STAT-615 Project I: Data Analysis

```
library(tidyverse)
```

Introduction

```
#Load Data

"britain_species.dat" %>%
  read_delim(delim = " ", col_names = FALSE) %>%
  rename(island = X1,
         area = X2,
         elevation = X3,
         soil = X4,
         latitude = X5,
         distance = X6,
         species = X7) %>%
  mutate(area = as.double(area),
         elevation = as.integer(elevation),
         soil = as.integer(soil),
         latitude = as.double(latitude),
         distance = as.double(distance),
         species = as.integer(species)) -> species


species
```

```
## # A tibble: 42 x 7
##    island       area elevation  soil latitude distance species
##    <chr>       <dbl>     <int> <int>    <dbl>    <dbl>   <int>
##  1 Ailsa         0.8       340     1     55.3     14        75
##  2 Anglesey    712.        127     3     53.3      0.2     855
##  3 Arran       429.        874     4     55.6      5.2     577
##  4 Barra        18.4       384     2     57       77.4     409
##  5 Bressay      31.1       226     1     60.1    202.      177
##  6 Britain  229850.       1343    16     54.3      0      1666
##  7 Canna        12.7       210     1     57.1     40.6     300
##  8 Coll         74.1       103     3     56.6     14.5     443
##  9 Colonsay     44.8       143     1     56.1     31.1     482
## 10 Eigg         29         393     1     56.9     12.3     453
## # ... with 32 more rows
```

For our data analysis project, we were interested in exploring the diversity of species. The importance of this question subject is nontrivial, as preserving species diversity is incredibly important to prevent extinction of them. By finding out what factors are related or reponsible for increased diversity, we can obtain general awareness of them, as well as

understand their role in diversity and how to manage them properly to preserve variety of species.

Because this is a very broad subject, we decided to narrow in on a particular dataset. The University of Florida contains a data repository which contained a dataset which included information about bird species diversity in the islands, as well as mainland Britain.

The following variables are found in this dataset: - island: name of the island - area: measured in squared kilometers - elevation: highest peak, measured in meters - soil: number of different soil types - latitude - distance: from mainland britain - species: total number of bird species

```
head(species, 5)
```

```
## # A tibble: 5 x 7
##   island    area elevation  soil latitude distance species
##   <chr>    <dbl>     <int> <int>    <dbl>    <dbl>   <int>
## 1 Ailsa      0.8       340     1     55.3     14        75
## 2 Anglesey 712.        127     3     53.3      0.2     855
## 3 Arran    429.        874     4     55.6      5.2     577
## 4 Barra     18.4       384     2     57       77.4     409
## 5 Bressay   31.1       226     1     60.1    202.      177
```

```
ncol(species)
```

```
## [1] 7
```

```
nrow(species)
```

```
## [1] 42
```

The dataset contains 7 variables, and 42 observations. Although there is no missing data it should be noted that a limitation of this dataset and following analyses is that it is not include a particularly large number of observations, particularly with respect to the number of variables considered.

#Limitations

Since `latitude` is not particularly useful, it will not be used in any analyses. This will place the observations to variable ratio at 42:6. However this would only be the case for conisdered models in which all other 6 variables are included. Some limitations of the small dataset are that it does not produce strong statistical power, thus, any conclusions, null or alternative, should be interpreted with caution.

#Preliminary Data Analysis

```
summary(species)
```

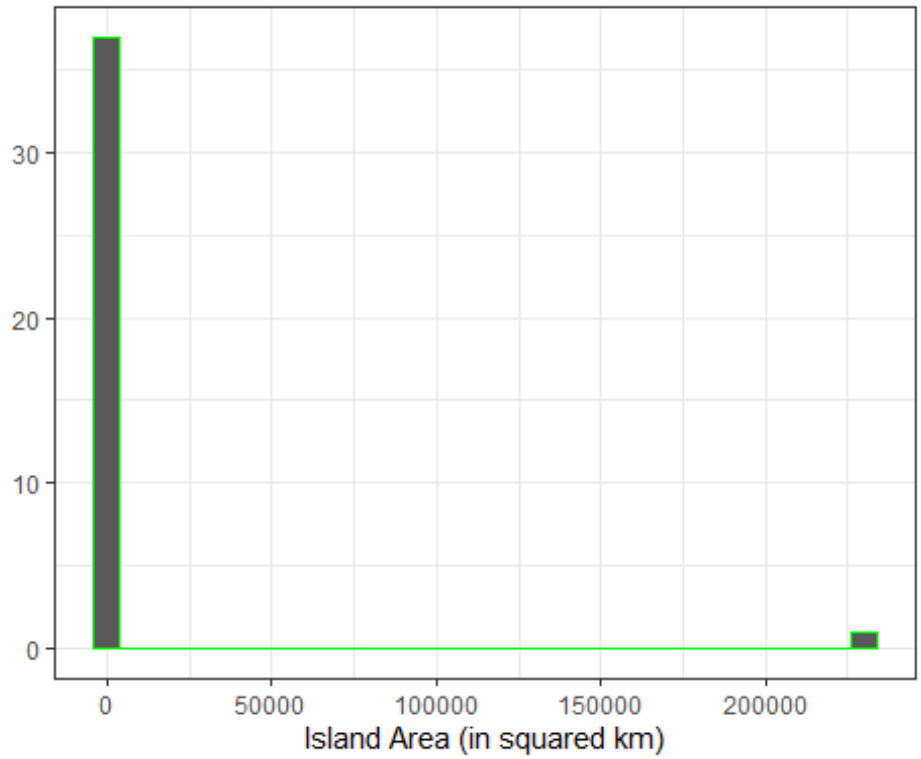
```
##     island               area            elevation           soil      
##  Length:42          Min.   :     0.5   Min.   :   7.0   Min.   :  1.00  
##  Class :character   1st Qu.:    14.0   1st Qu.: 129.5   1st Qu.:  1.00  
##  Mode  :character   Median :    52.8   Median : 232.0   Median :  2.00
```

```
##                     Mean   :  6324.9   Mean   : 346.4   Mean   : 28.62
##                     3rd Qu.:   417.2   3rd Qu.: 442.0   3rd Qu.:  3.75
##                     Max.   :229849.8   Max.   :1343.0   Max.   :620.00
##                     NA's   :4
##     latitude        distance          species
##  Min.   : 2.00   Min.   :  0.000   Min.   :   9.0
##  1st Qu.:55.62   1st Qu.:  9.825   1st Qu.: 159.5
##  Median :56.80   Median : 33.950   Median : 346.0
##  Mean   :51.97   Mean   : 63.119   Mean   : 368.2
##  3rd Qu.:59.08   3rd Qu.: 65.300   3rd Qu.: 450.8
##  Max.   :60.80   Max.   :258.100   Max.   :1666.0
##
```

#Variable Distributions

##Island Area

```
ggplot(species, aes(x = area)) +
  geom_histogram(color = "green") +
  theme_bw() +
  ylab("") +
  xlab("Island Area (in squared km)")
```



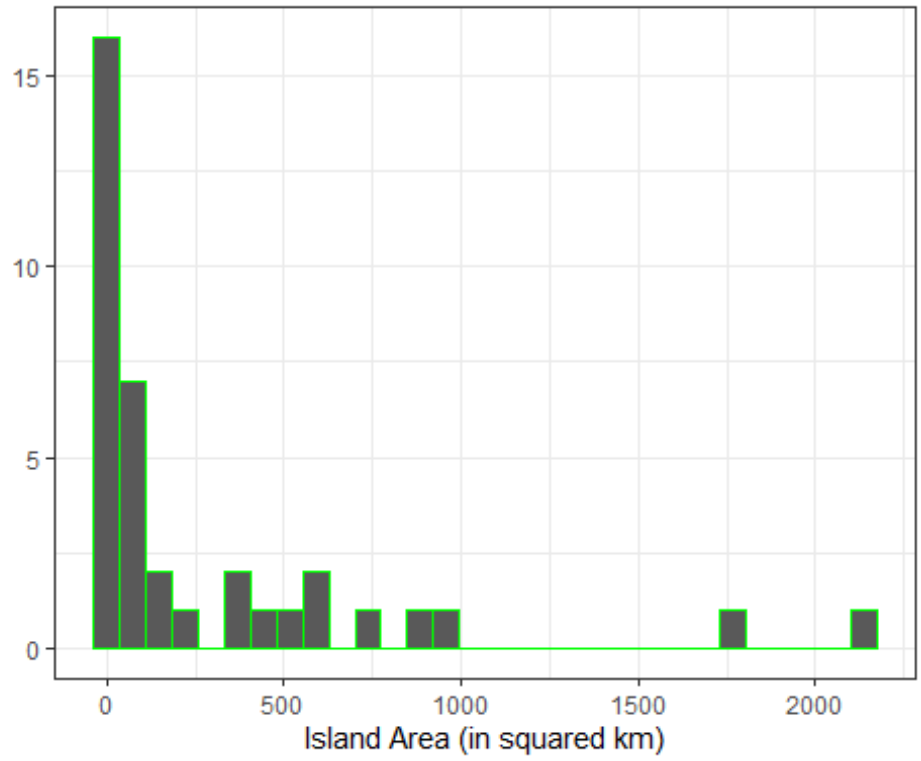
We note that for area there is a clear and extreme outlier.

```
species %>%
  filter(area > 200000)
```

```
## # A tibble: 1 x 7
##   island     area elevation  soil latitude distance species
##   <chr>     <dbl>     <int> <int>    <dbl>    <dbl>   <int>
## 1 Britain 229850.      1343    16     54.3        0    1666
```

Looking into the data we note that this outlier is clearly mainland Britain. To observe a more informative histogram, it is produced without this outlier.

```
species %>%
  filter(area < 200000) %>%
  ggplot(aes(x = area)) +
  geom_histogram(color = "green") +
  theme_bw() +
  ylab("") +
  xlab("Island Area (in squared km)")
```

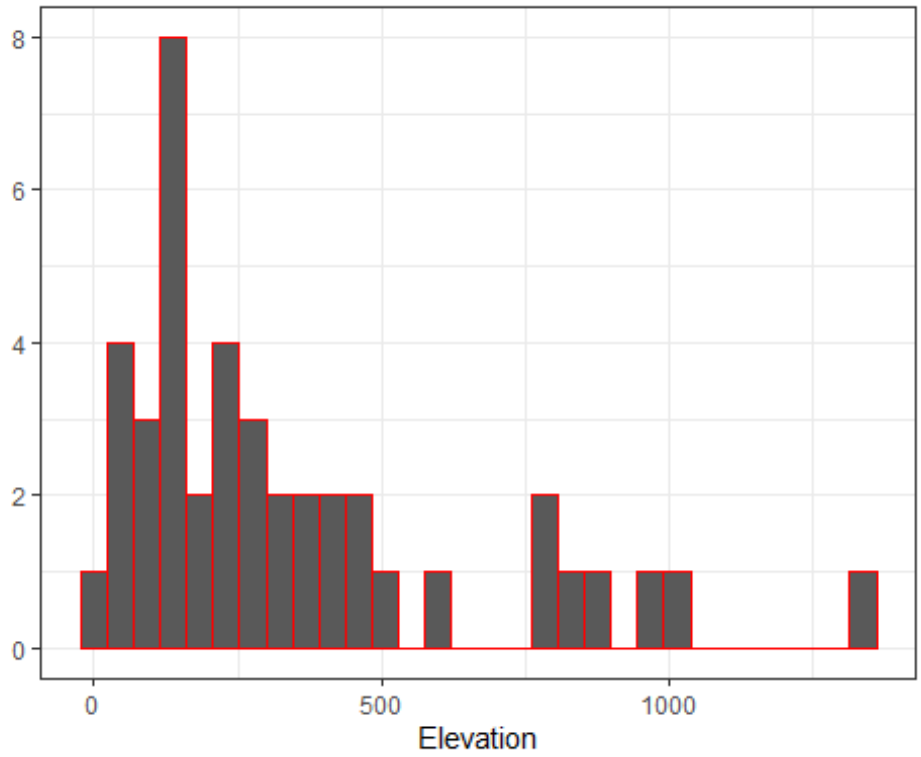


After exclusion of other the outlier, we note that there are still a few more towards higher areas, and this variable in general appears to be strongly right skewed. Tranformation of this variable might be nessesary. We also note for further data visualization that the observation of mainland Britain will likely be an outlier as well.

##Elevation

```
ggplot(species, aes(x = elevation)) +
  geom_histogram(color = "red") +
  theme_bw() +
```

```
  ylab("") +
  xlab("Elevation")
```



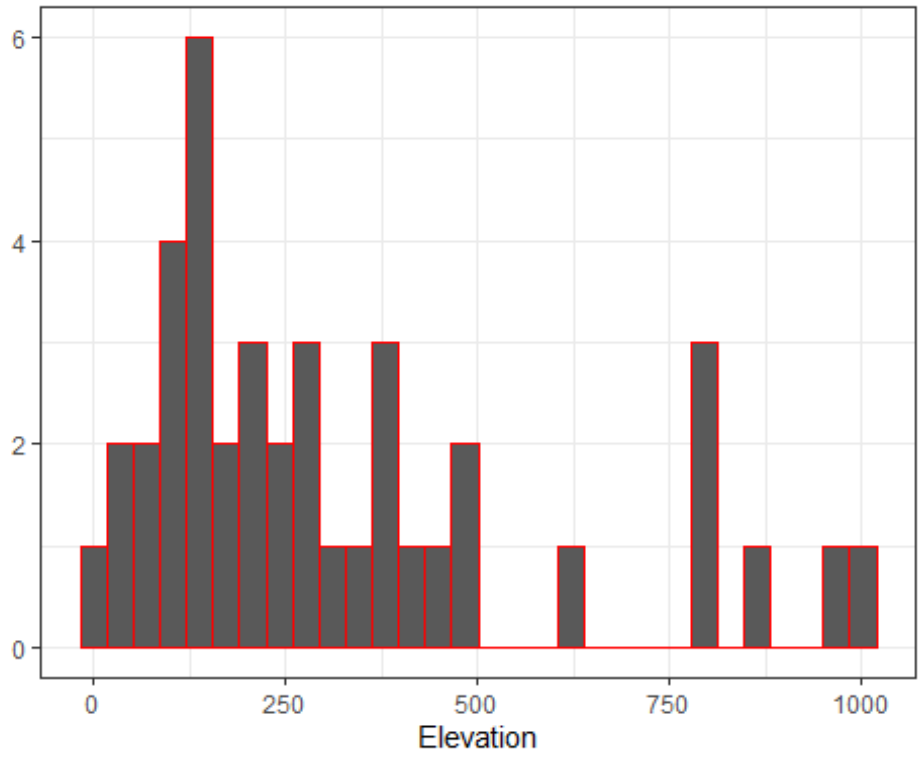
As clearly visible in the histogram above, there is a presence of an extreme outlier which could influence the predictions.

```
species%>%
  filter(elevation>1000)

## # A tibble: 2 x 7
##   island     area elevation  soil latitude distance species
##   <chr>     <dbl>     <int> <int>    <dbl>    <dbl>   <int>
## 1 Britain 229850.      1343    16     54.3      0      1666
## 2 Skye      1735.      1009     5     57.3      0.6     594
```

The table above denotes the outliers within this variable.

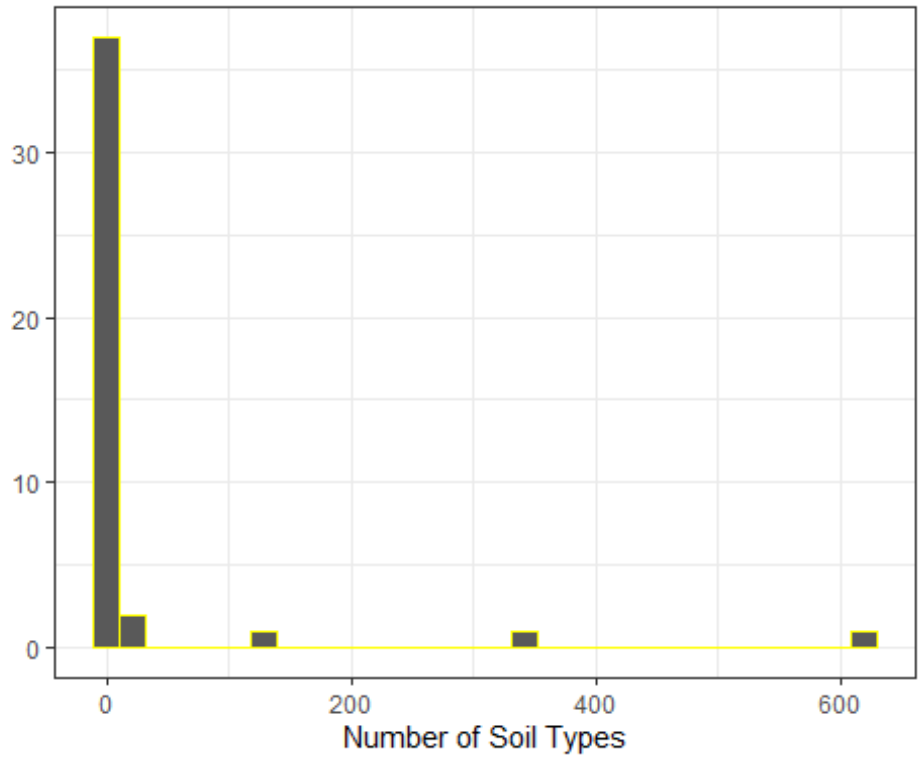
```
species %>%
  filter(elevation < 1010) %>%
  ggplot(aes(x = elevation)) +
  geom_histogram(color = "red") +
  theme_bw() +
  ylab("") +
  xlab("Elevation")
```



Upon rmeoving the outliers, there is still a few present towards higher elevations. However this variable seems to be skewed towrds the rigth as well. Data manipilation will be required to get accurate estimates from this dataste.

Number of Soil Types

```
ggplot(species, aes(x = soil)) +
  geom_histogram(color = "yellow") +
  theme_bw() +
  ylab("") +
  xlab("Number of Soil Types")
```



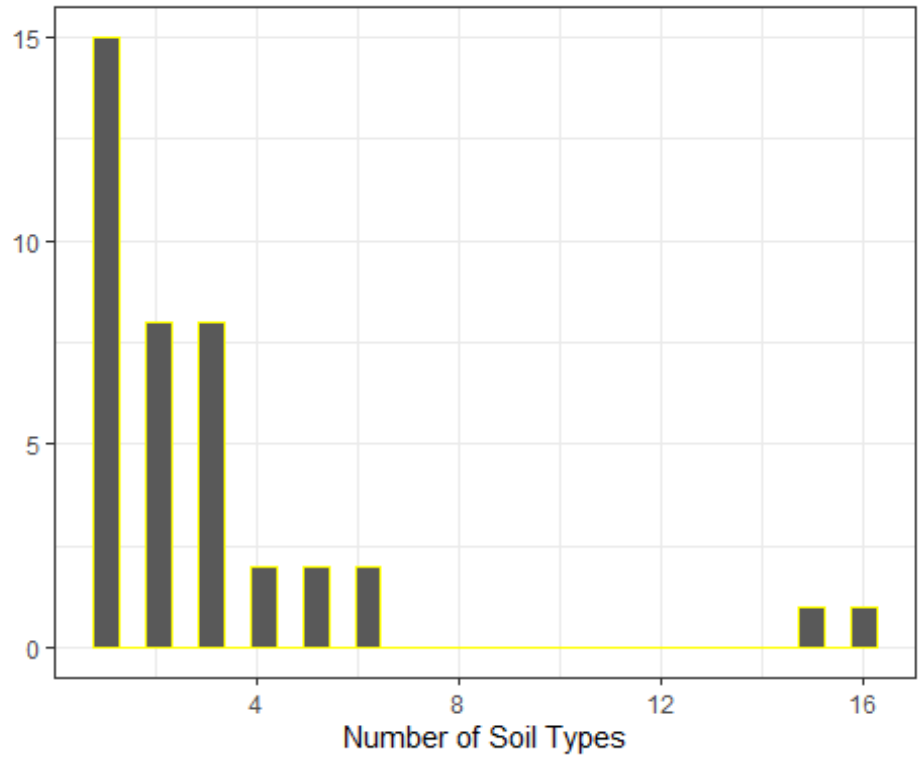
As clearly visible, there are 3 clear outliers present

```
species%>%
  filter(soil>20)
```

```
## # A tibble: 3 x 7
##   island  area elevation  soil latitude distance species
##   <chr>  <dbl>     <int> <int>    <dbl>    <dbl>   <int>
## 1 N.        NA       305   347        3     57.6      57
## 2 S.        NA        60   119        2     58.8       9
## 3 S.        NA       365   620        3     57.2      82
```

The table above denotes the extreme outliers within this variable.

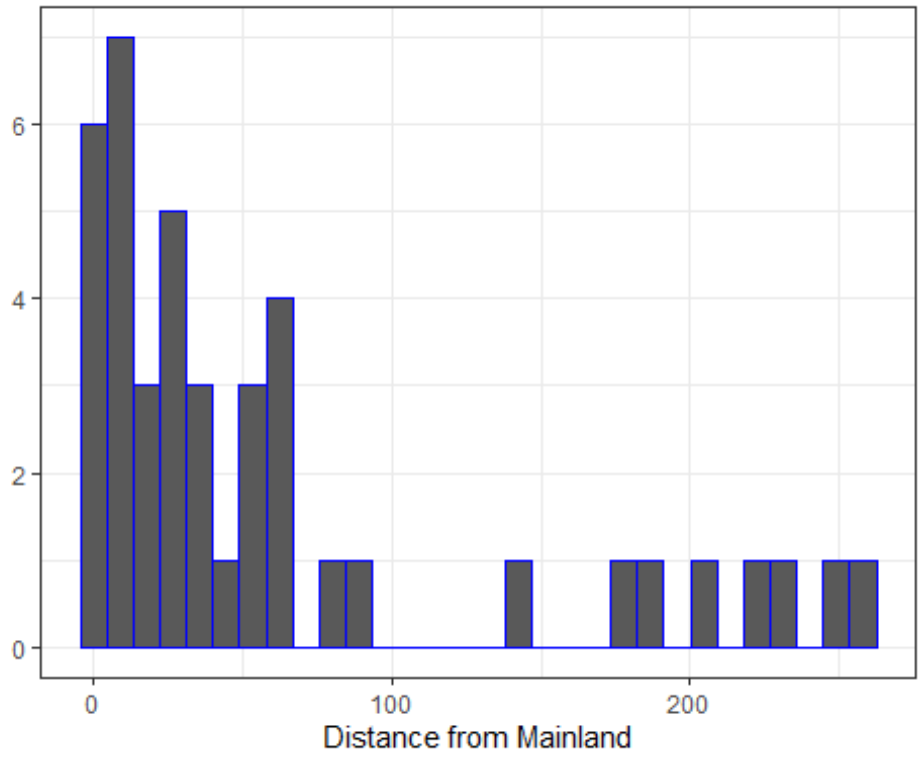
```
species %>%
  filter(soil < 100) %>%
  ggplot(aes(x = soil)) +
  geom_histogram(color = "yellow") +
  theme_bw() +
  ylab("") +
  xlab("Number of Soil Types")
```



Despite getting rid of the extreme outliers, there is still presence of a few on the right. As other variables before, even this this variable is not normally disctributed and is skewed towards the right which will imoact prediction accuracies.

Distance from Mainland

```
ggplot(species, aes(x = distance)) +
  geom_histogram(color = "blue") +
  theme_bw() +
  ylab("") +
  xlab("Distance from Mainland")
```



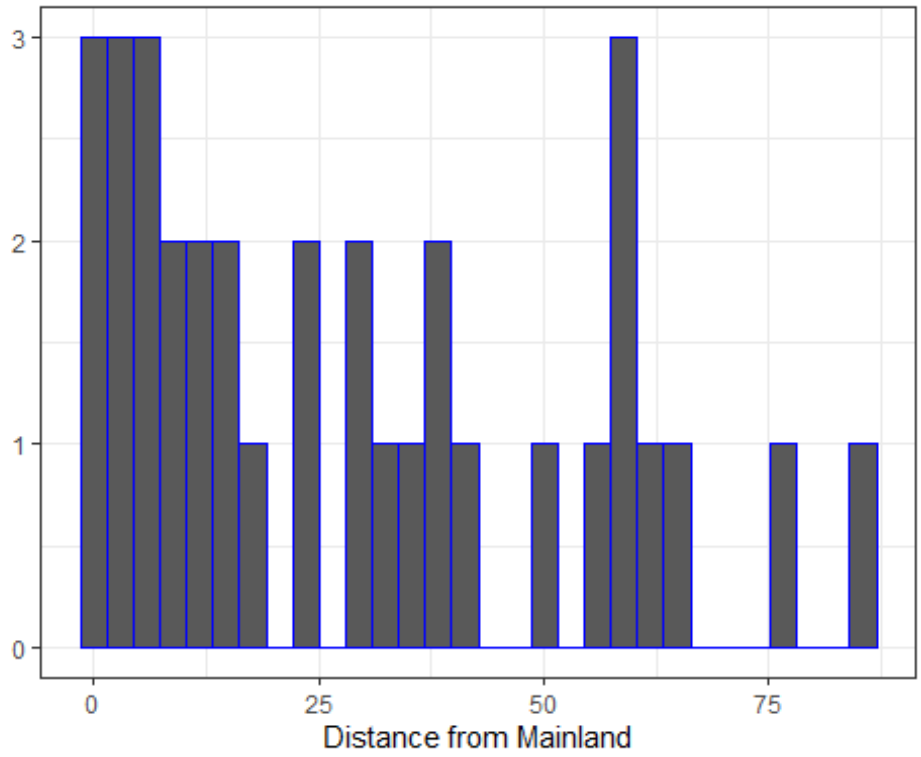
As clearly visible, there are a few outliers over 100 km .

```
species%>%
  filter(distance>100)
```

```
## # A tibble: 8 x 7
##   island    area elevation  soil latitude distance species
##   <chr>    <dbl>     <int> <int>    <dbl>    <dbl>   <int>
## 1 Bressay   31.1       226     1     60.1     202.     177
## 2 Fair       5.2       217     1     59.5     144.     174
## 3 Fetlar    40.9       159     2     60.6     247.     189
## 4 Foula     13.5       418     1     60.1     177.     149
## 5 Shetland 984.        450     6     60.3     189.     421
## 6 Unst     121.        285     2     60.8     258.     246
## 7 Whalsay   19.7       120     1     60.4     221      158
## 8 Yell     217.        205     2     60.6     236.     161
```

The table above denotes the extreme outliers within this variable.

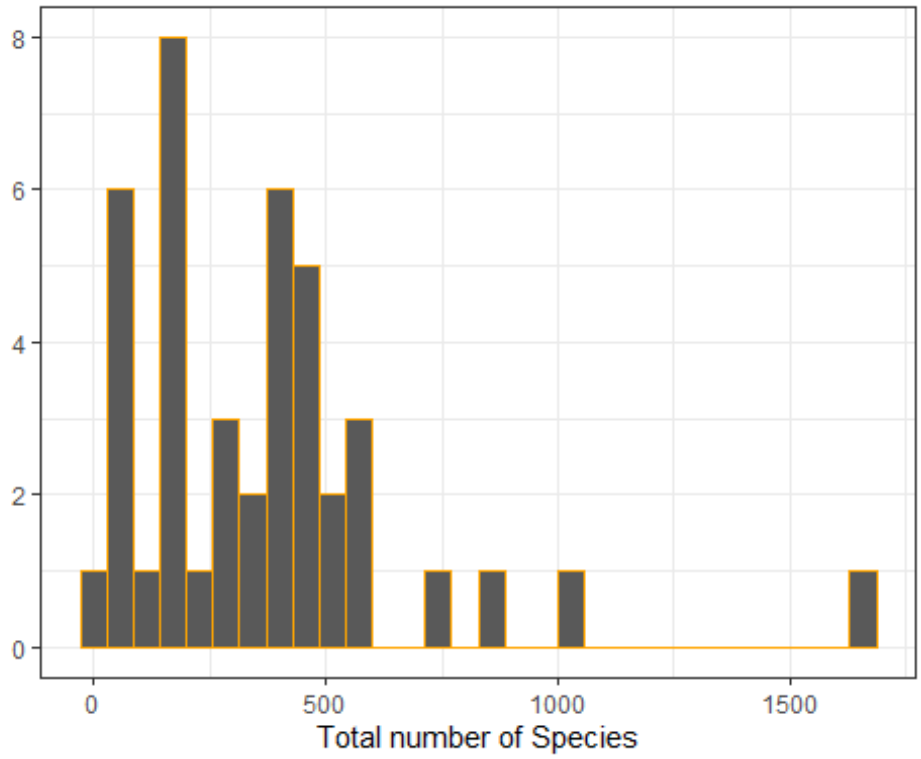
```
species %>%
  filter(distance < 100) %>%
  ggplot(aes(x = distance)) +
  geom_histogram(color = "blue") +
  theme_bw() +
  ylab("") +
  xlab("Distance from Mainland")
```



After removing teh outliers, we can see that the distribution is somewhat normal with some skewness towards the right. Slight data transformation will be required.

Distance from Mainland

```
ggplot(species, aes(x = species)) +
  geom_histogram(color = "orange") +
  theme_bw() +
  ylab("") +
  xlab("Total number of Species")
```



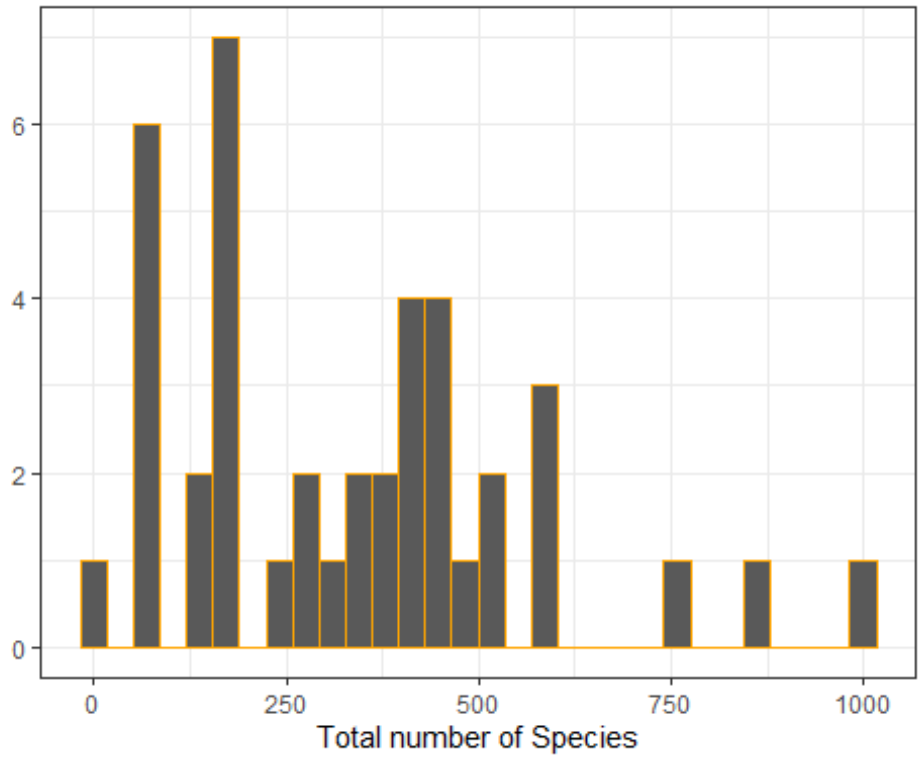
As clearly visible, there are a few outliers over 100 km .

```
species%>%
  filter(species>1500)
```

```
## # A tibble: 1 x 7
##   island     area elevation  soil latitude distance species
##   <chr>     <dbl>     <int> <int>    <dbl>    <dbl>   <int>
## 1 Britain 229850.      1343    16     54.3        0    1666
```

The table above denotes the extreme outliers within this variable.

```
species %>%
  filter(species < 1500) %>%
  ggplot(aes(x = species)) +
  geom_histogram(color = "orange") +
  theme_bw() +
  ylab("") +
  xlab("Total number of Species")
```



After removing teh outliers, we can see that the distribution is somewhat normal with some skewness towards the right and left. After soem data tranfromation, we will be able to being this normal distribution to be used in a prediction model.

Corelation Matrix

```
species_cor<-species%>%
  select(area,elevation,soil,distance,species)
species_cor
```

```
## # A tibble: 42 x 5
##        area elevation  soil distance species
##       <dbl>     <int> <int>    <dbl>   <int>
##  1      0.8       340     1     14        75
##  2    712.        127     3      0.2     855
##  3    429.        874     4      5.2     577
##  4     18.4       384     2     77.4     409
##  5     31.1       226     1    202.      177
##  6 229850.       1343    16      0      1666
##  7     12.7       210     1     40.6     300
##  8     74.1       103     3     14.5     443
##  9     44.8       143     1     31.1     482
## 10     29         393     1     12.3     453
## # ... with 32 more rows
```

```
cor(species_cor, use = "complete.obs")
```

```
##                 area  elevation       soil   distance    species
## area       1.0000000  0.5201509  0.8429887 -0.1333701  0.7028713
## elevation  0.5201509  1.0000000  0.7023121 -0.2541048  0.6126966
## soil       0.8429887  0.7023121  1.0000000 -0.1818096  0.7772743
## distance  -0.1333701 -0.2541048 -0.1818096  1.0000000 -0.4162782
## species    0.7028713  0.6126966  0.7772743 -0.4162782  1.0000000
```

#Scatter Plots

```
pairs(species_cor)
```