MGMA2024:

A Review of R for Microbial Data Analysis

Explore the atlas1006 microbiome dataset

Duy Dao
khuongduying@gmail.com

CONTENT

1. Working with sample metadata
2. The atlas1006 dataset exploration
3. Data manipulating with R
4. Statistical analysis with R

Working with metadata

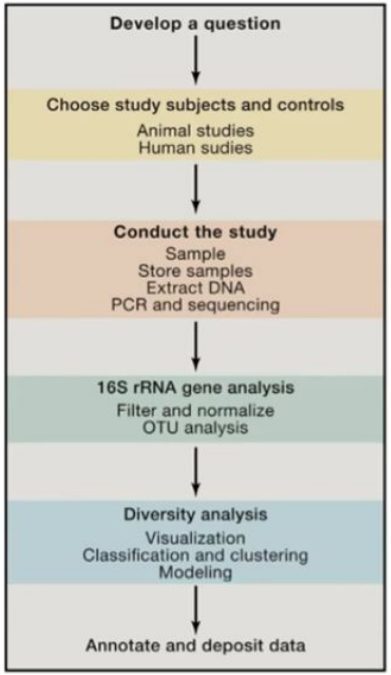
Performing a microbiome study

Study design
(power, randomization)
(metadata--standard + specific to your study)

Sample collection
(method, shipping/storage)
lab protocols
(include neg/pos controls)

Data processing

Data analysis



Sample Information

Sequence Data

Feature Table

Summary Statistics

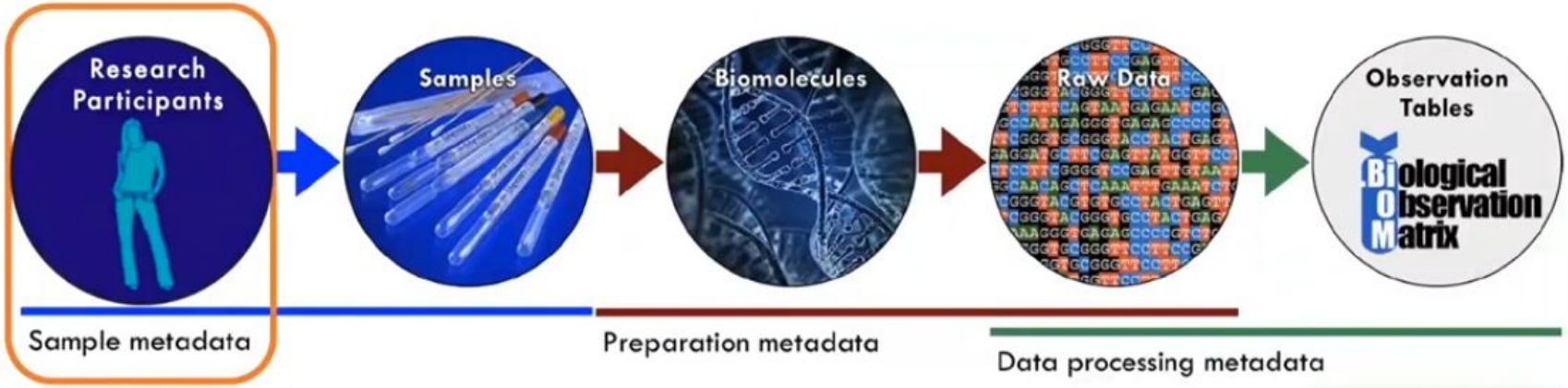
For more on these topics, see:
Conducting a Microbiome Study, by Goodrich et al. 2014
Reagent Contamination, Salter et al. 2014
Storage effects, by Song et al. 2016
Microbiome Quality Control (MBQC), by Sinha et al. 2017
MIMARKS, by Yilmaz et al. 2011
KatharoSeq low biomass workflow, by Minich et al. 2017

Other resources:
Earth Microbiome Project website
Human Microbiome Project website
American Gut Project website

What are Metadata ?

nmdc
National Microbiome Data Collaborative

Microbiome Metadata - Sample metadata



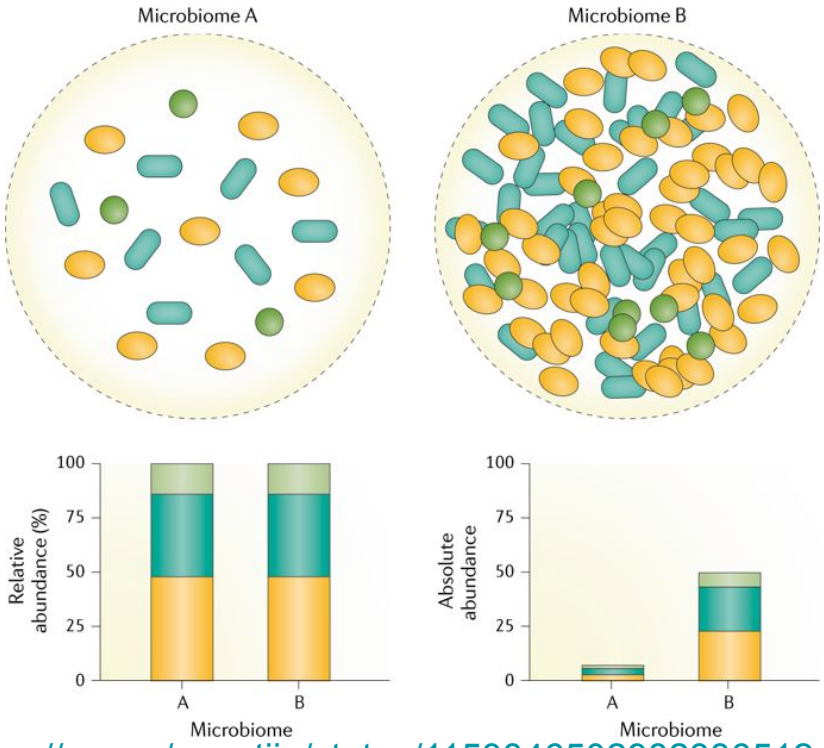
Adapted from: **Introduction to Metadata and Ontologies:** Everything You Always Wanted to Know About Metadata and Ontologies (But Were Afraid to Ask) DOI: 10.25979/1607365

https://www.youtube.com/watch?v=nCxZ8m7bcwY

Microbiome feature data & sample metadata

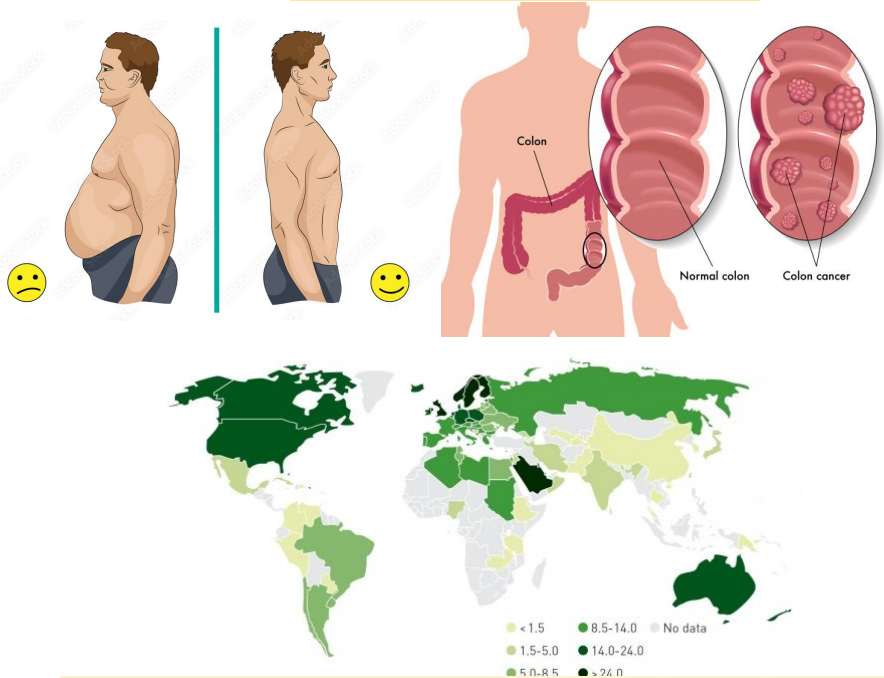
Combine between…

Feature data



https://x.com/econtijo/status/1153846502982336512

Sample metadata



→ This makes the microbiome analysis more meaningful

The atlas1006 dataset

Data from: Tipping elements in the human intestinal ecosystem

Lahti, Leo [1] ; Salojärvi, Jarkko [1] ; Salonen, Anne [1] ; Scheffer, Marten [2] ; de Vos, Willem M. [1]

Author affiliations

Published May 19, 2015 on Dryad. https://doi.org/10.5061/dryad.pk75d

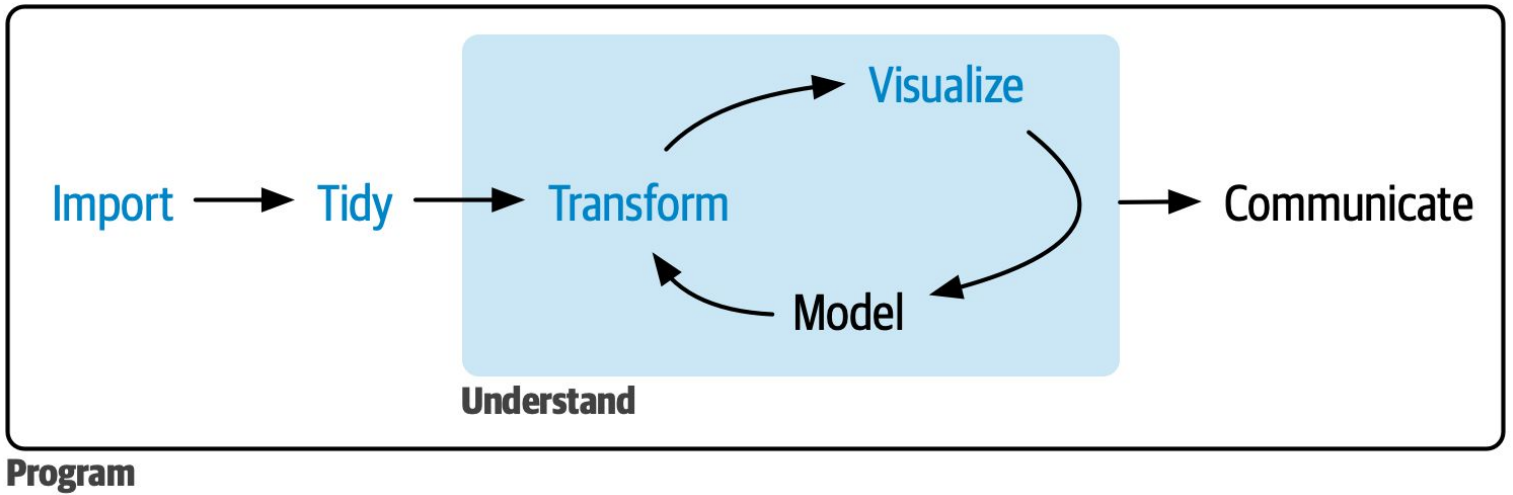
Cite this dataset

Lahti, Leo et al. (2015). Data from: Tipping elements in the human intestinal ecosystem [Dataset]. Dryad. https://doi.org/10.5061/dryad.pk75d

1006 individuals

https://doi.org/10.5061/dryad.pk75d

Data analysis workflow



https://r4ds.hadley.nz/whole-game

Import packages

```
# Install package
install.packages()

# Get the list of installed packages
installed.packages()

# Import package
library()
```

Import data

```
# Read CSV file
read.csv("file.csv")

#  Reads a file in table format more
flexibility in specifying the delimiter
read.table("file.txt")
```

Exploratory Data Analysis (EDA)

First look of our data

```
> atlas1006
phyloseq-class experiment-level object
otu_table()   OTU Table:          [ 130 taxa and 1151 samples ]
sample_data() Sample Data:        [ 1151 samples by 10 sample variables ]
tax_table()   Taxonomy Table:     [ 130 taxa by 3 taxonomic ranks ]
```

	age	sex	nationality	DNA_extraction_method	project	diversity	bmi_group	subject	time	sample
Sample-1	28	male	US	NA	1	5.76	severeobese	1	0	Sample-1
Sample-2	24	female	US	NA	1	6.06	obese	2	0	Sample-2
Sample-3	52	male	US	NA	1	5.5	lean	3	0	Sample-3
Sample-4	22	female	US	NA	1	5.87	underweight	4	0	Sample-4
Sample-5	25	female	US	NA	1	5.89	lean	5	0	Sample-5
…	…	…	…	…	…	…	…	…	…	…

Exploratory Data Analysis (EDA)

Explain the fields of atlas1006

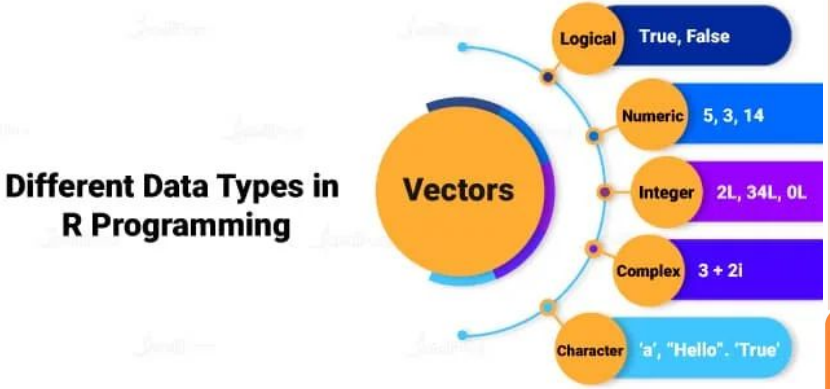
```
# First look of some first fields and values
head()
View()

# Explore the structure and variable type
str()
glimpse()
```

```
> str(df)
'data.frame':   1151 obs. of  10 variables:
 $ age                   : int  28 24 52 22 25 42 25 27 21 25 ...
 $ sex                   : Factor w/ 2 levels "female","male": 2 1 2 1 1 2 1 1 1 1 ...
 $ nationality           : Factor w/ 6 levels "CentralEurope",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ DNA_extraction_method: Factor w/ 3 levels "o","p","r": NA NA NA NA NA NA NA NA NA NA ...
 $ project               : Factor w/ 40 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ diversity             : num  5.76 6.06 5.5 5.87 5.89 5.53 5.49 5.38 5.34 5.64 ...
 $ bmi_group             : Factor w/ 6 levels "underweight",..: 5 4 2 1 2 2 1 2 2 2 ...
 $ subject               : Factor w/ 1006 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ time                  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sample                : chr  "Sample-1" "Sample-2" "Sample-3" "Sample-4" ...
```

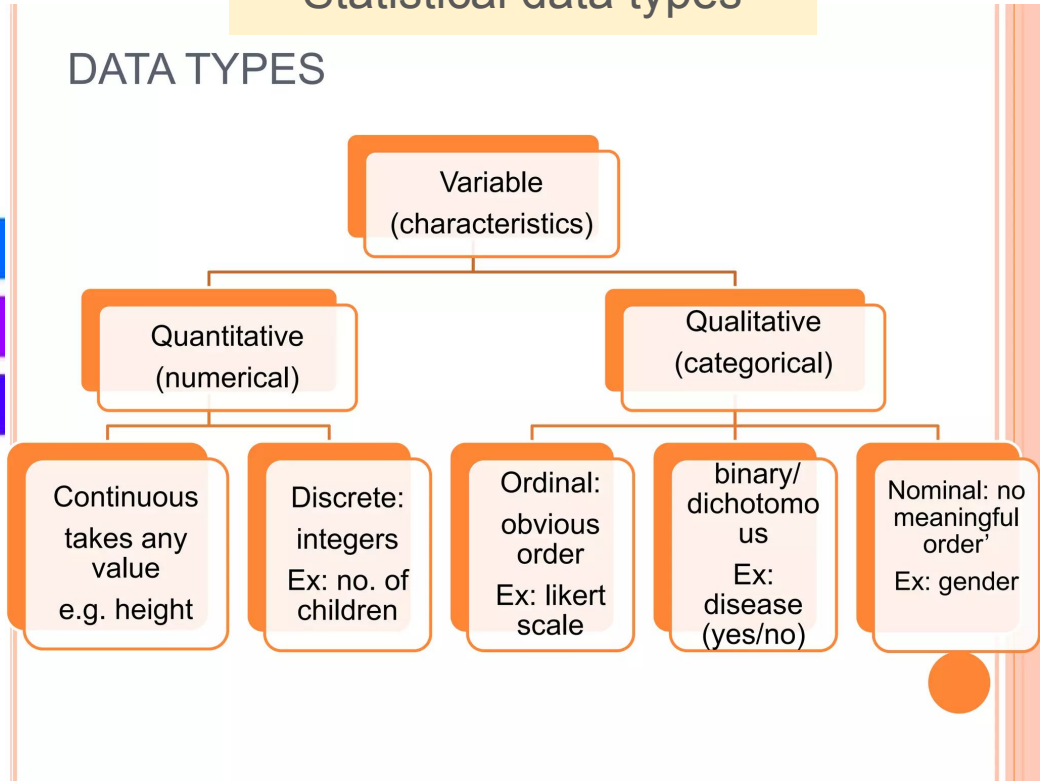
Exploratory Data Analysis (EDA)

A short mention about data type…

R Data types

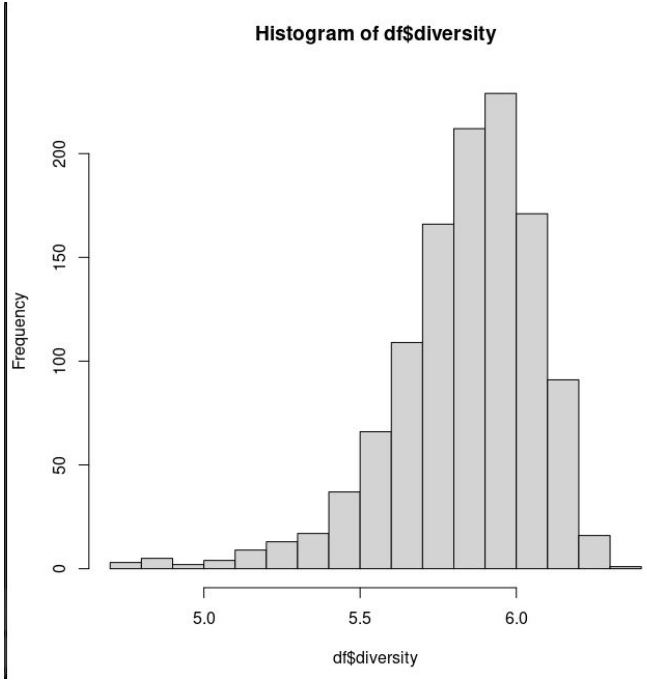


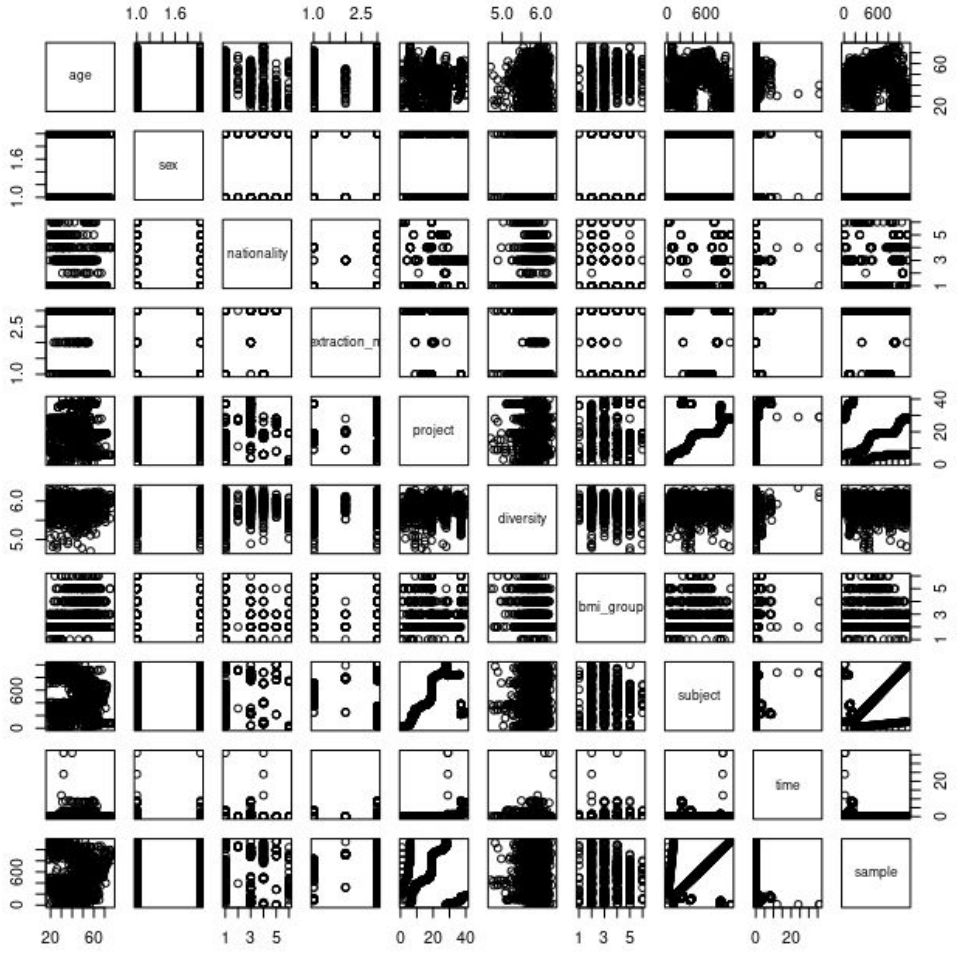
Statistical data types



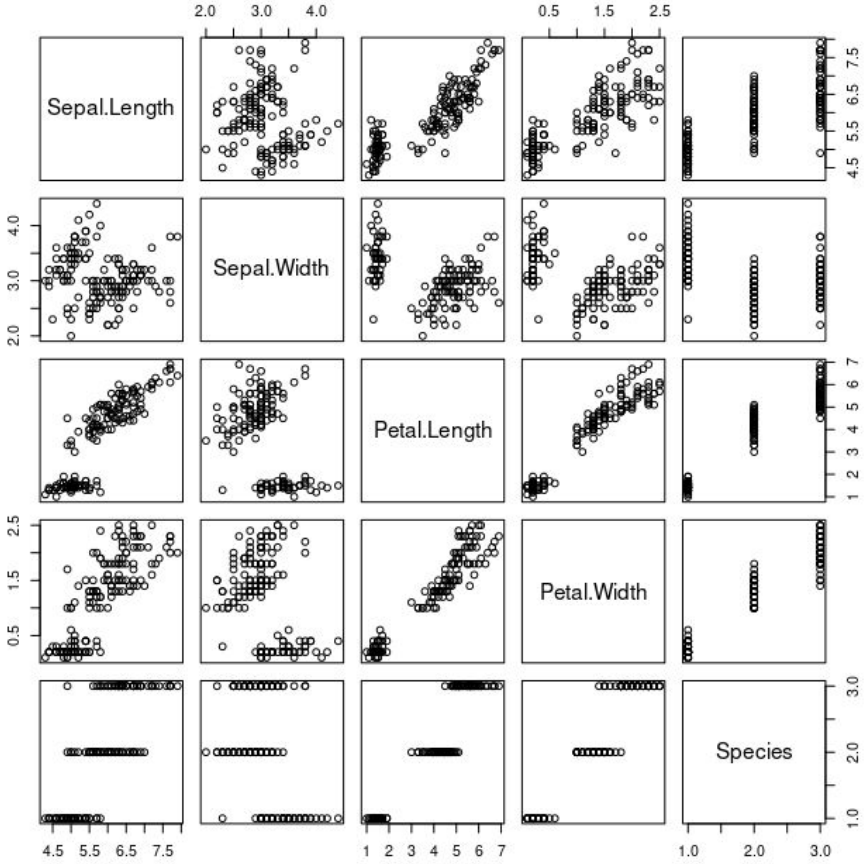
Exploratory Data Analysis (EDA)

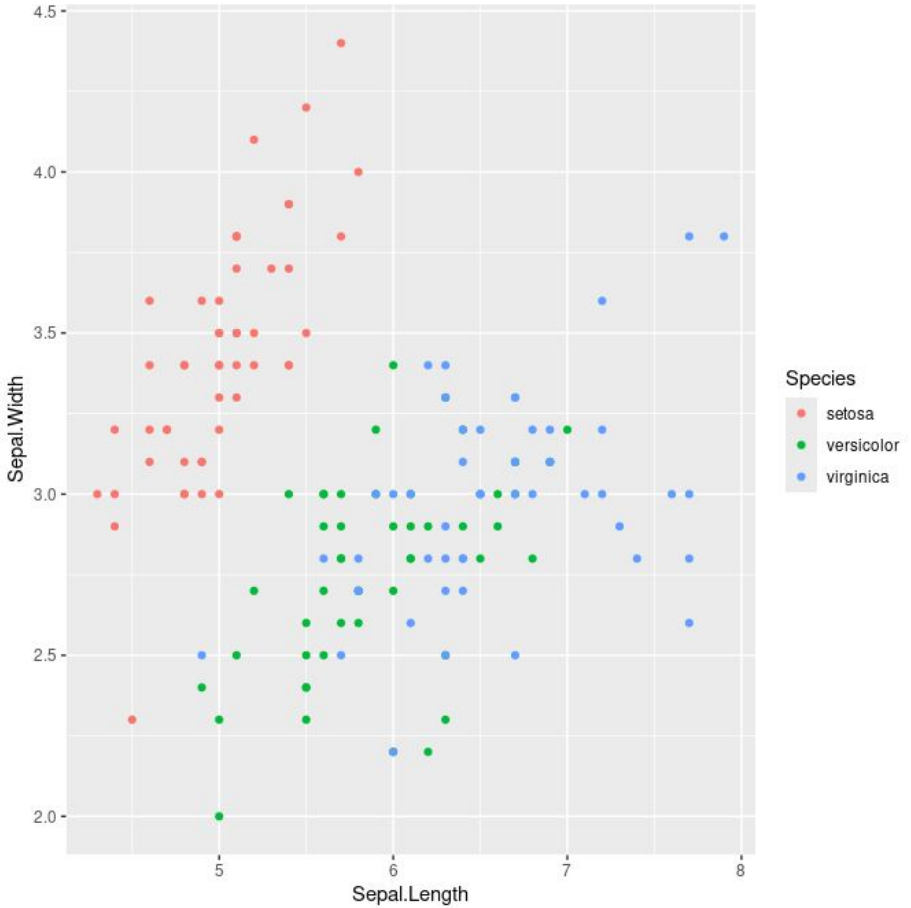
```
# Explore by visualizing
plot()
hist()
```





Explore iris





Exploratory Data Analysis (EDA)

```
# More advance
skim()
DataExplorer::create_report(df)
```

```
── Variable type: character ──────────────────────────────────────────────────
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 sample                0             1   8  11     0     1151          0

── Variable type: factor ─────────────────────────────────────────────────────
  skim_variable         n_missing complete_rate ordered n_unique
1 sex                          37         0.968 FALSE          2
2 nationality                  32         0.972 FALSE          6
3 DNA_extraction_method       209         0.818 FALSE          3
4 project                       0         1     FALSE         40
5 bmi_group                   106         0.908 FALSE          6
6 subject                       0         1     FALSE       1006
  top_counts
1 fem: 666, mal: 448
2 Cen: 650, Sca: 271, Sou: 89, UKI: 50
3 r: 510, o: 396, p: 36
4 19: 303, 7: 107, 28: 90, 9: 84
5 lea: 484, obe: 222, ove: 197, sev: 99
6 831: 5, 832: 5, 833: 5, 834: 5

── Variable type: numeric ────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean     sd   p0   p25   p50 p75  p100
1 age                  56         0.951 45.0   13.9   18   33    47     55 77
2 diversity             0         1      5.84   0.234  4.7  5.72  5.88   6  6.35
3 time                  0         1      0.438  2.12   0    0     0      0 36
  hist
```

Data Profiling Report

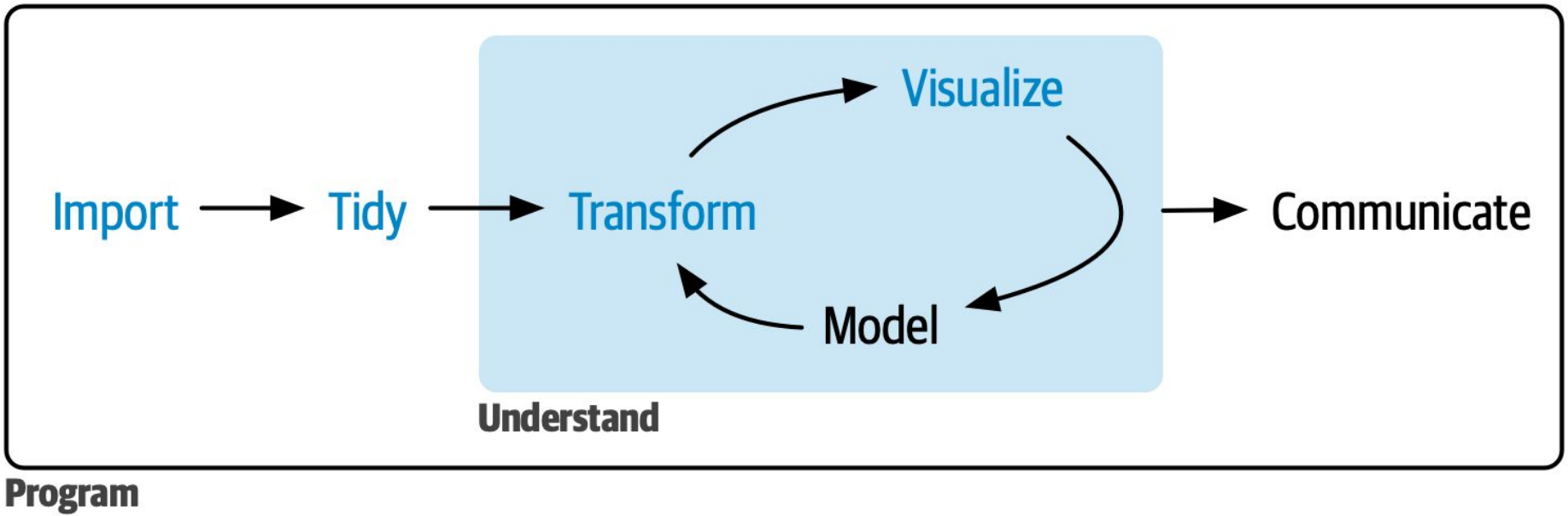
Basic Statistics

Raw Counts

Name	Value
Rows	1,151
Columns	10
Discrete columns	7
Continuous columns	3
All missing columns	0
Missing observations	440
Complete Rows	927
Total observations	11,510

Data manipulating with R

Data manipulation - The workflow



https://r4ds.hadley.nz/whole-game

Data manipulating with R

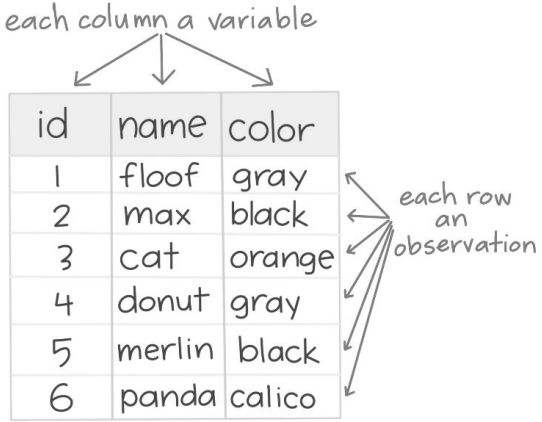
Introduction to data tidying

"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."

–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement

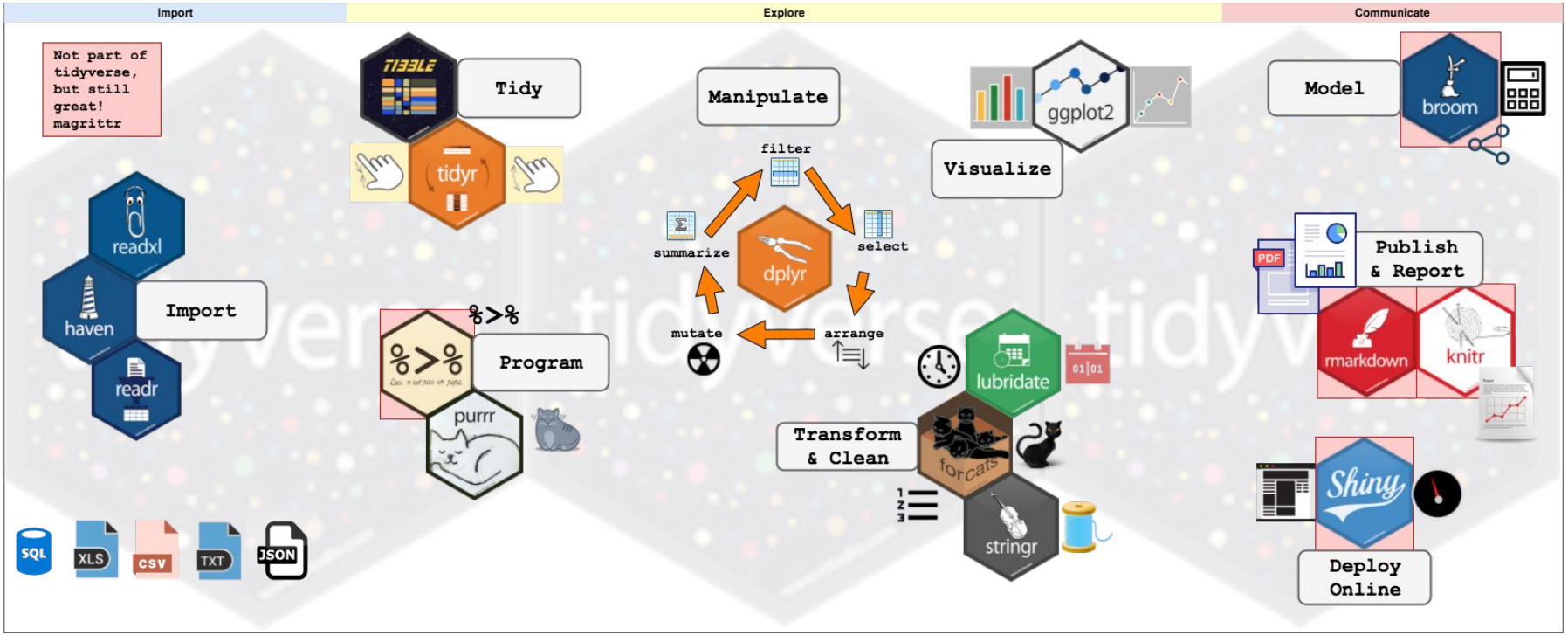


Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

https://openscapes.org/blog/2020-10-12-tidy-data/

Data manipulating with R

tidyverse package

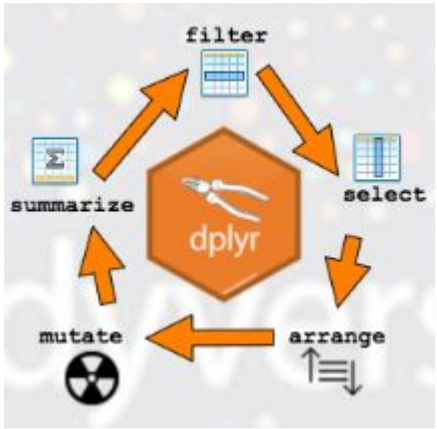


https://www.storybench.org/getting-started-with-tidyverse-in-r/

Data manipulating with R

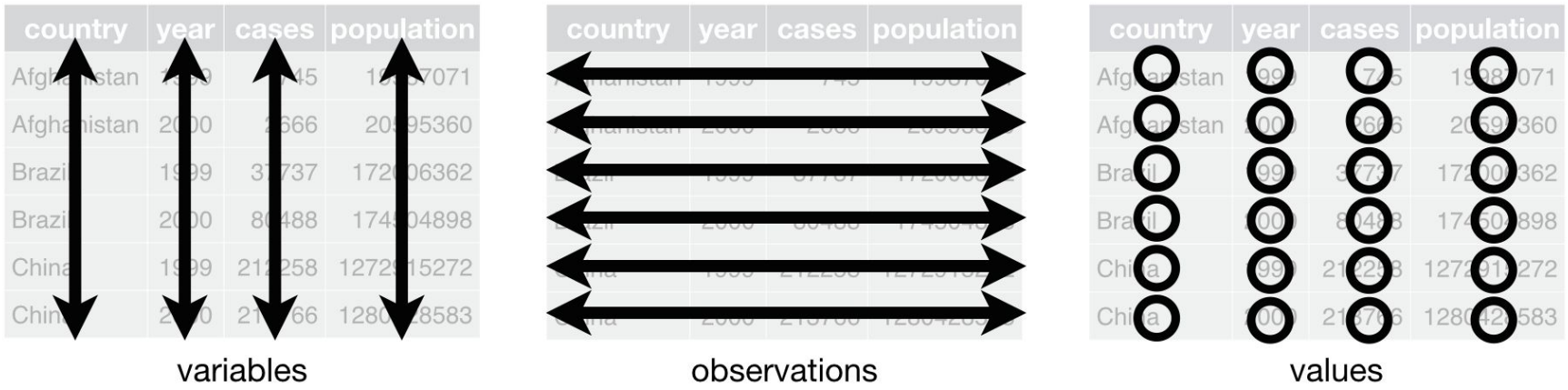
Data tidy - Cleaning and Filtering

```
# Select interested columns
select()

# Filter the row matched our interested
filter()

# Remove NAs
!is.na()
```



Data manipulating with R

Reshaping



The following three rules make a dataset tidy: variables are columns, observations are rows, and values are cells.

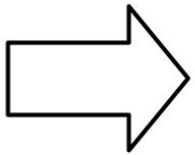
Data manipulating with R

Reshaping

```
# Reshape to a longer format
pivot_longer()

# Reshape to a wider format
pivot_wider()
```

id	bp1	bp2
A	100	120
B	140	115
C	120	125

Wide



id	measurement	value
A	bp1	100
A	bp2	120
B	bp1	140
B	bp2	115
C	bp1	120
C	bp2	125

Long

Data manipulating with R

From a wide table…

	Species	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	Sample-9	Sample-10
1	Actinomycetaceae	0	0	0	0	0	0	0	0	0	0
2	Aerococcus	0	0	0	0	0	0	0	0	0	0
3	Aeromonas	0	0	0	0	0	0	0	0	0	0
4	Akkermansia	21	36	475	61	34	14	27	21	78	169
5	Alcaligenes faecalis et rel.	1	1	1	2	1	1	1	1	1	1
6	Allistipes et rel.	72	127	34	344	50	57	21	9	9	23
7	Anaerobiospirillum	0	0	0	0	0	0	0	0	0	0
8	Anaerofustis	0	0	0	0	0	0	0	0	5	0
9	Anaerostipes caccae et rel.	176	108	27	203	232	166	32	1032	35	1309
10	Anaerotruncus colihominis et rel.	10	48	38	8	21	15	7	11	15	10

```
> dim(otu)
[1]  130 1151
```

Data manipulating with R

…to a long table

	Species	sample	count
1	Actinomycetaceae	Sample-1	0
2	Actinomycetaceae	Sample-2	0
3	Actinomycetaceae	Sample-3	0
4	Actinomycetaceae	Sample-4	0
5	Actinomycetaceae	Sample-5	0
6	Actinomycetaceae	Sample-6	0
7	Actinomycetaceae	Sample-7	0
8	Actinomycetaceae	Sample-8	0
9	Actinomycetaceae	Sample-9	0
10	Actinomycetaceae	Sample-10	0
11	Actinomycetaceae	Sample-11	0

```
# Reshape to a longer format
pivot_longer()
```

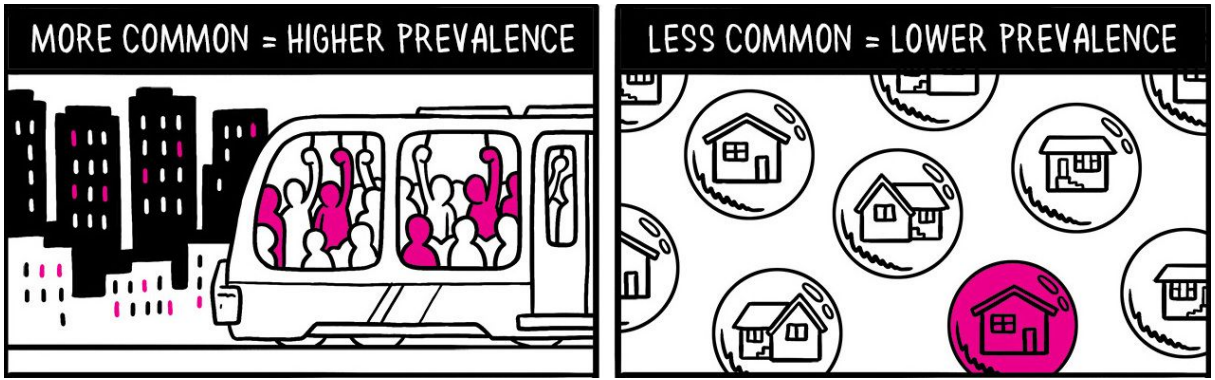
```
> dim(otu)
[1] 149630      3
```

Data manipulating with R - the practice

Examine the prevalence of microoranisms in atlas1006 cohort

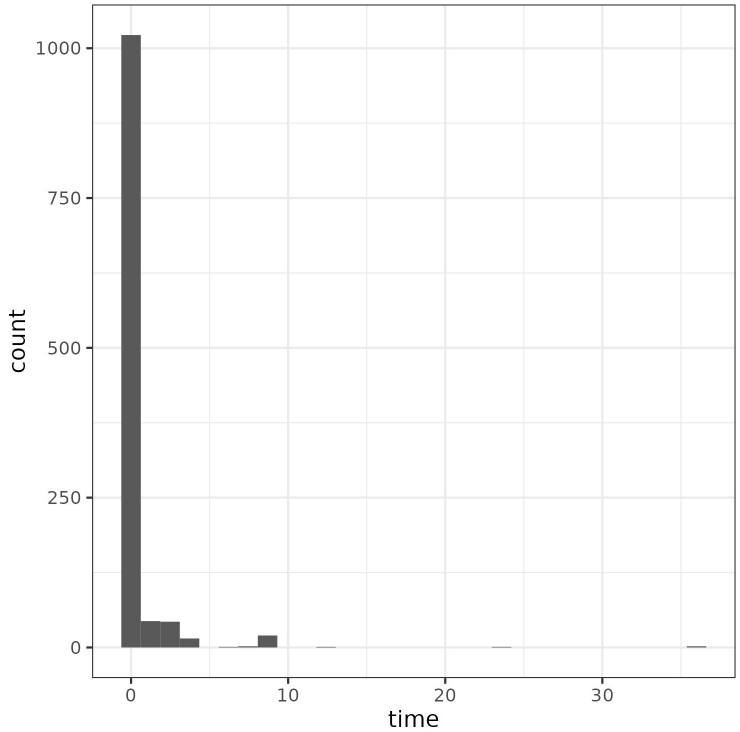
"The number of cases of a disease, number of infected people, or number of people with some other attribute present during a particular interval of time. It is often expressed as a rate (for example, the prevalence of diabetes per 1,000 people during a year)." - ***CDC***

$$\mathbf{P} = \frac{\text{Total number of cases}}{\text{Total population}} = \frac{C}{P}$$



Data manipulating with R - the practice

Our population has 1006 people, but why we have 1151 columns?



Based on time distribution,

→ There are some people whose samples collected more than 1 time

→ Filter it. Keep only the first time.

```
filter()
select()
```

```
> dim(meta)
[1] 1151   10
```

```
> dim(meta)
[1] 1006   10
```

Data manipulating with R - the practice

From count table…

	Species	Sample-1	Sample-2	Sample-3	Sample-4	Sample-5	Sample-6	Sample-7	Sample-8	Sample-9	Sample-10
1	Actinomycetaceae	0	0	0	0	0	0	0	0	0	0
2	Aerococcus	0	0	0	0	0	0	0	0	0	0
3	Aeromonas	0	0	0	0	0	0	0	0	0	0
4	Akkermansia	21	36	475	61	34	14	27	21	78	169
5	Alcaligenes faecalis et rel.	1	1	1	2	1	1	1	1	1	1
6	Allistipes et rel.	72	127	34	344	50	57	21	9	9	23
7	Anaerobiospirillum	0	0	0	0	0	0	0	0	0	0
8	Anaerofustis	0	0	0	0	0	0	0	0	5	0
9	Anaerostipes caccae et rel.	176	108	27	203	232	166	32	1032	35	1309
10	Anaerotruncus colihominis et rel.	10	48	38	8	21	15	7	11	15	10

Data manipulating with R - the practice

…to case table

	Species	Sample.1	Sample.2	Sample.3	Sample.4	Sample.5	Sample.6	Sample.7	Sample.8	Sample.9	Sample.10
1	Actinomycetaceae	0	0	0	0	0	0	0	0	0	0
2	Aerococcus	0	0	0	0	0	0	0	0	0	0
3	Aeromonas	0	0	0	0	0	0	0	0	0	0
4	Akkermansia	1	1	1	1	1	1	1	1	1	1
5	Alcaligenes faecalis et rel.	1	1	1	1	1	1	1	1	1	1
6	Allistipes et rel.	1	1	1	1	1	1	1	1	1	1
7	Anaerobiospirillum	0	0	0	0	0	0	0	0	0	0
8	Anaerofustis	0	0	0	0	0	0	0	0	1	0
9	Anaerostipes caccae et rel.	1	1	1	1	1	1	1	1	1	1
10	Anaerotruncus colihominis et rel.	1	1	1	1	1	1	1	1	1	1

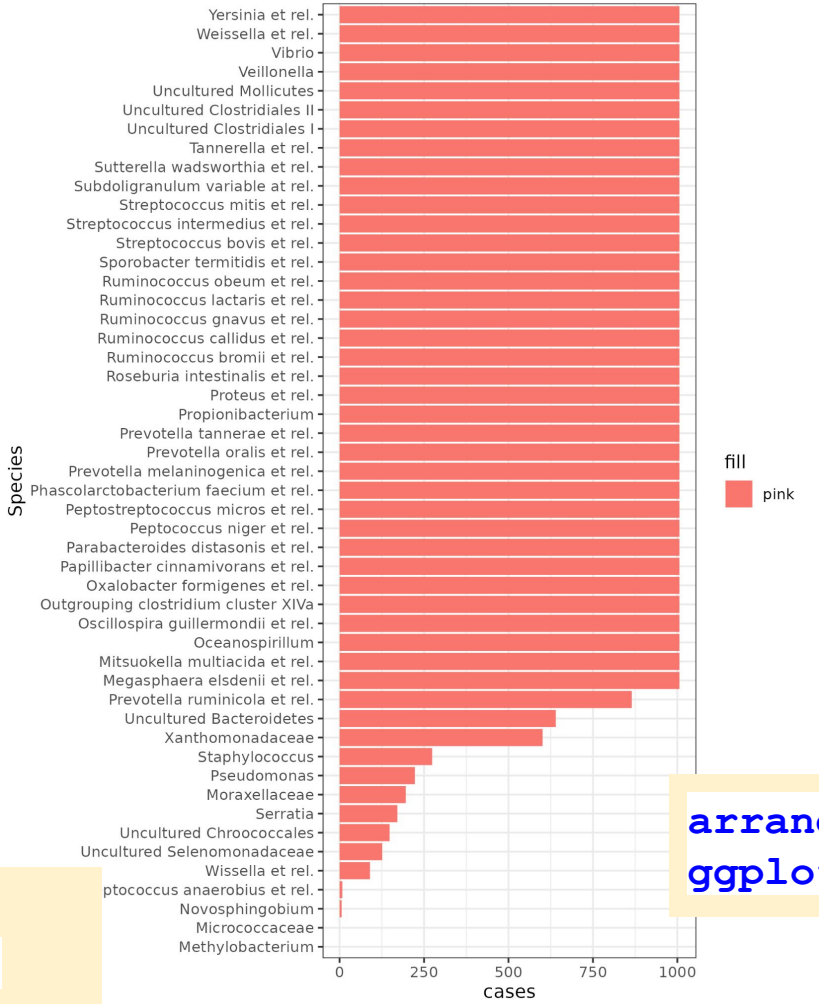
```
mutate()
if_else()
```

Data manipulating with R - the practice

…to prevalence table

	Species	cases	total
1	Actinomycetaceae	292	1006
2	Aerococcus	3	1006
3	Aeromonas	0	1006
4	Akkermansia	1006	1006
5	Alcaligenes faecalis et rel.	1006	1006
6	Allistipes et rel.	1006	1006
7	Anaerobiospirillum	2	1006
8	Anaerofustis	91	1006
9	Anaerostipes caccae et rel.	1006	1006
10	Anaerotruncus colihominis et rel.	1006	1006
11	Anaerovorax odorimutans et rel.	1006	1006
12	Aneurinibacillus	11	1006
13	Aquabacterium	93	
14	Asteroleplasma et rel.	0	

`sum()`
`across()`



`arrange()`
`ggplot()`

Statistical Analysis with R

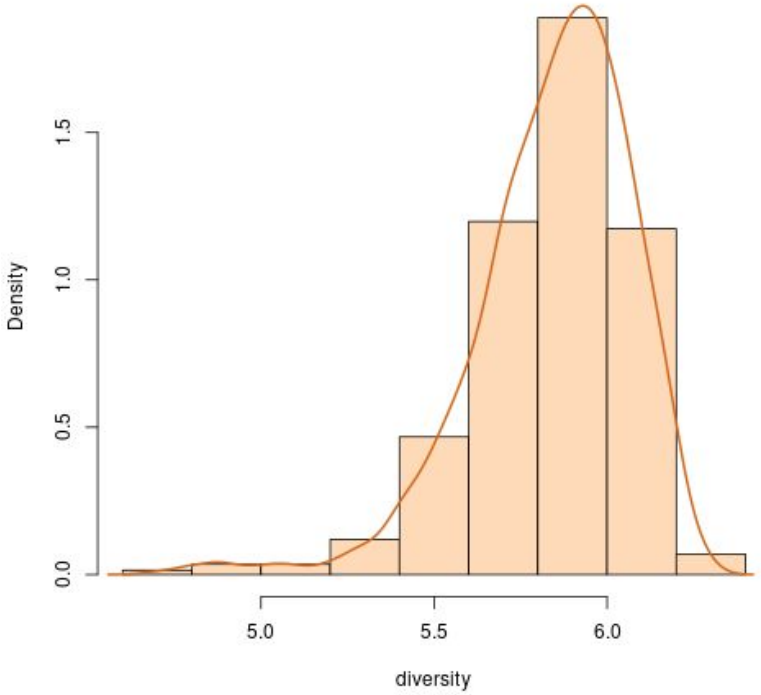
Ask a question

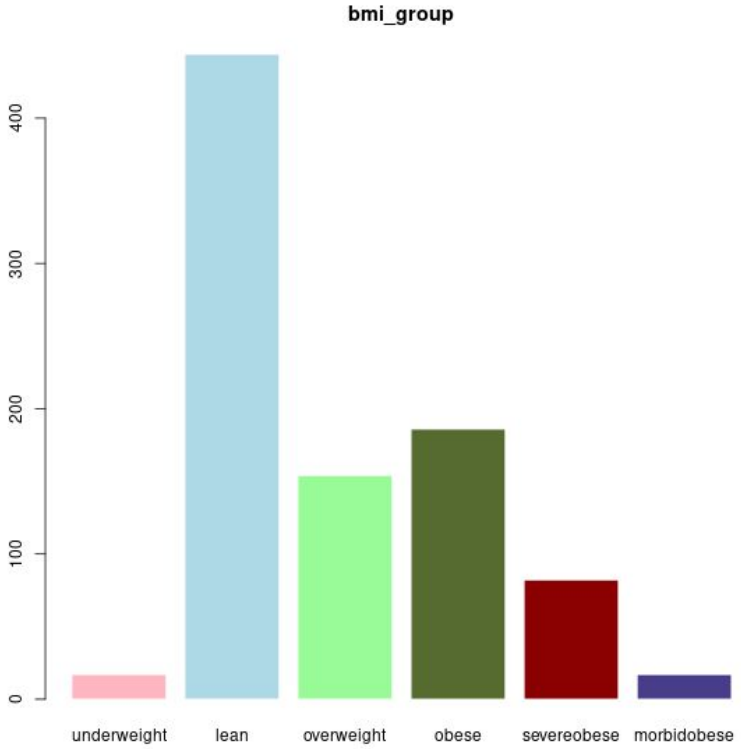
The question can be conducted before study design or after EDA

```
> str(df)
'data.frame':   1151 obs. of  10 variables:
 $ age                   : int  28 24 52 22 25 42 25 27 21 25 ...
 $ sex                   : Factor w/ 2 levels "female","male": 2 1 2 1 1 2 1 1 1 1 ...
 $ nationality           : Factor w/ 6 levels "CentralEurope",..: 6 6 6 6 6 6 6 6 6 6 ...
 $ DNA_extraction_method: Factor w/ 3 levels "o","p","r": NA NA NA NA NA NA NA NA NA NA ...
 $ project               : Factor w/ 40 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ diversity             : num  5.76 6.06 5.5 5.87 5.89 5.53 5.49 5.38 5.34 5.64 ...
 $ bmi_group             : Factor w/ 6 levels "underweight",..: 5 4 2 1 2 2 1 2 2 2 ...
 $ subject               : Factor w/ 1006 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ time                  : num  0 0 0 0 0 0 0 0 0 0 ...
 $ sample                : chr  "Sample-1" "Sample-2" "Sample-3" "Sample-4" ...
```

Ask a question

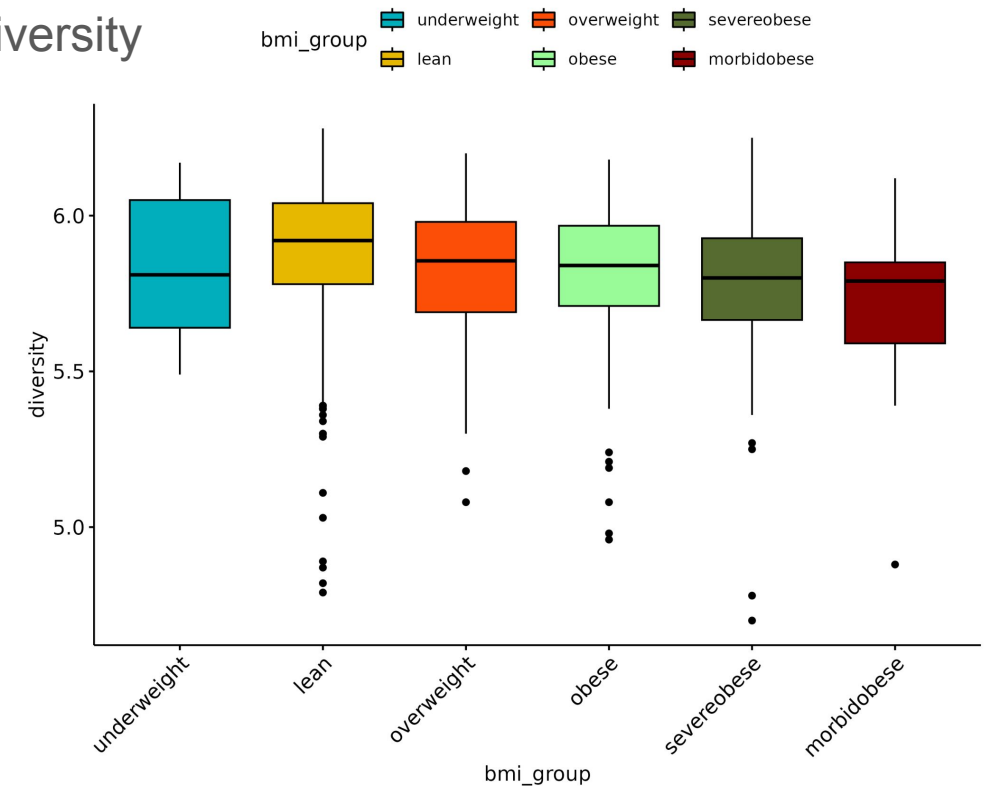
After doing EDA…





Ask a question

Based on the Shannon diversity



Do lean people have more diversity in interstitial microbiota?

Ask a question

Scientific question

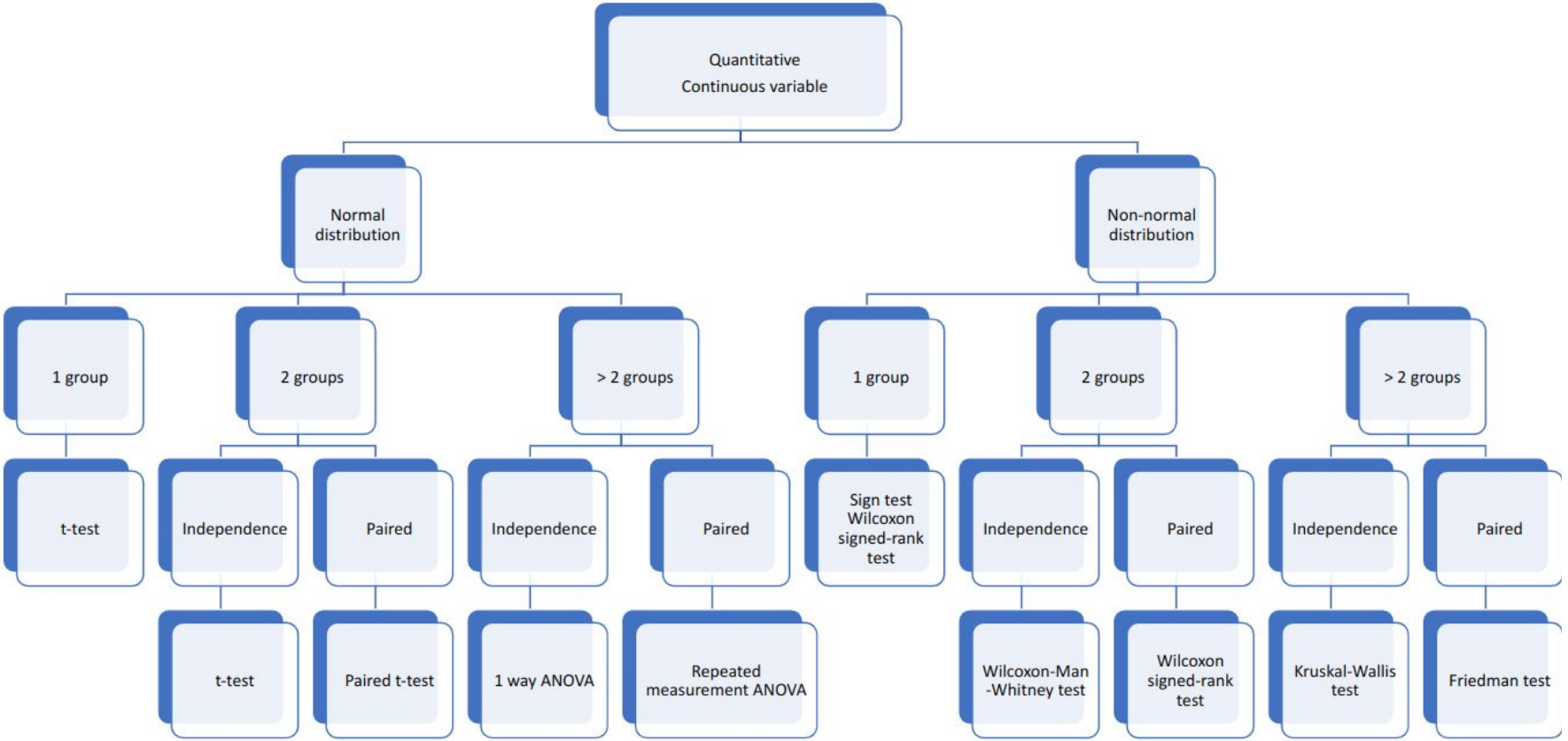
Is there any significant difference in **microbiome diversity** between those who have a **lean** body than those who are **overweight** or **obese**?

Hypothesis

Ho: No significant difference in microbiome diversity between individuals who have a lean body and those who are overweight or obese.

Ha: There is a significant difference in microbiome diversity between individuals who have a lean body and those who are overweight or obese.

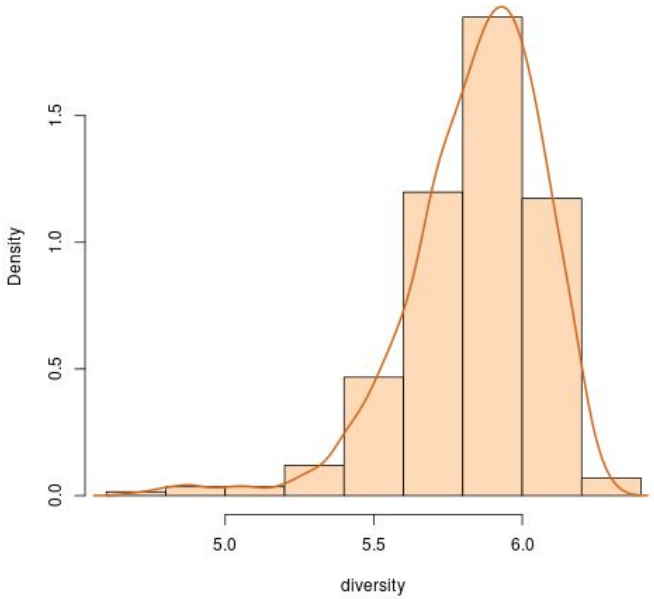
Conduct a hypothesis test

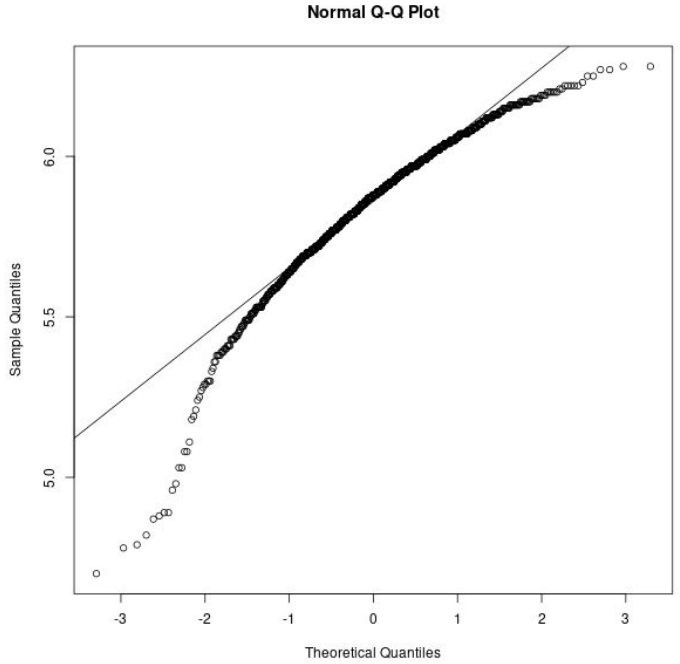


(*) This beautififul chart is from Tran Ba Thien

Conduct a hypothesis test

Normality check





Normaly distributed?

Choose hypothesis test

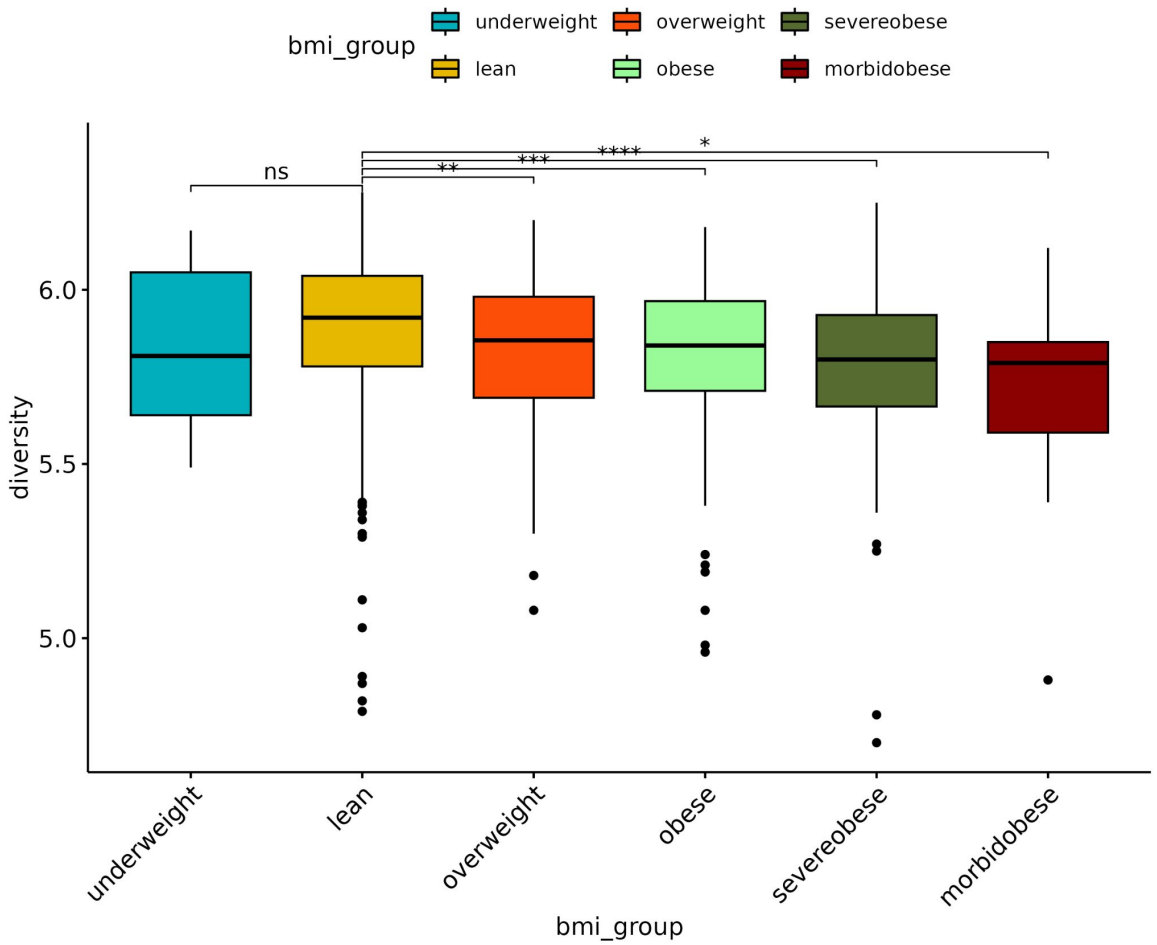
Normality check - Shapiro-Wilk test

```
> meta |> pull(diversity) |> shapiro.test()

        Shapiro-Wilk normality test

data:  pull(meta, diversity)
W = 0.93439, p-value < 2.2e-16
```

→ Not a normal distribution

→ Non-parametric test (Wilcoxon test)

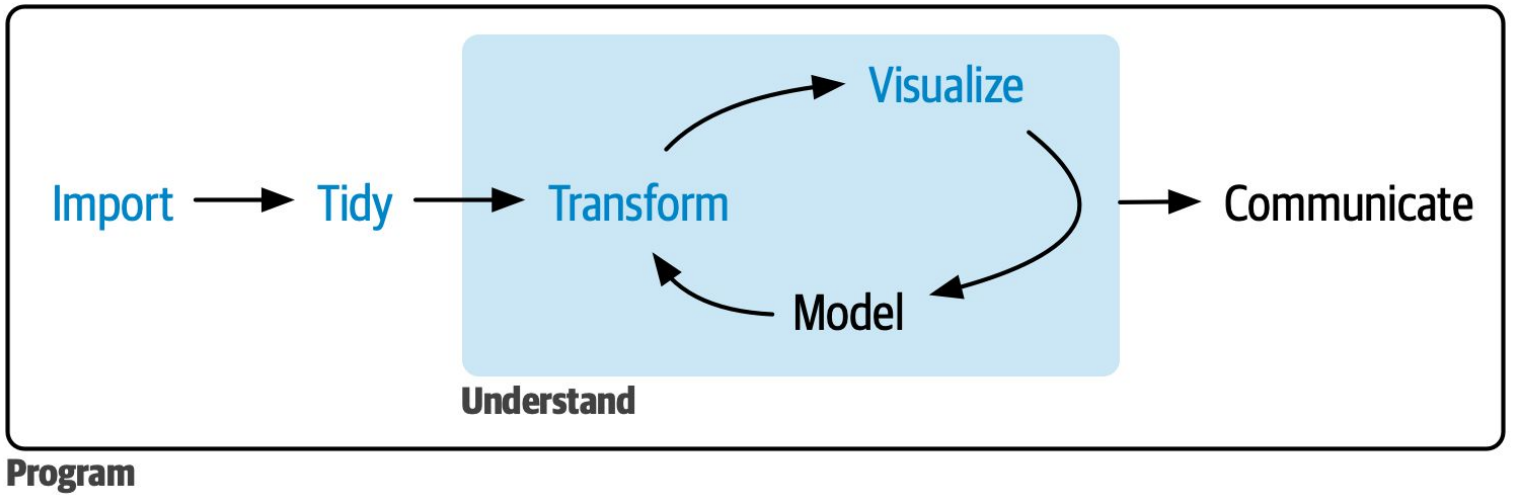
Conclusion



Reject Ho

→ Lean people have more diversity in intestinal microbiome

Conclusion


Import
Tidy
Transform
Visualize
Model
Communicate
Understand
Program

Further readings

Introduction to Urobiome Metadata Standards
https://www.youtube.com/watch?v=nCxZ8m7bcwY

Sample Metadata
https://www.youtube.com/watch?v=hh6pqmzJWds&t=22s

THANK YOU