

Introducing R and the RStudio IDE

Alana Schick & Hena Ramay

IMC Bioinformatics, University of Calgary



UNIVERSITY OF CALGARY
CUMMING SCHOOL OF MEDICINE



UNIVERSITY OF
CALGARY

International
Microbiome
Centre



What is R?

- R (since 1995) is a programming language developed to teach statistics
- R is open source (ie. free), widely used, flexible, and powerful



Packages: the power of R

A way for the R **community** to share functions and data sets

Importing & Exporting data
From
text files, excel,
stata, SPSS , and
databases

Data Modeling
Statistical tests
Linear & non-linear models
Machine learning
Survival analysis

Data Sharing
Plotting,
interactive plots,
reporting with
markdown and
shiny apps



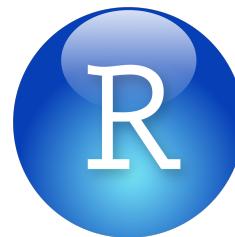
What is RStudio?

RStudio is an Integrated Development Environment (IDE) that allows users to run R in a more user-friendly way



R: Engine

+



RStudio: Dashboard



Let's open Rstudio and get to know it !!

RStudio looks like this

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Window Help

Project: (None)

index.Rmd * Schedule.Rmd * Session1.Rmd * jodi_btbr_phlyoseq.R

Source on Save Run Source

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keptaxa, rel)
125 psf <- prune_taxa(keptaxa, ps)
126
127
128 ###### Alpha diversity
129
130 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
131
132 ## Make table of alpha diversity calculations
133 alpha <- estimate_richness(ps)
134 alpha_info <- sample_data(ps)
135 aa <- cbind(alpha, alpha_info)
136
137 ## Check for outliers
138 qplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
139 qplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
140
141 ## Plot
142 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
143 a1
144
145 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")
146 a2
147
148 (Untitled): R Script
```

Console Terminal R Markdown

```
> s
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
> rownames(tax_table(ps)) <- asv_names
>
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
>
```

Files Plots Packages Help Viewer

Zoom Export Publish

Alpha diversity (Shannon)

base week3

treatment

- control
- pre
- pro
- syn

Screenshot

Environment History Connections

Import Dataset

Global Environment

Data

- a1 List of 9
- aa 95 obs. of 17 variables
- alpha 95 obs. of 9 variables
- alpha_info 95 obs. of 8 variables
- info 96 obs. of 7 variables
- ps Large phyloseq (1.5 Mb)
- seqtab Large matrix (208416 elements, 1.8 Mb)
- taxa Large matrix (15197 elements, 1.1 Mb)
- tree Large phylo (4 elements, 1 Mb)

Values

Mac OS X Dock icons: Mail, Music, Calendar, Google Chrome, Spotify, iMovie, iPhoto, iWork, R, RStudio, Tex, Arrows, etc.

RStudio screen

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

Project: (None)

index.Rmd Schedule.Rmd Session1.Rmd jodi_btbr_phlyoseq.R

Source on Save Run Source

Script

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keptaxa, rel)
125 psf <- prune_taxa(keptaxa, ps)
126
127
128 ###### Alpha diversity
129
130
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(ps)
136 alpha_info <- sample_data(ps)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 qplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
142 qplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
146 a1
147
148 o2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")
149 o2
150
```

Console Terminal R Markdown

```
> s
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
> rownames(tax_table(ps)) <- asv_names
>
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$tre
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "-"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
>
```

Files Plots Packages Help Viewer

Plot

Environment

Import Dataset

Global Environment

Data

- a1 List of 9
- aa 95 obs. of 17 variables
- alpha 95 obs. of 9 variables
- alpha_info 95 obs. of 8 variables
- info 96 obs. of 7 variables
- ps Large phyloseq (1.5 Mb)
- seqtab Large matrix (208416 elements, 1.8 Mb)
- taxa Large matrix (15197 elements, 1.1 Mb)
- tree Large phylo (4 elements, 1 Mb)

Values

Screenshot

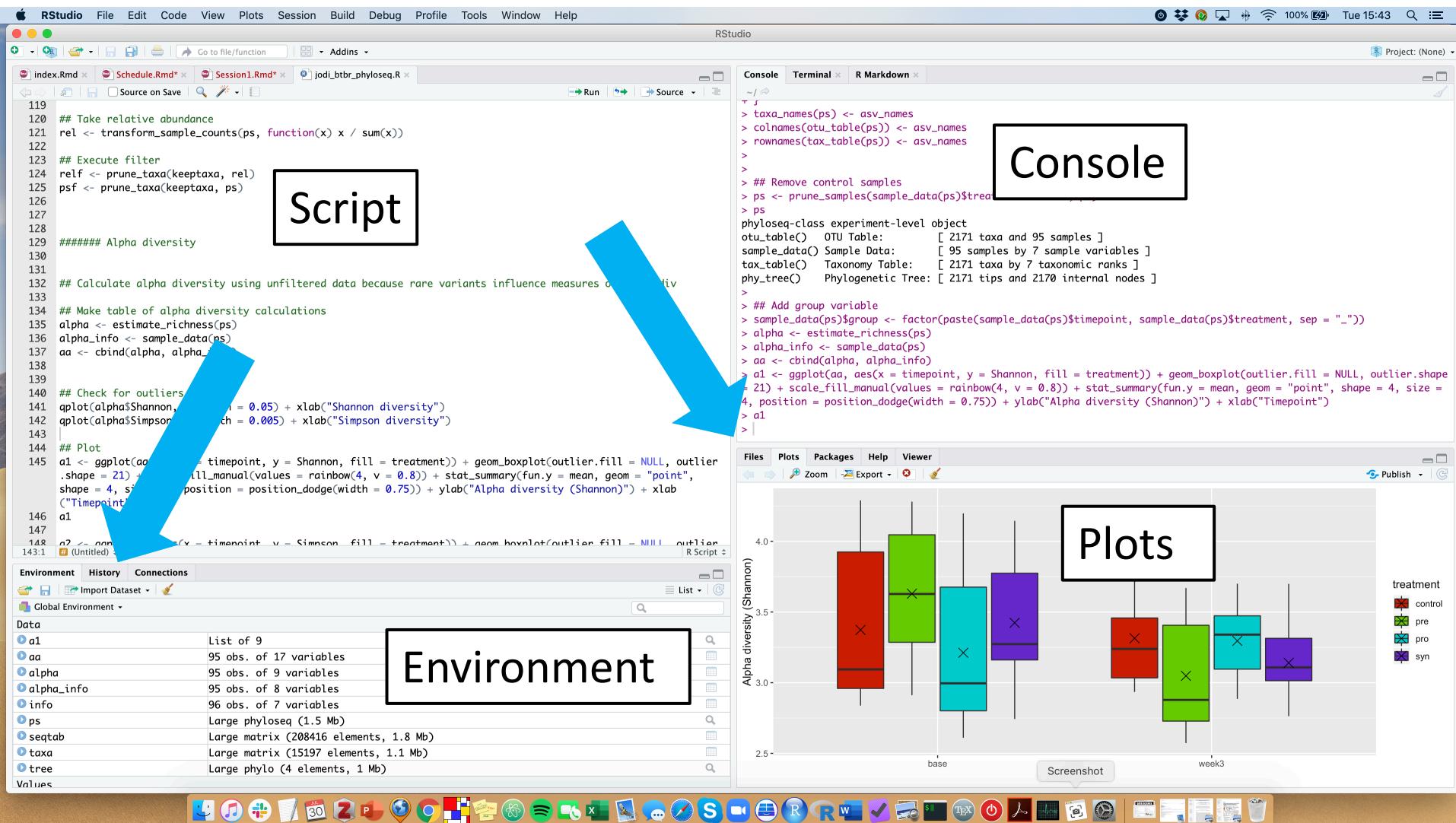
Console

Plots

treatment

- control
- pre
- pro
- syn

RStudio screen



RStudio screen

The screenshot shows the RStudio interface on a Mac OS X desktop. The window title is "RStudio". The menu bar includes "File", "Edit", "Code", "View", "Plots", "Session", "Build", "Debug", "Profile", "Tools", "Window", and "Help". The status bar at the bottom right shows "100% battery" and the date "Tue 16:27".

The interface is divided into several panes:

- Script**: The leftmost pane contains R code. A large black rectangle highlights the area from line 119 to 138.
- Console**: The top-right pane displays the R console output. A large black rectangle highlights the output from line 119 to 138.
- History**: The bottom-left pane shows the history of R commands run in the session. A large black rectangle highlights the output from line 119 to 138.
- Files**: The bottom-right pane shows the file system structure under "Home". A large black rectangle highlights the "Home" folder.

The R code in the Script pane:

```
119 ## Take relative abundance
120 rel <- transform_sample_counts(ps, function(x) x / sum(x))
121
122 ## Execute filter
123 relf <- prune_taxa(keptaxa, rel)
124 psf <- prune_taxa(keptaxa, ps)
125
126
127
128 ###### Alpha diversity
129
130
131
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(ps)
136 alpha_info <- sample_data(ps)
137 aa <- cbind(alpha, alpha_info)
138
138:1 (Untitled):
```

The R code in the History pane:

```
# MARCH Sample names
rownames(seqtab)
# Make a phyloseq object
ps <- phyloseq(cou_table(seqtab, taxa_are_rows=FALSE), sample_data(info), tax_table(taxa))
## Make a tree and add the tree to a new phyloseq object
tree <- rtree(ntaxa(ps), rooted = TRUE, tip.label = taxa_names(ps))
ps <- phyloseq(cou_table(seqtab, taxa_are_rows=FALSE), sample_data(info), tax_table(taxa), phy_tree(tree))
asv_names <- vector(dim(otu_table(ps))[2], mode = "character")
for (i in 1:dim(otu_table(ps))[2]){
  asv_names[i] <- paste("ASV", i, sep = "_")
}
taxa_names(ps) <- asv_names
colnames(otu_table(ps)) <- asv_names
rownames(tax_table(ps)) <- asv_names
## Remove control samples
ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
ps
## Add group variable
sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
alpha <- estimate_richness(ps)
alpha_info <- sample_data(ps)
aa <- cbind(alpha, alpha_info)
a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = ...)
```

The R code in the Console pane:

```
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
> rownames(tax_table(ps)) <- asv_names
>
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
> |
```

The file system in the Files pane:

```
Files Plots Packages Help Viewer
New Folder Delete Rename More ...
Home
Name
.Rhistory
Applications
Desktop
Documents
Downloads
Dropbox
Library
Movies
Music
Pictures
Public
Zotero
```

The desktop dock at the bottom of the screen contains icons for various applications, including Finder, Mail, Safari, and others.

How to R – 2 ways

The screenshot shows the RStudio interface with several panels:

- Script Editor:** Shows a code snippet for calculating alpha diversity. A box labeled "Script" highlights the first few lines:

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keeptaxa, rel)
125 psf <- prune_taxa(keeptaxa, ps)
126
127
128 ###### Alpha diversity
129
130 ## Calculate alpha diversity us
```
- Console:** Shows the R command history. A box labeled "Console" highlights the first few commands:

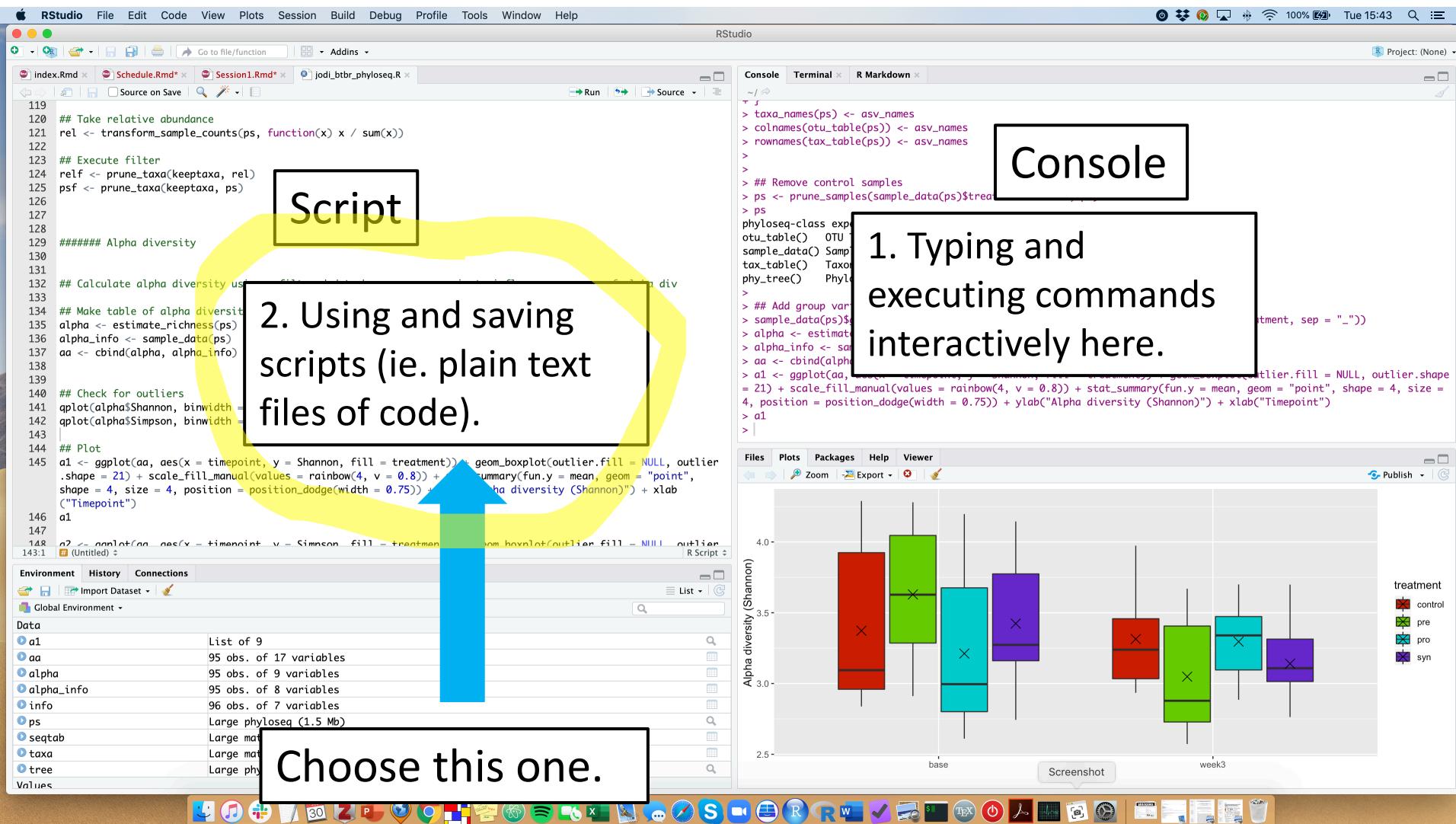
```
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
> rownames(tax_table(ps)) <- asv_names
>
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment)
> ps
```
- Plots:** A box labeled "1. Typing and executing commands interactively here." contains the following text:

1. Typing and executing commands interactively here.

A box plot visualizing alpha diversity (Shannon) across different time points (base and week3) and treatments (control, pre, pro, syn). The y-axis ranges from 2.5 to 4.0. The legend indicates: control (red), pre (green), pro (cyan), and syn (purple).

Timepoint	treatment	Median Alpha diversity (Shannon)	Approx. Range
base	control	~3.0	~2.8 - 3.8
	pre	~3.6	~3.3 - 4.0
	pro	~3.2	~2.8 - 3.8
	syn	~3.4	~3.0 - 3.8
week3	control	~3.2	~2.8 - 3.8
	pre	~3.3	~2.8 - 3.8
	pro	~3.4	~3.0 - 3.8
	syn	~3.2	~3.0 - 3.8

How to R – 2 ways



Scripts

```
1 # jodi_btbr project, Alana Schick, April 2019
2 # This is a script to analyze the output tables of the DADA2 workflow in phyloseq
3 # Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS() function
4 # The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"
5
6 library(phyloseq)
7 #packageVersion("phyloseq")
8 library(ggplot2)
9 #packageVersion("ggplot2")
10 library(ape)
11 library(viridis)
12 library(grid)
13 library(gridExtra)
14 library(reshape2)
15 library(DESeq2)
16 library(fields)
17 library(vegan)
18 library(ggpubr)
19 library(plyr)
20 library(RColorBrewer)
21
22 path_to_project <- "/Users/alanaschick/Dropbox/Jodi_BTBR"
23
24 # Read in files
25 seqtab <- readRDS(file.path(path_to_project, "seqtab.rds"))
26 taxa <- readRDS(file.path(path_to_project, "taxa.rds"))
27 info <- read.table(file.path(path_to_project, "jodi_btbr_metadata.txt"), header = TRUE)
28
29 # Match sample names
30 rownames(info) <- rownames(seqtab)
31
32 # Make a phyloseq object
```

(Untitled) ♦ R Script ♦

Everything in the console will be forgotten when you close the session.

Scripts are saved, keeping a complete record of the commands you ran so you can run them again (ie. completely reproducible).

Can execute parts of this or the entire script.

Scripts - commenting

```
1 # jodi_btbr project, Alana Schick, April 2019
2 # This is a script to analyze the output tables of the DADA2 workflow in phyloseq
3 # Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS()
4 # The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"
5
6 library(phyloseq)
7 #packageVersion("phyloseq")
8 library(ggplot2)
9 #packageVersion("ggplot2")
10 library(ape)
11 library(viridis)
12 library(grid)
13 library(gridExtra)
14 library(reshape2)
15 library(DESeq2)
16 library(fields)
17 library(vegan)
18 library(ggpubr)
19 library(plyr)
20 library(RColorBrewer)
21
22 path_to_project <- "/Users/alanaschick/Dropbox/time/projects/jodi_btbr"
23
24 # Read in files
25 seqtab <- readRDS(file.path(path_to_project, "results/seqtab_final.rds"))
26 taxa <- readRDS(file.path(path_to_project, "results/taxa_final.rds"))
27 info <- read.table(file.path(path_to_project, "jodi_btbr_metadata2.txt"), header = TRUE)
28
29 # Match sample names
30 rownames(info) <- rownames(seqtab)
31
32 # Make a phyloseq object
```

Comment out lines of your scripts by using the `#` symbol. R will not run these.

Be descriptive. You will not remember what you did a year later.

Packages

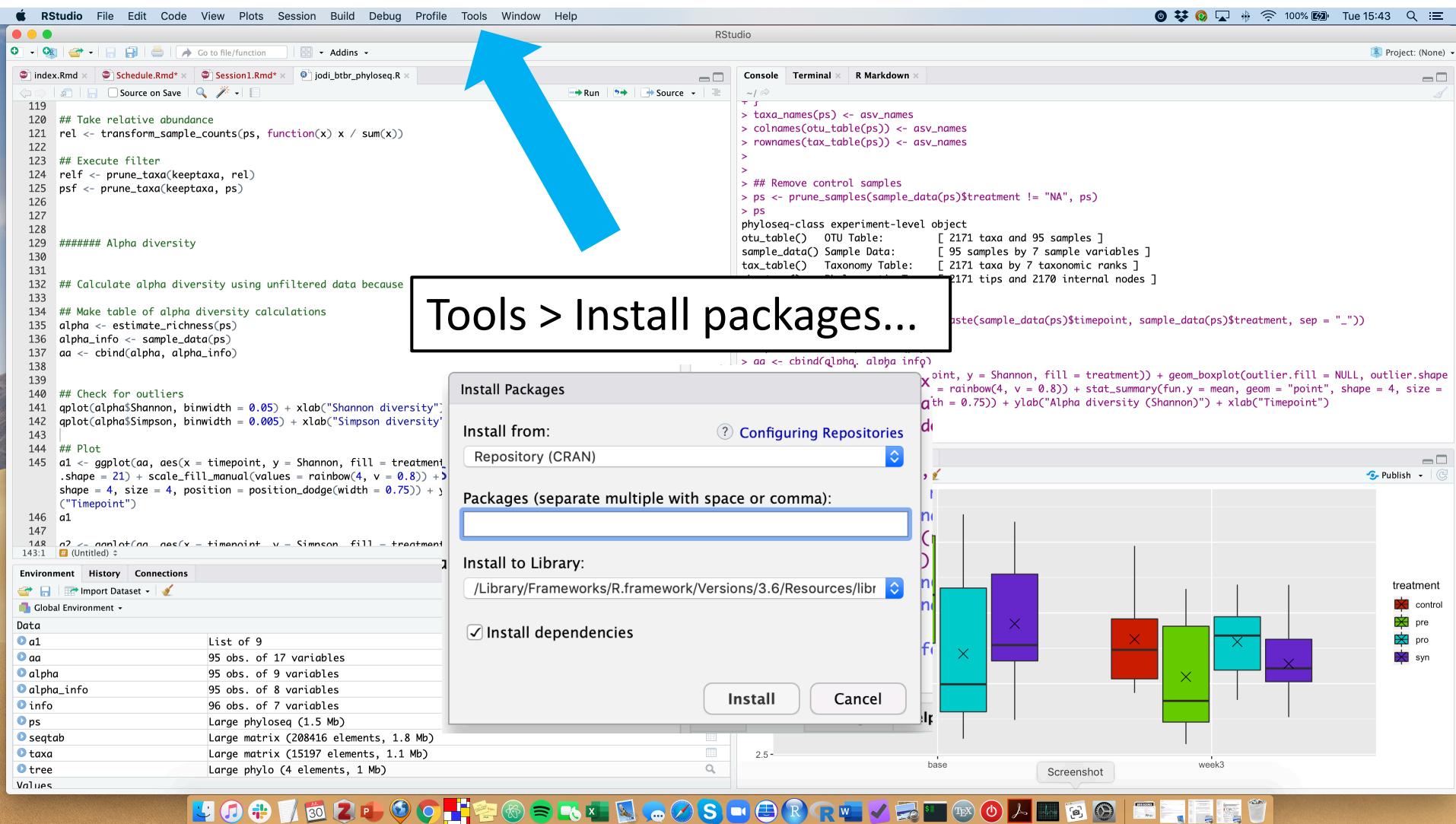
```
1 # jodi_btbr project, Alana Schick, April 2019
2 # This is a script to analyze the output tables of the DADA2 workflow in phyloseq
3 # Have two output files from dada2 - a sequence table and a taxonomy table, read them into R using the readRDS() function
4 # The formatted sample metadata is in a table called "jodi_btbr_metadata.txt"
5
6 library(phyloseq)
7 #packageVersion("phyloseq")
8 library(ggplot2)
9 #packageVersion("ggplot2")
10 library(ape)
11 library(viridis)
12 library(grid)
13 library(gridExtra)
14 library(reshape2)
15 library(DESeq2)
16 library(fields)
17 library(vegan)
18 library(ggpubr)
19 library(plyr)
20 library(RColorBrewer)
21
22 path_to_project <- "/Users/alanaschick/"
23
24 # Read in files
25 seqtab <- readRDS(file.path(path_to_project, "results/seqtan_final.rds"))
26 taxa <- readRDS(file.path(path_to_project, "results/taxa_final.rds"))
27 info <- read.table(file.path(path_to_project, "jodi_btbr_metadata2.txt"), header = TRUE)
28
29 # Match sample names
30 rownames(info) <- rownames(seqtan)
31
32 # Make a phyloseq object
38:1 # (Untitled) ▾
```

Packages are collections of R functions developed for a specific task.

Packages need to first be installed on your computer.

After installed, `library()` is the command used to load a package.

Packages



Pay close attention to the next few slides

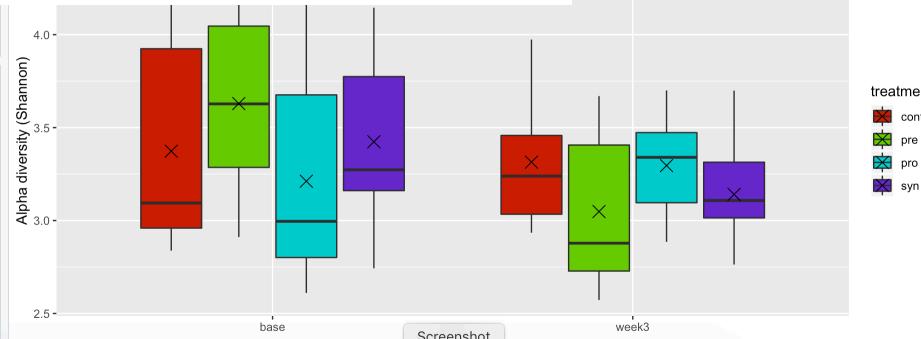
About half of the students in every workshop have problems
in understanding the concept of working directory!!

Working directory

Every time you open RStudio, it goes to a default directory, usually your home directory.

You can use the command **setwd()** to change the working directory.

```
setwd("home/aschick/projects/workshop")
```



Relative paths



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

Project: (None) Tue 16:27

index.Rmd Schedule.Rmd* Session1.RMd* jodi_btbr_phyloseq.R

Source on Save Run Source

```

119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keptaxa, rel)
125 psf <- prune_taxa(keptaxa, ps)
126
127
128 ###### Alpha diversity
129
130
131
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(ps)
136 alpha_info <- sample_data(ps)
137 aa <- cbind(alpha, alpha_info)
138
138:1 (Untitled):1 R Script

```

Environment History Connections To Console To Source

```

# MARCH Sample names
rownames(info) <- rownames(seqtab)
# Make a phyloseq object
ps <- phyloseq(otu_table(seqtab, taxa_are_rows=FALSE), sample_data(info), tax_table(taxa))
## Make a tree and add the tree to a new phyloseq object
tree <- rtree(ntaxa(ps), rooted = TRUE, tip.label = taxa_names(ps))
ps <- phyloseq(otu_table(seqtab, taxa_are_rows=FALSE), sample_data(info), tax_table(taxa), phy_tree(tree))
asv_names <- vector(dim(otu_table(ps))[2], mode = "character")
for (i in 1:dim(otu_table(ps))[2]){
  asv_names[i] <- paste("ASV", i, sep = "_")
}
taxa_names(ps) <- asv_names
colnames(otu_table(ps)) <- asv_names
rownames(tax_table(ps)) <- asv_names
## Remove control samples
ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
ps
## Add group variable
sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
alpha <- estimate_richness(ps)
alpha_info <- sample_data(ps)
aa <- cbind(alpha, alpha_info)
a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape =
a1

```

Console Terminal R Markdown

```

~/j
> taxa_names(ps) <- asv_names
> colnames(otu_table(ps)) <- asv_names
> rownames(tax_table(ps)) <- asv_names
>
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape =
= 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size =
4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
> a1

```

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

- Name
- R.history
- Applications
- Desktop
- Documents
- Downloads
- Dropbox
- Library
- Movies
- Music
- Pictures
- Public
- Zotero

Size Modified

1.1 KB Feb 11, 2019, 5:41 PM

Screenshot

The directory that you see here is not necessarily the same as your working directory. Please do not use this to find your files.

Working directory

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help RStudio Addins Go to file/function Addins ~ Run Source ~ Project: (None) 100% Tue 15:43

```
119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x)
122
123 ## Execute filter
124 relf <- prune_taxa(keptaxa, rel)
125 psf <- prune_taxa(keptaxa, ps)
126
127
128 ###### Alpha diversity
129
130 ## Calculate alpha diversity using unfiltered data
131
132 ## Make table of alpha diversity calculations
133 alpha <- estimate_richness(psf)
134 alpha_info <- sample_data(psf)
135 aa <- cbind(alpha, alpha_info)
136
137 ## Check for outliers
138 qplot(alpha$Shannon, binwidth = 0.05) + xlab("Timepoint")
139 qplot(alpha$Simpson, binwidth = 0.005) + xlab("Timepoint")
140
141 ## Plot
142 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon))
143 a1 + geom_boxplot(outlier.size = 0.75) + stat_summary(fun.y = mean, geom = "point",
144 .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point",
145 shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab(
146 ("Timepoint"))
147
148 o2 <- ggplot(aa, aes(x = timepoint, y = Simpson))
149 o2 + geom_boxplot(outlier.size = 0.75) + stat_summary(fun.y = mean, geom = "point",
150 .shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point",
151 shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab(
152 ("Timepoint"))
153
154
```

However: you may want to run your script on a different computer with a different directory structure where that directory does not exist.

Or you may want to work in multiple directories.

Alpha diversity (Shannon)

control pre pro syn

Screenshot

RStudio Project

File > New Project...

Clicking on New Directory will create an RStudio Project.

This directory will have all the data, files, plots, etc. for that project as well as a .Rproj file.

The screenshot shows the RStudio interface with a blue arrow pointing to the 'File' menu. A callout box highlights the 'File > New Project...' option. In the center, a 'New Project' dialog box is open, listing three options: 'New Directory', 'Existing Directory', and 'Version Control'. Below the dialog, a box plot plot is visible, showing 'Alpha div' on the y-axis and time points 'base' and 'week3' on the x-axis. The legend indicates four categories: 'control' (red), 'pre' (green), 'pro' (cyan), and 'syn' (purple).

```
## Take relative abundance
rel <- transform(ps, counts(ps, function(x) x / sum(x)))

## Execute filter
relf <- prune_taxa(is.sampled == TRUE, rel)
psf <- prune_taxa(is.sampled == TRUE, ps)

##### Alpha diversity
## Calculate alpha diversity
## Make table
alpha <- estim
alpha_info <- sample_data(ps)
aa <- cbind(alpha, alpha_info)

## Check for outliers
qplot(alpha$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
qplot(alpha$Simpson, binwidth = 0.005) + xlab("Simpson diversity")

## Plot
a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon) ("Timepoint")")
a1
a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson) ("Timepoint")")
a2
```

Error messages

Console Terminal R Markdown

~ / ↻

```
>
> ## Remove control samples
> ps <- prune_samples(sample_data(ps)$treatment != "NA", ps)
> ps
phyloseq-class experiment-level object
otu_table() OTU Table: [ 2171 taxa and 95 samples ]
sample_data() Sample Data: [ 95 samples by 7 sample variables ]
tax_table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]
phy_tree() Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]
>
> ## Add group variable
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "_"))
> alpha <- estimate_richness(ps)
> alpha_info <- sample_data(ps)
> aa <- cbind(alpha, alpha_info)
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
> a1
> ord1 <- ordinate(relf, method = "NMDS", distance = "bray")
Error in ordinate(relf, method = "NMDS", distance = "bray") :
  object 'relef' not found
> b1 <- plot_ordination(relf, ord1, color = "timepoint", shape = "treatment", title = "NMDS - Bray") + scale_colour_manual(values = viridis(3))
Error in plot_ordination(relf, ord1, color = "timepoint", shape = "treatment", :
  object 'relef' not found
> b1
Error: object 'b1' not found
> |
```

Error messages

Console Terminal × R Markdown ×

```
>  
> ## Remove control samples  
> ps <- prune_samples(sample_data(ps)$tr  
> ps  
phyloseq-class experiment-level object  
otu_table() OTU Table: [ 2171  
sample_data() Sample Data: [ 95 sa  
tax_table() Taxonomy Table: [ 2171  
phy_tree() Phylogenetic Tree: [ 2171  
>  
> ## Add group variable  
> sample_data(ps)$group <- factor(paste(  
> alpha <- estimate_richness(ps)  
> alpha_info <- sample_data(ps)  
> aa <- cbind(alpha, alpha_info)  
> a1 <- ggplot(aa, aes(x = timepoint, y  
= 21) + scale_fill_manual(values = rainb  
4, position = position_dodge(width = 0.7  
> a1  
> ord1 <- ordinate(relf, method = "NMDS"  
Error in ordinate(relf, method = "NMDS",  
object 'relf' not found  
> b1 <- plot_ordination(relf, ord1, col  
ual(values = viridis(3))  
Error in plot_ordination(relf, ord1, col  
object 'relf' not found  
> b1  
Error: object 'b1' not found  
>
```



```
tment, sep = "_"))  
  
tlier.fill = NULL, outlier.shape  
om = "point", shape = 4, size =  
"Timepoint")  
  
NMDS - Bray") + scale_colour_man
```

Getting help

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help RStudio 100% Tue 15:43 Project: (None)

```
index.Rmd * Schedule.Rmd * Session1.Rmd * jodi_btbr_phyloseq.R * Go to file/function Addins * Run Source ~ / s > taxa_names(ps) <- asv_names > colnames(otu_table(ps)) <- asv_names
```

119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keptaxa, rel)
125 psf <- prune_taxa(keptaxa, ps)
126
127
128
129 ##### Alpha diversity
130
131
132 ## Calculate alpha diversity using unfiltered data because rare vo
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(psf)
136 alpha_info <- sample_data(psf)
137 aa <- cbind(alpha, alpha_info)
138
139
140 ## Check for outliers
141 qplot(alpha\$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
142 qplot(alpha\$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment))
.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + s
shape = 4, size = 4, position = position_dodge(width = 0.75)) + yl
("Timepoint")
146
147
148 o2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4) + xlab("Timepoint")
149
150
151 (Untitled):

1) Search in Help tab

2) Type ? followed by the function name in the console (or ?? for installed packages)

> ?barplot
> ??geom_point

Variables]
ranks]
inal nodes]
eepoint, sample_data(ps)\$treatment, sep = "-"))
tat_summary(fun.y = mean, geom = "point", shape = 4, size =
iversity (Shannon") + xlab("Timepoint")

treatment
control
pre
pro
syn

Screenshot

Getting help

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help

Project: (None)

```
index.Rmd * Schedule.Rmd * Session1.Rmd * jodi_btbr_phlyoseq.R
```

Source on Save Run Source

119
120 ## Take relative abundance
121 rel <- transform_sample_counts(ps, function(x) x / sum(x))
122
123 ## Execute filter
124 relf <- prune_taxa(keptaxa, rel)
125 psf <- prune_taxa(keptaxa, ps)
126
127
128 ##### Alpha diversity
129
130
132 ## Calculate alpha diversity using unfiltered data because rare variants influence measures of alpha div
133
134 ## Make table of alpha diversity calculations
135 alpha <- estimate_richness(ps)
136 alpha_info <- sample_data(ps)
137 aa <- cbind(alpha, alpha_info)
138
139 ## Check for outliers
140 qplot(alpha\$Shannon, binwidth = 0.05) + xlab("Shannon diversity")
141 qplot(alpha\$Simpson, binwidth = 0.005) + xlab("Simpson diversity")
143
144 ## Plot
145 a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")
146 a1
147
148 a2 <- ggplot(aa, aes(x = timepoint, y = Simpson, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Simpson)") + xlab("Timepoint")
143:1 (Untitled) R Script

Console Terminal R Markdown

```
> /s  
> taxa_names(ps) <- asv_names  
> colnames(otu_table(ps)) <- asv_names  
>  
> ps  
phyloseq-class experiment-level object  
@ .table() OTU Table: [ 2171 taxa and 95 samples ]  
@ .sample_data() Sample Data: [ 95 samples by 7 sample variables ]  
@ .table() Taxonomy Table: [ 2171 taxa by 7 taxonomic ranks ]  
@ .tree Phylogenetic Tree: [ 2171 tips and 2170 internal nodes ]  
  
> s  
> sample_data(ps)$group <- factor(paste(sample_data(ps)$timepoint, sample_data(ps)$treatment, sep = "-"))  
> alpha <- estimate_richness(ps)  
> alpha_info <- sample_data(ps)  
> aa <- cbind(alpha, alpha_info)  
> a1 <- ggplot(aa, aes(x = timepoint, y = Shannon, fill = treatment)) + geom_boxplot(outlier.fill = NULL, outlier.shape = 21) + scale_fill_manual(values = rainbow(4, v = 0.8)) + stat_summary(fun.y = mean, geom = "point", shape = 4, size = 4, position = position_dodge(width = 0.75)) + ylab("Alpha diversity (Shannon)") + xlab("Timepoint")  
> a1
```

Files Plots Packages Help Viewer

Zoom Export Publish

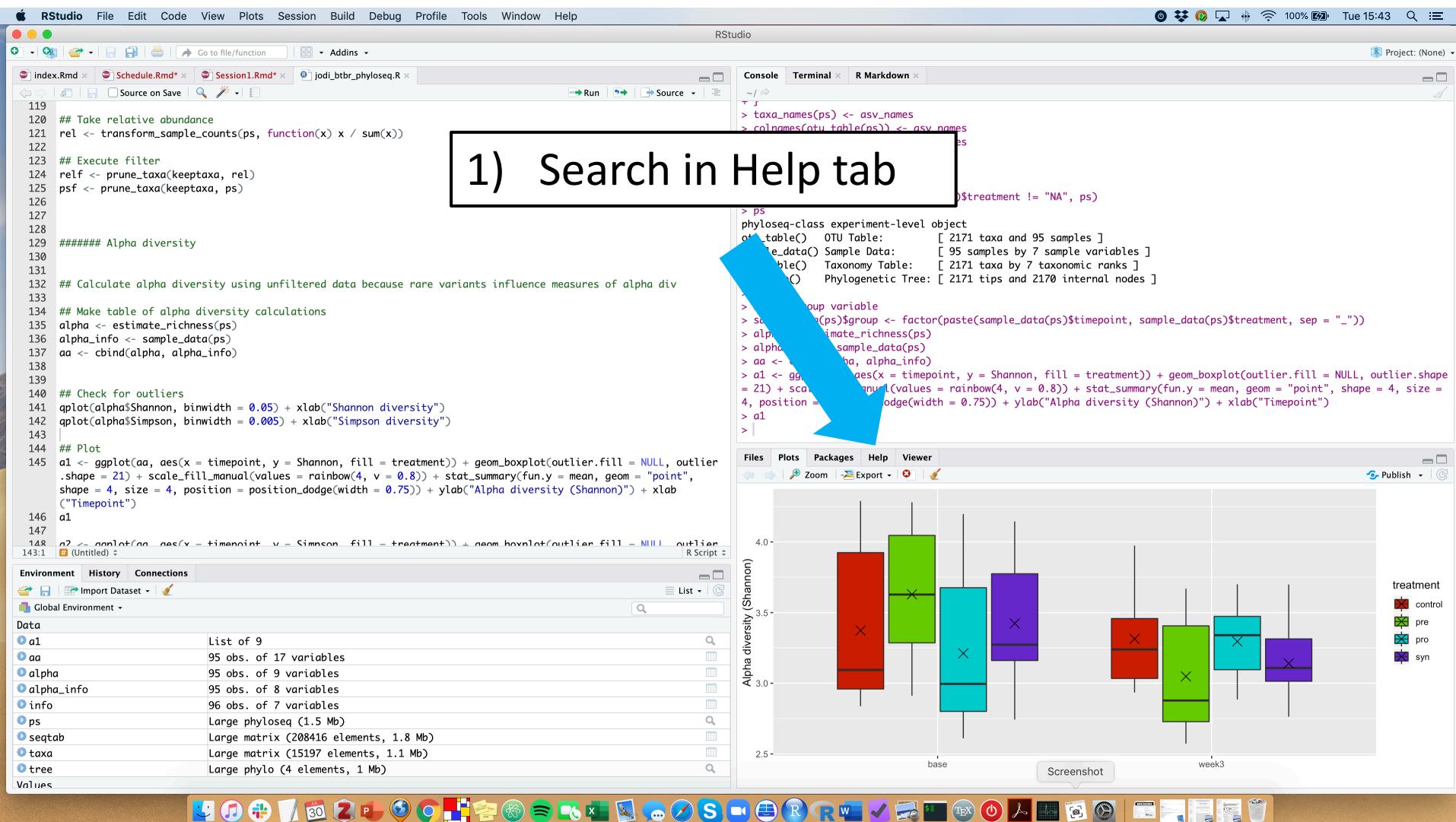
Alpha diversity (Shannon)

treatment

- control
- pre
- pro
- syn

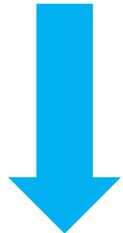
base week3

Screenshot



Getting help

- 1) Search in Help tab
- 2) Type ? followed by the function name in the console (or ?? for installed packages)
- 3) Google the error message



See website for tips and resources!

The internet will make those bad words go away



Essential

Googling the
Error Message

ORLY?

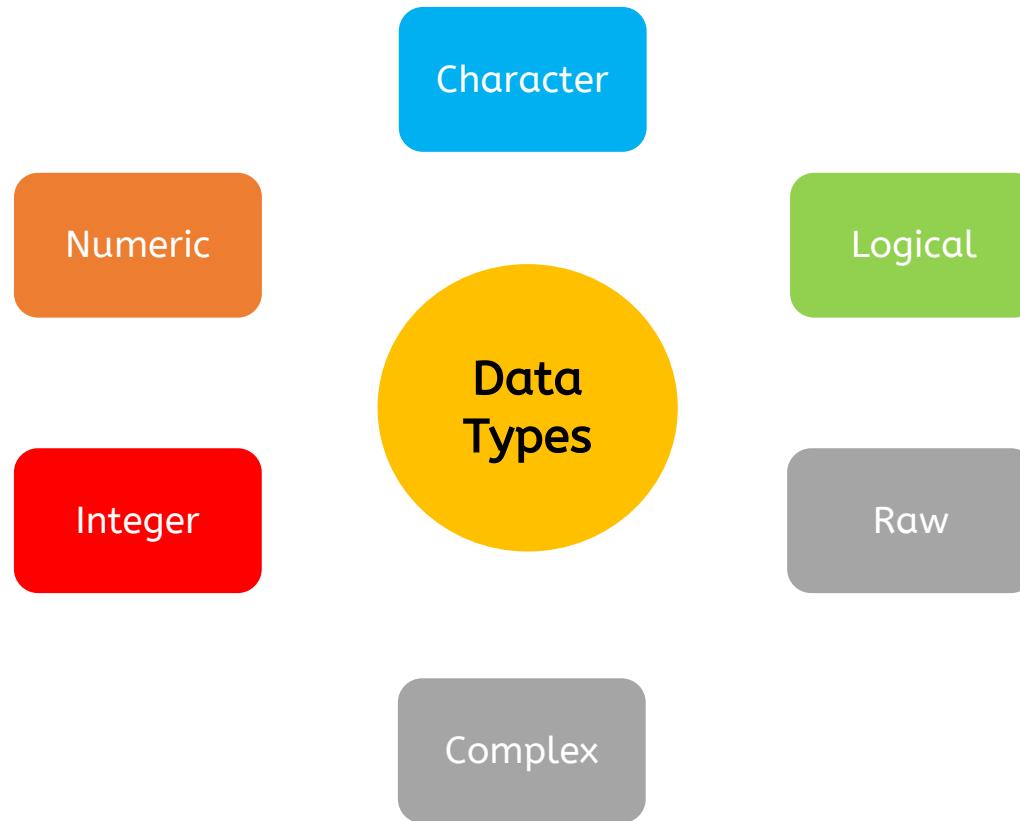
*The Practical Developer
@ThePracticalDev*

Summary and best practices

- Always save your code in R scripts
- Load packages using `library()` at the top of your script
- Write clear, readable code with comments*
- Be mindful of your working directory or location of files
- Use RStudio projects to organize scripts, data, and output

*See <http://adv-r.had.co.nz/Style.html> for tips.

Data Types

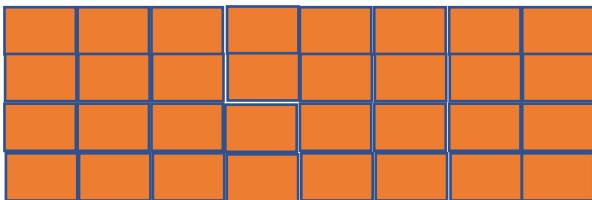


Data Structures

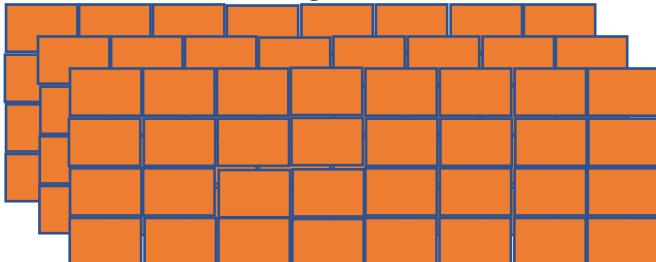
Vector
1d, homogeneous



Matrix
2d, homogeneous

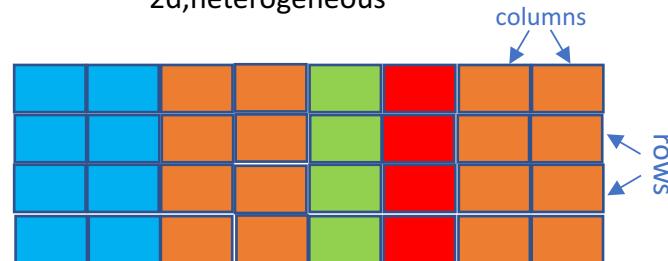


Array
3d, homogeneous



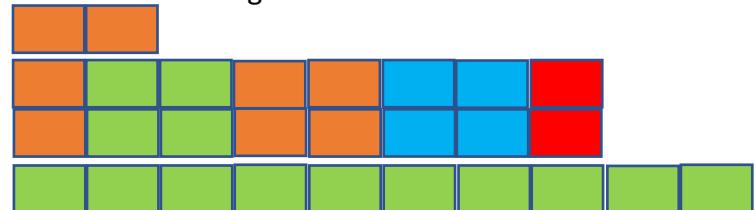
Dataframe & tibble

2d,heterogeneous



Lists

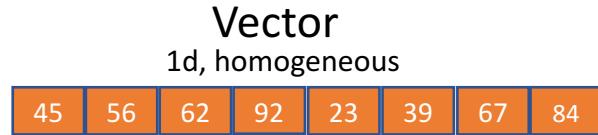
heterogeneous



Homogenous means that it can hold only one data type at a time.

Heterogeneous means it can hold multiple datatypes at a time.

Data Structures



Dataframe & tibble

2d,heterogeneous

a	cat	23					
b	dog	34					
c	bat	5					
d	bee	0.4					

columns

rows

Data Structures

Vector
1d, homogeneous

Index	45	56	62	92	23	39	67	84
	1	2	3	4	5	6	7	8

Dataframe & tibble
2d,heterogeneous

Index = row,col

	1,1	1,2	1,3					
1,1	a	cat	23					
2,1	b	dog	34					
3,1	c	bat	5					
4,1	d	bee	0.4					

columns

rows