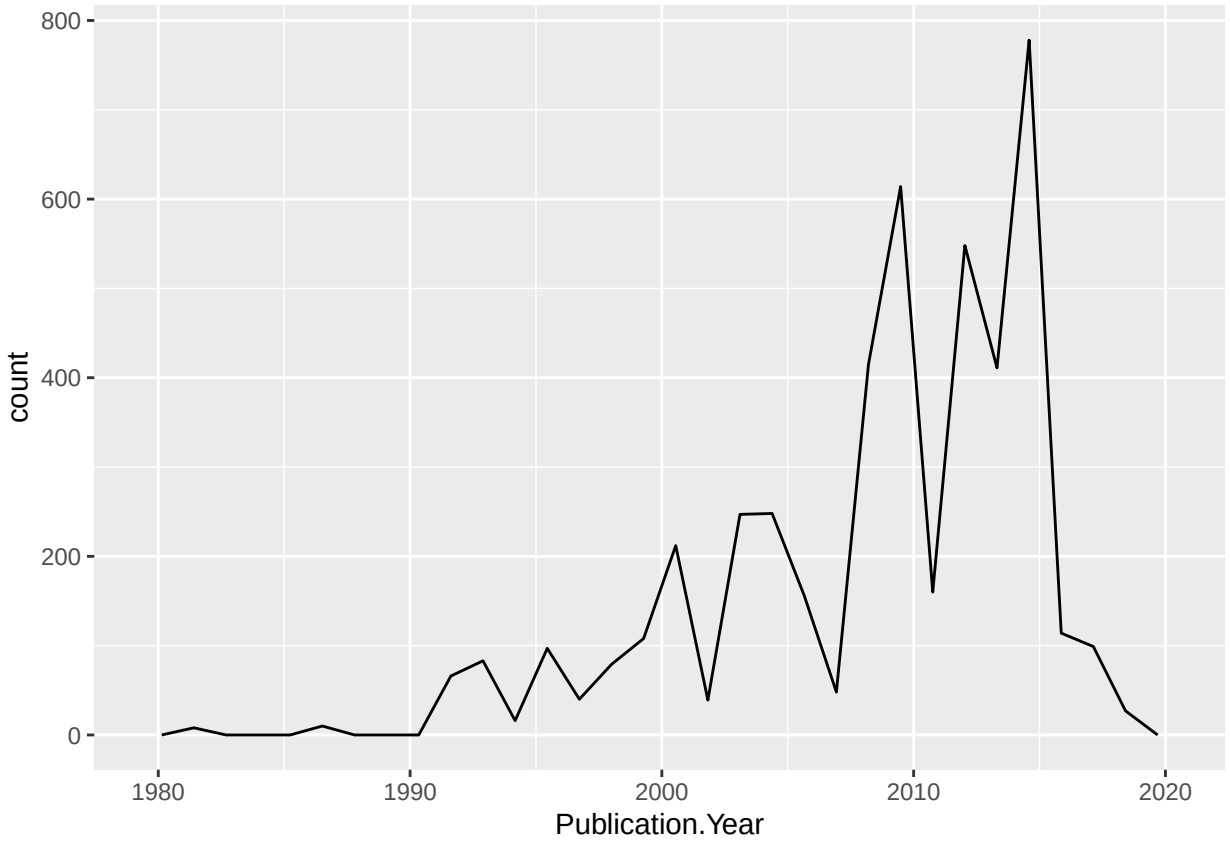
Answer: The class is a character. It is not numeric because it contains more than just numbers.

Explore your data graphically (Neonics)

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

```
library(ggplot2)
ggplot(Neonics, aes(Publication.Year)) +
  geom_freqpoly()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



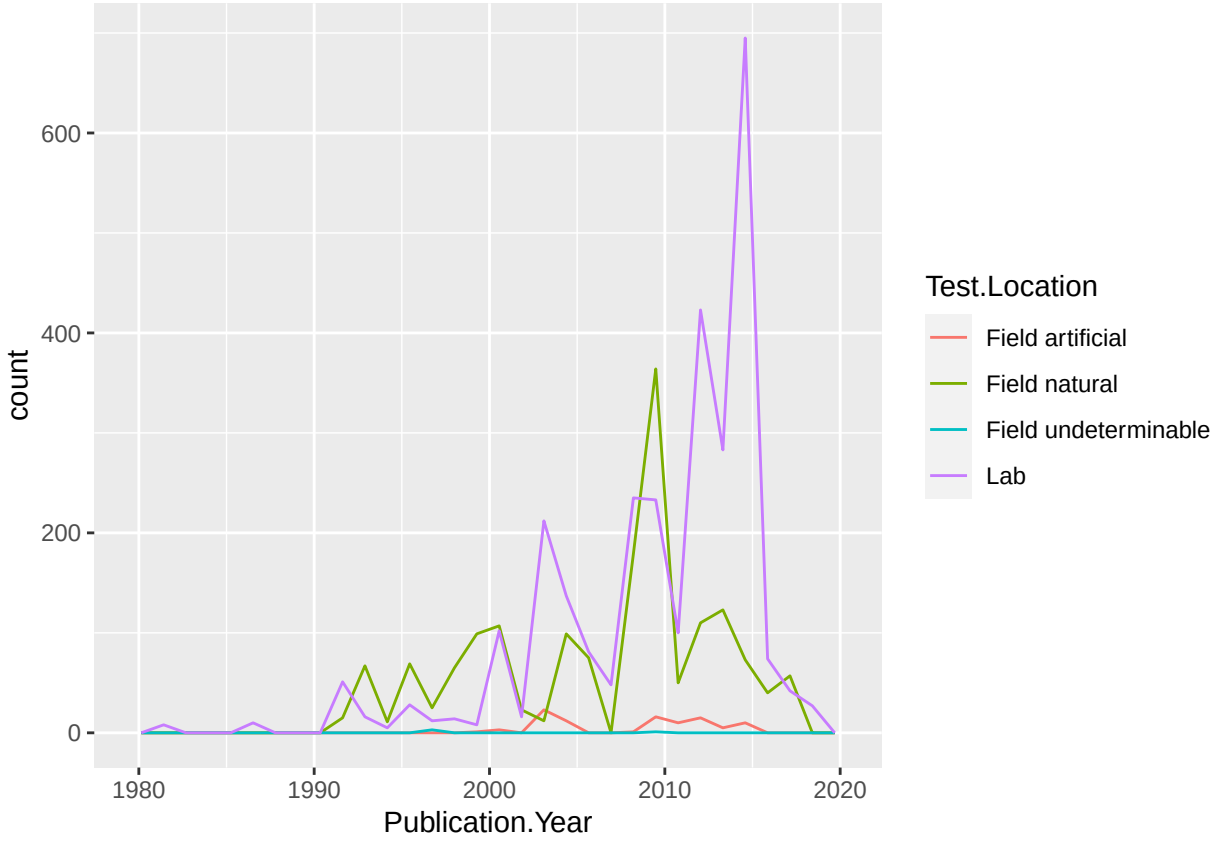

```
summary(Neonics$Publication.Year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1982    2005    2010    2008    2013    2019
```

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

```
ggplot(Neonics, aes(Publication.Year, colour=Test.Location)) +
  geom_freqpoly()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

> Answer: The most common test location is in the lab location, peaking in 2015. Historically natural field sites were also common, especially from the mid 1990s to 2010 where there was a major spike in field locations. Both lab and field sites have peaks and dips over time.There are very little publications using artifical sites and none using undeterminable sites. But recently the publications have dropped in all test locations. One reason for this could be the data hasn't been provided for 2019-2020 but also COVID-19 could have an effect on the amount of testing possible.

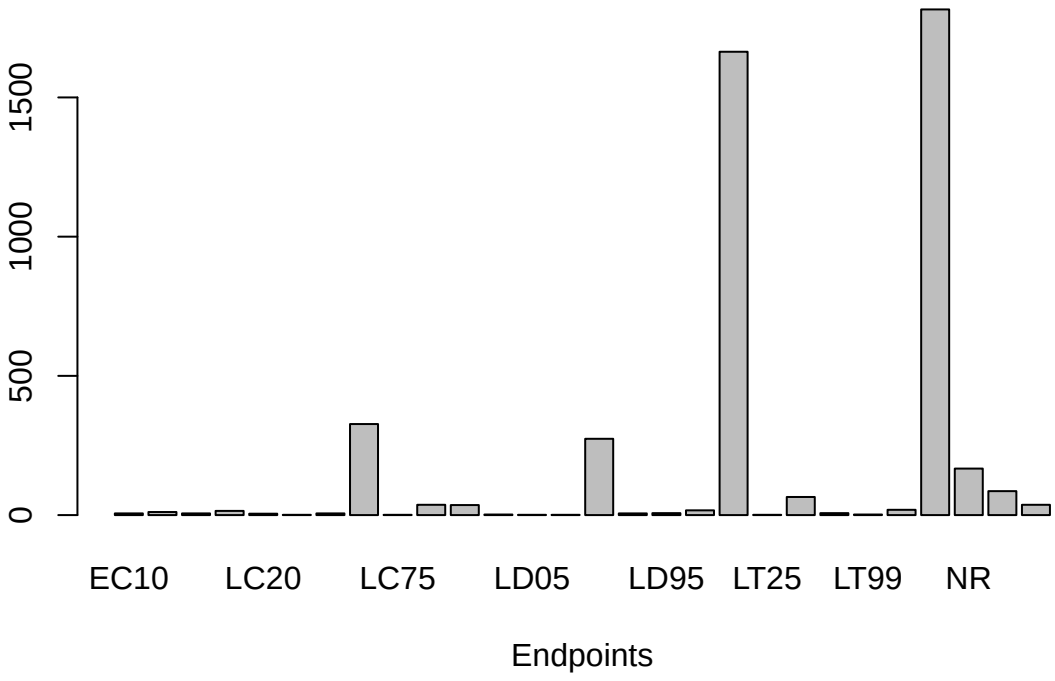
11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

```
summary(Neonics$Endpoint)
```

```
##    EC10    EC50    IC50    LC10    LC20    LC25    LC30    LC50    LC75    LC90
##       6      11       6      15       5       1       6     327       1      37
##    LC95    LC99    LD05    LD30    LD50    LD90    LD95    LOEC    LOEL    LT25
##      36       2       1       1     274       6       7      17    1664       1
##    LT50    LT90    LT99    NOEC    NOEL      NR NR-LETH NR-ZERO
##      65       7       2      19    1816     167      86      37
```

```
counts <- table(Neonics$Endpoint)

barplot(counts, main="Endpoint Counts",
        xlab="Endpoints")
```



Answer: The most common endpoints are LOEL and NOEL. LOEL is terrestrial and stands for Lowest-observable-effect-level: lowest dose that produced significantly different values. NOEL is also terrestrial and stands for No-observable-effect-level: highest dose producing effects not significantly different from responses.

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"
```

```
head(Litter$collectDate) #Yes it is a date.
```

```
## [1] 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02 2018-08-02
## Levels: 2018-08-02 2018-08-30
```

```
library("lubridate")
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
unique(Litter$collectDate)
```

```
## [1] 2018-08-02 2018-08-30
## Levels: 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)
```

```
##  [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
```

```
summary(Litter$plotID)
```

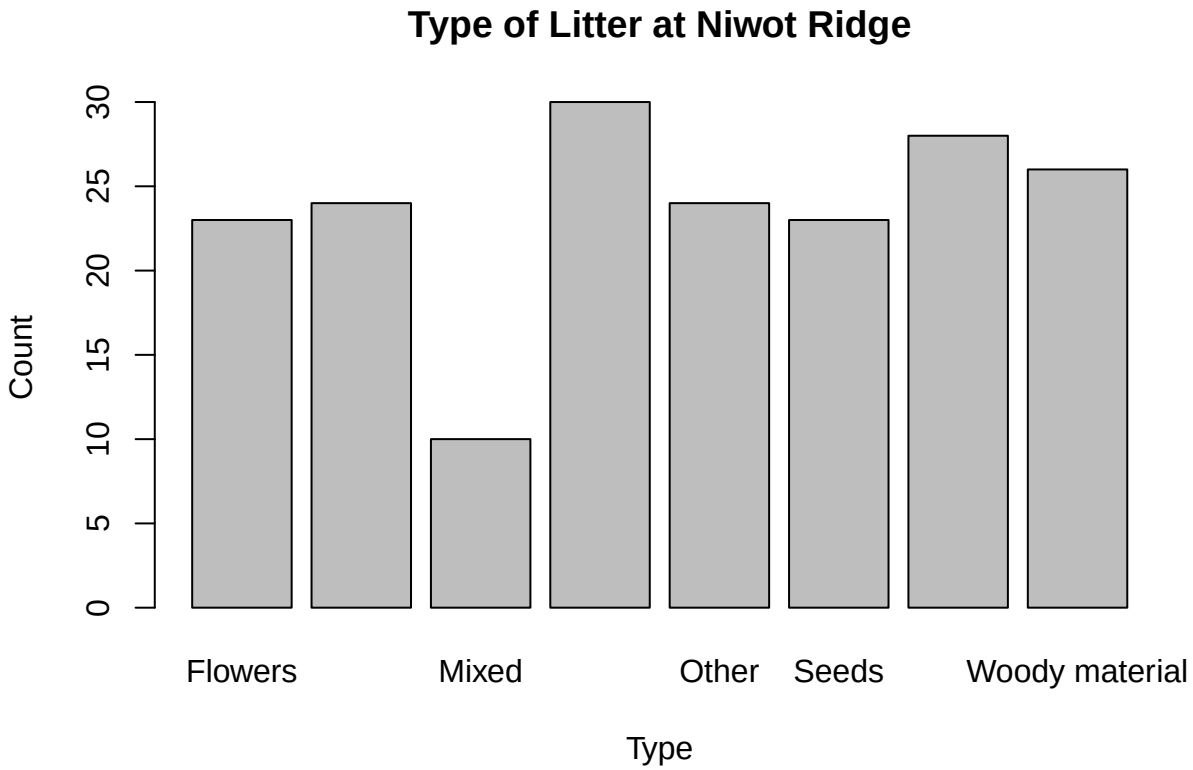
```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: There were 12 plots sampled at Niwot Ridge. The unique function gives you the value of the number of plots taken and summary gives you summary statistics on the variable. In this case, summary gave how many of each type of plot there were and unique gives the "unique value" without repeating.

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.

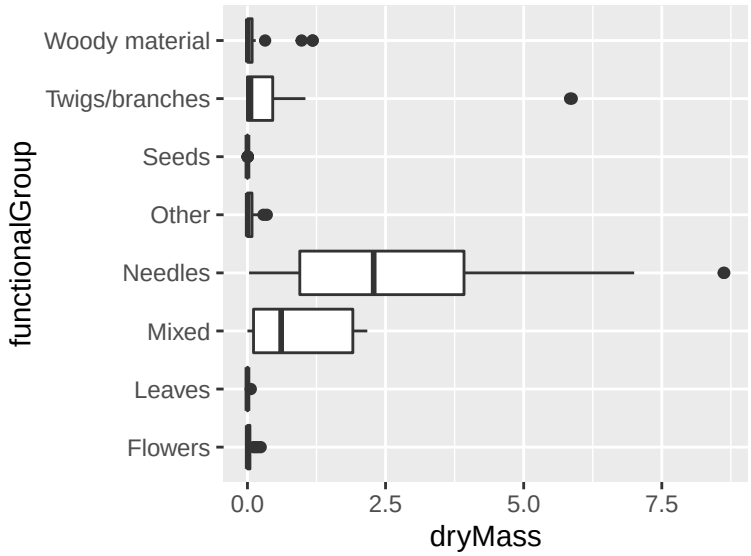
```
func_group_counts <- table(Litter$functionalGroup)

barplot(func_group_counts, main="Type of Litter at Niwot Ridge",
        xlab="Type", ylab="Count")
```

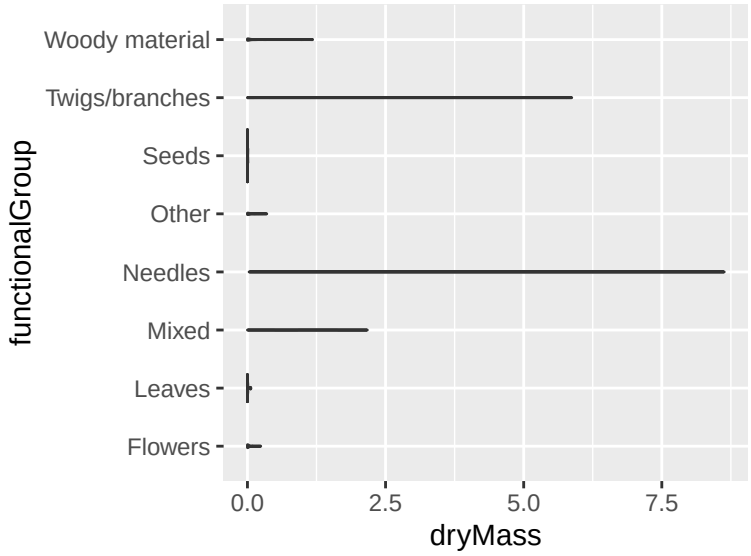


15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
ggplot(Litter) +
  geom_boxplot(aes(x=dryMass, y=functionalGroup))
```




```
ggplot(Litter) +
  geom_violin(aes(x = dryMass, y = functionalGroup))
```



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

> Answer: The boxplot shows more details on the range of variables. You can see the outliers and where the mean of the data fall. The violin plot does not accurately describe the fairly equal distribution across the group types.

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

> Answer: Needles has the highest biomass.