R-Programming Midterm

Alexandria Shonk

2023-03-09

R-Programming Midterm

Due 19Feb2023 11:55pm

Background, Research Question, & Hypothesis

Background:

The data gives binary values to the presence of squirrels, cats, and dogs that are active within 30 meters of bird feeders for at least 30 minutes a day, at least three times a week.

Research Question:

Do the locations that are in larger-sized towns generally have more types of animals active around bird feeders?

Expectation (hypothesis):

I expect that locations with a higher population will more often report all three different species in the vicinity of bird feeders. My reasoning is that a higher population of people will have more domestic cats and dogs.

I will examine the overall difference between town size and the presence of more species, and then also look at them broken down to four types of habitat: residential, industrial, agricultural, and mixed woods. I'm interested to see if the number of locations reporting the presence of cats and dogs differs between habitat type. Specifically, I am interested to see if the mixed woods habitat has fewer number of species than the residential, industrial, and agricultural habitats do.

Data Loading

I used the Bird FeederWatch data between 1989 and 2021 from Tidy Tuesday.

```
#Load packages
library(tidyverse)
library(readxl)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(readr)

#Import data, save to data memory
PFW_count_site_data_public_2021 <- read_csv("~/Desktop/PFW_count_site_data_public_2021.csv")
```

```
## Rows: 254355 Columns: 62

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): loc_id, proj_period_id
## dbl (60): yard_type_pavement, yard_type_garden, yard_type_landsca, yard_type...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
View(PFW_count_site_data_public_2021)
saveRDS(object = PFW_count_site_data_public_2021, file = "PFW2021.rds" )
```

There are some data missing from some of the observations, so I will omit those data points.

Data Transformation

```
#The code for transforming the data is in one long chunk, but below are the steps I take in it:
#(1) Categorizing town sizes into "Large Town", "Medium Town", "Small Town", and "Village or Small Comm
#(2) Removing habitat types not needed. I am keeping residential, agricultural, industrial, and mixed wo

#(3) I'm possibly curious in examining seasons, so I decided to section months into the seasons.
        # Winter = December, January, February
        # Spring = March, April, May
        # Summer = June, July, August
        # Autumn = September, October, November
#Removing variables not needed for research questions

pfw_new <- PFW_count_site_data_public_2021%>%

    mutate (town_size = case_when(
    population_atleast > 60000 ~ "Large town",
    (population_atleast) > 25000 & (population_atleast < 59999) ~ "Medium town",
    (population_atleast > 24999) & (population_atleast > 7500) ~ "Small town",
    population_atleast < 7500 ~ "Village or Small Community"))%>%

  mutate( total_animal_activity = squirrels + cats + dogs)%>%

    mutate( winter = fed_in_dec + fed_in_jan + fed_in_feb)%>%
    mutate( spring = fed_in_mar + fed_in_apr + fed_in_may)%>%
    mutate( summer = fed_in_jun + fed_in_jul + fed_in_aug)%>%
    mutate( autumn = fed_in_sep + fed_in_oct + fed_in_nov)%>%

select(-hab_dcid_woods, -hab_evgr_woods, -hab_orchard, -hab_park, -hab_water_fresh,    -hab_water_salt,
```

```
pfw_new<-pfw_new%>%
na.omit(hab_mixed_woods)%>%
  na.omit(hab_residential)%>%
  na.omit(hab_industrial)%>%
  na.omit(hab_agricultural)%>%
  na.omit(nearby_feeders)%>%
  na.omit(squirrels)%>%
  na.omit(cats)%>%
  na.omit(dogs)%>%
  na.omit(humans)%>%
  na.omit(housing_density)%>%
  na.omit(fed_yr_round)%>%
  na.omit(population_atleast)%>%
  na.omit(count_area_size_sq_m_atleast)%>%
  na.omit(winter)%>%
  na.omit(spring)%>%
  na.omit(summer)%>%
  na.omit(autumn)
```

Show your transformed table here. Use tools such as `glimpse()`, `skim()` or `head()` to illustrate your point.

```
pfw_new%>%
glimpse()
```

```
## Rows: 41,907
## Columns: 21
## $ loc_id                       <chr> "L100025", "L100025", "L100025", "L100032~
## $ proj_period_id               <chr> "PFW_2002", "PFW_2004", "PFW_2005", "PFW_~
## $ hab_mixed_woods              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1,~
## $ hab_residential              <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ hab_industrial               <dbl> 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1,~
## $ hab_agricultural             <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0,~
## $ nearby_feeders               <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ squirrels                    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1,~
## $ cats                         <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ dogs                         <dbl> 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ humans                       <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1,~
## $ housing_density              <dbl> 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 2, 2, 3,~
## $ fed_yr_round                 <dbl> 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ population_atleast           <dbl> 5001, 5001, 5001, 5001, 5001, 5001, 5001,~
## $ count_area_size_sq_m_atleast <dbl> 100.01, 100.01, 100.01, 375.01, 375.01, 1~
## $ town_size                    <chr> "Village or Small Community", "Village or~
## $ total_animal_activity        <dbl> 3, 3, 3, 2, 2, 1, 1, 1, 1, 1, 1, 0, 1, 1,~
## $ winter                       <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
## $ spring                       <dbl> 2, 1, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
## $ summer                       <dbl> 0, 0, 0, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
## $ autumn                       <dbl> 1, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
```

```
pfw_new%>%
head()
```

```
## # A tibble: 6 x 21
##   loc_id  proj_per~1 hab_m~2 hab_r~3 hab_i~4 hab_a~5 nearb~6 squir~7  cats  dogs
##   <chr>   <chr>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 L100025 PFW_2002         1       1       1       1       1       1     1     1
## 2 L100025 PFW_2004         1       1       1       1       1       1     1     1
## 3 L100025 PFW_2005         1       1       1       1       1       1     1     1
## 4 L100032 PFW_2018         1       1       0       0       1       1     0     1
## 5 L100032 PFW_2019         1       1       1       0       1       1     0     1
## 6 L100032 PFW_2020         1       1       1       0       1       1     0     0
## # ... with 11 more variables: humans <dbl>, housing_density <dbl>,
## #   fed_yr_round <dbl>, population_atleast <dbl>,
## #   count_area_size_sq_m_atleast <dbl>, town_size <chr>,
## #   total_animal_activity <dbl>, winter <dbl>, spring <dbl>, summer <dbl>,
## #   autumn <dbl>, and abbreviated variable names 1: proj_period_id,
## #   2: hab_mixed_woods, 3: hab_residential, 4: hab_industrial,
## #   5: hab_agricultural, 6: nearby_feeders, 7: squirrels
```

The values are rather similar to what I expected them to be. I was surprised at how many locations fell into the "Village or Small Community" category, but when examining the raw data closer, I saw that some population sizes were reported as "1". These data were reported by people watching bird feeders in their community, so it is possible that they either did not understand and report correctly, or that they did not report the number and the default value is 1.

Data Visualization and and Summation

Use `group_by()` and `summarize()` to make a summary of the data here. The summary should be relevant to your research question

```
pfw_new %>%
    group_by(winter) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

winter	squirrels	cats	dogs
0	223	115	137
1	221	107	141
2	234	121	127
3	34271	18036	19939

```
pfw_new %>%
    group_by(spring) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

spring	squirrels	cats	dogs
0	210	133	152
1	736	366	372

2	2382	1148	1335
3	31621	16732	18485

```
pfw_new %>%
    group_by(summer) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

summer	squirrels	cats	dogs
0	4109	2109	2301
1	746	375	414
2	240	108	143
3	29854	15787	17486

```
pfw_new %>%
    group_by(autumn) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

autumn	squirrels	cats	dogs
0	241	108	153
1	1875	917	1055
2	2042	1010	1113
3	30791	16344	18023

```
pfw_new %>%
    group_by(hab_residential) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

hab_residential	squirrels	cats	dogs
0	2737	1094	1579
1	32212	17285	18765

```
pfw_new %>%
    group_by(hab_industrial) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

hab_industrial	squirrels	cats	dogs

0	26968	13212	15182
1	7981	5167	5162

```
pfw_new %>%
    group_by(hab_agricultural) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

hab_agricultural	squirrels	cats	dogs
0	23672	12113	13431
1	11277	6266	6913

```
pfw_new %>%
    group_by(hab_mixed_woods) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

hab_mixed_woods	squirrels	cats	dogs
0	11105	6749	7146
1	23844	11630	13198

```
pfw_new %>%
    group_by(town_size) %>%
    summarize(squirrels = sum(squirrels, na.rm = TRUE),
              cats = sum(cats, na.rm = TRUE),
              dogs = sum(dogs, na.rm = TRUE)) %>% gt::gt()
```

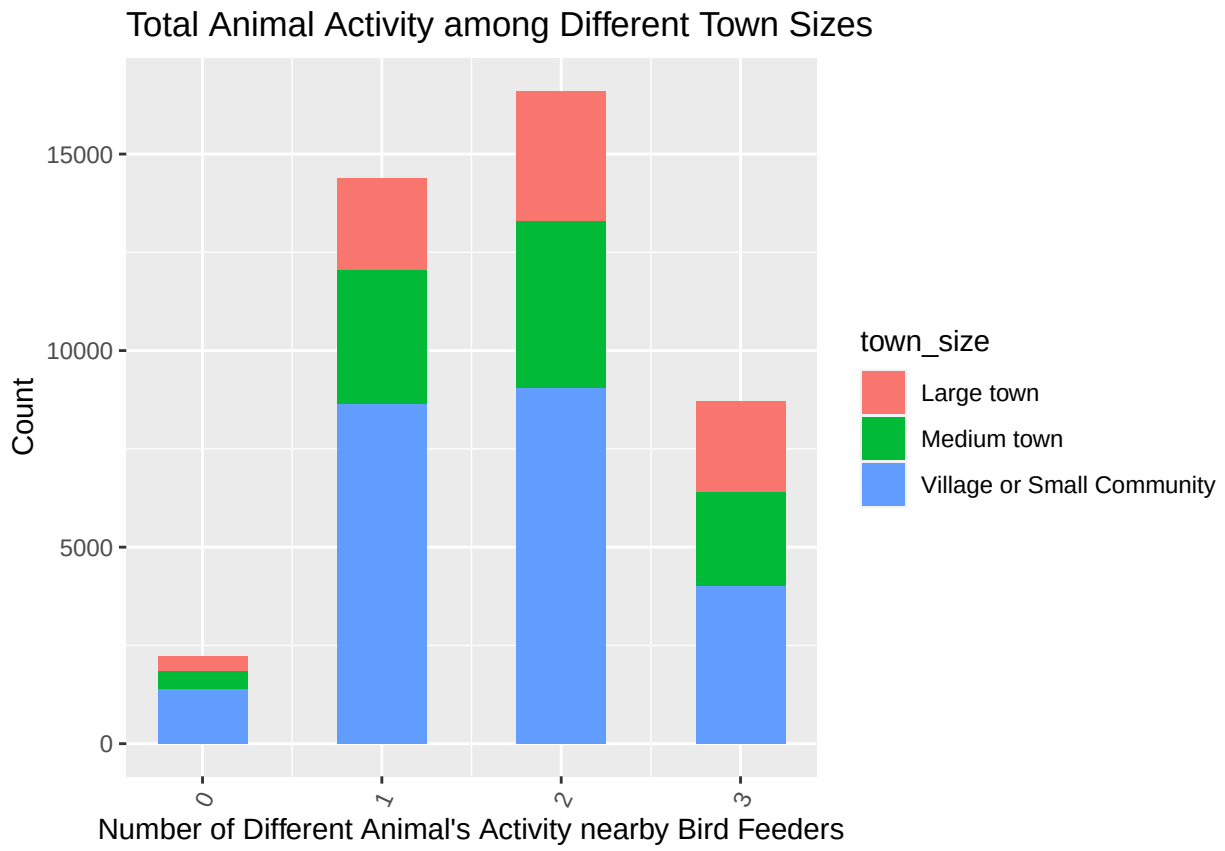
town_size	squirrels	cats	dogs
Large town	7106	4228	4509
Medium town	8980	4844	5279
Village or Small Community	18863	9307	10556

I decided to look at the summaries of animal species by season, but chose to move forward with my original question examining the relationship between town size and number of different species of animal in the vicinity of bird feeders.

I was surprised to see that the mixed woods habitat reported more binary sightings of cats and dogs than residential areas. I wonder if the mixed woods habitat includes parks and hiking trails where people bring their dogs, or if they are woods in residential areas.

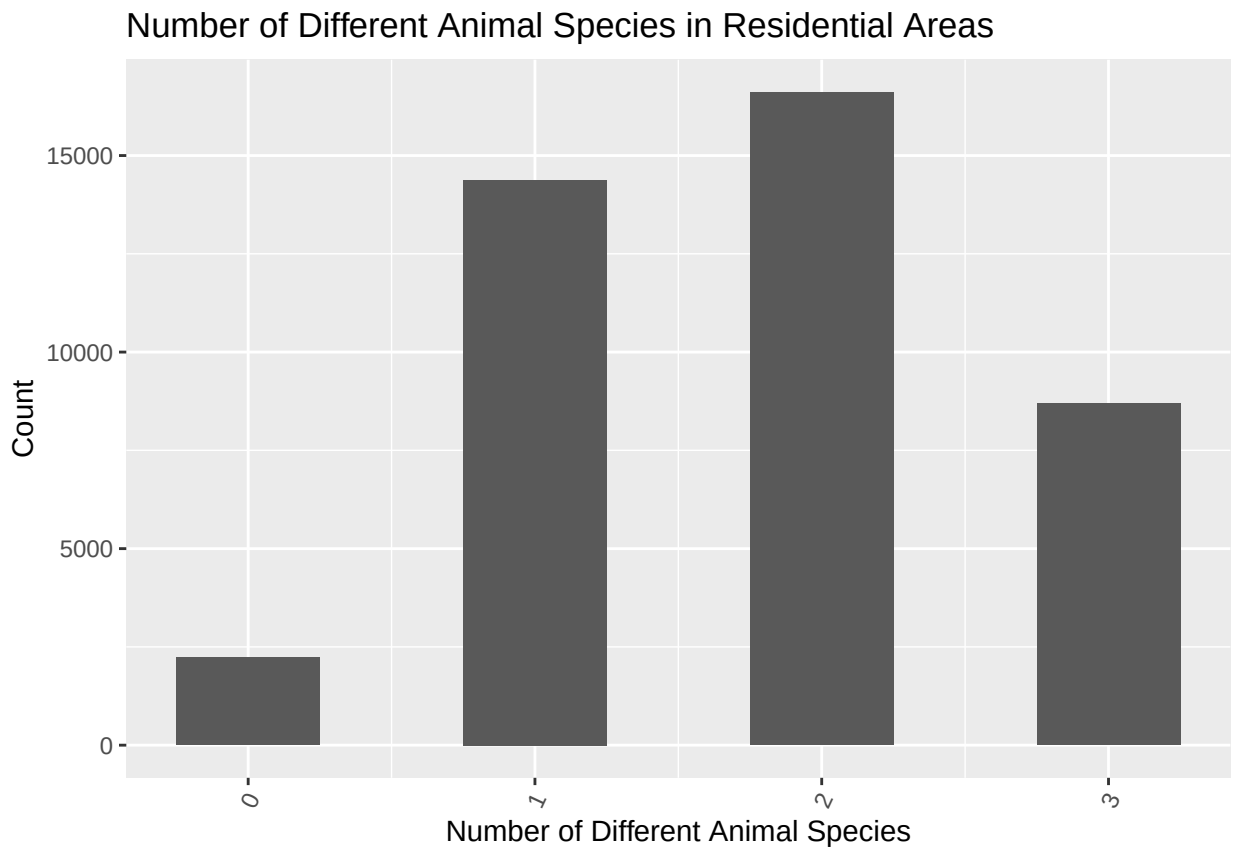
After putting the data into summary tables, I realized that there were no population sizes that corresponded with a “Small Town” size, which I thought was interesting.

```
#Examining the number of different species by town size
g <- ggplot(pfw_new, aes(total_animal_activity))
g + geom_bar(aes(fill=town_size), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Total Animal Activity among Different Town Sizes",
       x = "Number of Different Animal's Activity nearby Bird Feeders",
       y = "Count")
```



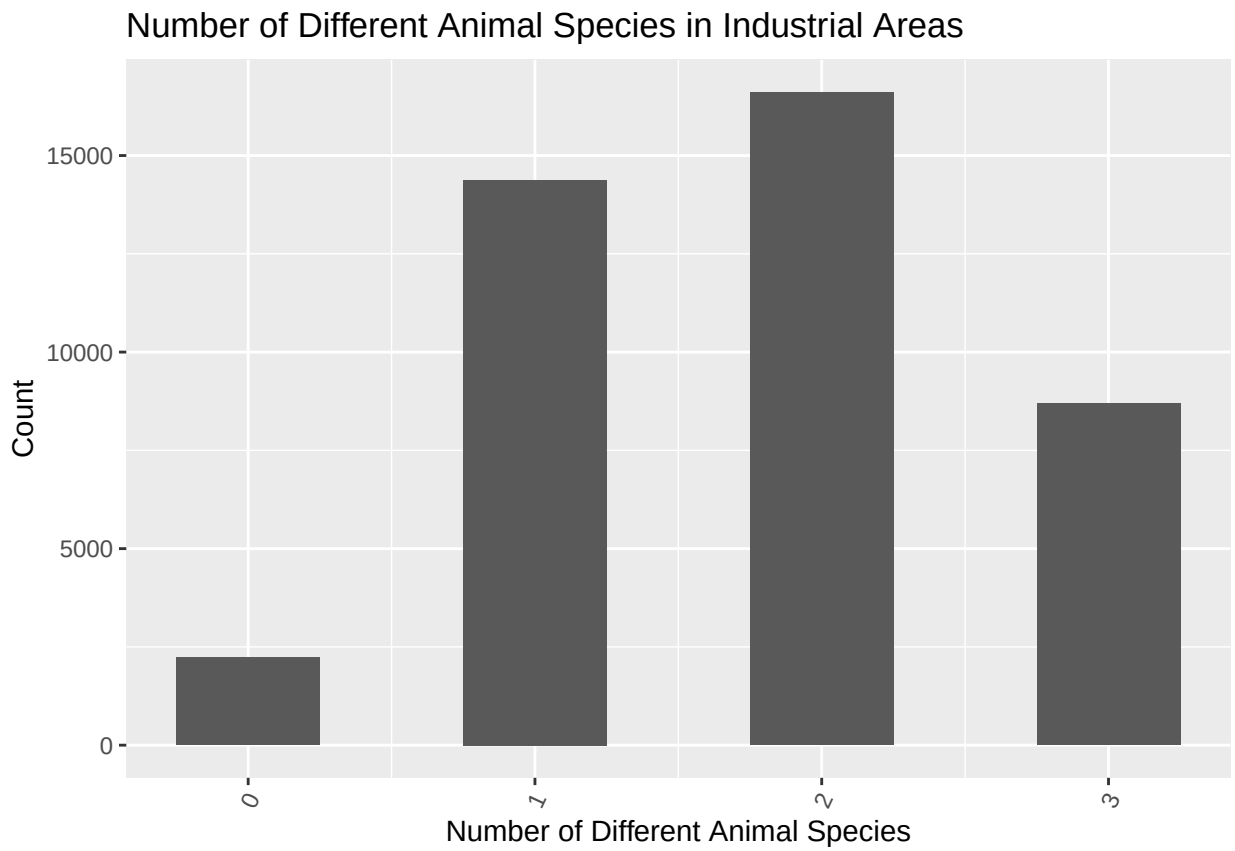

```
#Examining the number of different species between habitats
g <- ggplot(pfw_new, aes(total_animal_activity))
g + geom_bar(aes(fill=hab_residential), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Number of Different Animal Species in Residential Areas",
       x = "Number of Different Animal Species",
       y = "Count")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



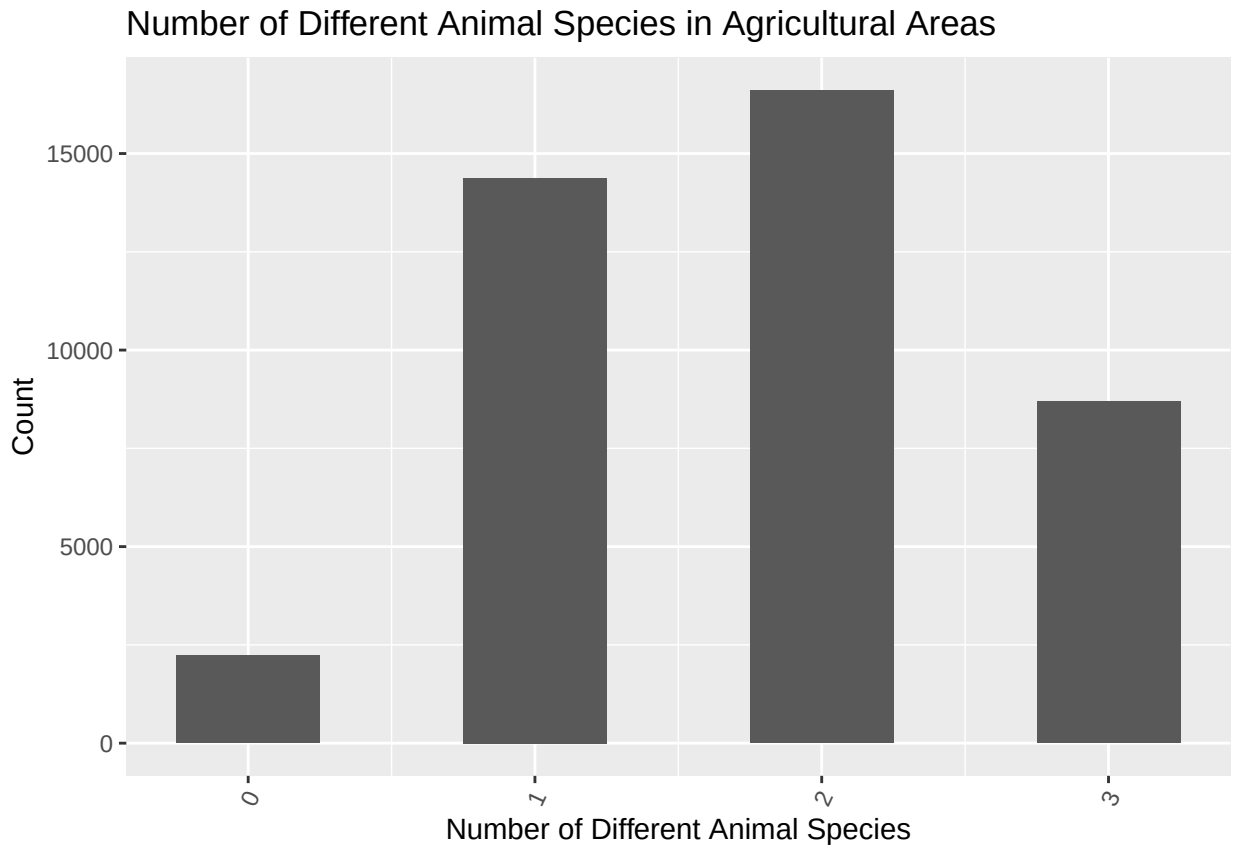

```
g <- ggplot(pfw_new, aes(total_animal_activity))
g + geom_bar(aes(fill=hab_industrial), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Number of Different Animal Species in Industrial Areas",
       x = "Number of Different Animal Species",
       y = "Count")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



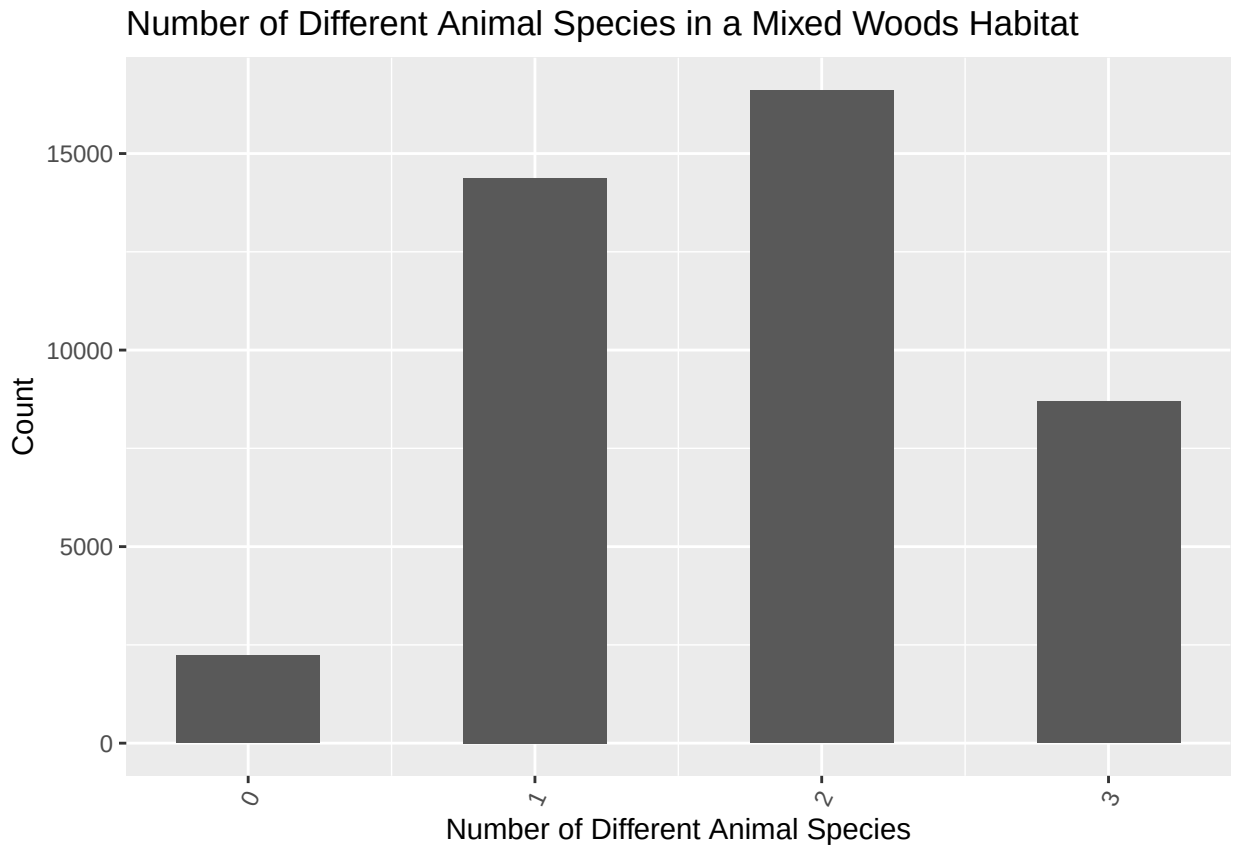

```
g <- ggplot(pfw_new, aes(total_animal_activity))
g + geom_bar(aes(fill=hab_agricultural), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Number of Different Animal Species in Agricultural Areas",
       x = "Number of Different Animal Species",
       y = "Count")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```




```
g <- ggplot(pfw_new, aes(total_animal_activity))
g + geom_bar(aes(fill=hab_mixed_woods), width = 0.5) +
  theme(axis.text.x = element_text(angle=65, vjust=0.6)) +
  labs(title="Number of Different Animal Species in a Mixed Woods Habitat",
       x = "Number of Different Animal Species",
       y = "Count")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fill
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



From the bar graph "Total Animal Activity among Different Town Sizes", we can see the trend of number of animal species active around bird feeders. There were not many locations that reported seeing zero species around bird feeders, and the trend continues upward for observing 1 and 2 species, and then decreases for 3 species. Since there were many more observations with "Village or Small Community" as their town size.

This graph may have not been the best way to see how the town sizes report different numbers of species, but I do think it shows important information about the most common numbers of species to see active around bird feeders.

The graphs examining the number of different animal species by habitat turned out all looking very similar. I tried to manipulate them a little more, but wasn't able to make them look any more distinct. I think next time I could limit the number of observations to the last 20 or so years, since this data encapsulates about 40 years.

Final Summary

I started out examining this data wondering if the number of different species active in the area of bird feeders differed between town sizes.

My findings are mostly what I expected, although I think with more training using R, I could produce more detailed summaries and graphs.

In the future, I think it would be interesting to be able to find the numbers of cats, dogs, and squirrels active in the area of bird feeders. Having a numerical value instead of a binary one would give us a lot more to work with. However, I do think this data exploration provides a good starting point.