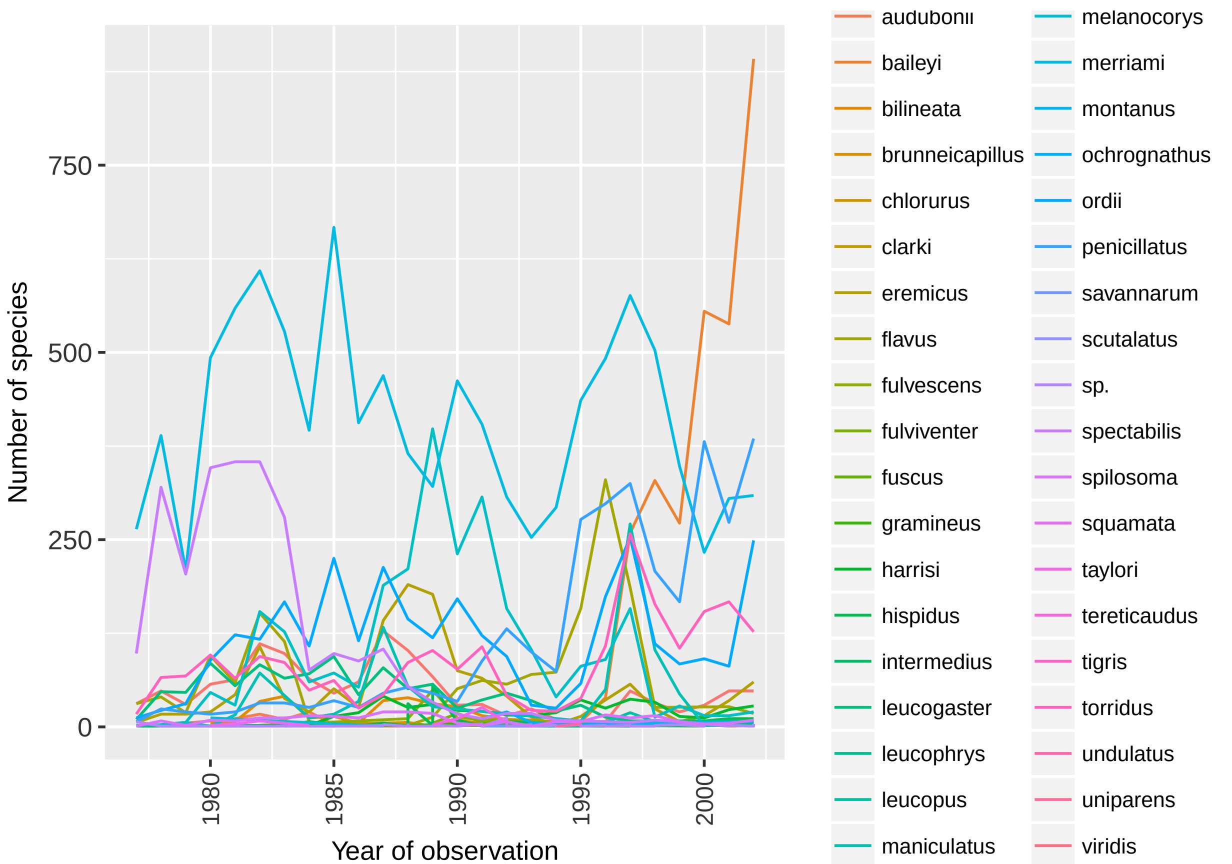
Big Data Analysis R Assignment

Kopsco and Ada

4/3/2018

The Portal, AZ rodent dataset provides information on all rodents sampled among 24 experimental plots in the Chihuahan Desert between 1977 and 2002. The following plot examines the total number of species over the entire observation time. While there is considerable change in total abundance among years, the most abundant species remained relatively stable. *C. baileyi* and *D. merriami* were the most abundant species present, however *C. baileyi* population exploded only in the late 1990s as *D. merriami*'s abundance began to decline.

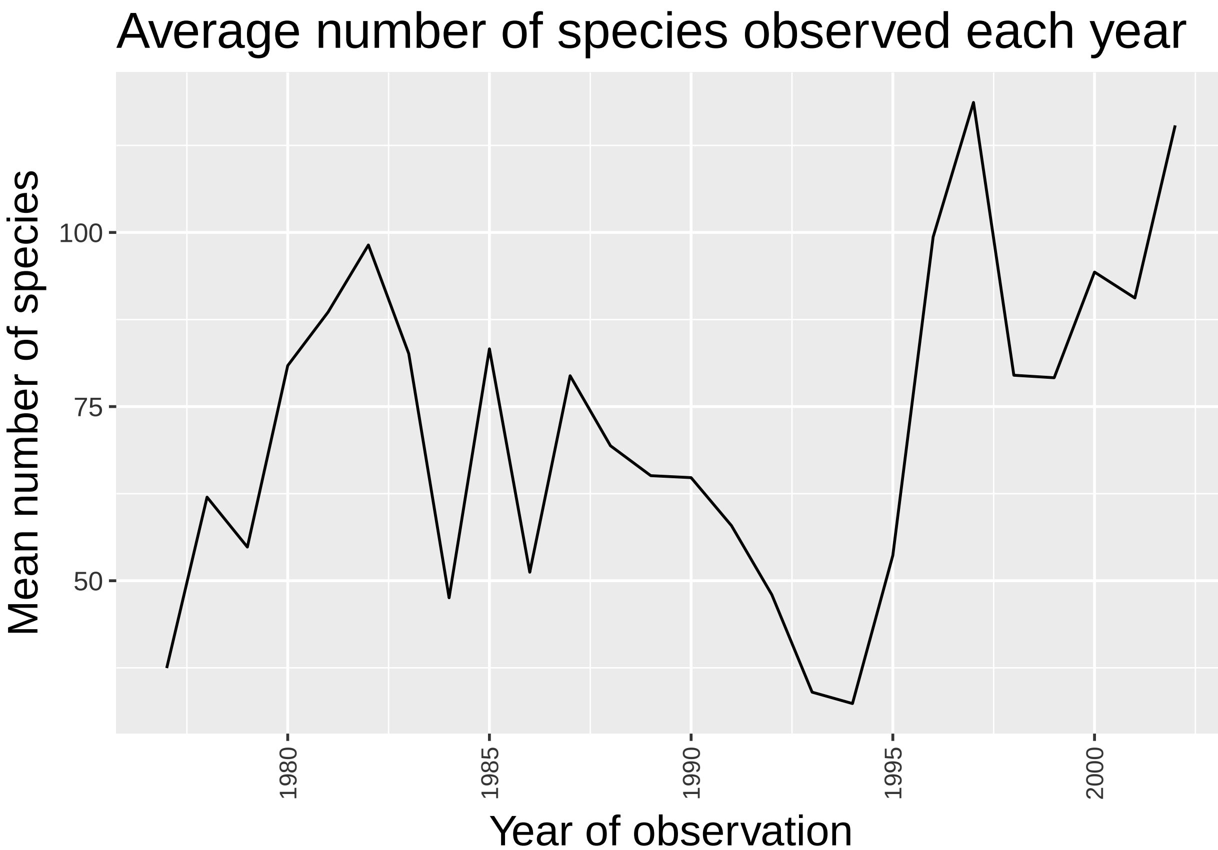
```
##       year         genus              species                n         
##  Min.   :1977   Length:509         Length:509         Min.   :  1.00  
##  1st Qu.:1984   Class :character   Class :character   1st Qu.:  3.00  
##  Median :1990   Mode  :character   Mode  :character   Median : 18.00  
##  Mean   :1990                                         Mean   : 68.34  
##  3rd Qu.:1996                                         3rd Qu.: 70.00  
##  Max.   :2002                                         Max.   :892.00  
```



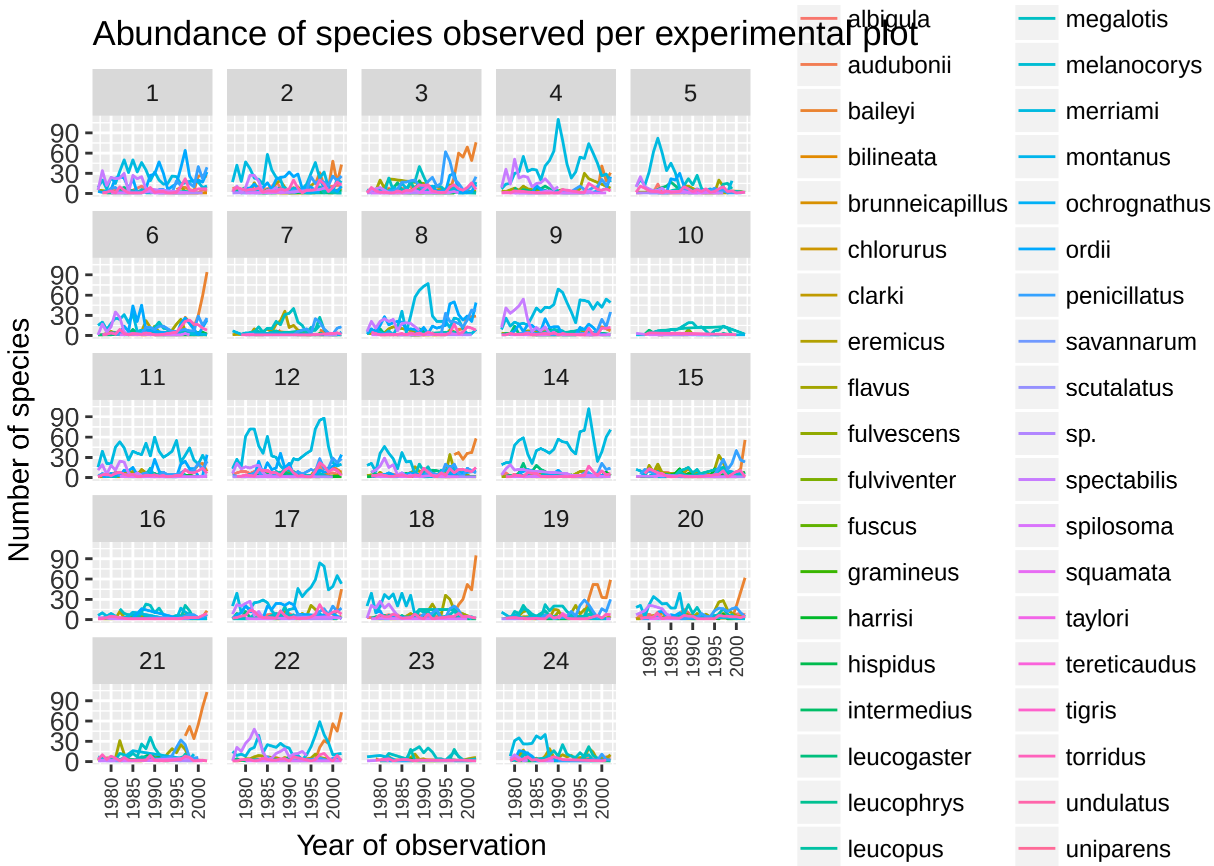
Overall, the average total number of species increased over the study period. The total average rodent population experienced a slow decline starting in the late 1980s, and then a sharp resurgence in the mid 1990s.

```
##       year      mean_abundance_yr
##  Min.   :1977   Min.   : 32.37   
```

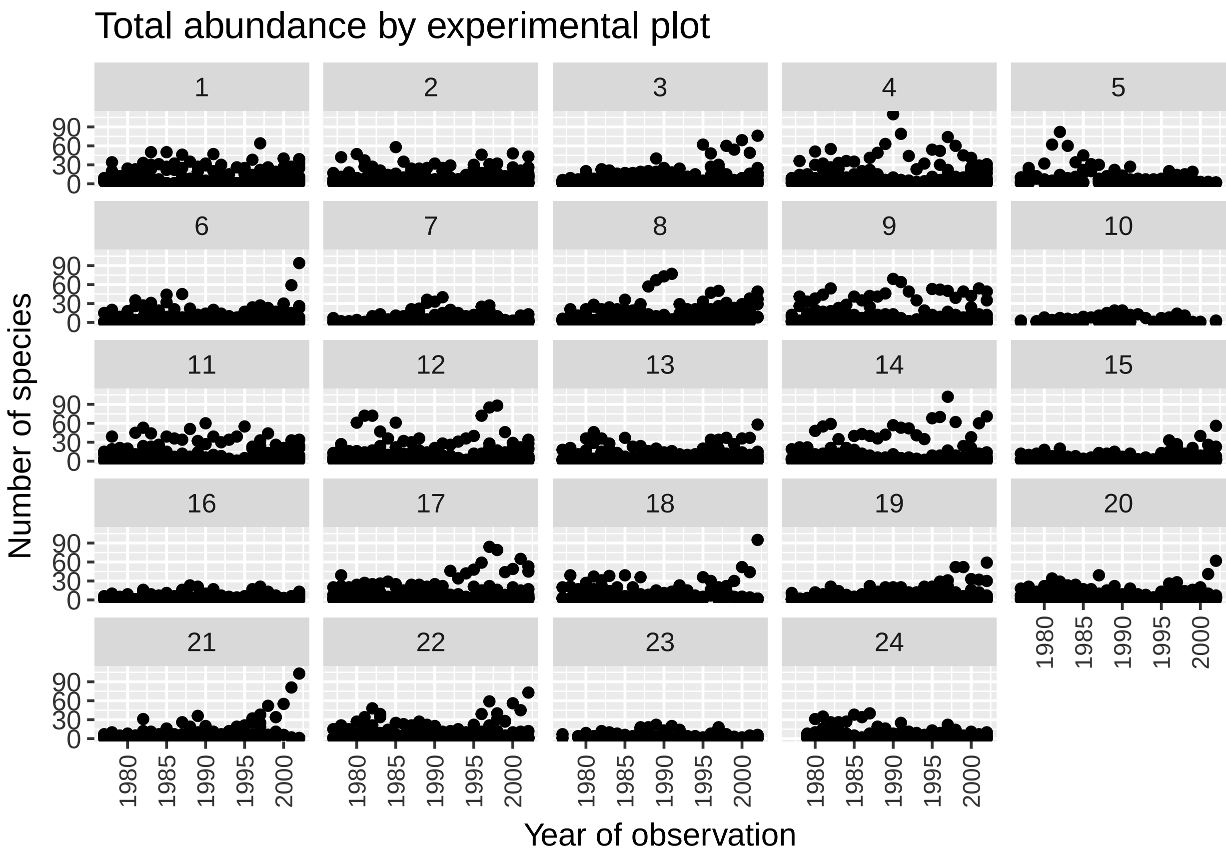
```
##  1st Qu.:1983   1st Qu.: 53.94
##  Median :1990   Median : 74.26
##  Mean   :1990   Mean   : 71.85
##  3rd Qu.:1996   3rd Qu.: 87.25
##  Max.   :2002   Max.   :118.67
```



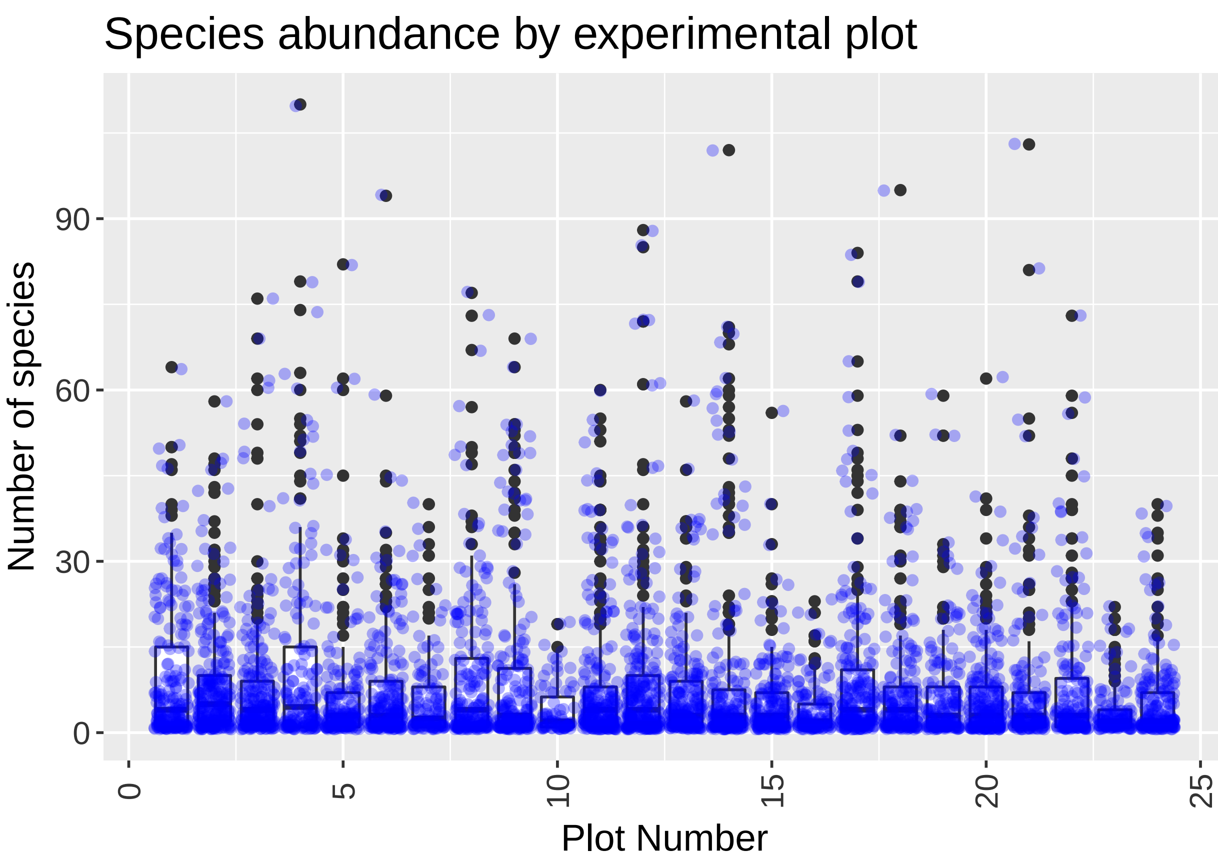
Most plots appear to contain fewer than 60 individuals of a particular species each year. However, *C. baileyi* and *D. merriami* exceed this threshold in several plots in various years. In particular, plot 4 had a surge of *D. merriami* in 1990, and *C. baileyi* increased in the late 1990s in plots 3, 6, 13, 18, 19, 20, 21, 22.



Plot 4 appears to host the largest abundance of species of all plots.

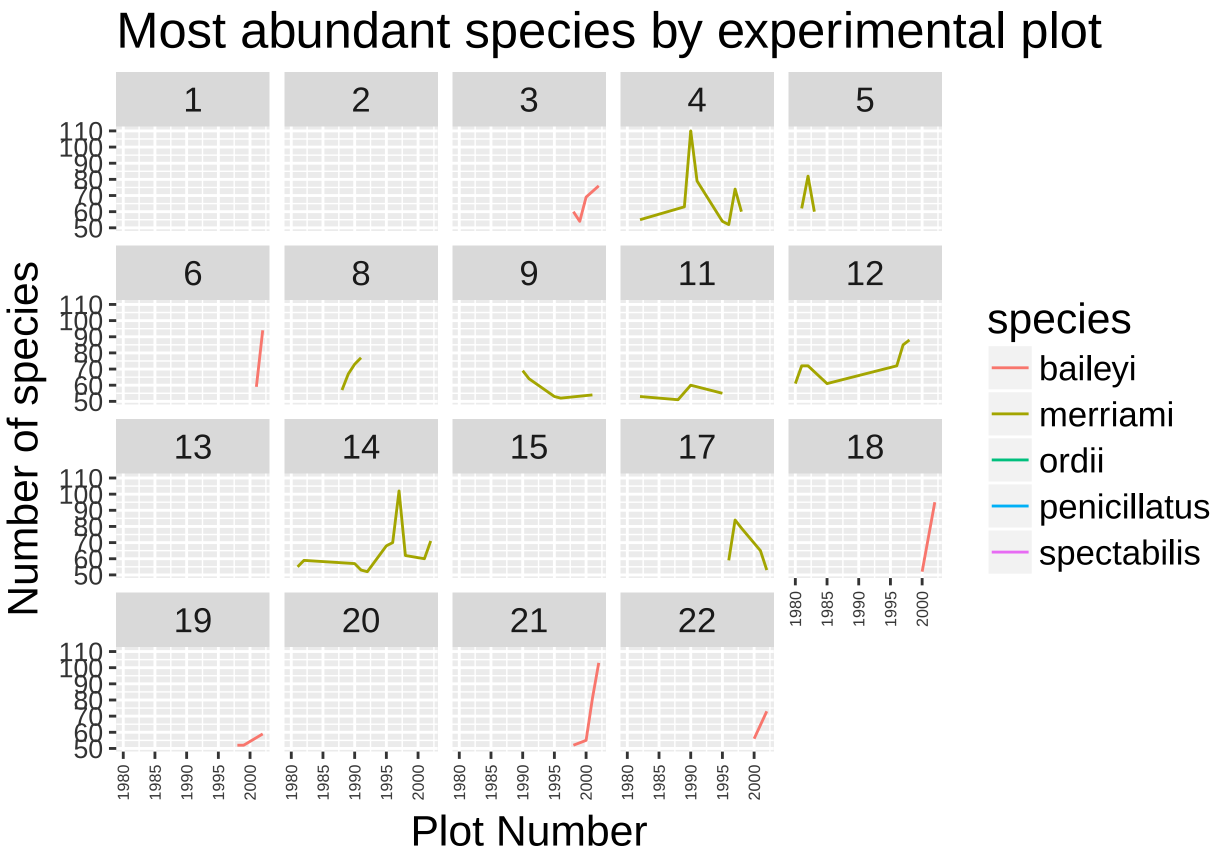


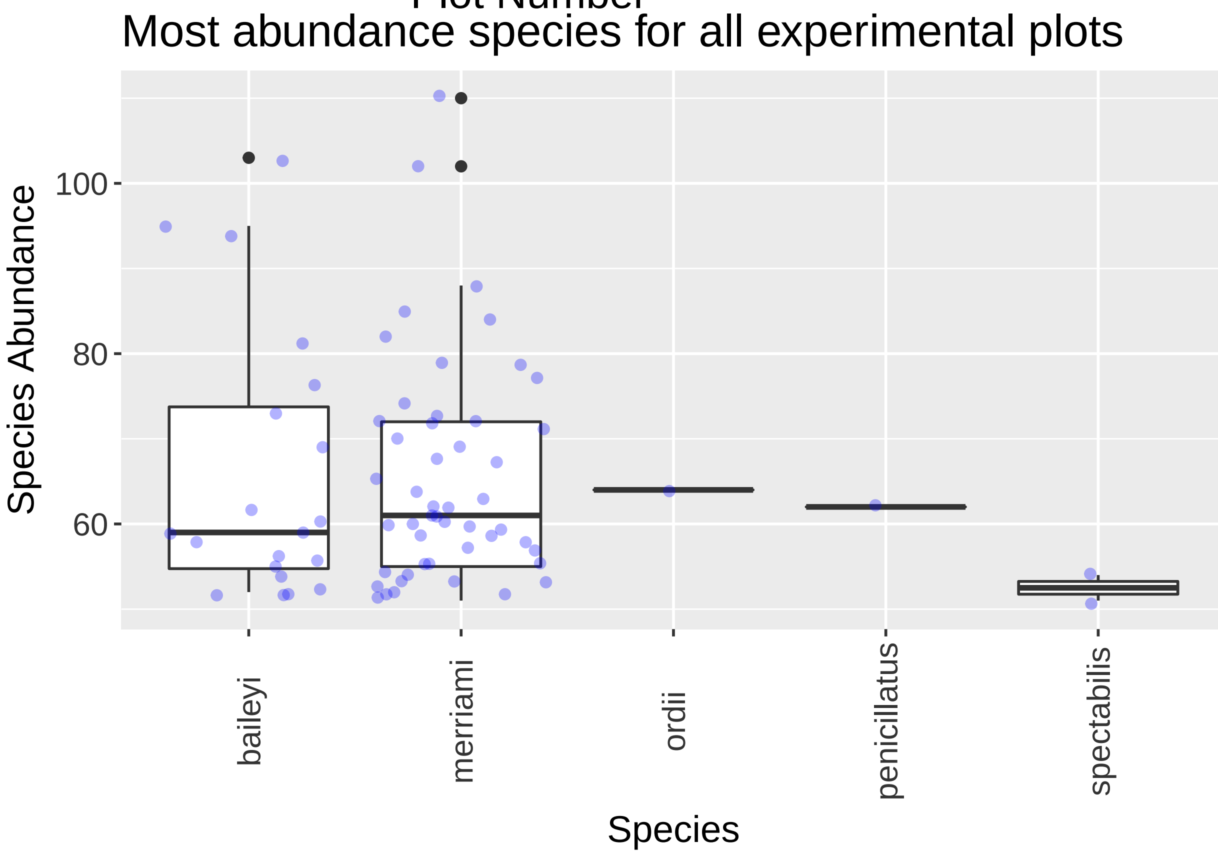
A boxplot of the mean abundance per experimental plot reveals a large number of outliers. A Kruskal-Wallis one way test of variance reveals that there is a significant difference among the mean species abundances per plot_id (p = 5.559e-14).




```
##
##  Kruskal-Wallis rank sum test
##
## data:  species by plot_id
## Kruskal-Wallis chi-squared = 113.71, df = 23, p-value = 5.559e-14
```

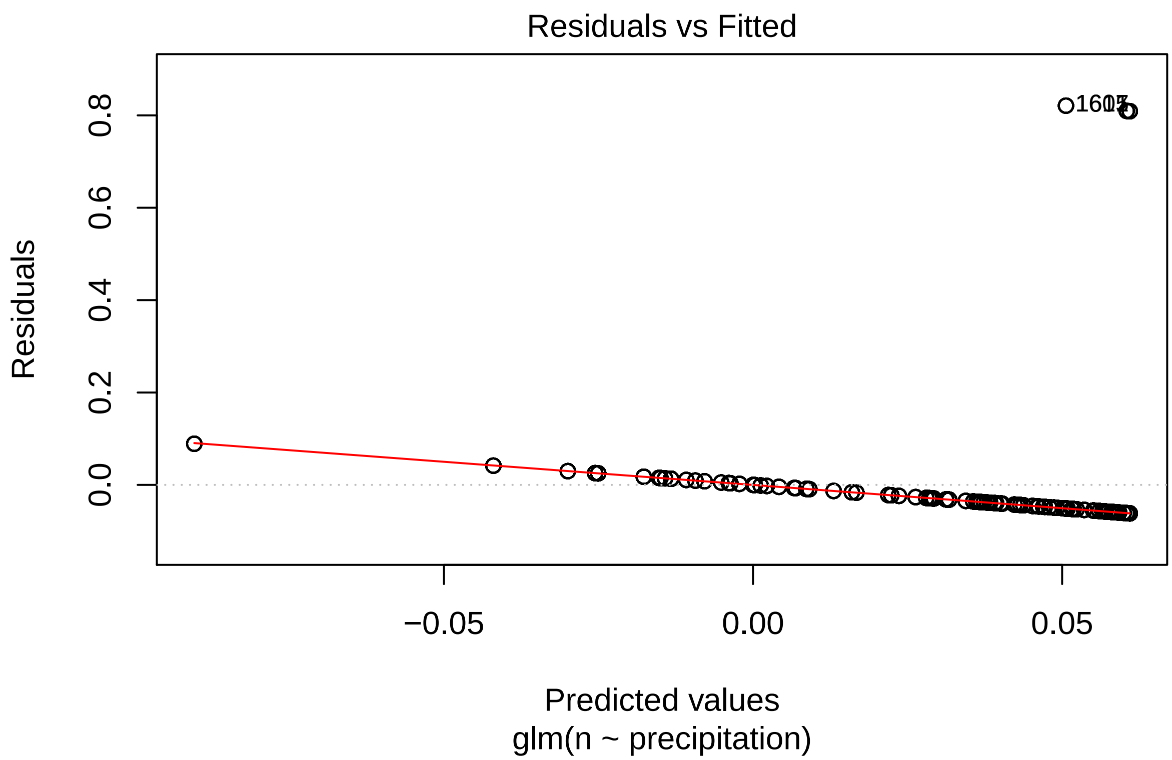
C. baileyi and *D. merriami* are the most abundant species for all experimental plots, but there is a lot of variance in their sample numbers throughout the years.

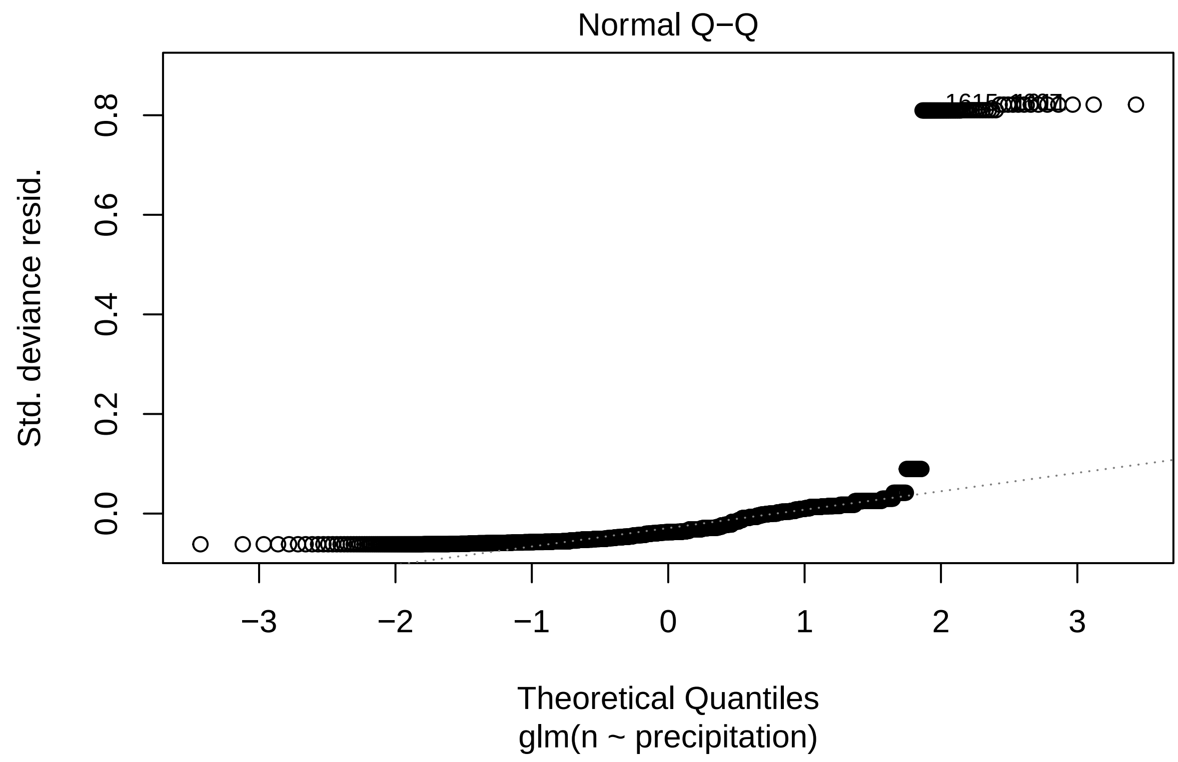


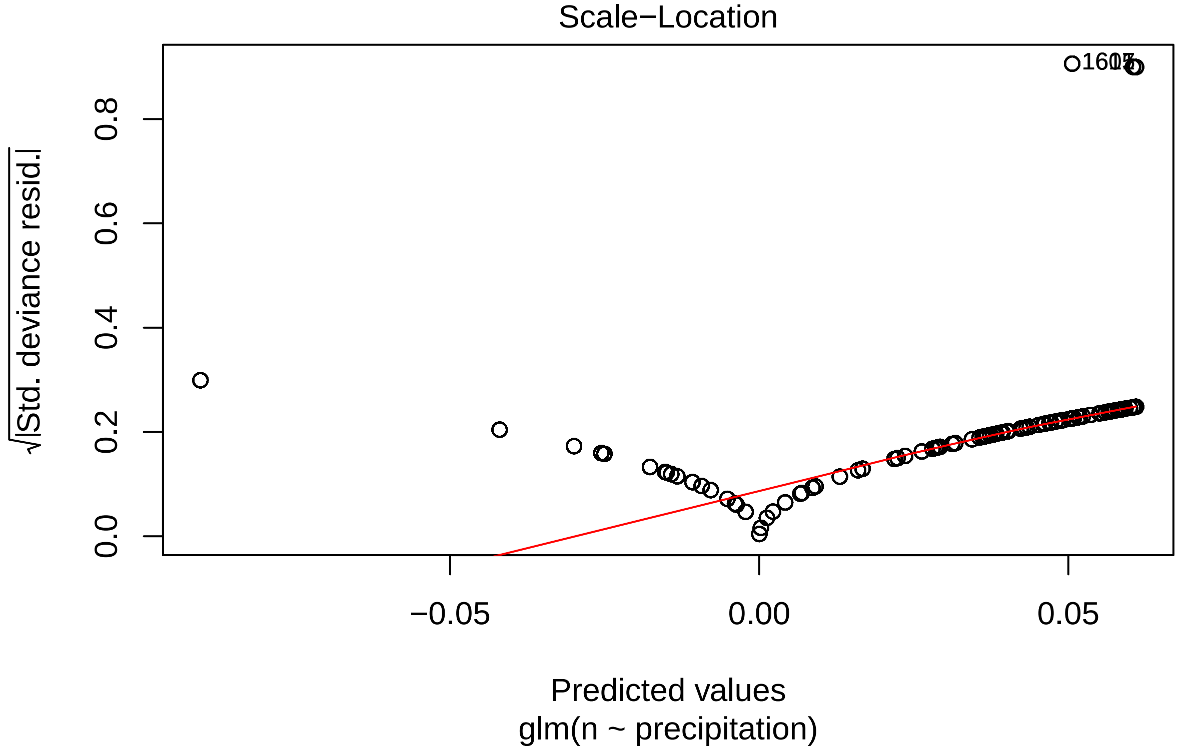


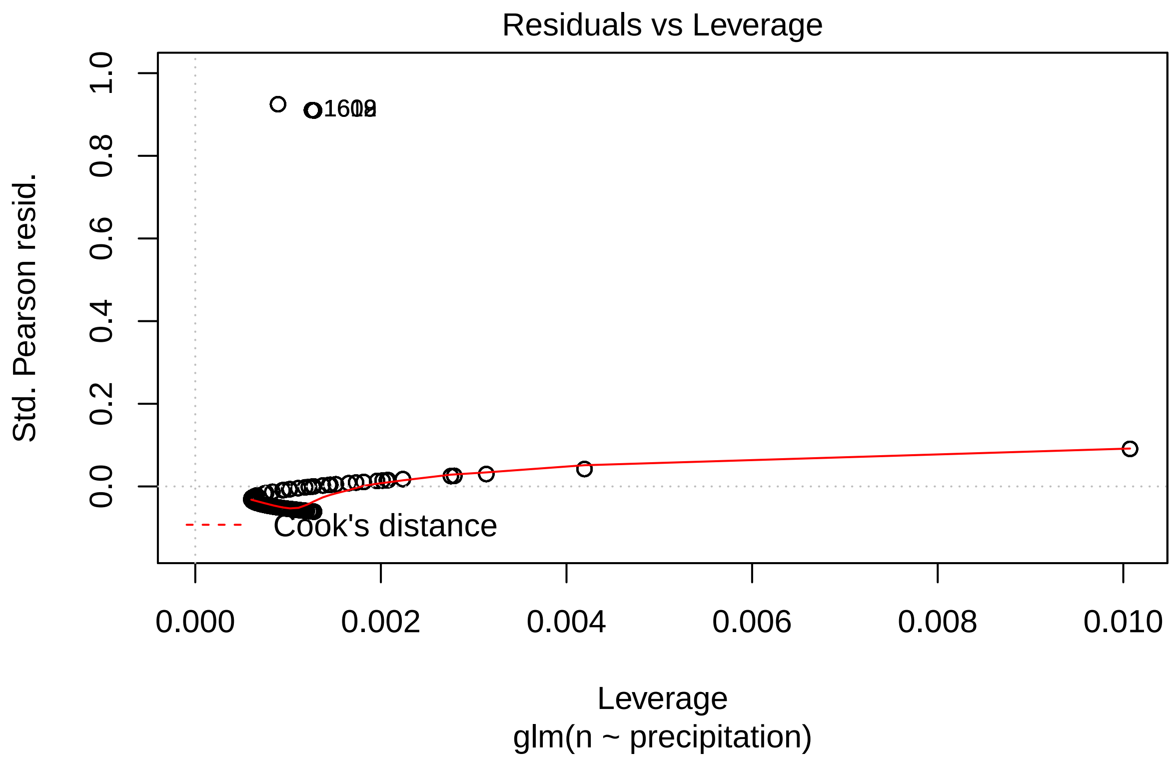
There does not appear to be a significant relationship between the annual species abundance and the annual precipitation (p = 0.234).

```
##
## Call:
## glm(formula = n ~ precipitation, family = "poisson", data = precip_survey)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.06157  -0.05406  -0.03710  -0.00420   0.82094
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.0609431  0.0346938   1.757    0.079 .
## precipitation -0.0009672  0.0008120  -1.191    0.234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 38.560  on 1656  degrees of freedom
## Residual deviance: 37.124  on 1655  degrees of freedom
## AIC: 3387
##
## Number of Fisher Scoring iterations: 4
```

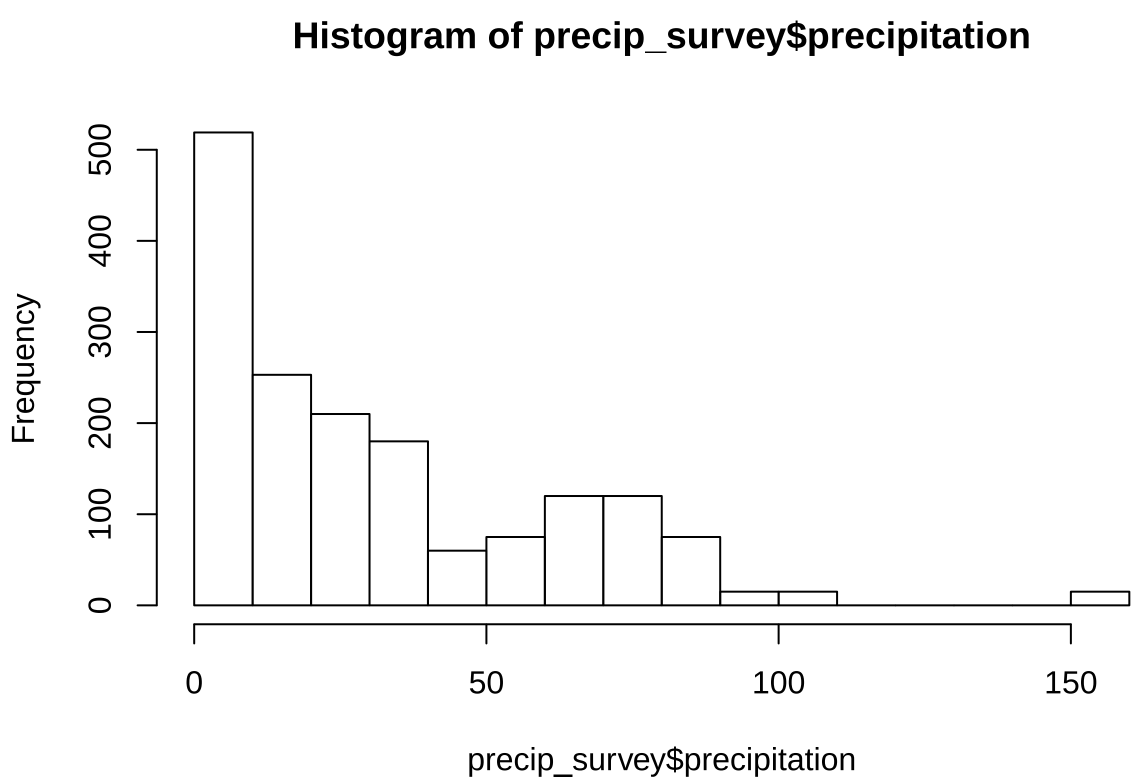



Normal Q−Q
Std. deviance resid.
0.0
0.2
0.4
0.6
0.8
−3
−2
−1
0
1
2
3
Theoretical Quantiles
glm(n ~ precipitation)


Scale−Location
√|Std. deviance resid.|
0.0
0.2
0.4
0.6
0.8
−0.05
0.00
0.05
Predicted values
glm(n ~ precipitation)



Likewise, there is no significant difference among sampling years and the amount of precipitation (p = 1).

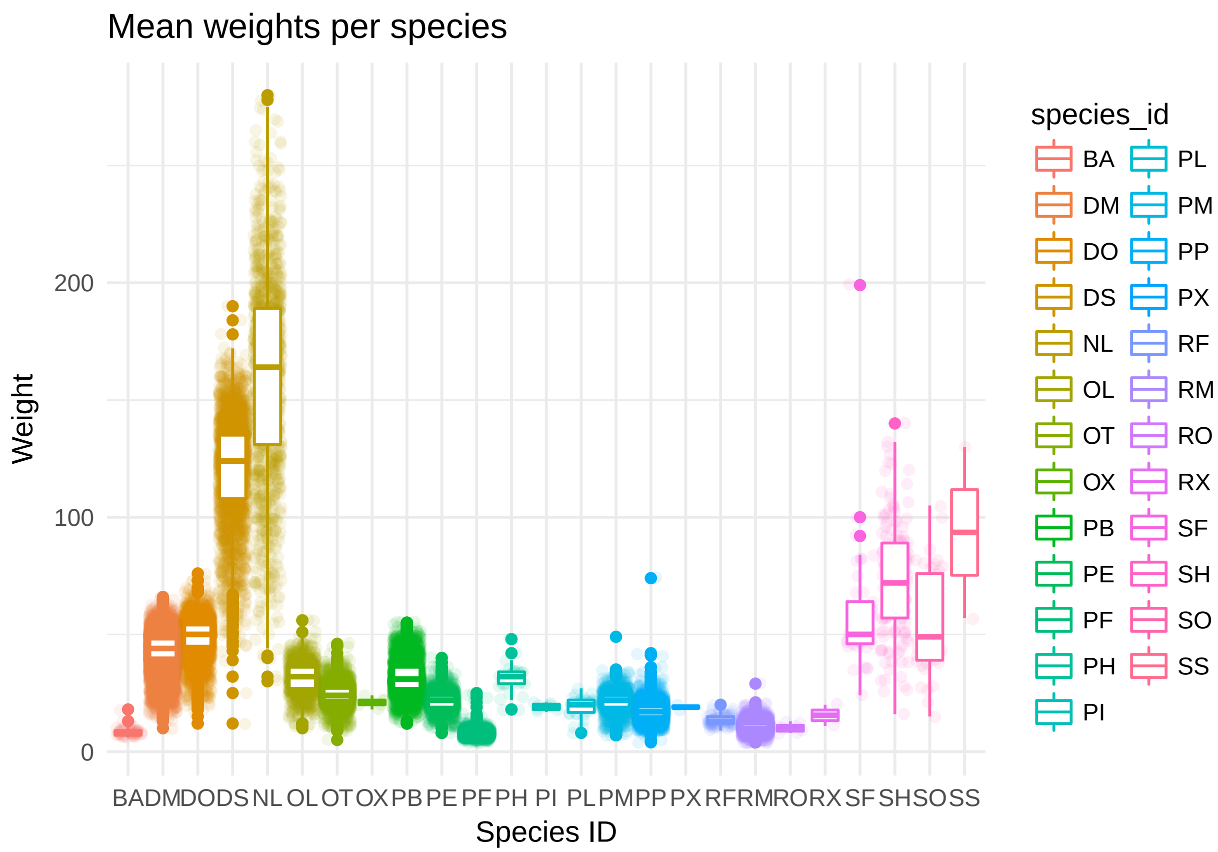



```
##       Year       precipitation          n
##  Min.   :1980   Min.   :  0.00   Min.   :1.000
##  1st Qu.:1982   1st Qu.:  6.10   1st Qu.:1.000
##  Median :1984   Median : 22.61   Median :1.000
```

```
## Mean   :1984   Mean   : 31.36   Mean   :1.036
## 3rd Qu.:1987   3rd Qu.: 53.84   3rd Qu.:1.000
## Max.   :1989   Max.   :156.46   Max.   :2.000
```

```
##
## Kruskal-Wallis rank sum test
##
## data:  precipitation by year
## Kruskal-Wallis chi-squared = 2.7048, df = 7, p-value = 0.9109
```

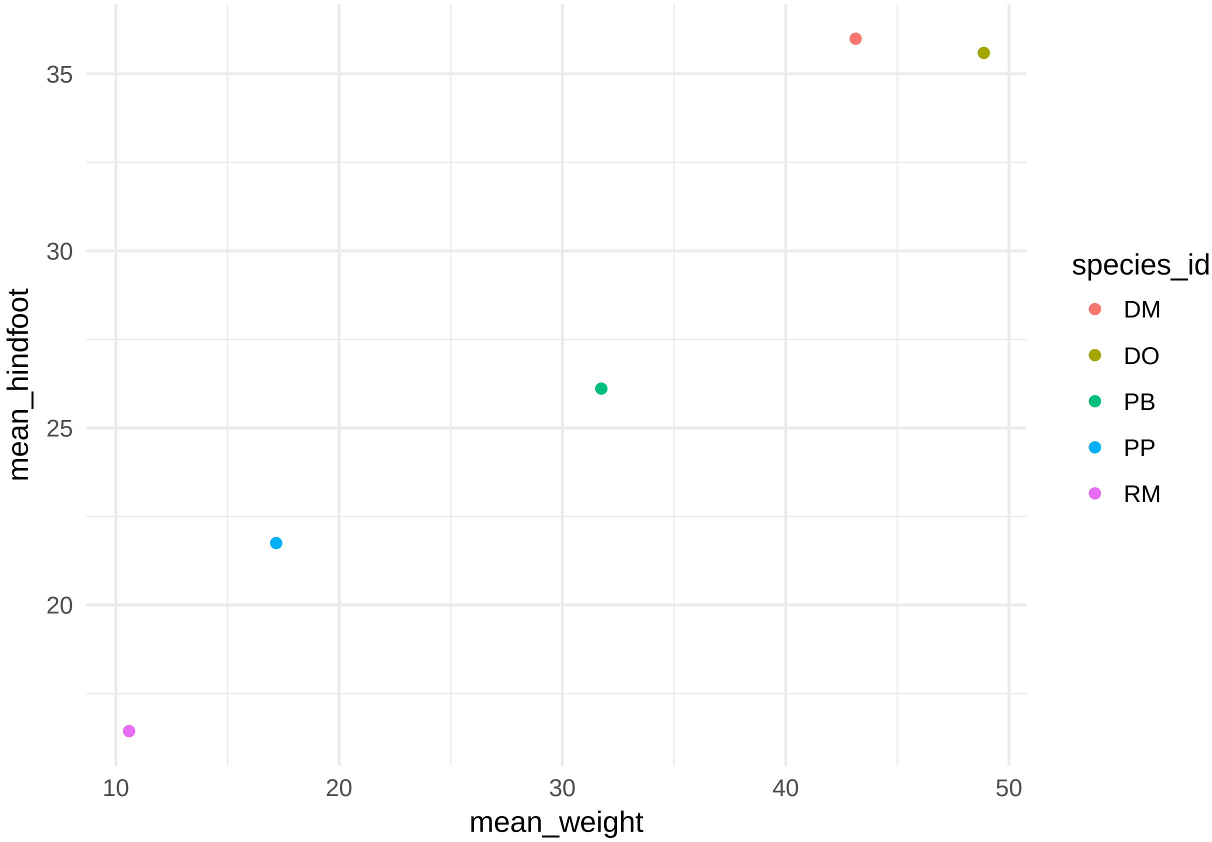
Avarage weight per species:



Summary the mean weights per species:

```
## # A tibble: 25 x 2
##    species_id mean_weight
##    <chr>            <dbl>
##  1 BA                8.60
##  2 DM               43.2
##  3 DO               48.9
##  4 DS              120.
##  5 NL              159.
##  6 OL               31.6
##  7 OT               24.2
##  8 OX               21.0
##  9 PB               31.7
## 10 PE               21.6
## # ... with 15 more rows
```

We choose top five most abundant species and compare mean hindfoot length and weight.



Summary of the relatinship between species' mean weight between mean hindfoot length:

```
##
## Call:
## glm(formula = mean_weight ~ mean_hindfoot, data = means_for_top_five)
##
## Deviance Residuals:
##      1       2       3       4       5
## -3.660   2.828   3.432  -2.971   0.372
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -20.5366     6.1693  -3.329  0.04476 *
## mean_hindfoot   1.8707     0.2185   8.563  0.00335 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 14.04789)
##
##     Null deviance: 1072.146  on 4  degrees of freedom
## Residual deviance:   42.144  on 3  degrees of freedom
## AIC: 30.848
##
## Number of Fisher Scoring iterations: 2

##
```

```
## Call:
## glm(formula = mean_weight ~ mean_hindfoot, data = means_for_top_five)
##
## Deviance Residuals:
##      1       2       3       4       5
## -3.660   2.828   3.432  -2.971   0.372
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -20.5366     6.1693  -3.329  0.04476 *
## mean_hindfoot   1.8707     0.2185   8.563  0.00335 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 14.04789)
##
##     Null deviance: 1072.146  on 4  degrees of freedom
## Residual deviance:   42.144  on 3  degrees of freedom
## AIC: 30.848
##
## Number of Fisher Scoring iterations: 2
```

There is not significant relationship between species' mean weight between mean hindfoot length ($p = 0.00335$).