MSB7102 Mini-project, semester I, 2021

OKWIR JULIUS

Student No: 2000720621 Reg No: 2020/HD07/20621U

Contents

1 Import data and provide descriptive summaries and statistics

1.1 Load packages to be used and set the plot theme

```
library(tidyverse)
library(phyloseq)
library(DESeq2)
library(ggthemes)
theme_set(theme_light())
```

1.2 Import the data

The datasets were imported using the read_csv for the csv file and read_tsv for the text file. The first row of the text file was skipped using the skip argument.

```
# subject data
sample_data <- read_csv("diabimmune_16s_t1d_metadata.csv")

# otu abundance and taxonomy data
otu_taxonomy_data <- read_tsv("diabimmune_t1d_16s_otu_table.txt", skip = 1)
```

1.3 Descriptive summaries of the subject data

Dimensions

```
dim(sample_data)
```

```
## [1] 777   6
```

Data structure

```
str(sample_data)
```

```
## spec_tbl_df [777 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Sample_ID        : chr [1:777] "G36449" "G36034" "G36993" "G35523" ...
##  $ Subject_ID       : chr [1:777] "E001463" "E001463" "E001463" "E001463" ...
##  $ Case_Control     : chr [1:777] "control" "control" "control" "control" ...
##  $ Gender           : chr [1:777] "male" "male" "male" "male" ...
```

```
##  $ Delivery_Route   : chr [1:777] "vaginal" "vaginal" "vaginal" "vaginal" ...
##  $ Age_at_Collection: num [1:777] 62 82 124 153 187 213 243 276 303 366 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   Sample_ID = col_character(),
##   ..   Subject_ID = col_character(),
##   ..   Case_Control = col_character(),
##   ..   Gender = col_character(),
##   ..   Delivery_Route = col_character(),
##   ..   Age_at_Collection = col_double()
##   .. )
```

Data summary

```
summary(sample_data)
```

```
##   Sample_ID          Subject_ID        Case_Control          Gender
##  Length:777         Length:777         Length:777         Length:777
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Delivery_Route     Age_at_Collection
##  Length:777         Min.   :   6.0
##  Class :character   1st Qu.: 229.0
##  Mode  :character   Median : 452.0
##                     Mean   : 482.9
##                     3rd Qu.: 702.0
##                     Max.   :1233.0
```

Number of subjects

```
length(unique(sample_data$Subject_ID))
```

```
## [1] 33
```

1.4 Exploring subject data

Head and Tail

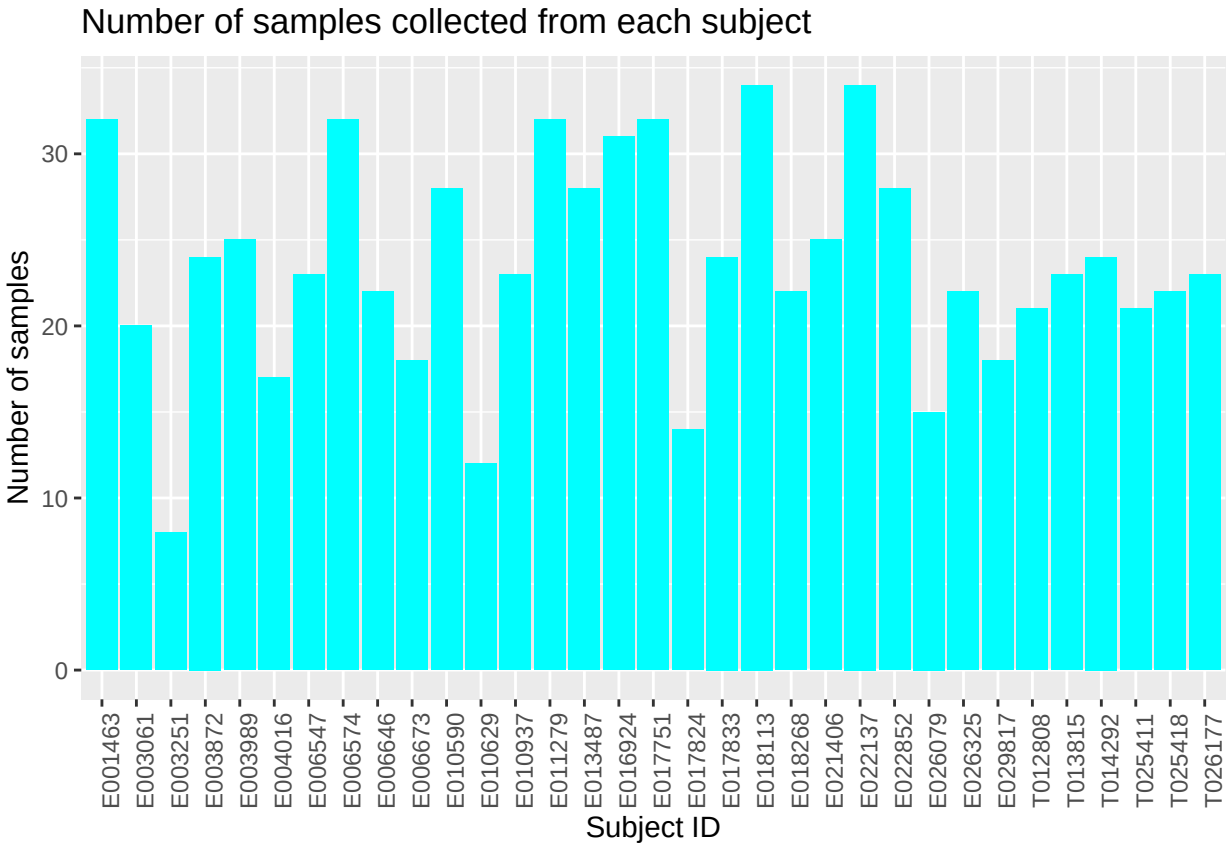
```
head(sample_data); tail(sample_data)
```

```
## # A tibble: 6 x 6
##   Sample_ID Subject_ID Case_Control Gender Delivery_Route Age_at_Collection
##   <chr>     <chr>      <chr>        <chr>  <chr>                       <dbl>
## 1 G36449    E001463    control      male   vaginal                        62
## 2 G36034    E001463    control      male   vaginal                        82
## 3 G36993    E001463    control      male   vaginal                       124
## 4 G35523    E001463    control      male   vaginal                       153
## 5 G36450    E001463    control      male   vaginal                       187
## 6 G36028    E001463    control      male   vaginal                       213
```

```
## # A tibble: 6 x 6
##   Sample_ID Subject_ID Case_Control Gender Delivery_Route Age_at_Collection
##   <chr>     <chr>      <chr>        <chr>  <chr>                       <dbl>
## 1 G36938    T026177    control      female vaginal                       570
## 2 G36936    T026177    control      female vaginal                       592
## 3 G36937    T026177    control      female vaginal                       646
## 4 G35535    T026177    control      female vaginal                       677
## 5 G35536    T026177    control      female vaginal                       703
## 6 G35537    T026177    control      female vaginal                       729
```

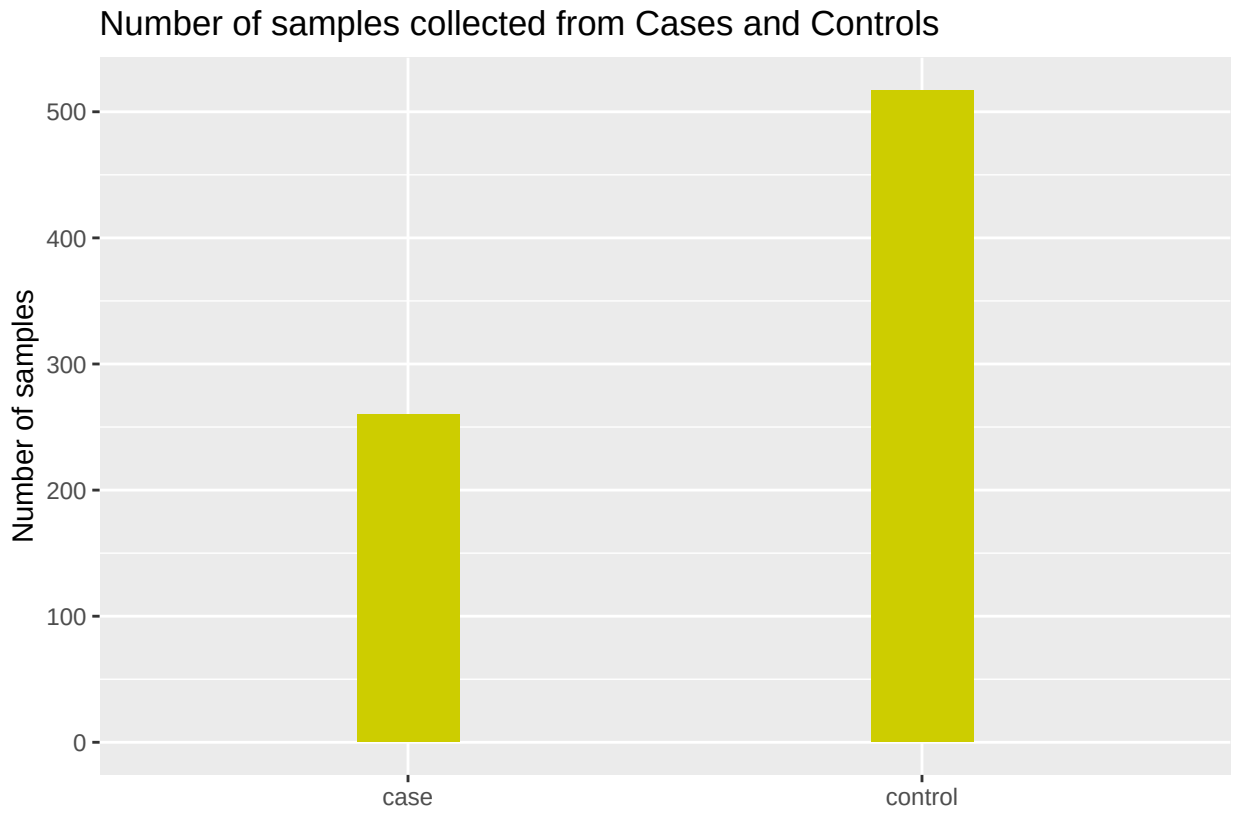
Number of samples collected by subject.
The plot indicates that frequency of sample collection from the subjects was not uniform.

```
ggplot(sample_data, aes(Subject_ID)) +
  geom_bar(fill = "cyan") +
  theme(axis.text.x = element_text(angle = 90)) +
  labs(title = "Number of samples collected from each subject", x = "Subject ID", y = "N
```



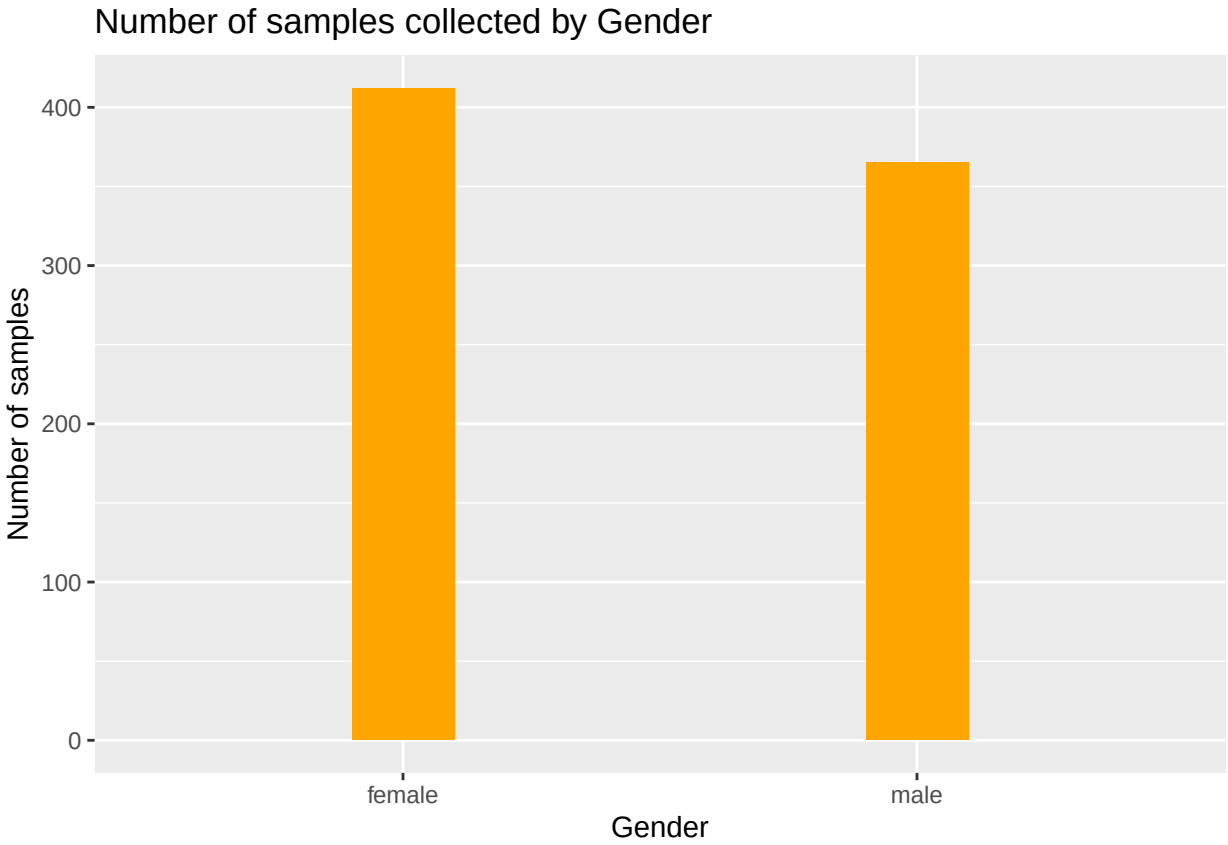
Number of samples from cases and controls.
The frequency of sample collection from the subjects that served as controls was higher than that from those that served as cases

```
ggplot(sample_data, aes(Case_Control)) +
  geom_bar(fill = "yellow3", width = 0.2) +
  labs(title = "Number of samples collected from Cases and Controls",
       x = "", y = "Number of samples")
```



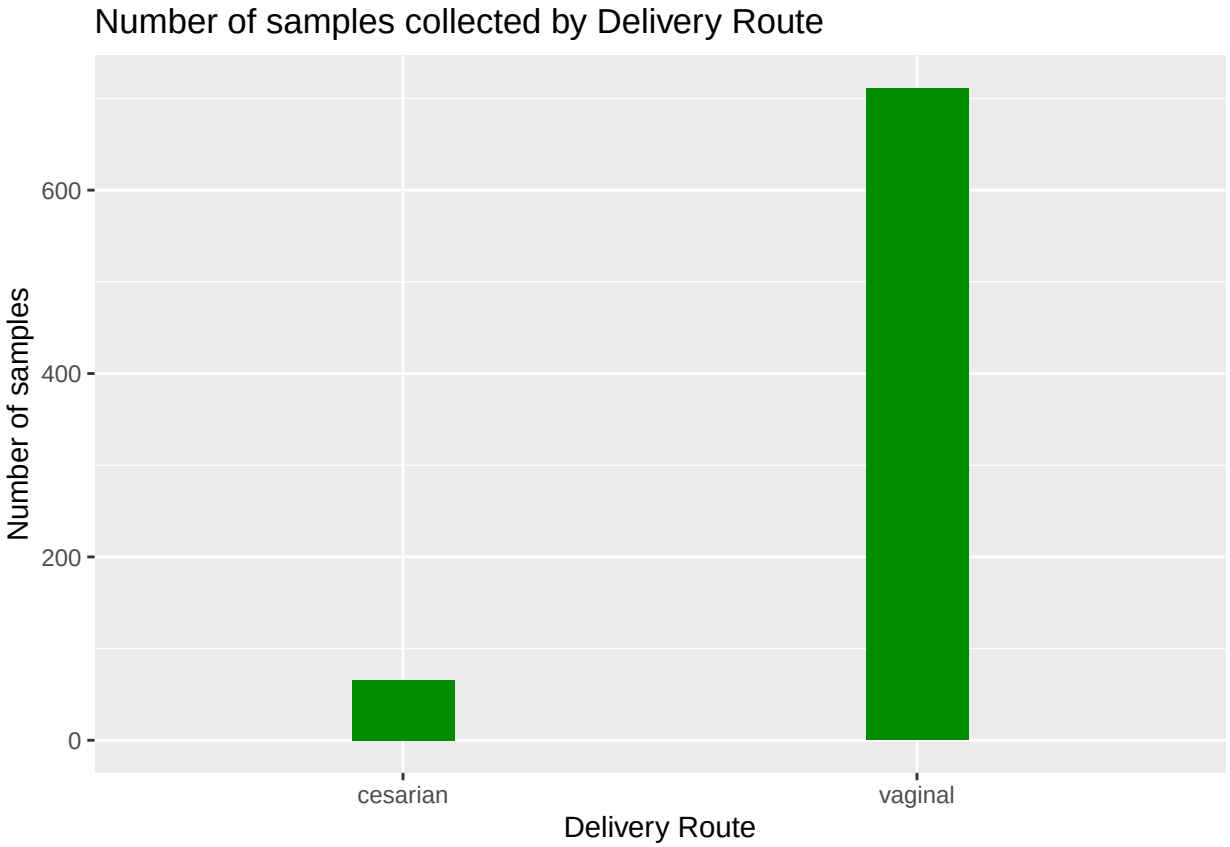
Number of samples collected by gender.
Slightly more samples were obtained from the female subjects than males.

```
ggplot(sample_data, aes(Gender)) +
  geom_bar(fill = "orange", width = 0.2) +
  labs(title = "Number of samples collected by Gender",
       x = "Gender", y = "Number of samples")
```



Number of samples by delivery route.
The frequency of samples obtained from subjects born via the vaginal canal was higher than those from subjects born via cesarian section.

```
ggplot(sample_data, aes(Delivery_Route)) +
  geom_bar(fill = "green4", width = 0.2) +
  labs(title = "Number of samples collected by Delivery Route",
       x = "Delivery Route", y = "Number of samples")
```



1.5 Chi-squared Test for Independence

Chi-squared test for Independence was used to examine whether there is significant association between disease status and other variables. Contingency tables were generated from these variables and used to perform the tests.
The hypotheses were set as follows:
Null hypothesis(H0): The two categorical variables are independent and there is no association between them. Alternative hypothesis(H1):The two categorical variables are dependent and there is an association between them.

Disease status and Gender

```
chisq.test(table(sample_data$Case_Control, sample_data$Gender))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(sample_data$Case_Control, sample_data$Gender)
## X-squared = 0.30687, df = 1, p-value = 0.5796
```

The p-value = 0.5796 and it is greater than 0.05. The null hypothesis was accepted. The disease status and Gender are independent and there is no significant relationship between them.

Disease status and Delivery mode

```
chisq.test(table(sample_data$Case_Control, sample_data$Delivery_Route))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(sample_data$Case_Control, sample_data$Delivery_Route)
## X-squared = 34.649, df = 1, p-value = 3.949e-09
```

The p-value = 3.949e-09 and is less than 0.05. The null hypothesis was rejected in favor of the alternative hypothesis. The disease status and Delivery mode are dependent. Therefore, there is a significant relationship between them.

1.6 Student's t.test for independence

Disease status vs age
A students t-test was performed to check for association between disease status and age.

```
t.test(table(sample_data$Case_Control, sample_data$Age_at_Collection))
```

```
##
##  One Sample t-test
##
## data:  table(sample_data$Case_Control, sample_data$Age_at_Collection)
## t = 32.788, df = 1085, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.6726527 0.7582865
## sample estimates:
## mean of x
## 0.7154696
```

The p-value < 2.2e-16. The null hypothesis was rejected in favor of the alternative hypothesis. The disease status Age are dependent. Therefore, there is a significant relationship between them.

2 Create a phyloseq object

Examine the otu_taxonomy_data
The `otu_taxonomy_data` contains both the OTU abundance data and the taxonomy data. The data was split up to create two data sets. The first containing the OTU abundance and the second containing the taxonomy information.

```
# dimensions
dim(otu_taxonomy_data)
```

```
## [1] 2240  779
```

2.1 Extract OTU abundance data

The OTU abundance data ranges from column 1 to column 778 of the `otu_taxonomy_data`.

```
# select only the otu abundance data
otu_data <- select(otu_taxonomy_data, 1:778)

# convert OTU IDs to row names
otu_data <- otu_data %>% column_to_rownames("#OTU ID")

# convert otu abundance data into a matrix
otu_data <- as.matrix(otu_data)

# class and dimensions
class(otu_data); dim(otu_data)
```

```
## [1] "matrix" "array"
```

```
## [1] 2240  777
```

2.2 Extract taxonomy data

The taxonomy data is present in the last column(779) of the `otu_taxonomy_data`.

```
# select the taxonomy data
taxonomy_data <- select(otu_taxonomy_data, 779)

# create a vector of taxa names
taxa_names <- c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species")
```

```r
# place all values under respective taxa names
taxonomy_data <- separate(taxonomy_data, col = "ConsensusLineage", into = taxa_names, se

# remove letters and underscores
taxonomy_data <- apply(taxonomy_data, 2, str_remove_all, "[a-z]__")

# remove white spaces
taxonomy_data <- apply(taxonomy_data, 2, str_remove_all, " ")

# create a tibble
taxonomy_data <- as_tibble(taxonomy_data)

# assign OTU row names to taxonomy data
rownames(taxonomy_data) <- rownames(otu_data)

# convert into a matrix
taxonomy_data <- as.matrix(taxonomy_data)

# class and head of the matrix
class(taxonomy_data); head(taxonomy_data)
```

```
## [1] "matrix" "array"
```

```
##         Kingdom    Phylum           Class                  Order
## 4333897 "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Enterobacteriales"
## 190162  "Bacteria" "Firmicutes"     "Clostridia"           "Clostridiales"
## 134726  "Bacteria" "Firmicutes"     "Bacilli"              "Lactobacillales"
## 679245  "Bacteria" "Firmicutes"     "Bacilli"              "Lactobacillales"
## 289734  "Bacteria" "Firmicutes"     "Clostridia"           "Clostridiales"
## 302049  "Bacteria" "Firmicutes"     "Clostridia"           "Clostridiales"
##         Family               Genus           Species
## 4333897 "Enterobacteriaceae" ""              ""
## 190162  "Lachnospiraceae"    "Blautia"       ""
## 134726  "Lactobacillaceae"   "Lactobacillus" ""
## 679245  "Lactobacillaceae"   "Lactobacillus" ""
## 289734  "Lachnospiraceae"    ""              ""
## 302049  "Lachnospiraceae"    "Blautia"       ""
```

2.3 Set Sample_ID as rownames for the sample data

```
sample_data <- sample_data %>%
  column_to_rownames("Sample_ID")
```

2.4 Create OTU table, taxonomy table, and sample table

```
# otu table
OTU <- otu_table(otu_data, taxa_are_rows = TRUE)

# taxonomy table
TAX <- tax_table(taxonomy_data)

# sample table
samples <- sample_data(sample_data)
```

2.5 Merge OTU table, taxonomy table, and sample table to create phyloseq object

```
diabimmune <- phyloseq(OTU, TAX, samples)
```

2.6 Explore the phyloseq object

```
# phyloseq object
diabimmune
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 2240 taxa and 777 samples ]
## sample_data() Sample Data:       [ 777 samples by 5 sample variables ]
## tax_table()   Taxonomy Table:    [ 2240 taxa by 7 taxonomic ranks ]
```

```
# class
class(diabimmune)
```

```
## [1] "phyloseq"
## attr(,"package")
## [1] "phyloseq"
```

```
# rank names
rank_names(diabimmune)
```

```
## [1] "Kingdom" "Phylum"  "Class"   "Order"   "Family"  "Genus"   "Species"
```

```
# variables
sample_variables(diabimmune)
```

```
## [1] "Subject_ID"        "Case_Control"      "Gender"
## [4] "Delivery_Route"    "Age_at_Collection"
```

3 Generate Alpha diversity plots and Ordination plots

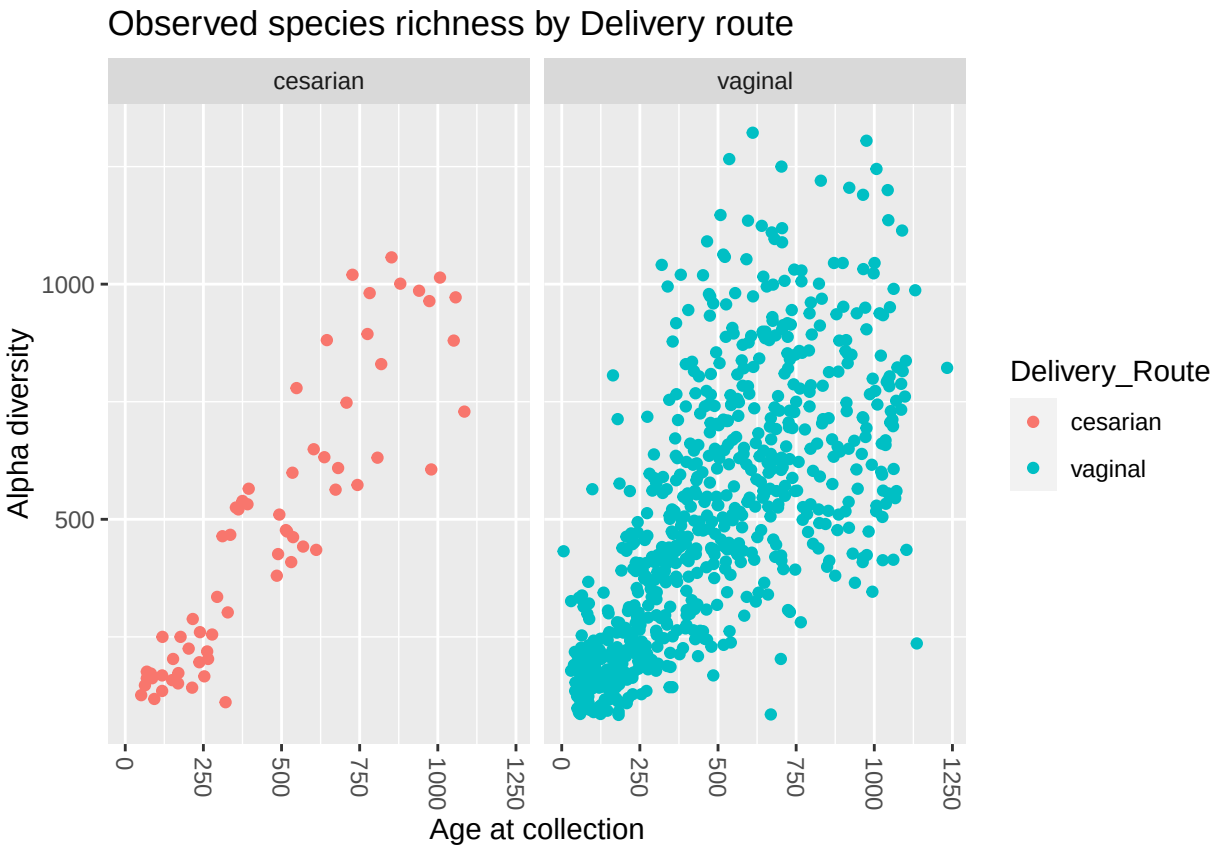
3.1 Alpha diversity plots

The Alpha diversity refers to the diversity within a particular area or ecosystem and is usually expressed by the number of species in that ecosystem. The `Observed` alpha diversity measure was used to examine the species richness by delivery route, gender and disease status.

3.1.1 Observed species richness by Delivery route

Generally there is a higher species richness observed in samples obtained from subjects born via the vaginal canal than those born via cesarian section. The species richness from these samples also increased with increase in the age of the study subjects.

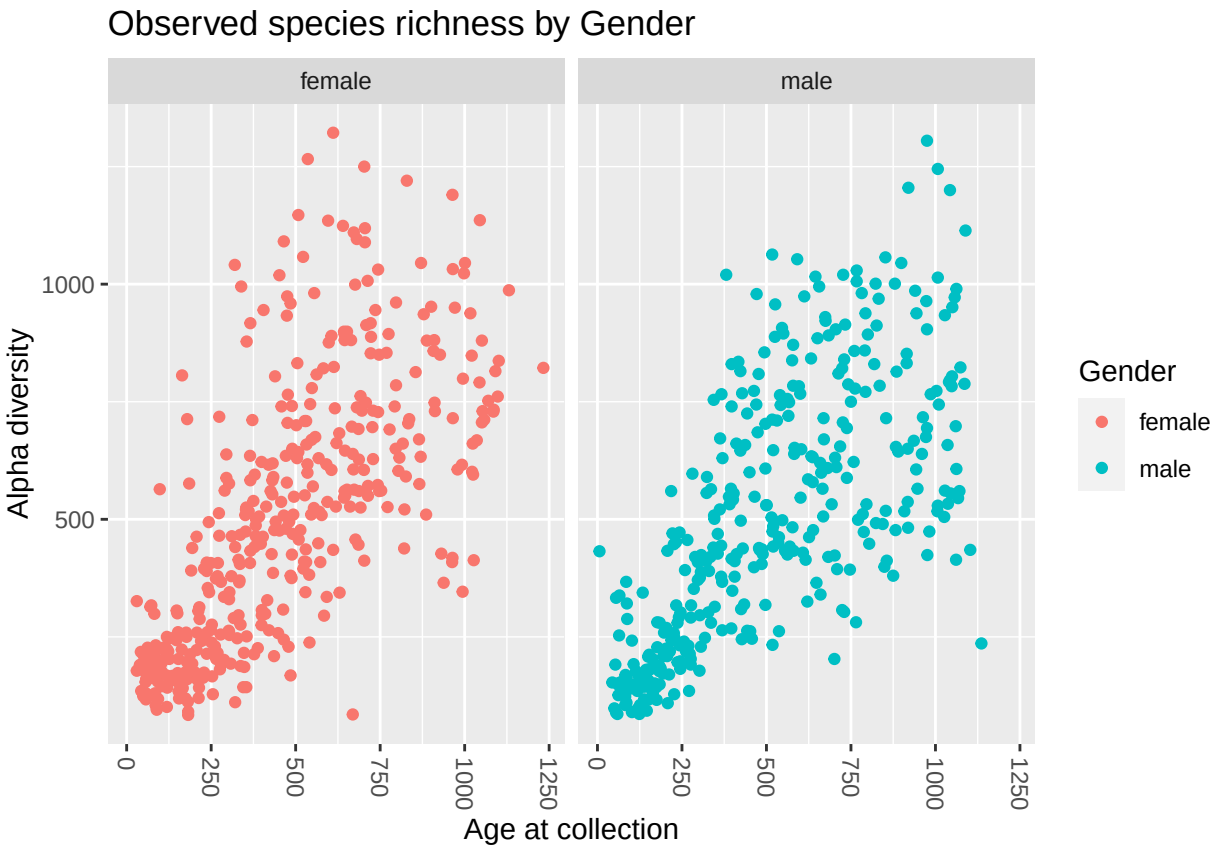
```
plot_richness(diabimmune, measures = "Observed",
              x = "Age_at_Collection", color = "Delivery_Route") +
       facet_grid(~Delivery_Route) +
       labs(title = "Observed species richness by Delivery route",
       x = "Age at collection",
       y = "Alpha diversity")
```



3.1.2 Observed species richness by Gender

No significant difference was observed in species richness of samples obtained from the male and female subjects. The observed species richness also increased in both males and females with increase in age.

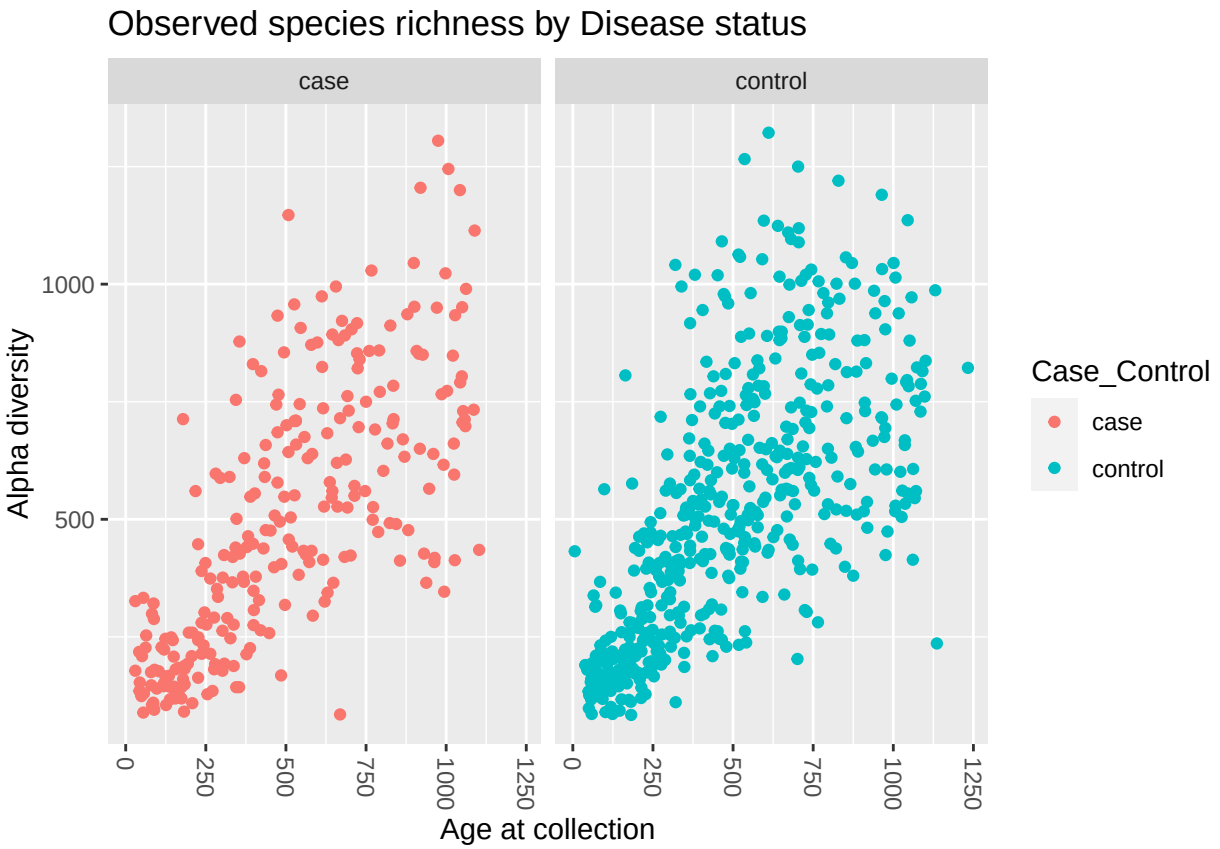
```
plot_richness(diabimmune, measures = "Observed",
              x = "Age_at_Collection", color = "Gender") +
       facet_grid(~Gender) +
       labs(title = "Observed species richness by Gender",
       x = "Age at collection",
       y = "Alpha diversity")
```



3.1.3 Observed species richness by disease status

There was a similar number of observed species richness of samples obtained from the cases and controls. The observed species richness also increased in both cases and controls with increase in age.

```
plot_richness(diabimmune, measures = "Observed",
              x = "Age_at_Collection", color = "Case_Control") +
       facet_grid(~Case_Control) +
       labs(title = "Observed species richness by Disease status",
       x = "Age at collection",
       y = "Alpha diversity")
```



3.2 ordination plots

OTUs were plotted to examine any observed patterns by delivery mode, gender and disease status.

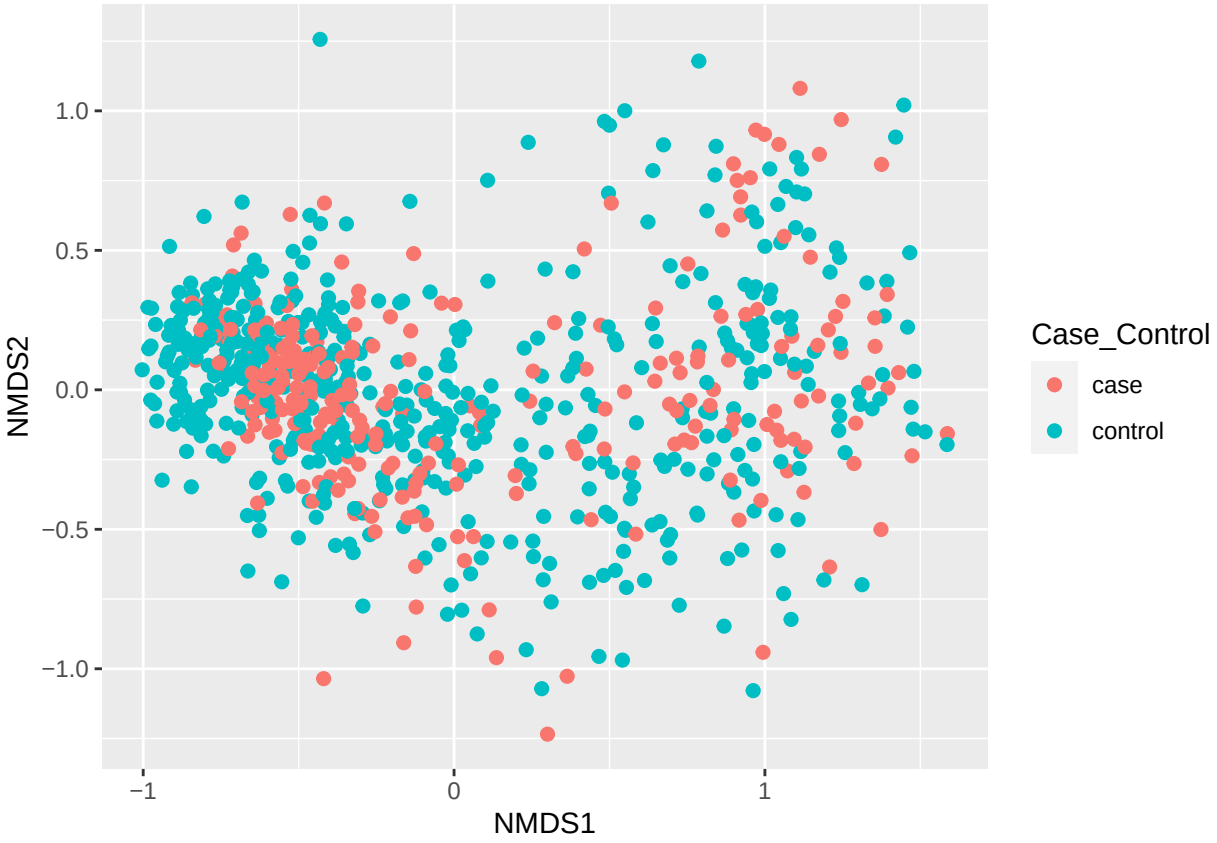
```
# ordinate the data
ord <- ordinate(diabimmune, "NMDS", "bray")
```

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.181114
## Run 1 stress 0.1906482
## Run 2 stress 0.1835989
## Run 3 stress 0.1869404
## Run 4 stress 0.190655
## Run 5 stress 0.1837589
## Run 6 stress 0.1843702
## Run 7 stress 0.1839133
## Run 8 stress 0.1836952
## Run 9 stress 0.4201126
```

```
## Run 10 stress 0.1856449
## Run 11 stress 0.1856611
## Run 12 stress 0.1923885
## Run 13 stress 0.1858836
## Run 14 stress 0.1908216
## Run 15 stress 0.1828309
## Run 16 stress 0.1845049
## Run 17 stress 0.4201046
## Run 18 stress 0.190725
## Run 19 stress 0.1906339
## Run 20 stress 0.1870571
## *** No convergence -- monoMDS stopping criteria:
##      7: no. of iterations >= maxit
##      6: stress ratio > sratmax
##      7: scale factor of the gradient < sfgrmin
```

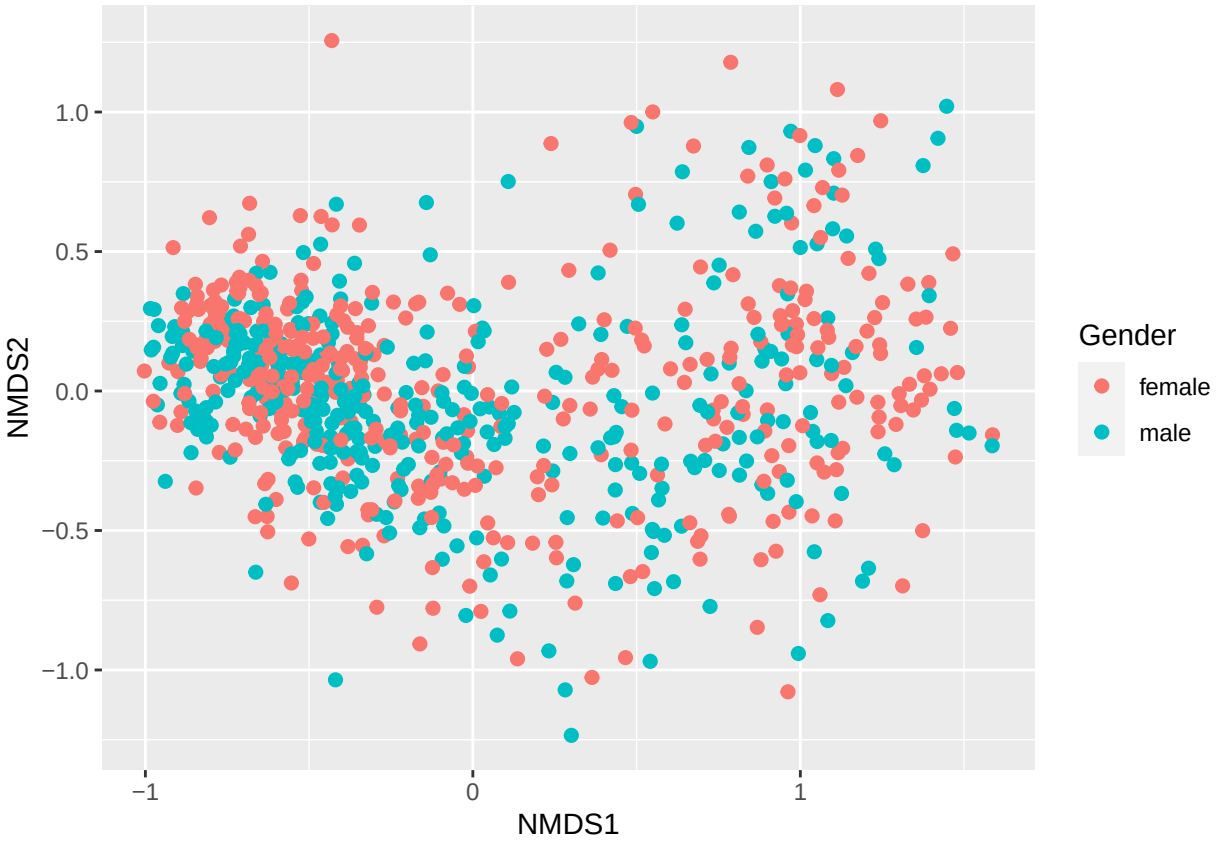
3.2.1 Delivery mode

```
plot_ordination(diabimmune, ord, type="samples",           color="Case_Control") +
  geom_point(size = 2)
```



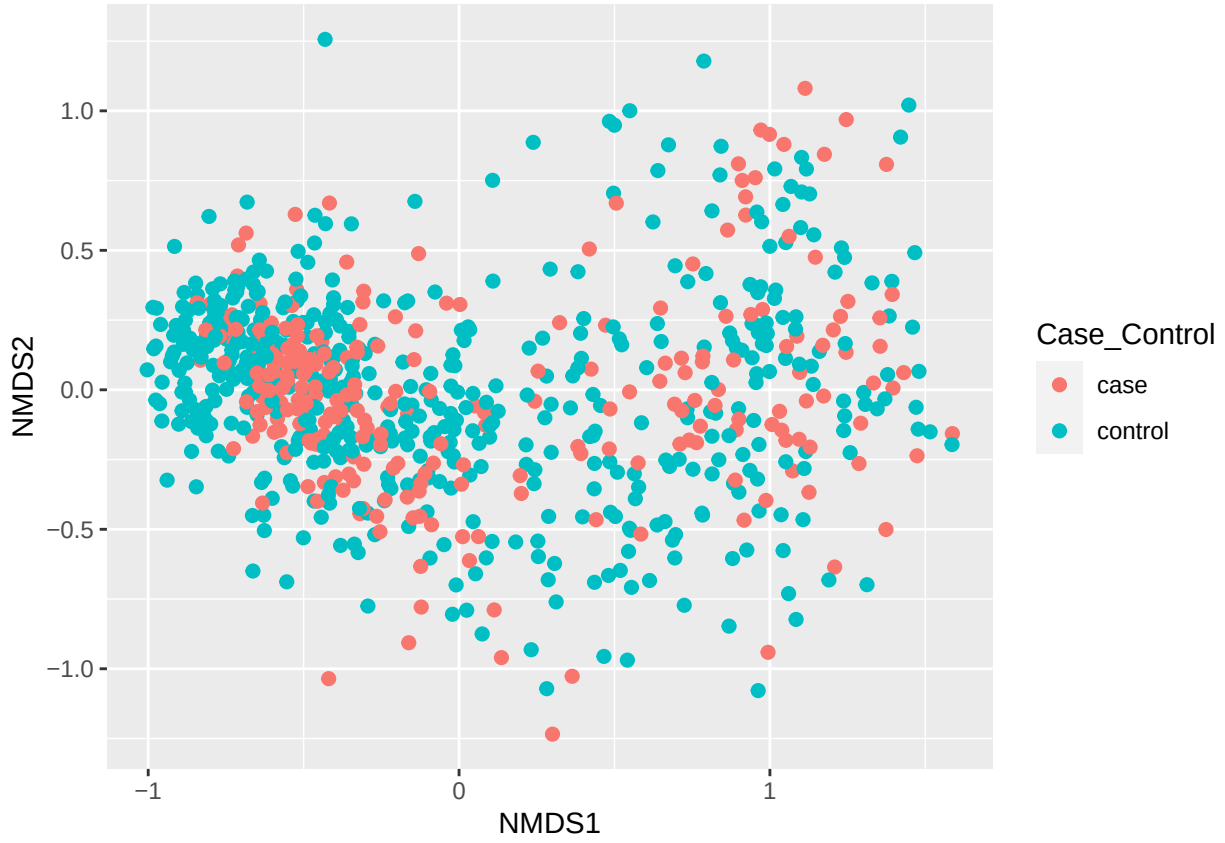
3.2.2 Gender

```
plot_ordination(diabimmune, ord, type="samples", color="Gender") +
  geom_point(size = 2)
```



3.2.3 Disease status

```
plot_ordination(diabimmune, ord, type="samples", color="Case_Control") +
  geom_point(size = 2)
```



4 Differential Abundance using DESeq2

The phyloseq object was converted to DEseqDataset class using the Case_Control variable as the study design factor. This was then followed by the geometricMeans and estimation of size factors. The differential expression analysis was then done using the DESeq function

```
# Convert data from class phyloseq to DESeq2's DESeqDataSet class
deseq <- phyloseq_to_deseq2(diabimmune, ~ Case_Control)

# function to calculate the geometric mean
gm_mean <- function(x, na.rm=TRUE){
  exp(sum(log(x[x > 0]), na.rm=na.rm) / length(x))
}

# geometricMean
geoMeans <- apply(counts(deseq), 1, gm_mean)

# size factor estimation
deseq <- estimateSizeFactors(deseq, geoMeans = geoMeans)
```

```
# perform the differential expression analysis
deseq <-  DESeq(deseq, fitType="local")
```

4.1 Construct the differential results table

```
# test results table
res <- results(deseq, cooksCutoff = FALSE)

res
```

```
## log2 fold change (MLE): Case Control control vs case
## Wald test p-value: Case Control control vs case
## DataFrame with 2240 rows and 6 columns
##           baseMean log2FoldChange     lfcSE      stat      pvalue       padj
##          <numeric>      <numeric> <numeric> <numeric>   <numeric>  <numeric>
## 4333897   14.68854     -0.0368163  0.239861 -0.153490 0.878011790 0.93286739
## 190162     8.12512     -0.3882871  0.170899 -2.272027 0.023084893 0.07759447
## 134726     5.65090      2.1313224  0.566587  3.761684 0.000168773 0.00174078
## 679245     8.96095     -0.5747490  0.462537 -1.242602 0.214014712 0.37835118
## 289734  1052.54282     -0.1550911  0.194789 -0.796201 0.425915446 0.59700370
## ...           ...            ...       ...       ...         ...        ...
## 842596   0.1281748       0.832916   1.87446 0.4443495    0.656790         NA
## 144395   0.1630050       1.035407   1.88611 0.5489644    0.583030         NA
## 229348   0.0533851       0.353280   3.06135 0.1153999    0.908128         NA
## 187121   0.0498920       0.399092   3.06135 0.1303644    0.896278         NA
## 208972   0.0973446       0.101569   1.49153 0.0680975    0.945708         NA
```