#PART 1: Introduction, hypothesis, question

#I chose Bombus (Bumblebee) as my taxonomic group for this project. This early summer, I just came across a blog called “Montréal, officially a Bee City!” and a community science project called “Bumble Bee Watch”, bumblebees have drawn my attention since then. Besides, I barely saw a bee in my home city (Hangzhou, China), but recently, I have seen some bumblebees in Ontario and Quebec. So I wondered if any geographical or environmental factor results in the different distribution of one species between areas. I came up with a hypothesis “environmental diversity between different climate zones leads to contrasting geographic densities of Bombus species”. The eight climate zones are mainly demarcated by latitudes. Therefore, for this hypothesis, I will need to determine whether the latitude is correlated with the distribution of Bombus species.
#Bombus, as a pollinator, plays an essential role in the ecosystem; 265 species of Bombus were identified worldwide (Williams& Jepsen, 2021). Bombus genus is most numerous in the northern cool subtropical and temperate areas, where the latitude is between approximately 23.5° north and 66.5° north (Willams, 1998 & Cameron et al., 2007). Bombus can survive in relatively low temperatures with the heat produced by body metabolism (Heinrich, 2004). By analyzing the Bombus data from the BOLD dataset with R, and comparing the median latitude of each Bombus species to the total count of each species, we should be able to see the distribution patterns (i.e. Latitude) of species.

#PART 2: BOLD data exploration----

```
#First, load the package "tidyverse" and "vegan" that is needed for this project
library("tidyverse")
library("vegan")
library("ggplot2")
```

#I chose genus Bombus as my group of interested for this assignment. I accessed the publicly available data for Bombus genus from BOLD (the Barcode of Life Data Systems), and I downloaded the correspondent Bombus BOLD data as tsv file on Sept 23, 2021.

```
dfBOLDB <- read_tsv(file =
"http://www.boldsystems.org/index.php/API_Public/combined?taxon=Bombus&format=tsv")
```

```
#Then set the file to the right working directory
write_tsv(dfBOLDB, "Bombus_BOLD_data.tsv")
setwd("~/Desktop/6210 Software Tools/Assignement 1")
```

#Here, I used function read_tsv() to read in the Bombus BOLD data file and assigned to a new tidyverse-style data frame called "dfBombus" for downstream data exploration.

```
dfBombus <- read_tsv(file = "Bombus_BOLD_data.tsv")
```

#We can view the "dfBombus" by clicking "dfBombus" in the environment pane of RStudio or with command View(dfBombus). It will generate another window with a data frame being displayed. The data frame contains 9849 observations in rows and 80 variables in columns. Each cell represents a specific value.

```
View(dfBombus)
```

#In order to check the class of this data object, I used command class(dfBombus)

```
class(dfBombus)
```

#Result "tbl" means "tibble", it is a tidyverse style of data frame. The "data.frame" also indicates that it is the correct data structure that I set before, and data frame contains various type of data (numeric, character, etc.).

#After having the basic idea of how this data object structured, we then could check the particular data type of each variable, and have a look at the examples of values by using str(dfBombus)

```
str(dfBombus)
```

#Results show there are mainly 3 data types for variable in this data frame: character, numeric, and logical. Every single variable contains 9849 values in cells.

#Like str(), summary() has the similar function, but it displays a relatively clearer summary on the numeric data and logical data.

```
summary(dfBombus)
```

#For example, variable "lat" stands for the latitude of the location where specimen was collected. "lat" variable contains numeric data, other than the length of variable, summary() will give the statistical distribution information for numeric data: mean, median, maximum and minimum, etc.

#For the logical data, summary() will give the total count of "TURE", "FALSE", and "NA".

#To view the name and its position of each variable in the data frame, I used names(dfBombus) and data.frame(colnames(dfBombus)). There are 80 different variables, each indicates the different character of data collected.

```
names(dfBombus)
data.frame(colnames(dfBombus))
```

#By getting the exact position of each variable, it is much easier for me to subset certain variables into a new object for further study. During subsetting, there is no need to type all the variable names completely (even R has auto-completion, it can still be time consuming), I only need to use the position of each variable.

#For better processing the data required to evaluate my hypothesis, I extra several variables column by indexing by position, and create a new data frame called "dfBdata". "processid" at column 1, "bin_uri" at column 8, "species_name" at column 22, "identification_method" at column 26, "lifestage" at column 39, "lat" at column 47, "lon" at column 48, "country" at column 55, "province_state" at column 56.

```
dfBdata <- dfBombus[, c(1, 8, 22, 26, 39, 47, 48, 55, 56)]
```

```
dfBdata
```

#Because my hypothesis mainly is to test the correlation between the count of species and climate zones (which are classified by latitude). So latitude data is the first that need to be focused on. To ensure the accuracy of latitude data entered, I will check if any unexpected data fall out the range. (the South pole is at -90 depress, North pole is at 90 degrees).

```
Undata <- between(dfBdata$lat, -90, 90)
summary(Undata)
```

#Results shows 4758 data is within the right range, no unexpected data appears (FALSE), and there are 5091 observations have missing value for latitude.

#Histogram is a helpful plot to look at the distribution of latitude. No unexpected data was found in plot either, data range in between 40 degrees and 60 degrees have the higher frequency.

```
hist(dfBdata$lat, main= "Histogram of Latitude", xlab = "Latitude")

summary(dfBdata$lat)
dfBdata$country[which.max(dfBdata$lat)]
dfBdata$country[which.min(dfBdata$lat)]
```

#Above lines show the statistical distribution of latitude value. The mean of latitude is 48.12, and median latitude is 50.21; geographic range of specimens collected is from -43.60 (New Zealand) to 81.83 degrees (Canada).

#To observe which countries are presented, and how many records are there for each country.

```
tBcountry <- table(dfBdata$country)
dim(tBcountry)
sort(tBcountry, decreasing = TRUE)
sort(tBcountry, decreasing = TRUE)[1:10]
```

#Or using piping to reduce the temporary data frame in the R studio environment.

```
dfBdata %>%
  count(country, sort = TRUE)
```

#BOLD data for dombus includes specimens from 64 countries.The country has the most record is Canada with 2359 records.

#I remove all the records that have missing value in "bin_uri" and "species_name", and get the total number of unique BIN and species name.Overall, I found 7230 records have BIN， 231 of them are unique BIN; 8976 records have species name and represent 248 unique species.

```
dim(filter(dfBdata,!is.na(bin_uri)))
length(unique(dfBdata$bin_uri))
dim(filter(dfBdata,!is.na(species_name)))
length(unique(dfBdata$species_name))
```

#It is an interesting observation that there are more species than BINs. One of the reasons is that there are more records with no sequence than records with no species name. In addition, as Sally mentioned in Script5, using interim species names would be another explanation. To point out the interim species name used in BOLD data, we will need to count the number of spaces, dots, and digits (i.e. 0-9) in each species name. A new column "scientific_species_name" is added to the end of "dfBdata", this column contains the count result correspondingly. str_count() in stringr package is used to count the number of matches

```
dfBdata$scientific_species_name <- str_count(string = dfBdata$species_name, pattern =
"[\\s\\.\\d]")
table(dfBdata$scientific_species_name)
sum(dfBdata$scientific_species_name == 1, na.rm = TRUE)
sum(!dfBdata$scientific_species_name == 1, na.rm = TRUE)
```

#A standard scientific species name only has 1 space, so any species name with a count greater than 1 is interim. In total, 8937 records of the species names are scientific species names, and 39 specimens used interim species names when entered into the BOLD system. This result can explain why there are more species than BINs originally.

```
dfB_species_name <- dfBdata[dfBdata$scientific_species_name == 1, ]
unique(dfB_species_name$scientific_species_name)
dim(dfB_species_name)
```

#To remove all the records with interim species names, I use indexing by condition. If the count of variable "scientific_species_name" is 1, all the columns of that row will be kept, and set to a new data frame "dfB_species_name". (Records has NA as species name are still kept into new data frame)

```
dfBombus2 <- dfB_species_name %>%
  filter(!is.na(scientific_species_name)) %>%
  filter(!is.na(bin_uri)) %>%
  print()
```

#All the records that have missing values for species names and BINs are filtered out. Current data frame "dfBombus2" only contains records that has both scientific species name and BIN, which is ideal for downstream analysis.

```
dim(dfBombus2)
length(unique(dfBombus2$species_name))
length(unique(dfBombus2$bin_uri))
```

#After removing all interim species names, I have 206 unique species names and 224 BINs from 6475 specimens records.

```
tBSpecies <- table(dfBombus2$species_name)
sort(tBSpecies, decreasing = TRUE) [1:10]
tBB <- table(dfBombus2$bin_uri)
sort(tBB, decreasing = TRUE) [1:10]
```

#Above lines give me the idea of the top 10 species and BINs. *Bombus terrestris* is the most abundance among 206 species, it has 435 effective records. BOLD:AAB1062 is the most abundance 224 BINs, it has 439 records.

#PART 3: Data analysis and visualization----

#Sally mentioned Barcode Index Numbers (BINs) are unique identifiers for DNA sequence data, and is more promising to use when morphological identification is not easy to employ. Bombus genus is especially difficult to identify, its various patterns on colored band and its tongue structure could lead to misidentification. So I will start to use BINs "bin_uri" as the proxy of species for the following analysis.

#Here, I am setting records into groups of countries, and counting the number of BINs per country, then assigning the count results to "dfB.by.country".

```
dfB.by.country <- dfBombus2 %>%
  filter(!is.na(country)) %>%
  group_by(country) %>%
  count(bin_uri)
```

#As mentioned in the introduction part, literature shows Bombus genus mainly occurs in the area where latitude range from 23.5 to 66.5 degrees. So I retain rows that are located in between the target latitude range, and get the BINs count of each country, then assign them to a new object "dfB.by.country.n".

```
dfB.by.country.n <- dfBombus2 %>%
  filter(!is.na(country)) %>%
  filter(!is.na(lat)) %>%
  filter(lat >= 23.5) %>%
  filter(lat <= 66.5) %>%
  group_by(country) %>%
  count(bin_uri)
```

```
all.equal(dfB.by.country, dfB.by.country.n)
length(unique(dfB.by.country$country))
length(unique(dfB.by.country.n$country))
```

#Comparing the both objects "dfB.by.country" and "dfB.by.country.n", In total 236 records, 28 countries were eliminated from "dfB.by.country". One of the reasons is that some specimens were not collected within the latitude range of northern temperate zone; and another reason is that some records only have country, but not the coordinate information (as Sally mentioned in Script6, these records are mainly mined from GenBank or NCBI directly)

#I want to create a barplot showing the total number of BINs by each country globally. So I decide to using geom_bar()function from the "ggplot2" package, also combine the knowledge learnt from Script7 to better present the plot visually.

```
ggplot(data = dfB.by.country, aes(x = country, y = n, fill = bin_uri)) +
  geom_bar(stat = "identity") +
  labs(x = "Country", y = "Number of BINs", title = "Bombus BINs by Country") + coord_flip() +
theme(plot.title = element_text(hjust = 0.5)) + scale_fill_discrete("BINs") +
  theme(legend.key.size = unit(0.1, 'mm'), legend.text = element_text(size = 3))
```

```
#+ theme(legend.position = "none")
```

#Each species is represented by different colors in the barplot, but because there are over 100 BINs, so the code "theme(legend.key.size = unit(0.1, 'mm'), legend.text = element_text(size = 3))" or "theme(legend.position = "none")" is optional to run. I recommend running the first to change the size of legend, otherwise the plot or legend itself is going to be too small to read.

#In order to observe the BINs difference between whole world and northern temperate zone, I create another bar graph with "dfB.by.country.n" to show the number of BINs per country which is only in the northern temperate zone.

```
ggplot(data = dfB.by.country.n, aes(x = country, y = n, fill = bin_uri)) +
  geom_bar(stat = "identity") +
  labs(x = "Country", y = "Number of BINs", title = "Bombus BINs by County in Northern
Temperate Zone ") + coord_flip() + theme(plot.title = element_text(hjust = 0.5)) +
scale_fill_discrete("BINs") +
  theme(legend.key.size = unit(0.1, 'mm'), legend.text = element_text(size = 3))
```

#When looking at both plots, it is so obvious that Canada has the most BINs richness with numerous BINs. And the top 1 BINs BOLD:AAB1062 is in dark orange, we can see this BINs has appeared in many other countries as well.

#After checking the number of each unique BINs in each country, I want to generate an rarefaction curve and an accumulation curve of Bombus community, to see how well sampled is Bombus in the BOLD database.

#Firsty, I create a new data frame without the latitude restriction and generate the counts of each BIN. Then I use pivot_wider() to reshape data frame into the right format for further analysis with "vegan".

```
dfBINs.count <- dfBombus2 %>%
  group_by(bin_uri) %>%
  count(bin_uri)
dfBINs <- pivot_wider(data = dfBINs.count, names_from = bin_uri, values_from = n)
```

#I create a rarefaction curve by using rarecurve() function in "vegan" package, and treat all records as one site globally.

```
RC <- rarecurve(dfBINs, xlab = "Number of Individuals Barcoded", ylab = "BIN Richness")
```

#In order to draw accumulation curve, I create and reconstruct a new data frame "dfBINs.country" with counts of each unique BIN in each country

```
dfBINs.country.count<- dfBombus2 %>%
  filter(!is.na(country)) %>%
  group_by(country, bin_uri) %>%
  count(bin_uri)
dfBINs.country <- pivot_wider(data = dfBINs.country.count, names_from = bin_uri, values_from
= n)
```

```
#The NA in cell need to be replaced with 0, which represent the specific BIN has not been collected in the specific country yet.
dfBINs.country <- dfBINs.country %>%
  replace(is.na(.), 0)

#"Vegan"'s specaccum() function cannot run on character data, so the first column "country" need to be set as row names.
dfBINs.country <- dfBINs.country %>%
  remove_rownames %>%
  column_to_rownames(var = "country")

#Right now, it is all set the plot a site-based species accumulation curve, we are able to see how BINs accumulate as more country being sampled.
AC <- specaccum(dfBINs.country)
plot(AC, xlab = "Number of Countries Sampled", ylab = "BIN Richness", ci.col = "blue")
#Both rarefaction curve and accumulation curve are nicely plotted, plots will be discussed in part 4.

#Next, to determine the correlation between latitude and species (BINs) for Bombus, I will need to get the median latitude of each BIN first. Here I use median instead of the mean of latitudes because literature has shown that several species of Bombus were introduced from its origin to another country, and two locations could be half-world apart. In this situation, the mean is not as accurate as the median, which can demonstrate the latitude information of each BIN without many human factors.
#I summary the latitude median among specimens of each species, and set to a new data object "dfMedianLat'
dfMedianLat<- dfBombus2 %>%
  group_by(bin_uri) %>%
  summarise(MedianLat = median(lat, na.rm = TRUE)) %>%
  print()

#We have the number of each BIN in the data frame "dfBINs.count", I will combine the "dfMedianLat" and "dfBINs.count" together by the following steps, that we will have both the latitude and count information of each BIN in one data frame "dfLat.BIN"
#Firstly subset the number of species column as new data frame "dfno"
dfno<- dfBINs.count[2]
dfLat.BIN <- bind_cols(dfMedianLat, dfno)
dfLat.BIN <- dfLat.BIN %>%
  arrange(desc(MedianLat))
#BOLD:AAF6202 is the northernmost BIN that has coordinate, there are 6 valid records, and the median latitude among them is 69.04 degrees.

#Many BINs in "dfLat.BIN" do not have latitude, mainly because these data are from GenBank or NCBI, unfortunately they have to be removed.
```

```
dfLat.BIN.na.rm<- filter(dfLat.BIN, !is.na(dfLat.BIN$MedianLat))
summary(dfLat.BIN.na.rm)
#After removing BINs that median of latitude is missing, only 141 BINs are remained. Final executable data frame is called "dfLat.BIN.na.rm".

#I am going to create a scatterplot for median latitude and BIN richness (both are continuous variables), to see their degree of correlation. I use geom_point for the scatterplot; use geom_rug() to show the distribution of dots on both x and y axis.
ggplot(data = dfLat.BIN.na.rm, mapping = aes(x = MedianLat, y = n , color = bin_uri)) +
  geom_point(size = 3) +
  geom_rug() +
  labs(title = "Median Latitude of BIN vs. BIN", x = "Median Latitude", y = "Number of BIN") +
  scale_colour_discrete("Barcode Index Number (BIN)") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "none")
 #+ theme(legend.key.size = unit(0.5, 'mm'), legend.text = element_text(size = 3)) this line is optional to run if need to see the legend.

#I want to draw another scatter plot for latitude and BINs which are in the northern hemisphere, and I use stat_smooth() to add  possible linear regression line for latitude and BIN's count.
dfLat.BIN.na.rm1 <- dfLat.BIN.na.rm[1:134,]

ggplot(data = dfLat.BIN.na.rm1 , mapping = aes(x = MedianLat, y = n , color = bin_uri)) +
  geom_point(size = 3) +
  geom_rug() +
  stat_smooth(method='lm', formula = y ~ x, colour = 'black') +
  labs(title = "Latitude vs. BIN", x = "Median Latitude", y = "BIN count") +
  scale_colour_discrete("BINs") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "none")

#Here, I'll repeat the procedure from 178 to 216, instead of using BIN, this time is to check the median latitude among all the records of each species. I create a scatter plot to check the correlation between latitude and number of Bombus species.
dfMedianLat.S<- dfBombus2 %>%
  group_by(species_name) %>%
  summarise(MedianLat = median(lat, na.rm = TRUE))

dfB.count <- dfBombus2 %>%
  group_by(species_name) %>%
  count(species_name)

dfno.s<- dfB.count[2]
```

```
dfLat.S <- bind_cols(dfMedianLat.S, dfno.s)
dfLat.S <- dfLat.S %>%
  arrange(desc(MedianLat))

dfLat.S.na.rm<- filter(dfLat.S, !is.na(dfLat.S$MedianLat))

ggplot(data = dfLat.S.na.rm, mapping = aes(x = MedianLat, y = n , color = species_name)) +
  geom_point(size = 3) +
  geom_rug() +
  labs(title = "Median Latitude of Bombus Species vs. Number of Species", x = "Median Latitude", y = "Number of Species") +
  scale_colour_discrete("Species Name") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "none")
#+ theme(legend.key.size = unit(0.5, 'mm'), legend.text = element_text(size = 3)) this line is optional to run if need to see the legend.
```

#As go over studies, I found *Bombus atratus, Bombus morio, and Bombus brasiliensis* belong to the very few Bombus species found in South America; *Bombus atratus, Bombus morio* are Argentine native species; *Bombus brasiliensis* is more common on mountaintops of southeastern Brazil. (Plischuk, Lange, 2009 & Santos et al., 2015) So I want to have another scatter plot that only has Bombus species from the northern hemisphere.

```
dfLat.S.na.rm1 <- dfLat.S.na.rm%>%
  filter(!species_name == "Bombus atratus") %>%
  filter(!species_name == "Bombus brasiliensis") %>%
  filter(!species_name == "Bombus morio")

ggplot(data = dfLat.S.na.rm1, mapping = aes(x = MedianLat, y = n , color = species_name)) +
  geom_point(size = 3) +
  geom_rug() +
  stat_smooth(method='lm', formula = y ~ x, colour = 'black') +
  labs(title = "Latitude vs. Species Richess in Northern Hemisphere", x = "Median Latitude", y = "Number of Species") +
  scale_colour_discrete("Species Name") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme(legend.position = "none")
```

#Other than looking at graphs, I also want to conduct linear regression analysis on the latitude and count of each BIN/species. Because latitude is a special data, "+" is for the northern hemisphere, "-" is used to indicate the degrees for the southern hemisphere. To study my hypothesis and question of Bombus, I prefer only use positive latitude from the northern hemisphere to do linear regression analysis.

```
cor.test(dfLat.BIN.na.rm1$MedianLat, dfLat.BIN.na.rm1$n)
lrB <- lm(n ~ MedianLat, data = dfLat.BIN.na.rm1)
```

```
lrB
summary(lrB)
#Results show p-value: 0.0158, it is less than 5%. In the northern hemisphere, the correlation between latitude and the number of BINs is significant. y = -13.274 + 1.145 x

cor.test(dfLat.S.na.rm1$MedianLat, dfLat.S.na.rm1$n)
lrS <- lm(n ~ MedianLat, data = dfLat.S.na.rm1)
lrS
summary(lrS)
#The p-value is 0.0632, no significant correlation is found between latitude and the number of each Bombus species.
```

#PART 4: Results and discussion----

For this assignment, my question is whether the latitude is correlated with the distribution of Bombus species. I analyzed with Bombus species and used Barcode Index Number (BIN) as the proxy of Bombus species since Bombus is not easy to identify morphologically. Candan has contributed the most abundance of Bombus records to the BOLD database, and the top 1 species *Bombus terrestris*. However, many records were removed from the subsequent studies because they were mined from GeneBank and NCBI directly to BOLD; they don't have coordinates (latitude). After removing all the invalid data, I plotted the rarefaction curve using the global level as one site and the accumulation curve using countries as sites to determine if the sample size and sample completeness are still appropriate for my question. Both curves show a steeper slope as the Bombus was just being collected. As more individuals of species/BIN were collected, the rate in the rarefaction curve started to slow down but not level to asymptote. The accumulation curve shows records from different countries are still needed; as more countries are sampled, more unique species will be found (Figure 1). Especially after filtering out BOLD data without coordinates information, the sample size is not large enough for a comprehensive conclusion. However, it is still possible to have a general concept of the correlation between the latitude and species/BIN. Literature shows Bombus mainly appear in the north temperate zone (Willams, 1998 & Cameron et al., 2007). Several species are found in South America (Plischuk, Lange, 2009 & Santos et al., 2015), and very few species were introduced to Australia and New Zealand artificially for pollination. (Stewart et al., 2010 & Ashley et al., 2019). Latitude is an interesting numeric data, a positive number indicates the northern hemisphere, and a negative number indicates the southern hemisphere. During the study, I created scatterplots and conducted linear regression analysis for Bombus species/BIN in the whole world and Bombus species/BIN in the northern hemisphere, respectively. Both scatter plots (Figure 2, Figure 3) show the unique Bombus species/BIN is more intensive between 40 degrees north and 60 degrees north, and the richness of each species/BIN is higher in this latitudinal range. Meanwhile, the linear regression analysis indicates the correlation between latitude and the number of BINs is significant in the northern hemisphere (p-value: 0.0158).

In order to get a more comprehensive and more persuasive conclusion in the future study, more countries will need to be sampled. Especially when entering specimen data into BOLD, the coordinates information must be included; this will help to increase study sample size and minimized bias.

#PART 5: Figures----

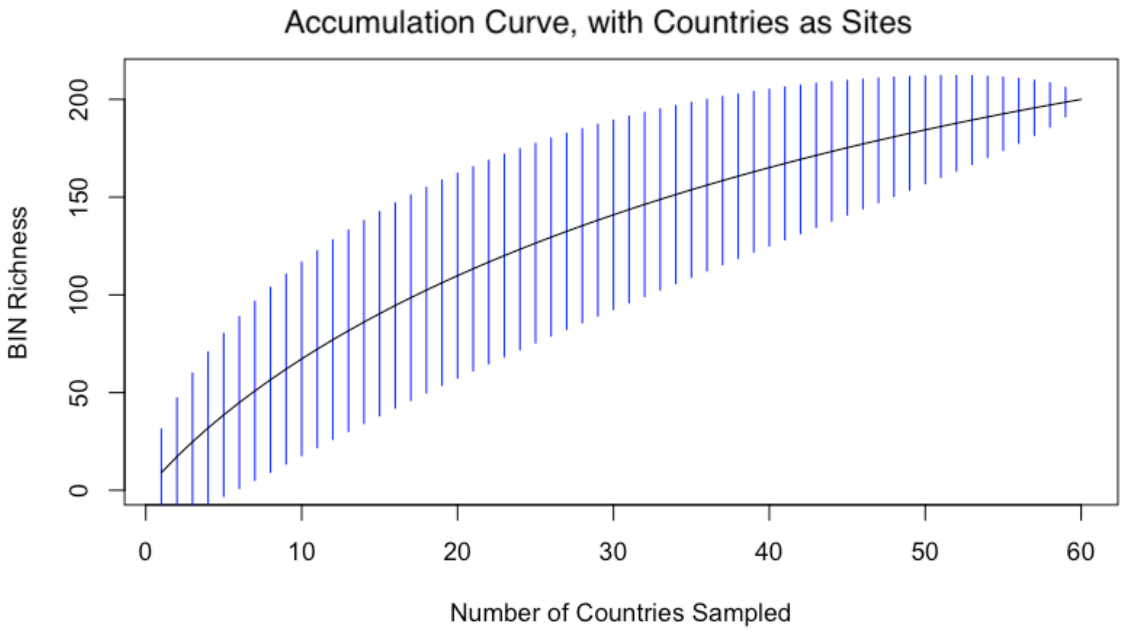

Figure 1.
Bombus BINs accumulation curve using countries as sites. As more countries will be sampled in the future, more unique Bombus BIN/species will be determined.

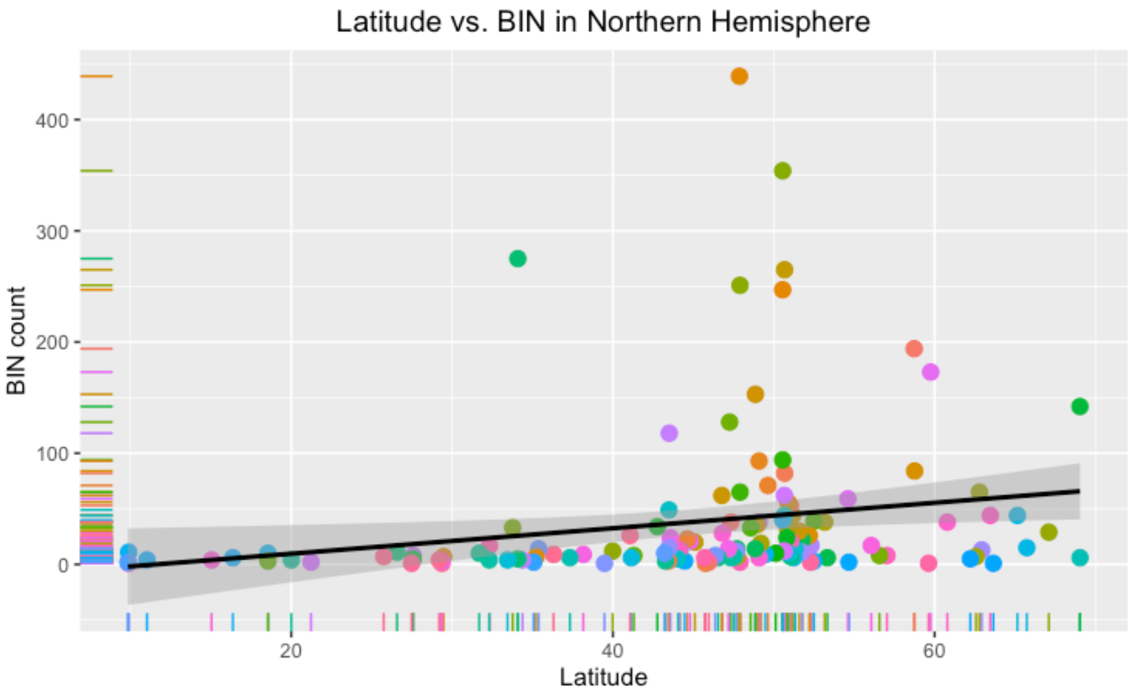

Figure 2.
The scatterplot of latitude and the count of each unique BIN in the northern hemisphere, different colored dots represent the different BINs, rug on the x and y-axis shows the distribution on dots. (Legend section is removed from the plot, there are over 100 unique BINs. If you want to see the list of BINs, there is a line of code in the R script, line 205.)

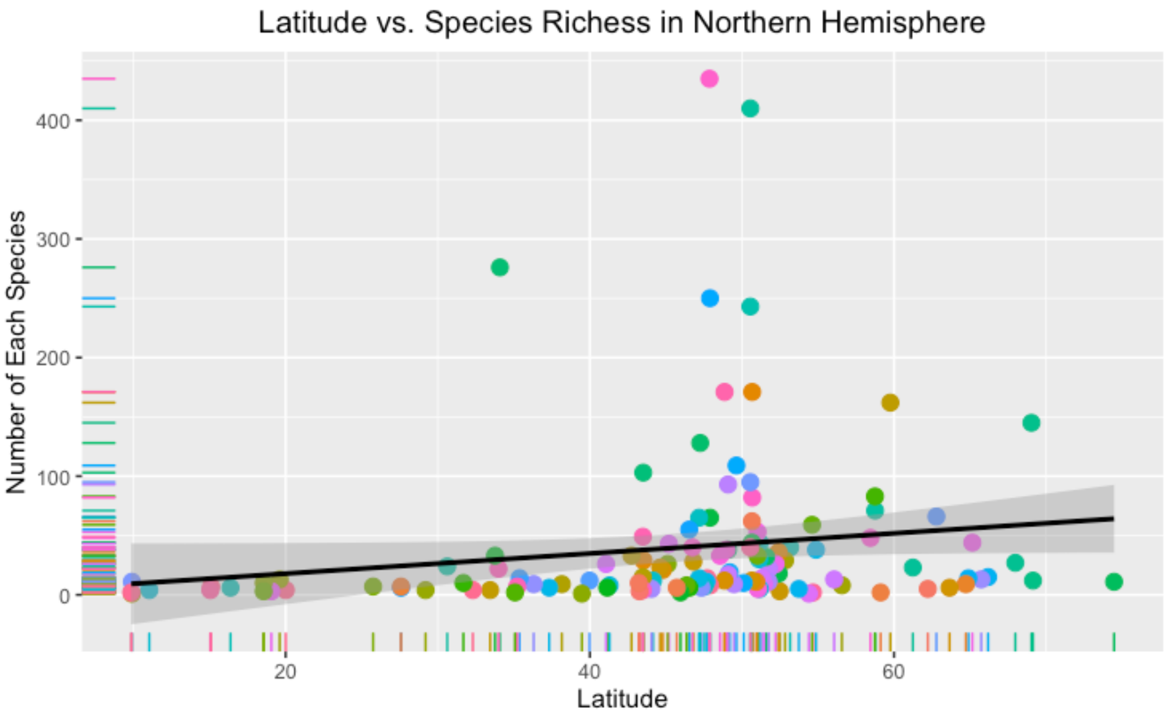

Figure 3.
The scatterplot of latitude and the count of each unique species in the northern hemisphere, different colored dots represent the different BINs, rug on the x and y-axis shows the distribution on dots.
The dots are much denser between 40 degrees and 60 degrees north of latitude, and the count of each species is also higher.

#PART 6: Acknowledgements----

#Abdallah Meknas
#Exchanged the general idea of hypothesis and question of the assignment.
#Provided me information on how to change the setting of ggplot legend, with very useful link https://www.datanovia.com/en/blog/ggplot-legend-title-position-and-labels/

#Jacqueline May
#Really appreciate the help session and 1 on 1 opportunity, helped me with geom_smooth() function and liner regression analysis.

#PART 7: References----

#https://github.com/perlatex/R_for_Data_Science/blob/master/index.Rmd

#https://medium.com/@smitpate08/how-to-easily-find-column-and-row-index-numbers-in-r-f720c435730

#https://stackoverflow.com/questions/24201222/compute-correlation-in-r-between-two-columns-from-different-data-frame

#https://www.datanovia.com/en/blog/ggplot-legend-title-position-and-labels/

#https://www.geeksforgeeks.org/check-if-a-numeric-value-falls-between-a-range-in-r-programming-between-function/

#https://www.jianshu.com/p/9f87025e8ab5

#Ashley, K., Sainsbury, J., McBrydie, H., Robertson, A. W., & Pattemore, D. (2019). A scientific note on range expansion of a sedentary bumble bee (Bombus hortorum) in New Zealand. Apidologie, 50(1), 11–13. https://doi.org/10.1007/s13592-018-0613-z

#Cameron, S.A., Hines, H.M. and Williams, P.H. (2007), A comprehensive phylogeny of the bumble bees (Bombus). Biological Journal of the Linnean Society, 91: 161-188. https://doi.org/10.1111/j.1095-8312.2007.00784.x

#Plischuk, S., & Lange, C. E. (2009). Invasive Bombus terrestris (Hymenoptera: Apidae) parasitized by a flagellate (Euglenozoa: Kinetoplastea) and a neogregarine (Apicomplexa: Neogregarinorida). Journal of Invertebrate Pathology, 102(3), 263–265. https://doi.org/10.1016/j.jip.2009.08.005

#Santos Júnior J. E., Santos F.R., Silveira F. A. (2015) Hitting an Unintended Target: Phylogeography of Bombus brasiliensis Lepeletier, 1836 and the First New Brazilian Bumblebee Species in a Century (Hymenoptera: Apidae). PLOS ONE 10(5): e0125847. https://doi.org/10.1371/journal.pone.0125847

#Stewart, L. C., Hale, R. J., & Hale, M. L. (2010). Species-specific primers for the molecular identification of cryptic bombus species in new zealand.Conservation Genetics, 11(3), 1207-1209. doi:http://dx.doi.org.subzero.lib.uoguelph.ca/10.1007/s10592-009-9920-2

#Willams PH (1998) An annotated checklist of bumble bees with an analysis of patterns of description (Hymenoptera: Apidae, Bombini). Bulletin of The Natural History Museum (Entomology) 67: 79–152.

#Williams, P. & Jepsen, S. (2021). IUCN Bumblebee Specialist Group Annual Report 2020.