Alisha Soares
BINF 6210 – Assignment 1

######BINF6210 Assignment 1 ----

####Introduction

###In the midst of the novel COVID-19 virus, other natural disasters such as forest fires and a very active hurricane season have exacerbated the stresses that countries around the world face. During the spring and summer months of 2020, news reports sent out from Africa and South Asia, highlighted another disaster; large locust swarms. Locusts belong to the family of organisms called Acrididae. In Pakistan, Acrididae are considered pests, destroying vast acres of crops , in a country where the GDP relies on the agriculture sector. Studies such as the one conducted by Hussain et. al.[1], are routinely performed in the country throughout the year to study the species diversity, richness and seasonal habits.

###As climate change continues to change the weather patterns in the habitats and migratory routes of Acrididae, concern grows over the long term affects of devastation caused in their wake. The World Bank Organization has declared the situation in Africa and South Asia dire[2].

###This topic i s interesting as it is close to home, as well as something I have been curious about, though without the resources to examine it closer. As a novice of the R program, the scope of this project is limited while I explore the data. I hypothesize that the species between South Africa and Pakistan will not overlap greatly as it has been as long time since the last migration came towards Asia from Africa [3]. I expect that the species between China and Pakistan will show some overlap. While exploring the data I hope to develop a better understanding of the limitations in data sets if there are missing data points, and to come to a conclusion of how studies on small Family groups such as Acrididae may be better studied in the future.

##Below begins the coding used for this project:

```
# First obtain all the necessary libraries for analyzing the Acrididae dataset

library(readr)

library(tidyverse)

library(dplyr)

library(ggplot2)

library(VennDiagram)

library(grid)

library(ggmap)

library(maps)
library(vegan)



######Obtain Acrididae data from BOLD for South Africa, Pakistan and China ----


#Access Acrididae data from BOLD for countries. Data obtained on September 18 2020 at 3:51PM

Acrididae.BOLD <- read_tsv("http://www.boldsystems.org/index.php/API_Public/combined?taxon=Acrididae&geo=Pakistan|China|South%20Africa&format=tsv")


#Write the data to hard disk

write_tsv(Acrididae.BOLD, "Acrididae_BOLD_data.tsv")


#Read the saved .tsv file form the hard disk

Acrididae.BOLD <- read_tsv("Acrididae_BOLD_data.tsv")


######Explore/Organize the data ----

#Check the class of the data set to become familiar with the data

class(Acrididae.BOLD)


#Summarize the data to see what it contains

summary(Acrididae.BOLD)


#Confirm that country data needed has been acquired

unique(Acrididae.BOLD$country)


#Sort by country to see number of BOLD records per country

Acrididae.BOLD %>%

  count(country, sort = TRUE)
```

```
sort(table(Acrididae.BOLD$country), decreasing = TRUE)


#Make a bar plot of the number of BOLD records per country. This is to see what the distribution of the
raw data is.See Figure 1 a in .pdf.

records <- c(1014,520,401)

countries <- c("'Pakistan", "South Africa", "China")


barplot(records,names.arg=countries,xlab="Countries",ylab="Number of BOLD records",col="blue",
    main="Number of BOLD records per Country",border="green")


#View the variable names in the main data set Acrididae.BOLD, noting the column names of interest

names(Acrididae.BOLD)


#Make a subset of the data Acrididae.sub for easier manipulation

Acrididae.sub <- Acrididae.BOLD[,c(8,55,47,48,22)]


###Clear up the data of missing values (NAs)


#Create the variable blank at the end of Acrididae.sub to count the characters in the species names. This
will assign each species name with a value of 1 or 0. This makes it easier to view the data and view
where missing data points are.

Acrididae.sub$blank <- str_count(string=Acrididae.sub$species_name, pattern="[\\s\\.\\d]")


#Check how many records have species name populated

sum(Acrididae.sub$blank==1, na.rm=TRUE)


#Remove records that are missing both species name and BIN. I would like to see how much of the data
set is removed from the data set, and how this would affect the distribution of the BOLD data records
```

```
Acrididae.core <- subset(Acrididae.sub,blank==1 & is.na(blank)==F & is.na(species_name)==F & is.na(bin_uri)==F)


#Count the unique species names before and after the removal of NA data
#Before for BINS
unique(Acrididae.sub$bin_uri)
sum(is.na(Acrididae.sub$bin_uri))
length(unique(Acrididae.sub$bin_uri))


#After for BINS. Note that NA count is zero
unique(Acrididae.core$bin_uri)
sum(is.na(Acrididae.core$bin_uri))
length(unique(Acrididae.core$bin_uri))


#Before for species name
unique(Acrididae.sub$species_name)
sum(is.na(Acrididae.sub$species_name))
length(unique(Acrididae.sub$species_name))


#After for species name. Note that NA count is zero
unique(Acrididae.core$species_name)
sum(is.na(Acrididae.core$species_name))
length(unique(Acrididae.core$species_name))


###As expected there is a large change in the values for unique BINs and species names after removing all the NA values from the data set. Unique BIn count decreased from 218 BINs to 114 BINs. While named species decreased to 106 names.


#Make a new bar plot to show the results of removing the NAs. Expect to see a lot of data points missing.See Figure 1 b in .pdf.
```

```
Acrididae.core %>%
  count(country, sort = TRUE)


records <- c(669,19,360)

countries <- c("'Pakistan", "South Africa", "China")


barplot(records,names.arg=countries, width=1, xlab="Countries", ylab="Number of BOLD records",
col="green",
        main="Number of BOLD records per Country (after NAs from BIN and species removed)",
border="blue")


rm(Acrididae.sub, Acrididae.core, countries, records)


##Clear up the data. My focus is on the geographical distribution (by country) of BINs representing
species

#remove rows that have missing BINs. So will study a data set where only BINs with data will be
considered.

Acrididae.bin <- drop_na(Acrididae.sub,bin_uri)


#Make a comparison bar plot to the bar plots done previously. This will have only the NAs from the BIN
column removed. See Figure 1 c in .pdf.

Acrididae.bin %>%
  count(country, sort = TRUE)


records <- c(990,515,373)

countries <- c("'Pakistan", "South Africa", "China")


barplot(records,names.arg=countries, width=1, xlab="Countries", ylab="Number of BOLD records",
col="red",
        main="Number of BOLD records per Country (after NAs from BIN removed)", border="blue")
```

###The bar plots provide a good illustration of how missing data points whether due to lack of sequence or ambiguous phenotypic or morphological identification, can alter the data set.

######Analyzing the data----

#Check the ratio of BINs to species.This is for the data set to be studied in this project. I expect the value to be above 1. This is due to the geographic isolation of the South African species from the Asian species for a long time period.

```
length(unique(Acrididae.bin$bin_uri))/length(unique(Acrididae.bin$species_name))
```

###The value was much higher than expected at 1.94. However, this is a relatively small sample of the global Acrdidae population. When comparing the species in this study they seem to have diverged.

###Prepare the data for the vegan package so that one or two statistical tests may be performed and to obtain Rarefaction and Species Abundance Curves.

#Count the number of species per BIN per country

```
BINcount.country <- Acrididae.bin %>%
  group_by(country, bin_uri) %>%
  count(bin_uri)
```

#Spread the data to create a data set that is readable by the vegan package

```
BINspread.country <- spread(data = BINcount.country, key = bin_uri, value = n)
```

#Convert the NAs to zeroes so that the table has numerical values for species counts

```
BINspread.country[is.na(BINspread.country)] <- 0
```

#Prepare data for vegan, setting countries as row names instead of as part of a column

```
BINspread.country <- BINspread.country %>%
  remove_rownames %>%
  column_to_rownames(var="country")
```

```
##Plot species accumulation curve analysis

AccumCurve <- specaccum(BINspread.country)

plot(spa, ci.type="poly", col="blue", lwd=2.5, ci.lty=0, ci.col="lightblue", main = "Accumulation Curve",xlab = "Countries", ylab = "BIN Richness"  )
```

###Species accumulation curves compare the species sampling based on country or site sampling data. AS more samples are added with additional countries the slope of the curve is more linear. This can be seen by the bar plots that were charted earlier. The more countries and samples there are, the more likely it is that more species will be found. I would also like to check the rarefaction curve to see how it compares to this country based sampling. See Figure 2 a

```
##Plot rarefaction curves for the country data to compare with  the species accumulation curve

#count the number of records by BIN

BINcount <- Acrididae.bin %>%

  group_by(bin_uri) %>%

  count(bin_uri)


#To reorganize the bin data

BINspread <- spread(data = BINcount, key  = bin_uri, value = n)


#Create rarefaction cure to see the species richness off all 3 countries combined

Rarefaction.curve <- rarecurve(BINspread, main = "Rarefaction curve", xlab = "Individuals Barcoded", ylab = "BIN Richness")
```

###The rarefaction curve starts of linearly but slopes off slightly towards the end. This is an individual based curve. As more samples are added it appears that there are fewer species left to discover. However, that is not what the accumulation curve suggests with its linear curve. The sample set being studied here is of only 3 countries after the BINs with NA have been removed. This could explain the trend. See Figure 2 b

```
###Comparing the species diversity of each country. Used R documentation to understand these indices in more detail [4].

#Calculate Shannon's index to see the species diversity within each country. Values range from 0 to 5.

shannon <- diversity(BINspread.country, index = "shannon")


#Calculate Simpson's index to compare the results to Shannon's index. Here values range from 0 to 1.

simpson <- diversity(BINspread.country, index = "simpson")


###For both Shannon's (3.35 to 3.8) and Simpson's (0.95 to 0.97) indices the values obtained indicate a high species diversity within the geographic populations. while Shannon's index assumes that species were obtained randomly, Simpson's index places greater importance on the abundance of species, especially those with more counts.Perhaps for my data these indices are not the best choice but they show that within each country, the species in and of itself is diverse.


###Compare species (BINs) between countries. The goal is to answer my question about whether there is any overlap between the species of each country.

#Create a dataset to pull country data points from

Country.comp <- BINcount.country %>%

  group_by(bin_uri) %>%

  count(country)


#Use the unique function with conditions to find the BINs unique to the countries

Pakistan <- unique(Country.comp$bin_uri[Country.comp$country=="Pakistan"])

China <- unique(Country.comp$bin_uri[Country.comp$country=="China"])

South.Africa <- unique(Country.comp$bin_uri[Country.comp$country=="South Africa"])


#Rearrange the Country.comp dataset to see a table that shows the distribution as well

Country.comp.table <- spread(Country.comp,country, n)


#Use the datasets above to create a Venn diagram showing the overlap in BINs between the 3 countries.

v <- venn.diagram(list(Pakistan = Pakistan, China = China, South.Africa = South.Africa),
          fill = c("yellow", "red","blue"),

          main = "Species Overlap", main.cex = 2.5,

          cat.cex = 1.5, cex=1.5,

          lwd = 2,

          lty = 'blank',

          filename=NULL)

grid.newpage()

grid.draw(v)
```

###The Venn diagram results were as expected though it was interesting to see how few species intersected between the countries. Only one was amongst all 3! See Figure 3 in .pdf

###Draw a plot to see the BINs mapped in a grid.

#How many records are contain geographic coordinate data?

```
sum(!is.na(Acrididae.bin$lat))

sum(!is.na(Acrididae.bin$lon))
```

#Subset the Acrididae.bin dataset for the country, longitude and latitude columns

```
Country.map <- Acrididae.bin[,c(2,3,4)]
```

#Notice that China has a lot of NA values. If this were to be mapped as is, all the data points would appear on the equator. To prevent that, we need to remove the lines that have NA in them.

```
Country.map <- drop_na(Country.map,lon,lat)
```

#Plot the Values on a graph to see the location

```
plot(Country.map, main = "Species plotted by Country")
```

###As expected the locations of the data points were close together on the whole due to the Pakistan data set. The outlier values belong to South Africa.

###Results and Discussion.

#My main question regarding the species distribution by country and species overlap were proven correct. Pakistan has a lot of species data with all the values populated, allowing for better analysis and understanding of the data. I enjoyed exploring the data. The results of Shannon's and Simpson's indices are still surprising, though it could be due to the small data set size. The ratio of unique BINs to unique species names suggests and verifies my hypothesis that there is geographical separation between the species of the countries with little to no interaction between them. The results found here agree with the papers I cited by Hussain et al [1] and Bam , Addison and Conlong [2]. Acrididae have not been as well studied and categorized by researchers in Africa. However, the country is larger and the habitats for the locusts may be more varied than for the species located in Pakistan.

#For further study on this topic, I would look at the rest of the data concerning Acrididae and see what the spread of species looks like then. It may also be better to draw a graph instead of plotting the latitudes and longitudes.

###Conclusion

#I had hoped to be able to draw a map plotting the species in their geographical locations using longitude and latitude. Unfortunately, I was having trouble with R and was unable to get maps working again. I had spoken to Jacqueline about my earlier R troubles, but they resolved quickly though this persisted. I have discovered that what I thought I knew about R is, not as much as I had thought. I am truly a beginner. This project allowed me to explore the data of an organism of interest and I hope to improve my performance and skills for the next project. Thank you very much

###Acknowledgments

#Dr. Sarah Adamowicz for the excellent class notes and code which were used for the basis of the assignment

# The R cheat cheats which were provided in class were also a valuable resource.

#Jacqueline May for helping me with technical issues which keep on cropping up especially at the last minute. They are to do with the R packages themselves and not the code. She also helped me adapt the API code to download the data by country

###References

#[1] Hussain, Mubashar. (2017). Diversity, distribution and seasonal variations of grasshopper populations in Sialkot, Punjab, Pakistan. Pure and Applied Biology. 6. 10.19045/bspab.2017.600148.

#[2]The Locust Crisis: The World Bank's Response. (2020, July 1). Retrieved September 27, 2020, from https://www.worldbank.org/en/news/factsheet/2020/04/27/the-locust-crisis-the-world-banks-response

#[3] Bam A, Addison P, Conlong D (2020) Acridid ecology in the sugarcane agro-ecosystem in the Zululand region of KwaZulu-Natal, South Africa. Journal of Orthoptera Research 29(1): 9-16. https://doi.org/10.3897/jor.29.34626

#[4]R documentation for vegan package.
https://www.rdocumentation.org/packages/vegan/versions/2.4-2/topics/diversity

#[5]R help for Venn diagram

Figures:

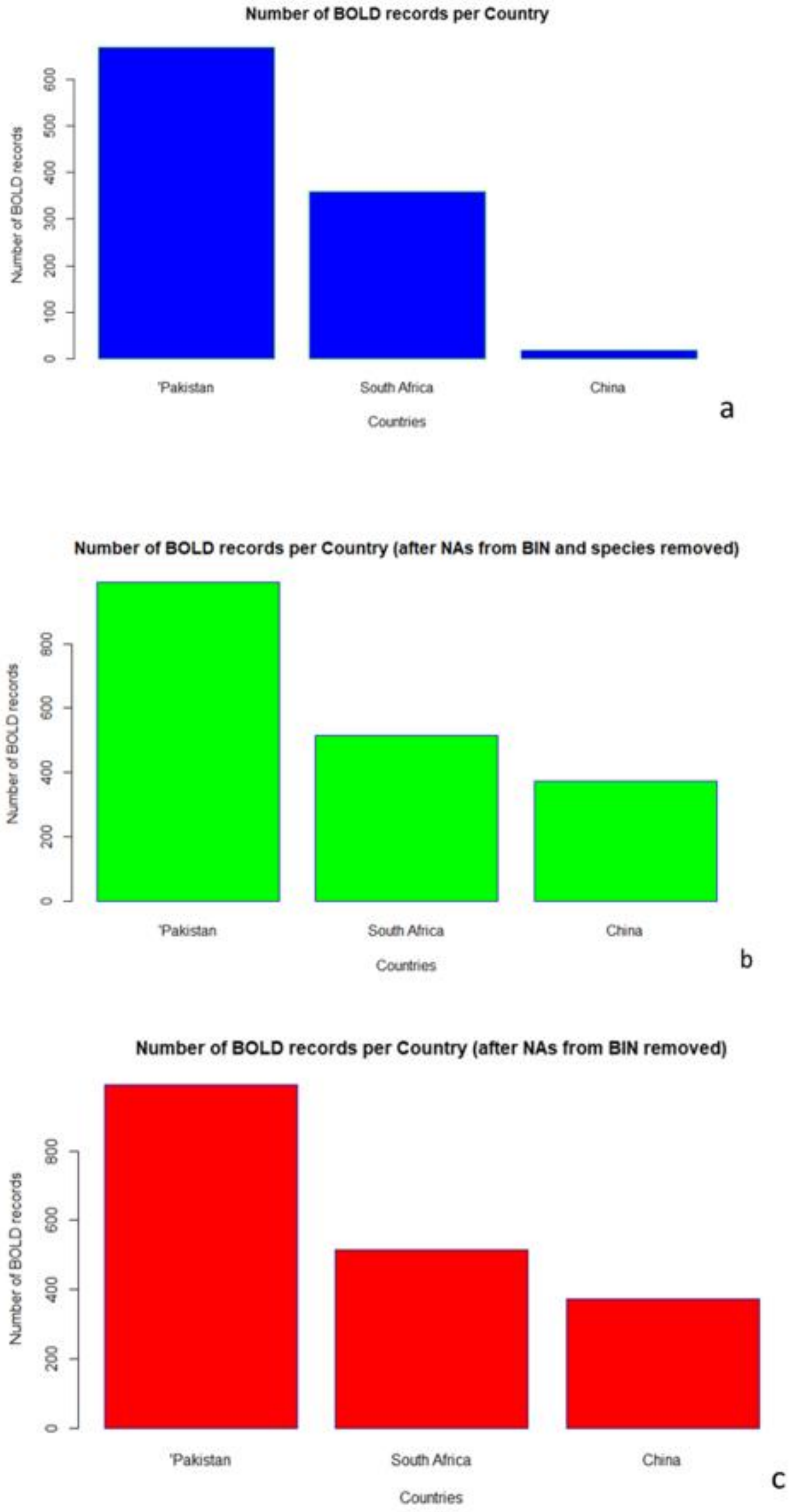
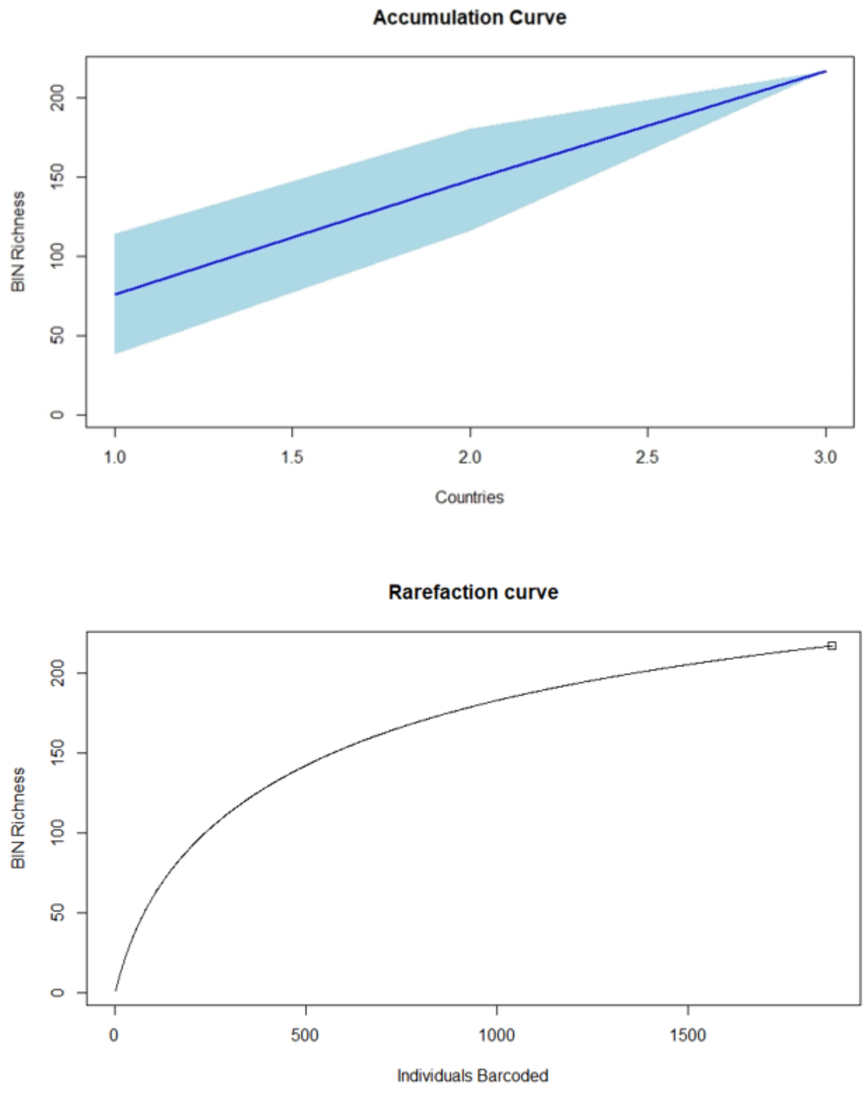


Figure 1: Bar plots of the number of BOLD records per country. (a) Raw data, in blue.(b) After data with NAs were removed from both species names and BINs, in green. (c) After data points where only BIN was NA were removed, in red.



Accumulation Curve

Rarefaction curve

Figure 2: Accumulation curve and Rarefaction curve for the dataset showing the species richness by country and individual respectively

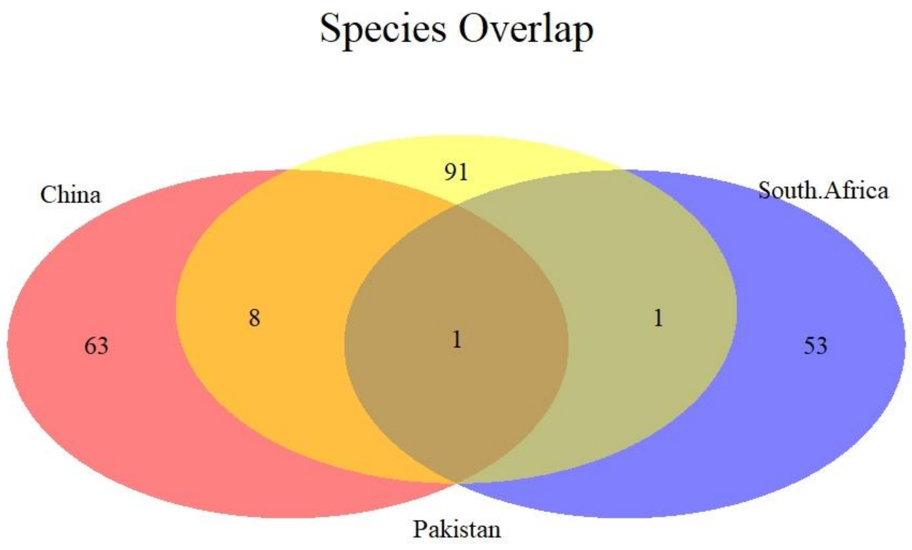


Figure 3: Species Overlap shown by Venn diagram. Very low species overlap. Highest diversity of species in Pakistan. This may be due to rigorous research and data collection of Acrididae in the region.

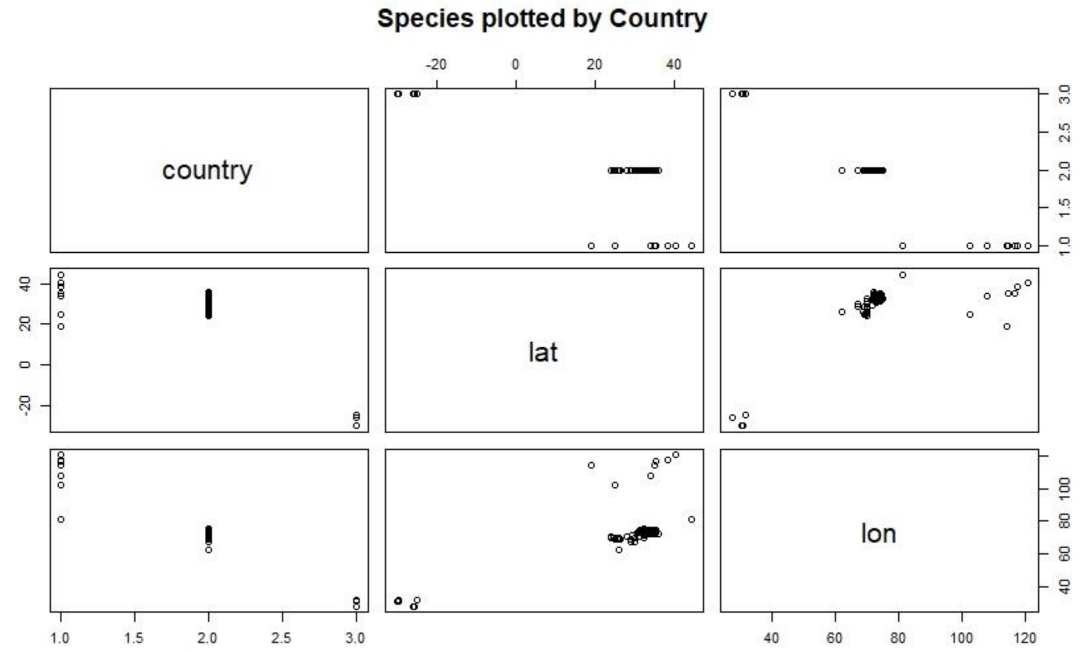


Figure 4: Species plotted by longitude and latitude. A large number of these geographical data points were missing from the data for China. In the Figure most of the points are clustered at the position for Pakistan