```r
library(tidyverse)
```

```r
birds <- as_tibble(birdabundance)
#write_csv(birds, file = "./data/birdabundance.csv")
```

Look at the structure of the data

What are the variables? What are the values?

First look at the summary and structure of the data

- only numeric columns
- response column is ABUND
- Summary shows not missing values

```r
str(birds)
```

```
## tibble [56 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Site   : int [1:56] 1 2 3 4 5 6 7 8 9 10 ...
##  $ ABUND  : num [1:56] 5.3 2 1.5 17.1 13.8 14.1 3.8 2.2 3.3 3 ...
##  $ AREA   : num [1:56] 0.1 0.5 0.5 1 1 1 1 1 1 1 ...
##  $ DIST   : int [1:56] 39 234 104 66 246 234 467 284 156 311 ...
##  $ LDIST  : int [1:56] 39 234 311 66 246 285 467 1829 156 571 ...
##  $ YR.ISOL: int [1:56] 1968 1920 1900 1966 1918 1965 1955 1920 1965 1900 ...
##  $ GRAZE  : int [1:56] 2 5 5 3 5 3 5 5 4 5 ...
##  $ ALT    : int [1:56] 160 60 140 160 140 130 90 60 130 130 ...
```

```r
summary(birds)
```

```
##       Site           ABUND            AREA              DIST
##  Min.   : 1.00   Min.   : 1.50   Min.   :   0.10   Min.   :  26.0
##  1st Qu.:14.75   1st Qu.:12.40   1st Qu.:   2.00   1st Qu.:  93.0
##  Median :28.50   Median :21.05   Median :   7.50   Median : 234.0
##  Mean   :28.50   Mean   :19.51   Mean   :  69.27   Mean   : 240.4
##  3rd Qu.:42.25   3rd Qu.:28.30   3rd Qu.:  29.75   3rd Qu.: 333.2
##  Max.   :56.00   Max.   :39.60   Max.   :1771.00   Max.   :1427.0
##      LDIST           YR.ISOL         GRAZE            ALT
##  Min.   :  26.0   Min.   :1890   Min.   :1.000   Min.   : 60.0
##  1st Qu.: 158.2   1st Qu.:1928   1st Qu.:2.000   1st Qu.:120.0
##  Median : 338.5   Median :1962   Median :3.000   Median :140.0
##  Mean   : 733.3   Mean   :1950   Mean   :2.982   Mean   :146.2
##  3rd Qu.: 913.8   3rd Qu.:1966   3rd Qu.:4.000   3rd Qu.:182.5
##  Max.   :4426.0   Max.   :1976   Max.   :5.000   Max.   :260.0
```

Step 0: organize data

First, I want to rename some of the columns to make the data set easier to work with:

- Change all column headers to lower case
- rename column `yr.isol` to `isol_since`
- add a new column with years since isolation
- change `graze` variable to a factor

```r
birds <- birds %>%
  rename_with(tolower, everything()) %>%
  rename(isol_since = yr.isol) %>%
```

```
  mutate(isol_years = 2021 - isol_since) %>%
  mutate(graze = factor(graze))
```

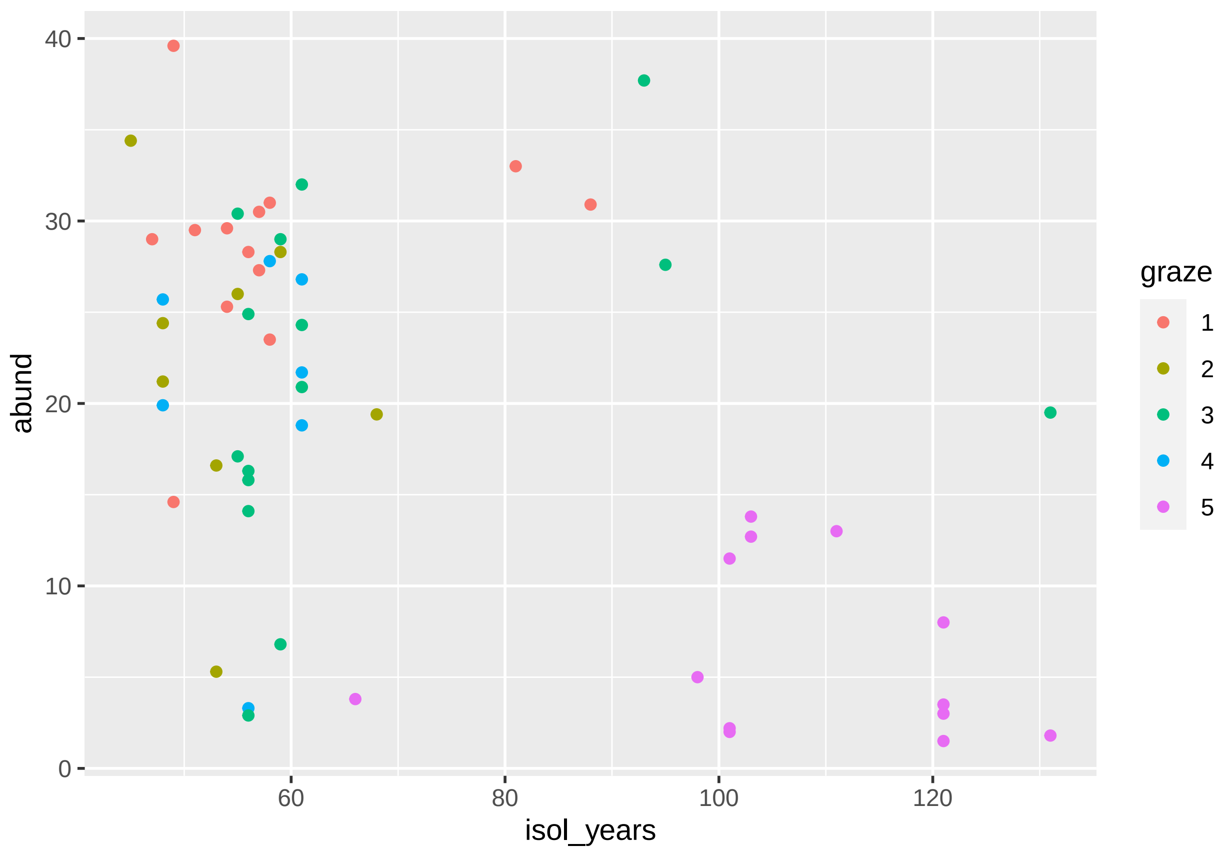
Step 1: Exploratory plotting

I will start with some exploratory data analysis using `ggplot2`.

Question: Which factors influence bird abundance most?

Isolation time and grazing Looks like: - the longer a site is isolated the higher the grazing intensity is - the longer a site is isolated the lower the abundance - grazing intensity does not affect abundance but years of isolation do

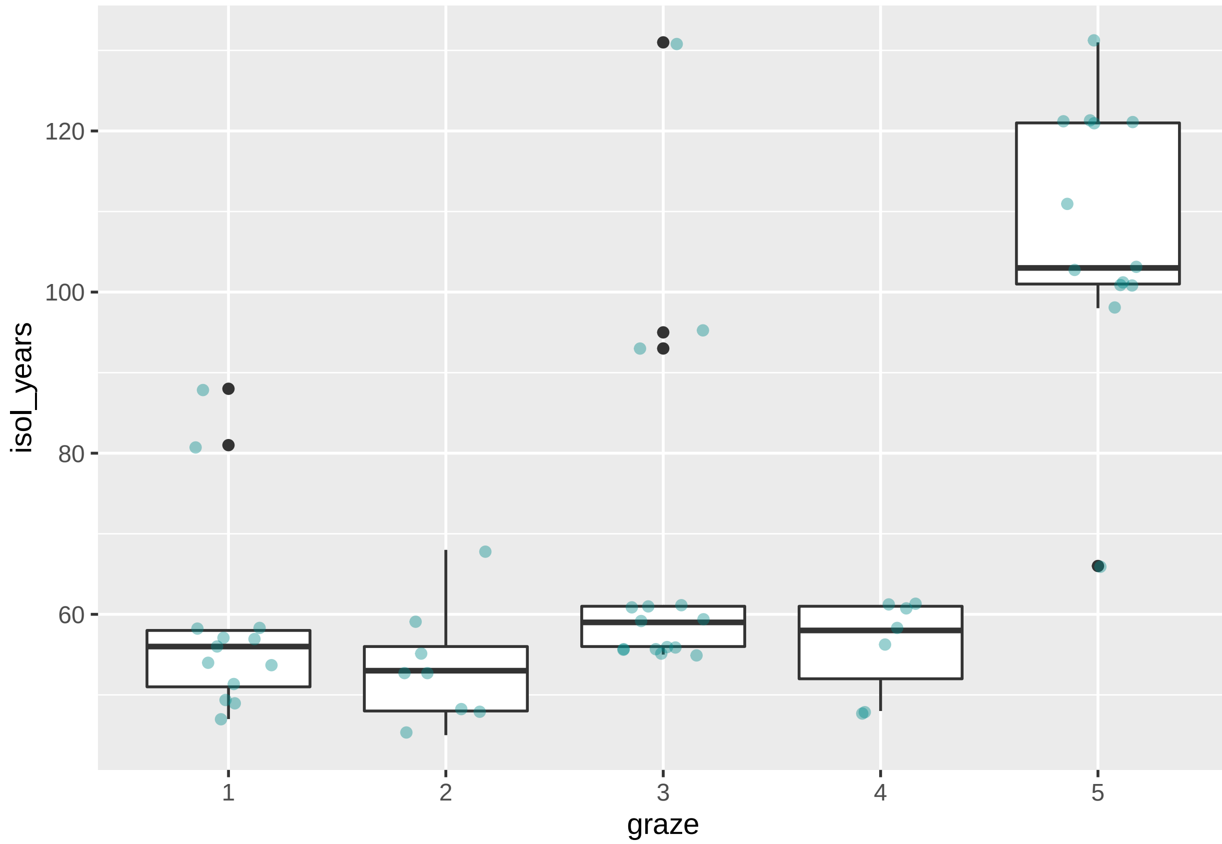
```
ggplot(birds, aes(x = isol_years, y = abund, color = graze)) +
  geom_point()
```



I there an interaction between grazing intensity and time since isolation?

Does not look like it. But the sites with a very high grazing intensity seem to be isolated since a very long time.

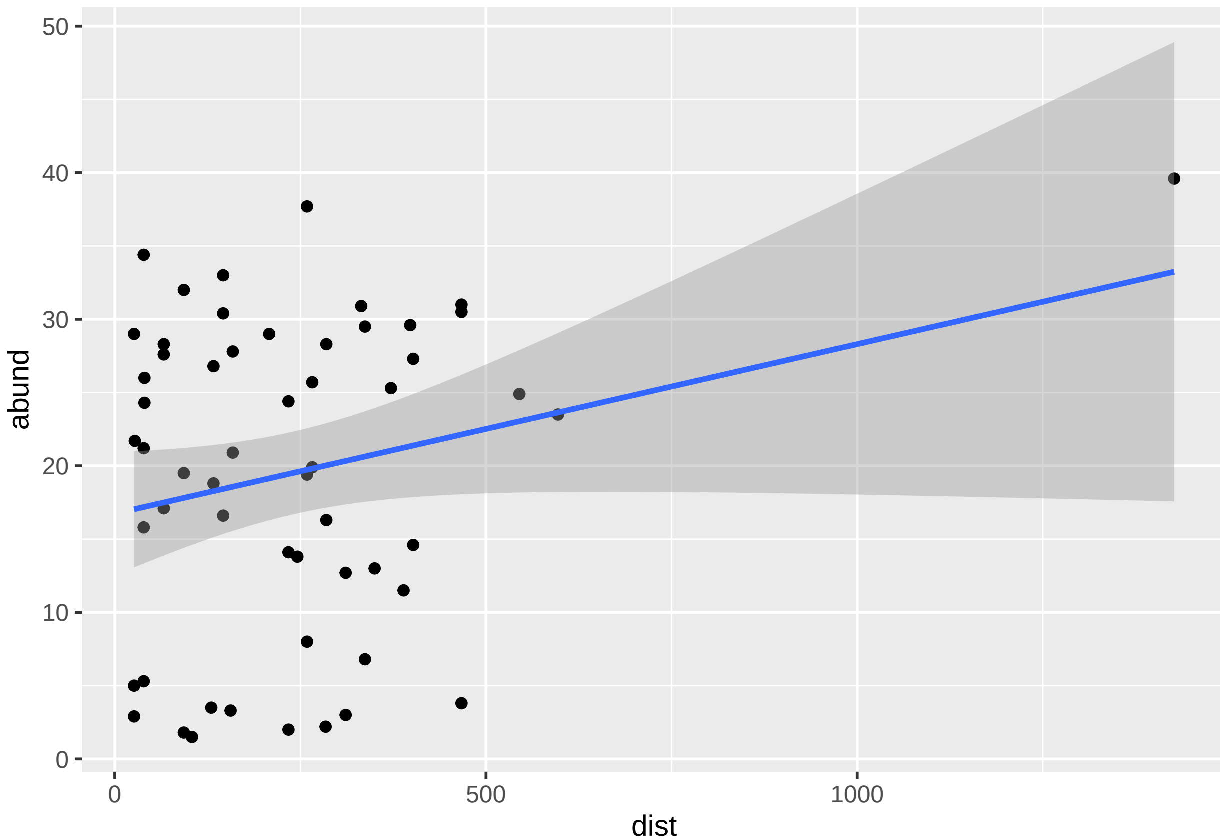
```
ggplot(birds, aes(x = graze, y = isol_years)) +
  geom_boxplot() +
  geom_point(position = position_jitter(seed = 123, width = 0.2), alpha = 0.4, color = "cyan4")
```



Distance to forest Does the bird abundance depend on the distance to the nearest forest patch?

Does not seem like it. There is no clear pattern in showing that the distance to the nearest forest affects bird abundance. However, the distances are all quite small considering the radius of bird movements

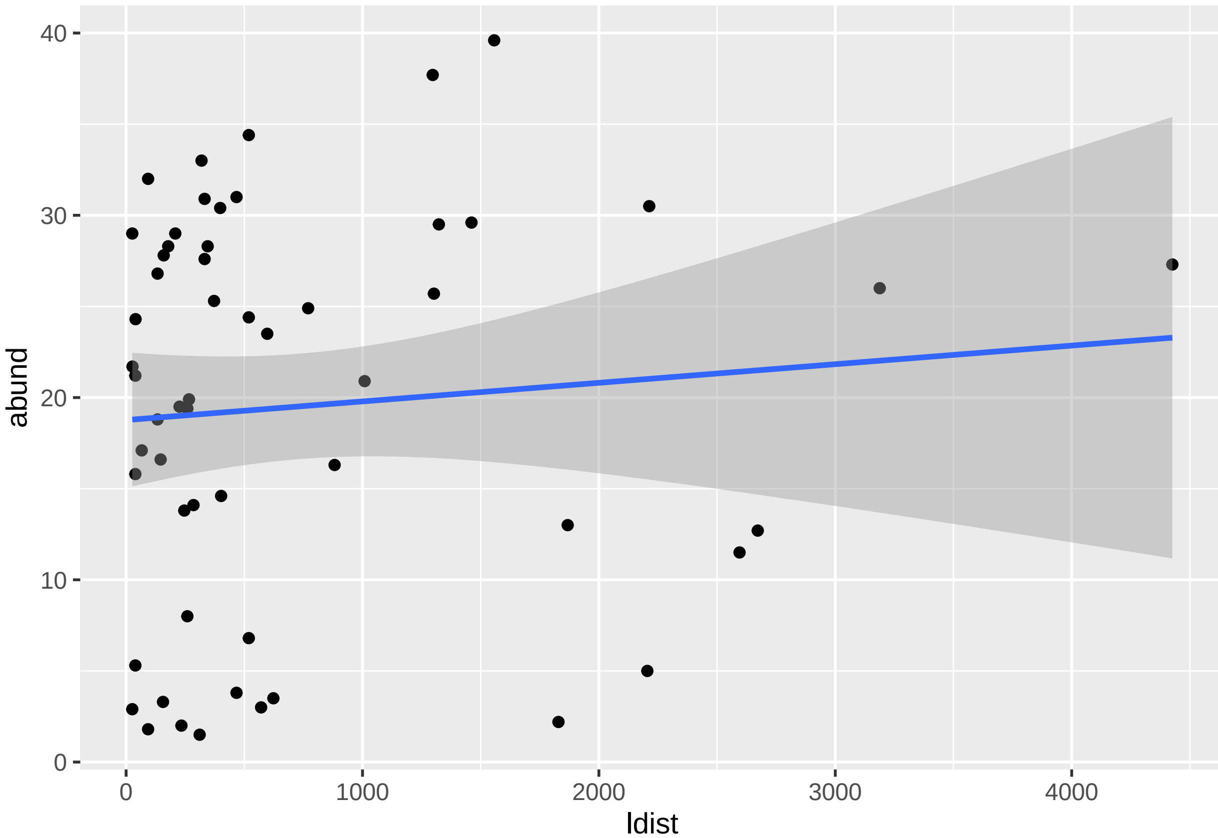
```
ggplot(birds, aes(x = dist, y = abund)) +
  geom_point()+
  geom_smooth(method="lm")
```



How about the distance to the nearest large forest patch?

Also here, there does not seem to be a clear pattern

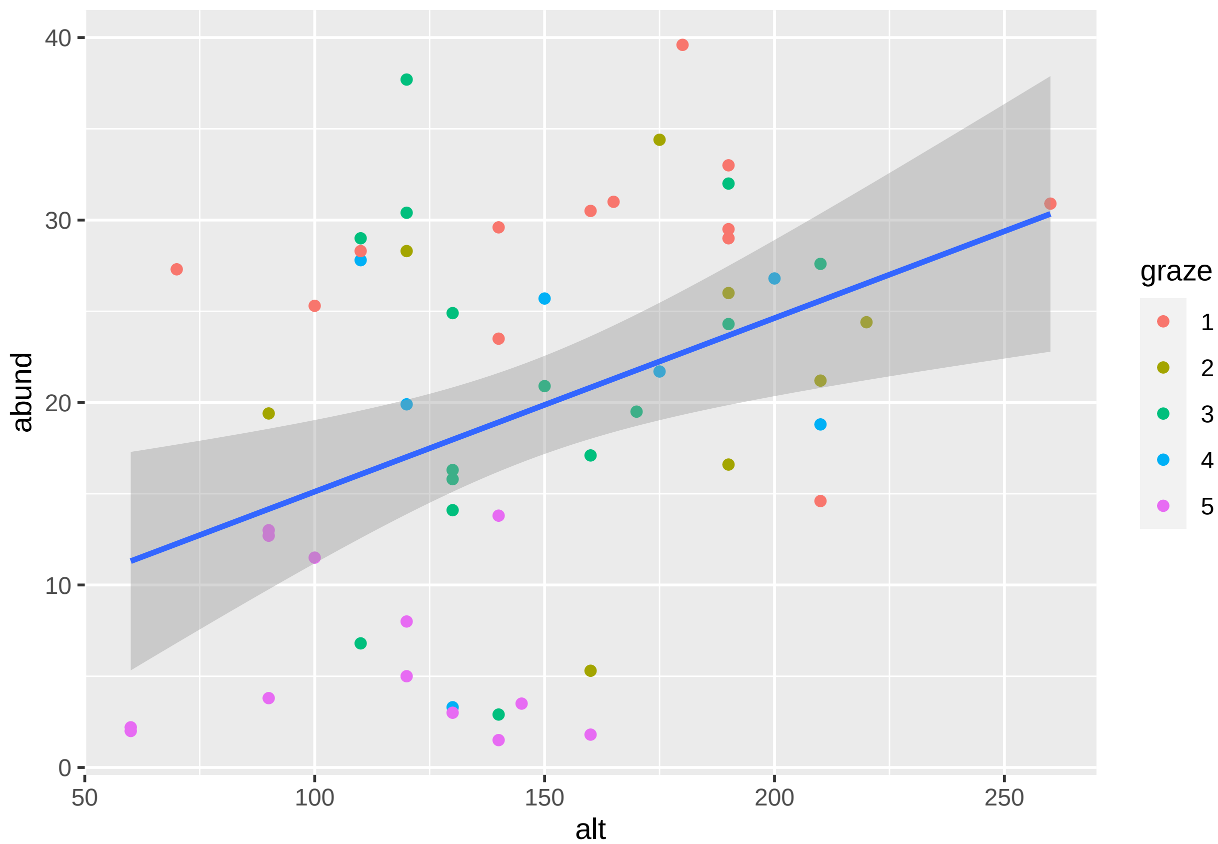
```
ggplot(birds, aes(x=ldist, y=abund))+
  geom_point()+
  geom_smooth(method = "lm")
```



Altitude Is there an effect of altitude?

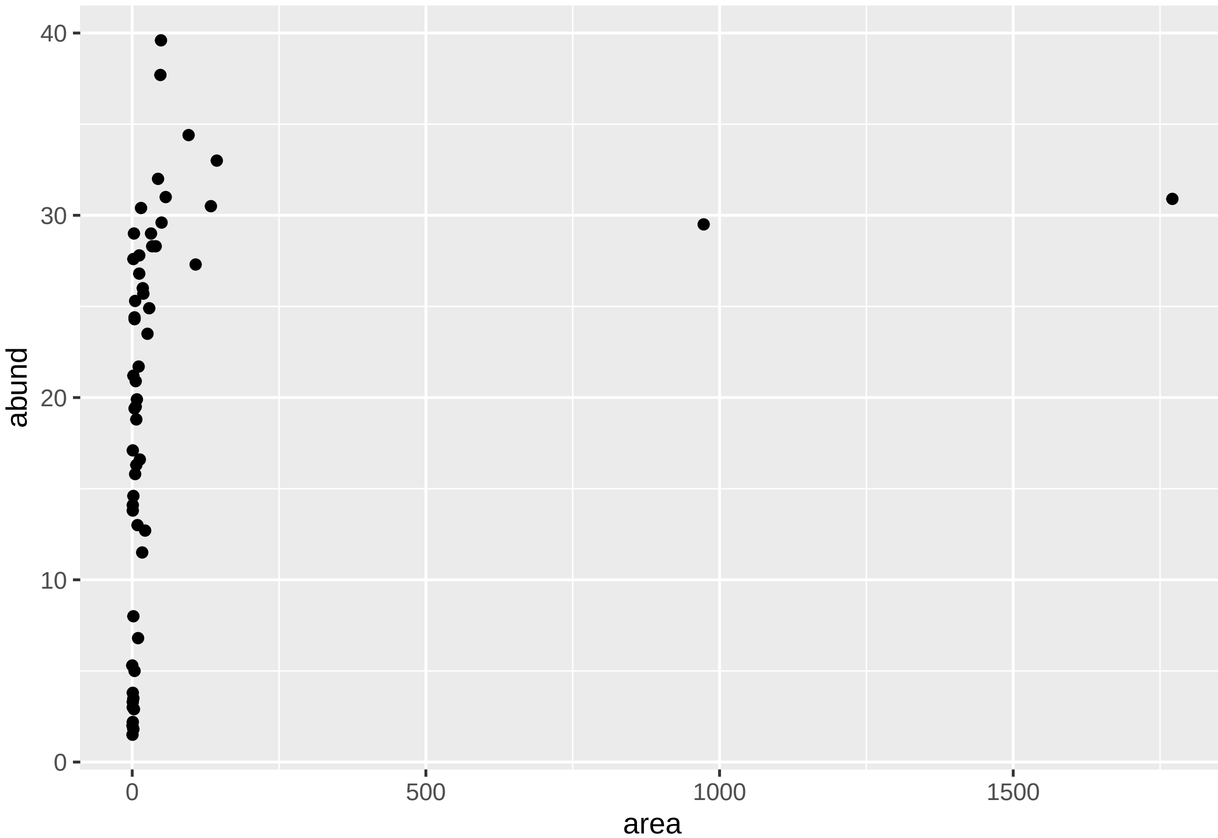
It seems like the higher the altitude, the higher the bird abundance.

```
ggplot(birds, aes(x = alt, y = abund)) +
  geom_point(aes(color = graze)) +
  geom_smooth(method = "lm")
```



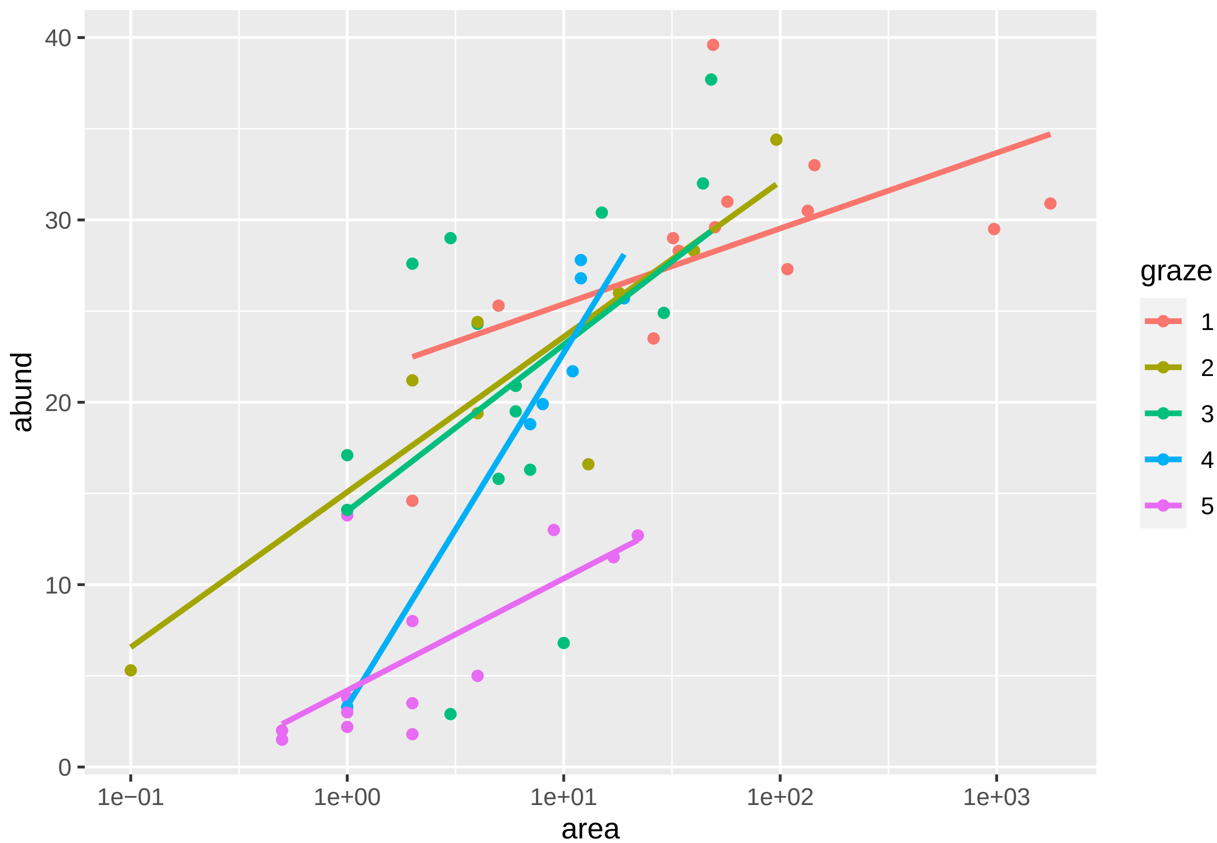
Area Is there an effect of the area of the nearest fragment?

```
ggplot(birds, aes(x=area, y = abund))+
  geom_point()
```



Looks like a log transformation of area could help

```
ggplot(birds, aes(x = area, y = abund, color = graze)) +
  geom_point() +
  scale_x_log10() +
  geom_smooth(method = "lm", se = FALSE)
```



Looks like there is a clear relationship here.

Step 2: Some statistical tests and models

First test a model with interaction between grazing intensity and area of the forest fragement.

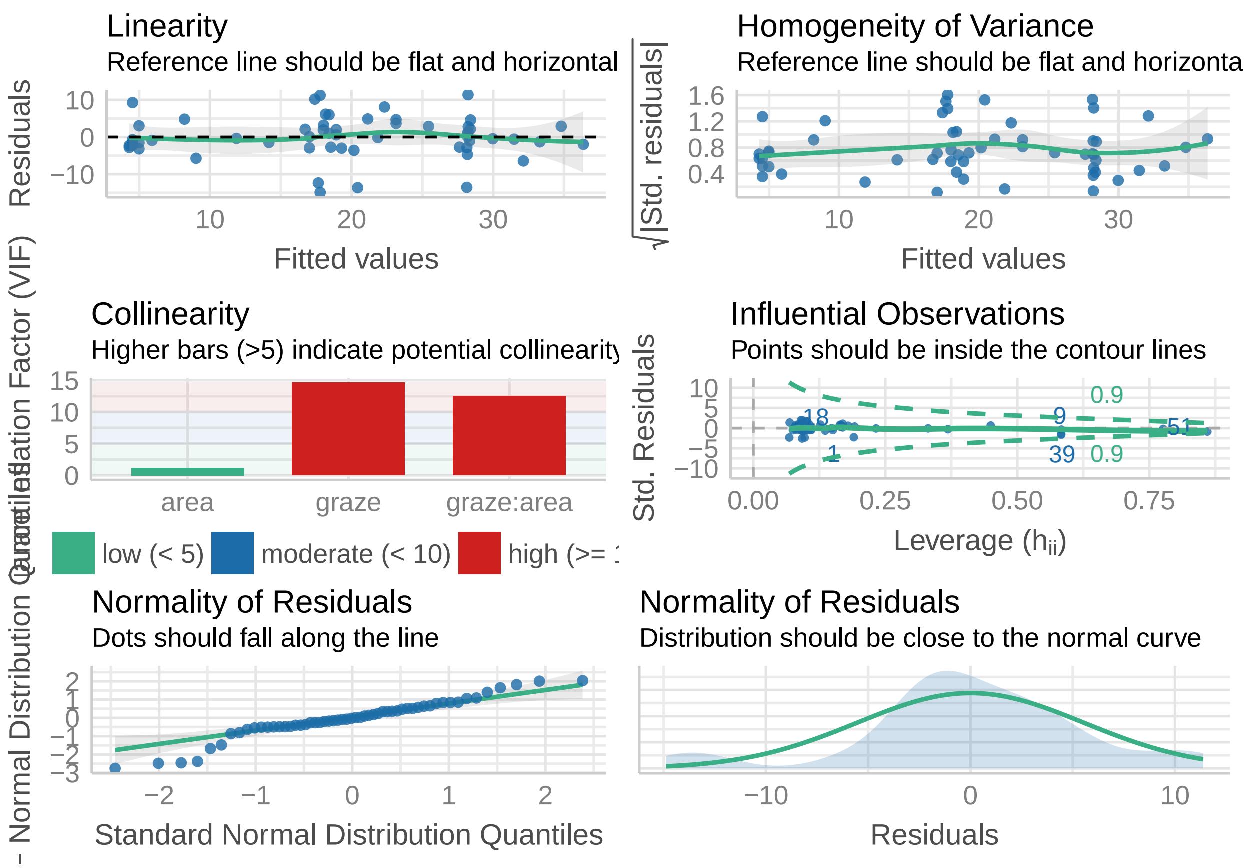
```
lm1 <- lm(abund ~ graze * area, data = birds)
drop1(lm1, test = "F")
```

```
## Single term deletions
##
## Model:
## abund ~ graze * area
##            Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                  1686.6 210.69
## graze:area  4    1170.7 2857.3 232.21  7.9825 5.653e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is a significant interaction between the grazing intensity and the area of the remaining forest fragment.

But are the assumptions of a linear model fulfilled?

```
performance::check_model(lm1)
```



It looks like the normality of residuals is not fulfilled. Looking back at the plot from before, it might make sense to test a model with log-transformed area and a sqrt transfomed abundance.

(Sometimes the square root transformation can help with count data).

```
lm1b <- lm(abund~graze*log(area), data = birds)
drop1(lm1b, test = "F")
```

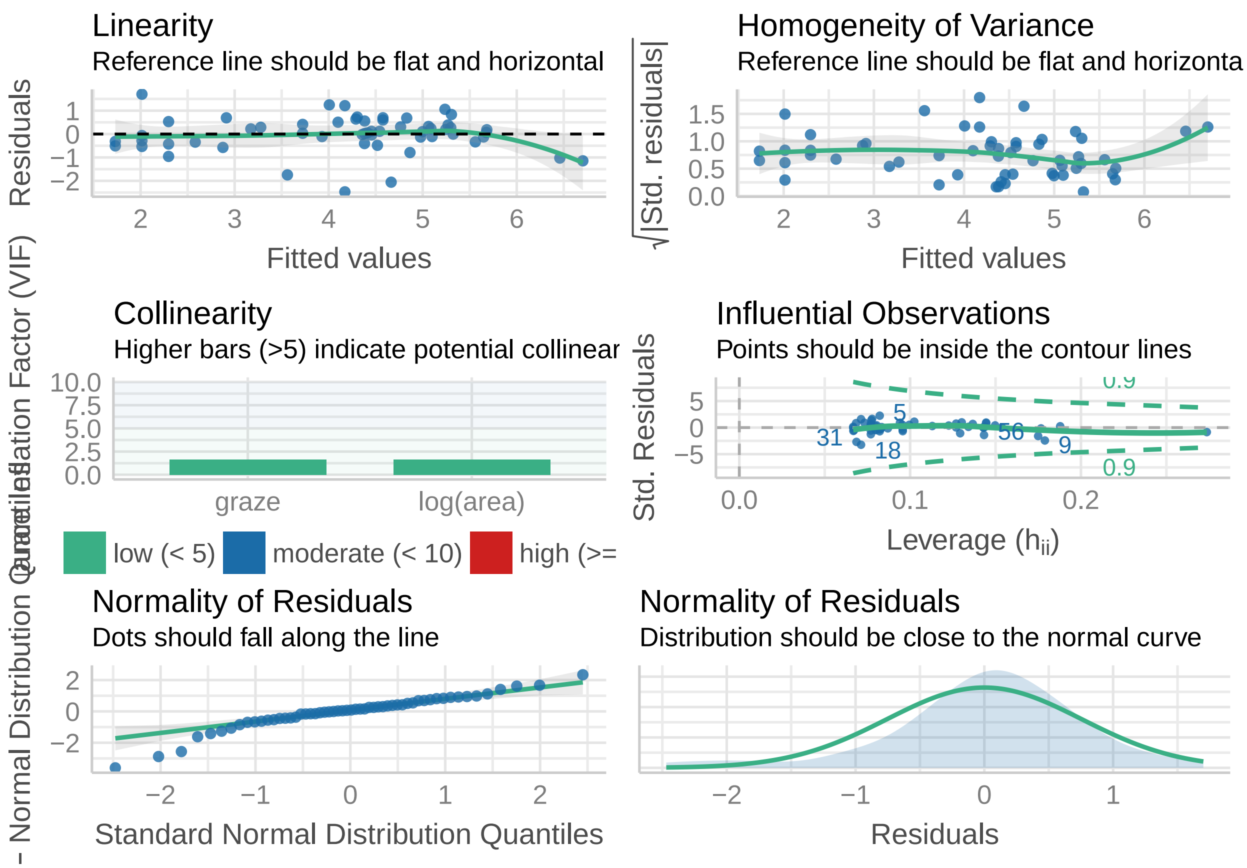
```
## Single term deletions
##
## Model:
## abund ~ graze * log(area)
##                 Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                       1476.6 203.24
## graze:log(area)  4    253.77 1730.4 204.12  1.9764 0.1139
```

```
# update model without interaction
lm1c <- lm(sqrt(abund)~graze + log(area), data = birds)
drop1(lm1c, test = "F")
```

```
## Single term deletions
##
## Model:
## sqrt(abund) ~ graze + log(area)
##           Df Sum of Sq    RSS      AIC F value    Pr(>F)
## <none>                 31.387 -20.4213
## graze      4    24.220 55.607   3.6054  9.6455 7.341e-06 ***
## log(area)  1    19.459 50.846   4.5933 30.9982 1.020e-06 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
performance::check_model(lm1c)
```

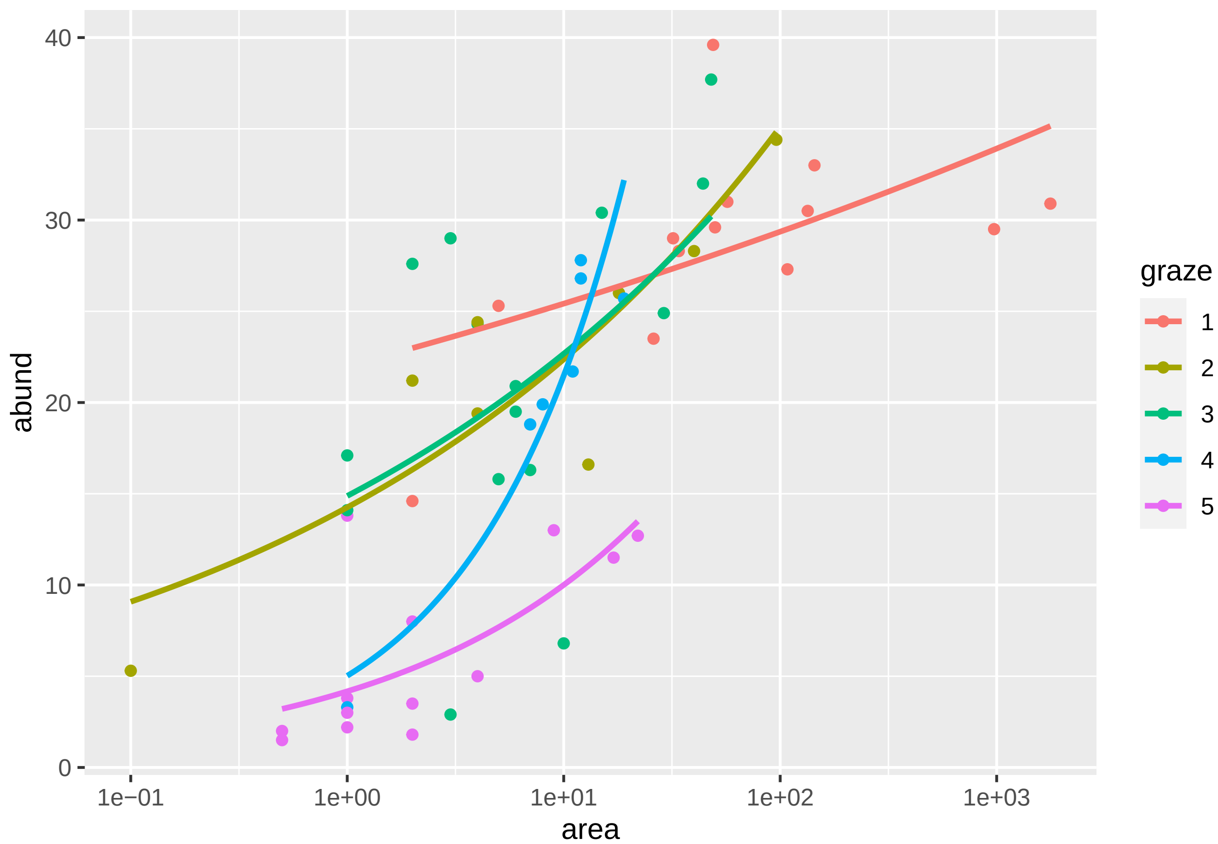


This looks much better. However, it might also be a good idea to use a poisson glm in this case

```
glm1 <- glm(abund ~ graze + log(area), data  =birds, family = "poisson")
drop1(glm1, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## abund ~ graze + log(area)
##           Df Deviance AIC    LRT  Pr(>Chi)
## <none>          129.18 Inf
## graze      4   227.19 Inf 98.016 < 2.2e-16 ***
## log(area)  1   183.60 Inf 54.422 1.617e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

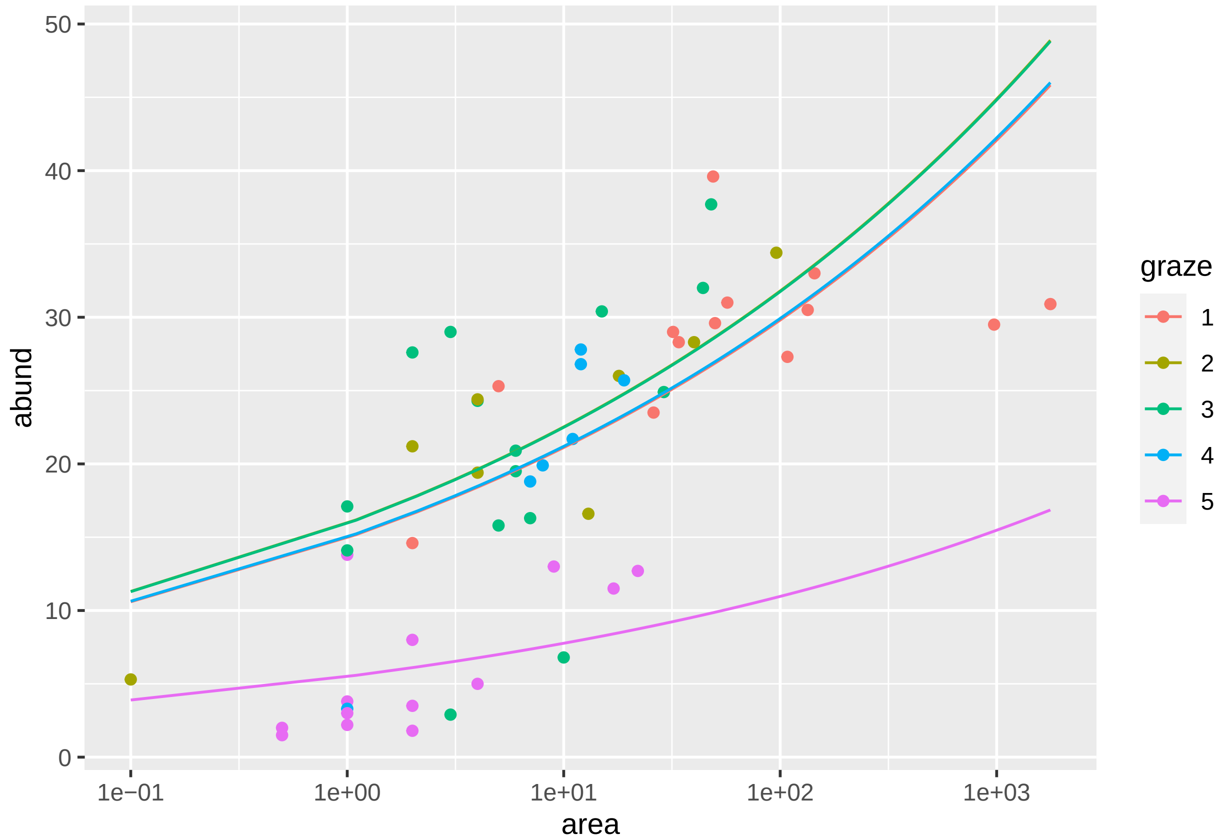
```
ggplot(birds, aes(x = area, y = abund, color = graze)) +
  geom_point() +
  scale_x_log10() +
  geom_smooth(method = "glm", se = FALSE, method.args = list(family = "poisson"))
```



Or using the predict function:

```
pred_dat <- expand_grid(
  graze = factor(1:5),
  area = min(birds$area):max(birds$area)
)
pred_dat$abund <- predict(glm1, newdata = pred_dat, type = "response")

ggplot(birds, aes(x = area, y = abund, color = graze)) +
  geom_point() +
  scale_x_log10() +
  geom_line(data = pred_dat)
```



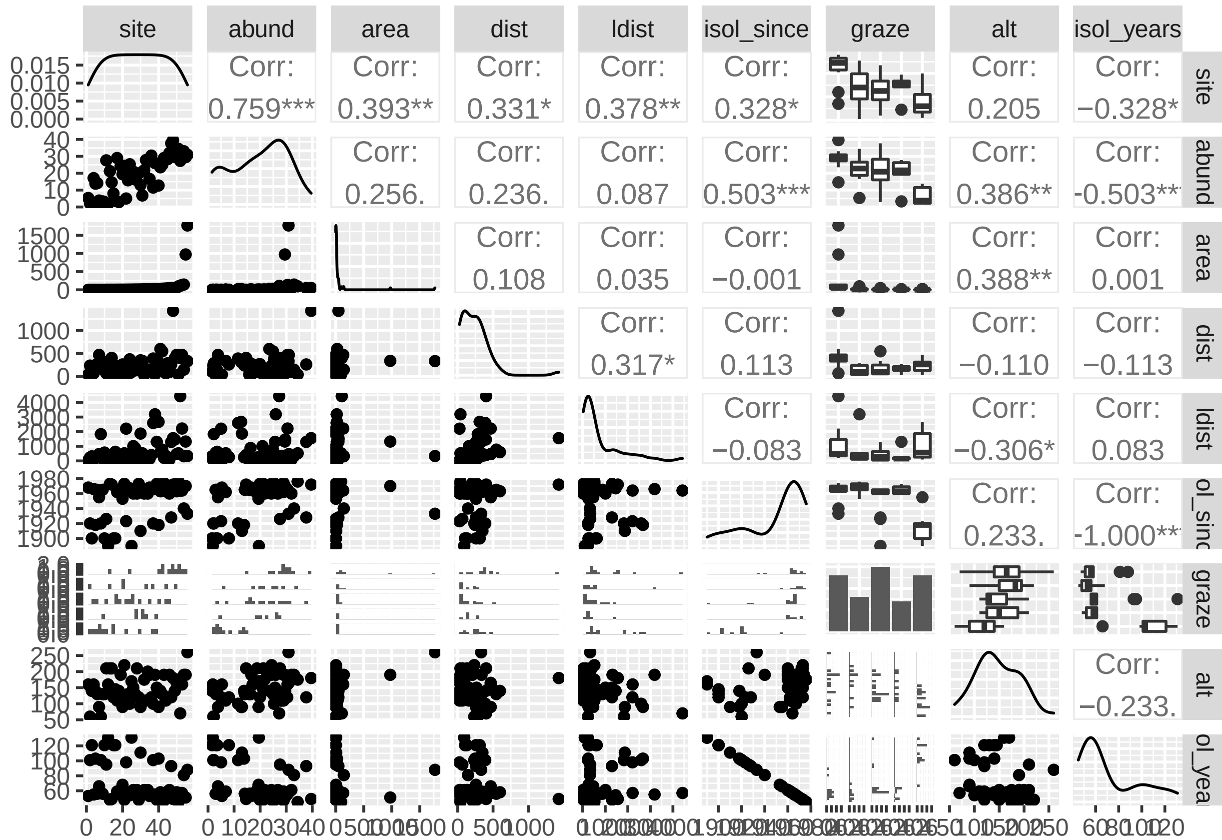
Correlation between variables Are there correlated variables?

```
birds %>%
  select(area, dist, ldist, isol_years, alt) %>%
  cor()
```

```
##                   area       dist       ldist   isol_years        alt
## area       1.000000000  0.1083429  0.03458035  0.001494192  0.3877539
## dist       0.108342870  1.0000000  0.31717234 -0.113217524 -0.1101125
## ldist      0.034580346  0.3171723  1.00000000  0.083316857 -0.3060222
## isol_years 0.001494192 -0.1132175  0.08331686  1.000000000 -0.2327154
## alt        0.387753885 -0.1101125 -0.30602220 -0.232715406  1.0000000
```

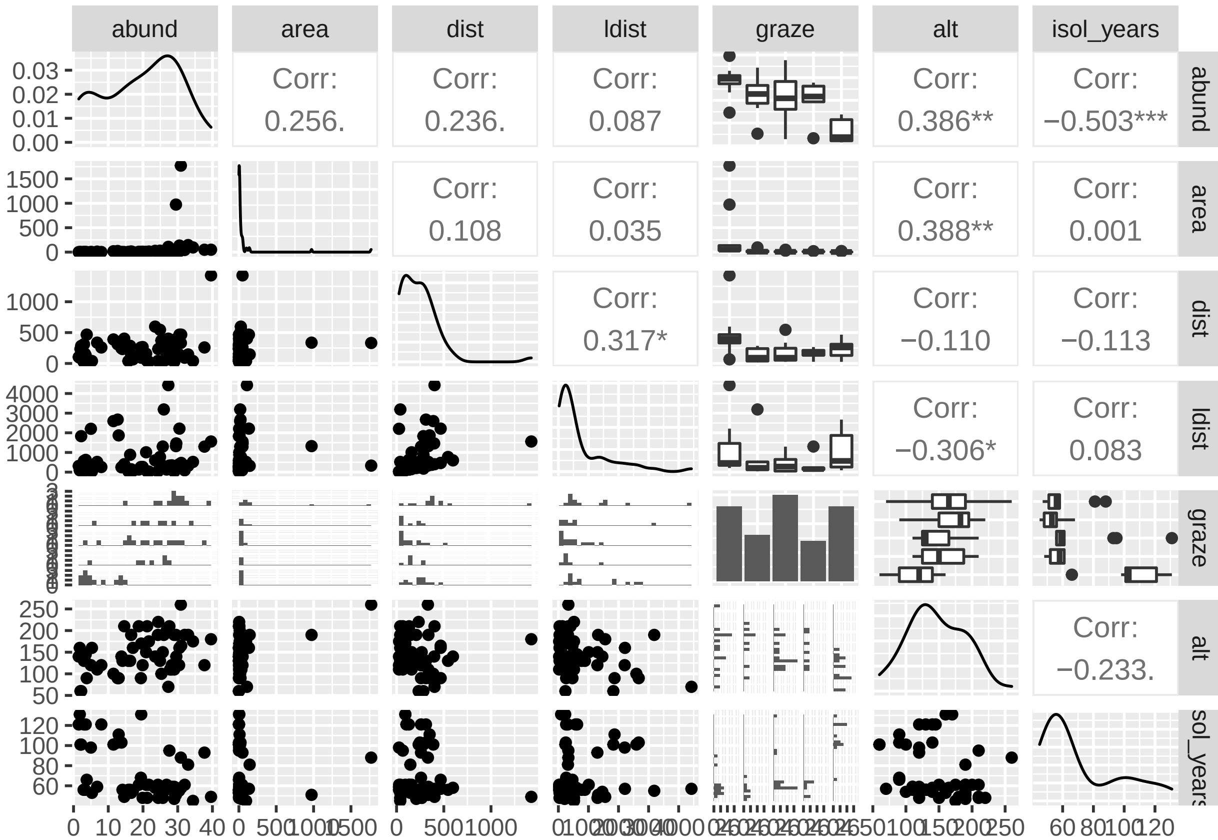
Use the `ggpairs` function from the `GGally` package to plot a matrix plot of all variables from the data and look at possible correlations.

```
GGally::ggpairs(birds)
```

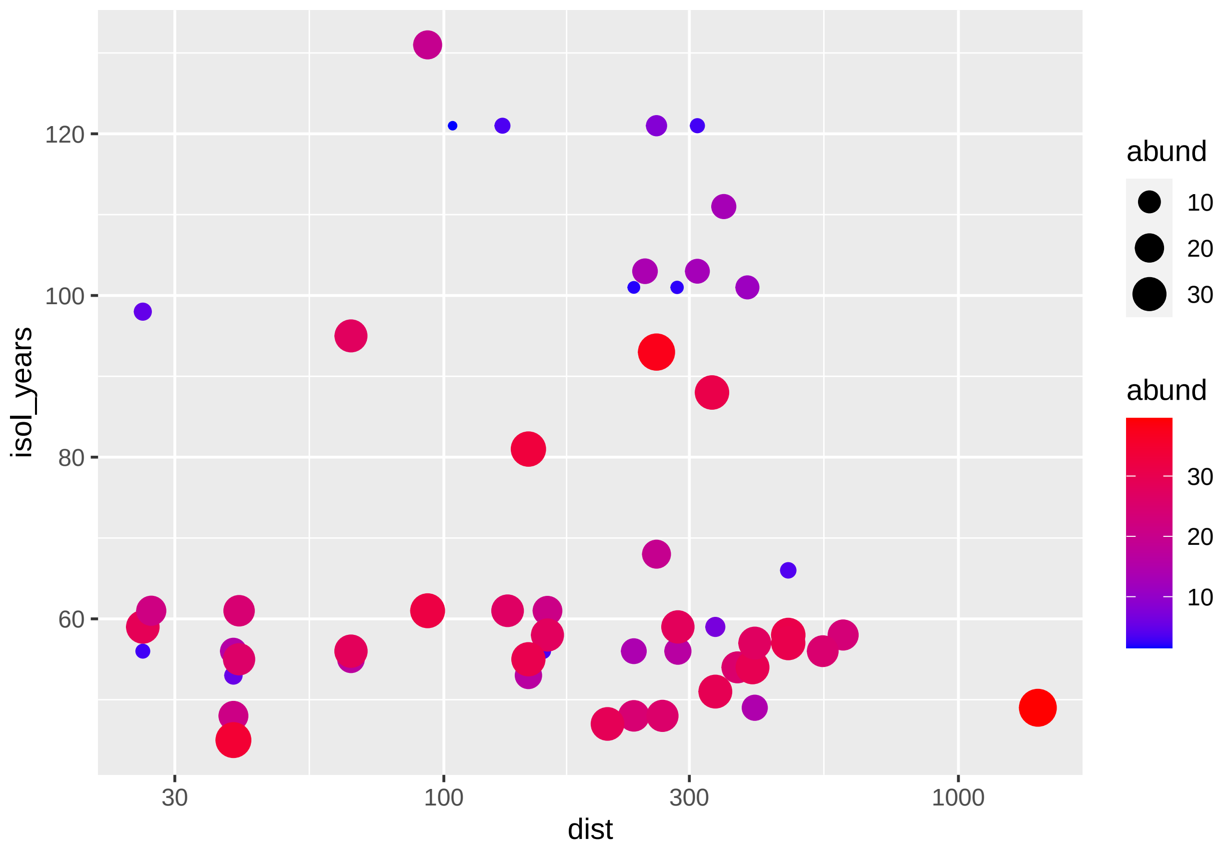


To get a better overview, I now select some independent variables that I am interested in:

```
birds %>%
  select(-isol_since, -site) %>%
  GGally::ggpairs()
```




```
ggplot(data = birds,
mapping = aes(
x = dist,
y = isol_years,
size = abund,
color = abund)) +
geom_point() +
scale_color_gradient(low="blue",high="red")+
scale_x_log10()
```



Including some categorical variables for exploratory plotting

```
q_10 <- quantile(birds$area, 0.1)
q_25 <- quantile(birds$area, 0.25)
q_80 <- quantile(birds$area, 0.8)

birds %>%
  mutate(
    area_class = case_when(
      area < q_10  ~ "tiny",
      between(area, q_10, q_25) ~ "small",
      between(area, q_25, q_80) ~ "medium",
      area > q_80 ~ "large"
    )
  )
```

```
## # A tibble: 56 x 10
##     site abund  area  dist ldist isol_since graze   alt isol_years area_class
##    <int> <dbl> <dbl> <int> <int>      <int> <fct> <int>      <dbl> <chr>
##  1     1   5.3   0.1    39    39       1968 2       160         53 tiny
##  2     2   2     0.5   234   234       1920 5        60        101 tiny
##  3     3   1.5   0.5   104   311       1900 5       140        121 tiny
##  4     4  17.1   1      66    66       1966 3       160         55 small
##  5     5  13.8   1     246   246       1918 5       140        103 small
##  6     6  14.1   1     234   285       1965 3       130         56 small
##  7     7   3.8   1     467   467       1955 5        90         66 small
##  8     8   2.2   1     284  1829       1920 5        60        101 small
```

```
## 9      9   3.3   1     156   156        1965 4        130          56 small
## 10     10   3     1     311   571        1900 5        130         121 small
## # ... with 46 more rows
```