talk06 练习与作业

目录

0.1 练习和作业说明

将相关代码填写入以 “‘{r} “‘ 标志的代码框中，运行并看到正确的结果；

完成后，用工具栏里的”Knit” 按键生成 PDF 文档；

将 PDF 文档改为：姓名-学号-`talk06` 作业`.pdf`，并提交到老师指定的平台/钉群。

0.2 Talk06 内容回顾

1. 3 个生信任务的 R 解决方案
2. factors 的更多应用 (forcats)
3. pipe

0.3 练习与作业：用户验证

请运行以下命令，验证你的用户名。

如你当前用户名不能体现你的真实姓名，请改为拼音后再运行本作业！

```
Sys.info()[["user"]]
```

```
## [1] "s56hh"
```

```
Sys.getenv("HOME")
```

```
## [1] "C:/Users/s56hh/Documents"
```

0.4 练习与作业 1：作图

0.4.1 用下面的数据作图

1. 利用下面代码读取一个样本的宏基因组相对丰度数据

```
abu <-
  read_delim(
    file = "../data/talk06/relative_abundance_for_RUN_ERR1072629_taxonlevel_species.txt
    delim = "\t", quote = "", comment = "#");
```

2. 取前 5 个丰度最高的菌，将其它的相对丰度相加并归为一类 `Qita`；
3. 用得到的数据画如下的空心 pie chart:

```
## 代码写这里，并运行；
library(tidyverse)
```

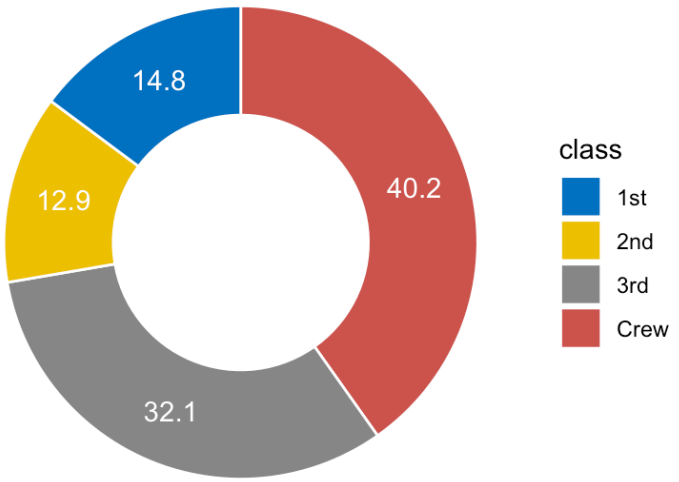


图 1: make a pie chart like this using the meteagenomics data

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dbplyr)
```

```
##
## 载入程辑包: 'dbplyr'
##
## The following objects are masked from 'package:dplyr':
##
##     ident, sql
```

```
library(ggplot2)
library(ggforce)
    abu <-
      read_delim(
        file = "../data/talk06/relative_abundance_for_RUN_ERR1072629_taxonlevel_species
        delim = "\t", quote = "", comment = "#");
```

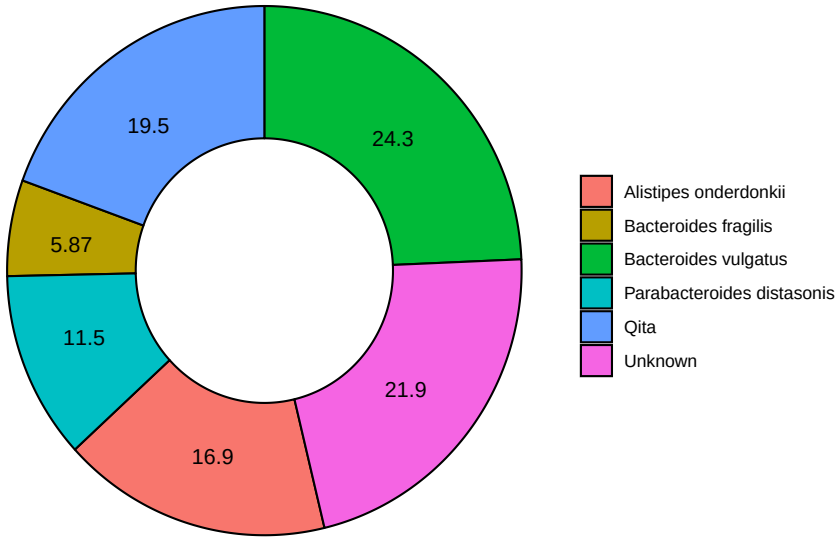
```
## Rows: 122 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (1): scientific_name
## dbl (2): ncbi_taxon_id, relative_abundance
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
fengdu<-abu %>% summarise(scientific_name=scientific_name,
                  relative_abundance=relative_abundance)%>%  arrange(-relative_abundanc
shenme<-head(fengdu,n=5L)
kkk<-abu %>% summarise(zonghe=mean(relative_abundance)*122)
lll<-shenme %>% summarise(zonghe=mean(relative_abundance)*5)
Qita=kkk-lll
shenme<-shenme%>%add_row(scientific_name="Qita",relative_abundance=as.numeric(Qita))
shenme
```

```
## # A tibble: 6 x 2
##   scientific_name            relative_abundance
##   <chr>                                   <dbl>
## 1 Bacteroides vulgatus                     24.3
## 2 Unknown                                  21.9
## 3 Alistipes onderdonkii                    16.9
## 4 Parabacteroides distasonis               11.5
## 5 Bacteroides fragilis                      5.87
## 6 Qita                                     19.5
```

```
A<-as.data.frame(shenme)
ggplot()+
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.ticks = element_blank(),
        axis.text.y = element_blank(),
        axis.text.x = element_blank(),
        legend.title=element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank())+
  xlab("")+ylab('')+
  geom_arc_bar(data=A,
               stat = "pie",
               aes(x0=0,y0=0,r0=1,r=2,
```

```
                    amount=relative_abundance,fill=scientific_name)
)+
annotate("text",x=1,y=1,label="24.3",angle=0)+
annotate("text",x=1.1,y=-0.9,label="21.9",angle=0)+
annotate("text",x=-0.4,y=-1.4,label="16.9",angle=0)+
annotate("text",x=-1.4,y=-0.5,label="11.5",angle=0)+
annotate("text",x=-1.5,y=0.2,label="5.87",angle=0)+
annotate("text",x=-0.9,y=1.1,label="19.5",angle=0)
```

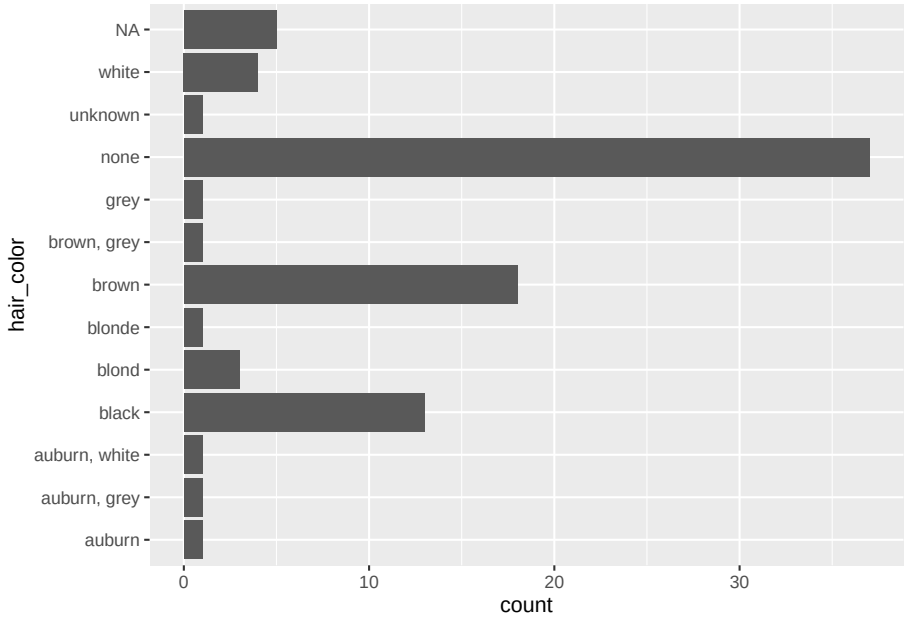


0.4.2 使用 starwars 变量做图

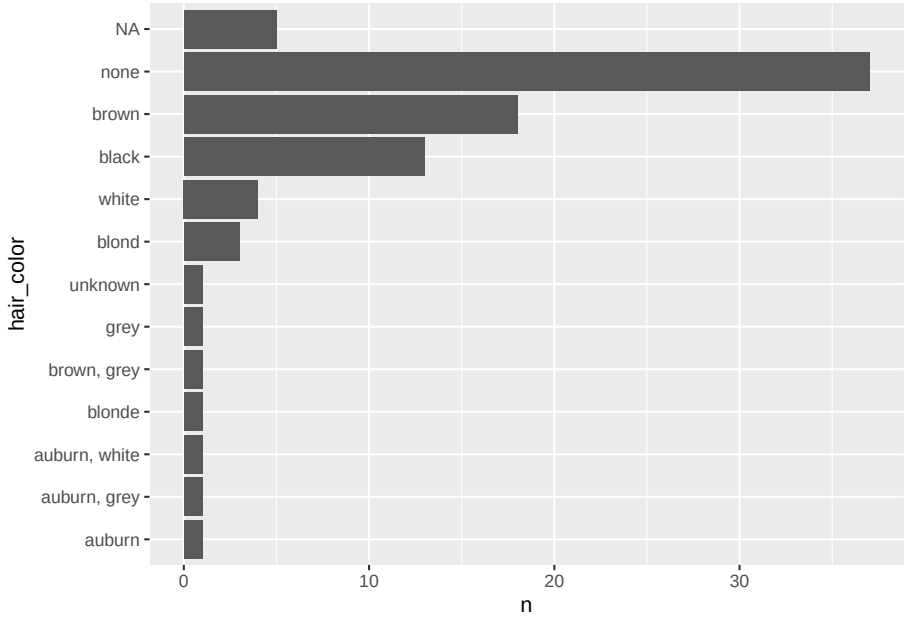
1. 统计 starwars 中 hair_color 的种类与人数时，可用下面的代码：

但是，怎么做到**按数量从小到大**排序？

```
library(dplyr)
library(ggplot2)
library(forcats)
ggplot(starwars, aes(x = hair_color)) +
  geom_bar() +
  coord_flip()
```

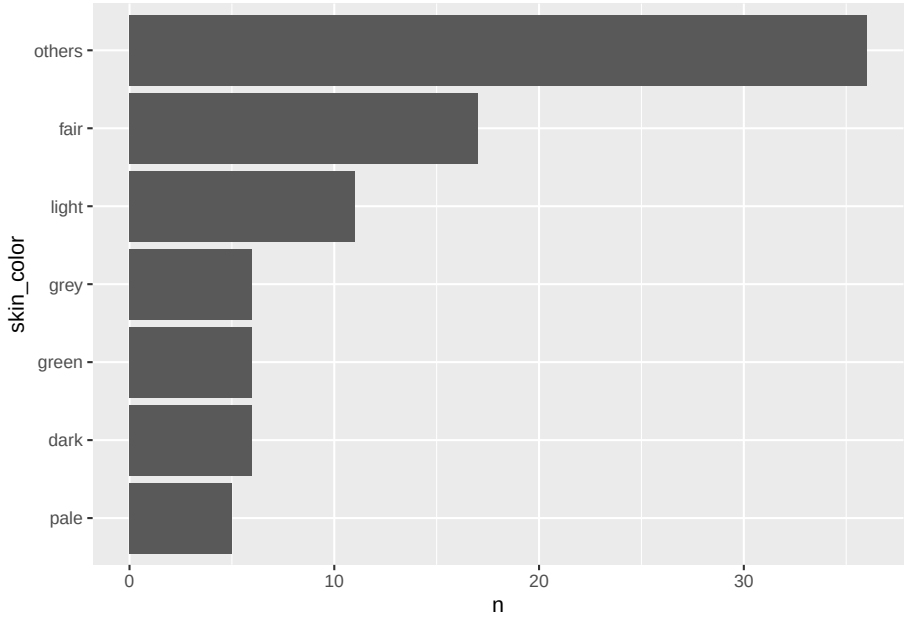



```
## 代码写这里，并运行；
library(dplyr)
library(ggplot2)
library(forcats)
starwars %>%
  count(hair_color) %>%
  mutate(hair_color=fct_reorder(hair_color,n)) %>%
  ggplot(aes(hair_color, n)) +
  geom_col()+coord_flip()
```



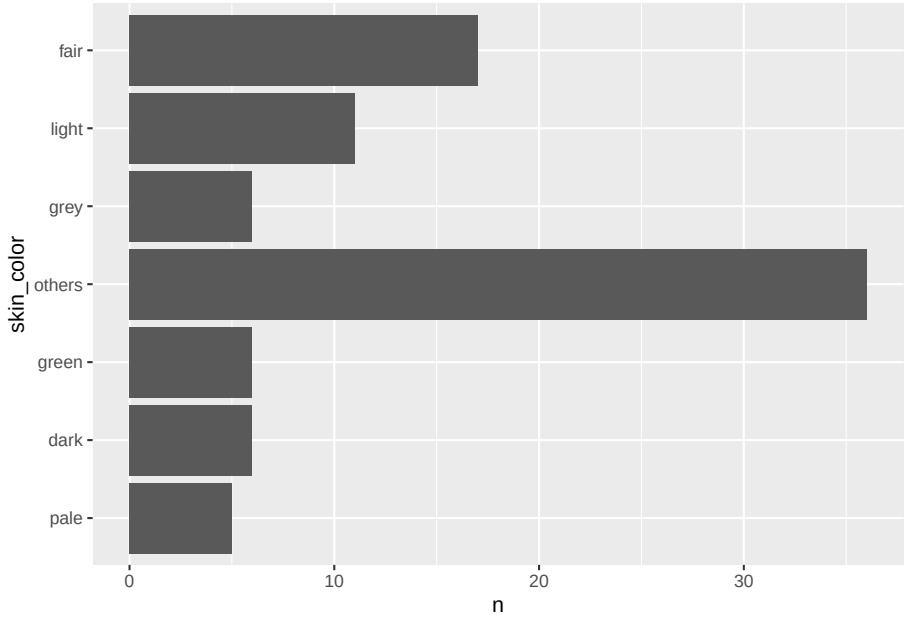
2. 统计 skin_color 时，将出现频率小于 0.05（即 5%）的颜色归为一类 Others，按出现次数排序后，做与上面类似的 barplot；

```
## 代码写这里，并运行；
library(dplyr)
library(ggplot2)
library(forcats)
dat_skin<-starwars %>%
  mutate(skin_color = fct_lump(skin_color, prop = .05, other_level = "others")) %>%
  count(skin_color, sort = TRUE)
dat_skin %>%
mutate(skin_color=fct_reorder(skin_color,n)) %>%
ggplot(aes(skin_color, n)) +
geom_col()+coord_flip()
```



3. 使用 2 的统计结果，但画图时，调整 bar 的顺序，使得 `Others` 处于第 4 的位置上。提示，可使用 `fct_relevel` 函数；

```
## 代码写这里，并运行；
library(dplyr)
library(ggplot2)
library(forcats)
dat_skin<-starwars %>%
  mutate(skin_color = fct_lump(skin_color, prop = .05, other_level = "others")) %>%
  count(skin_color, sort = TRUE)
dat_skin %>%
mutate(skin_color=fct_reorder(skin_color,n)) %>%
  mutate(skin_color=fct_relevel(skin_color,"others",after = 3)) %>%
ggplot(aes(skin_color, n)) +
geom_col()+coord_flip()
```



0.5 练习与作业 2：数据分析

0.5.1 使用 STRING PPI 数据分析并作图

1. 使用以下代码，装入 PPI 数据；

```
ppi <- read_delim( file = "../data/talk06/ppi900.txt.gz", col_names = T,
                   delim =  "\t", quote = "" );
```

2. **随机挑选**一个基因，得到类似于本章第一部分的互作网络图；

```
## 代码写这里，并运行；
ppi <- read_delim( file = "../data/talk06/ppi900.txt.gz", col_names = T,
                       delim =  "\t", quote = "" );
```

```
## Rows: 504436 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (2): gene1, gene2
## dbl (1): score
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
toppart <- ppi %>% filter( gene1 == "SALL4" ) %>%
arrange( desc( score ) ) %>% slice( 1:10 );

genes <- unique( c( "SALL4", toppart$gene2 ) );
netdata <- ppi %>% filter( gene1 %in% genes & gene2 %in% genes );
nrow(netdata);
```

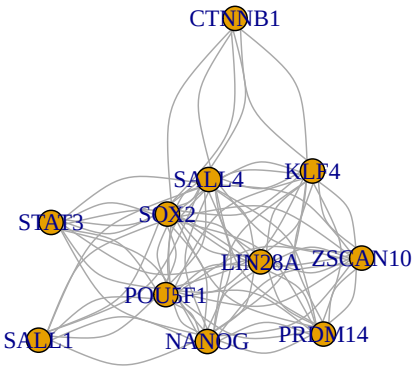
```
## [1] 80
```

```
if (!require("igraph")){
chooseCRANmirror();
install.packages("igraph");
}
```

```
## 载入需要的程辑包: igraph
##
## 载入程辑包: 'igraph'
##
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
##
## The following objects are masked from 'package:purrr':
##
##     compose, simplify
```

```
##
## The following object is masked from 'package:tidyr':
##
##     crossing
##
## The following object is masked from 'package:tibble':
##
##     as_data_frame
##
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum
##
## The following object is masked from 'package:base':
##
##     union
```

```
library( igraph );
netnet <- graph_from_data_frame( netdata, directed = FALSE );
plot(netnet);
```



0.5.2 对宏基因组相对丰度数据进行分析

1.data/talk06 目录下有 6 个文本文件，每个包含了一个宏基因组样本的分析结果：

```
relative_abundance_for_curated_sample_PRJEB6070-DE-073_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-074_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-075_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-076_at_taxonlevel_species.txt
relative_abundance_for_curated_sample_PRJEB6070-DE-077_at_taxonlevel_species.txt
```

2. 分别读取以上文件，提取 `scientific_name` 和 `relative_abundance` 两列；
3. 添加一列为样本名，比如 `PRJEB6070-DE-073`, `PRJEB6070-DE-074` … ；
4. 以 `scientific_name` 为 `key`，将其内容合并为一个 `data.frame` 或 `tibble`，其中每行为一个样本，每列为样本的物种相对丰度。注意：用 `join` 或者 `spread` 都可以，只要能解决问题。

5. 将 NA 值改为 0。

```
## 代码写这里，并运行；
library(tidyverse)
txt1<-read_table('data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-073_at
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ncbi_taxon_id = col_double(),
##   relative_abundance = col_character(),
##   taxon_rank_level = col_double(),
##   scientific_name = col_character()
## )

## Warning: 73 parsing failures.
## row col  expected    actual
##   1  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   2  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   3  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   4  -- 4 columns 6 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   5  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
## ... ... ......... ......... ..........................................................
## See problems(...) for more details.
```

```
txt2<-read_table('data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-074_at
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ncbi_taxon_id = col_double(),
##   relative_abundance = col_character(),
##   taxon_rank_level = col_double(),
##   scientific_name = col_character()
```

```
## )

## Warning: 77 parsing failures.
## row col  expected    actual
##   1  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   2  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   3  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   4  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   5  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
## ... ... ......... ......... ..........................................................
## See problems(...) for more details.
```

```
txt3<-read_table('data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-075_at
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ncbi_taxon_id = col_double(),
##   relative_abundance = col_character(),
##   taxon_rank_level = col_double(),
##   scientific_name = col_character()
## )

## Warning: 97 parsing failures.
## row col  expected    actual
##   1  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   2  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   3  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   4  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   5  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
## ... ... ......... ......... ..........................................................
## See problems(...) for more details.
```

```
txt4<-read_table('data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-076_at
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ncbi_taxon_id = col_double(),
##   relative_abundance = col_character(),
##   taxon_rank_level = col_double(),
##   scientific_name = col_character()
## )

## Warning: 88 parsing failures.
## row col  expected    actual
##   1  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   2  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   3  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   4  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   5  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
## ... ... ......... ......... ...........................................................
## See problems(...) for more details.
```

```
txt5<-read_table('data/talk06/relative_abundance_for_curated_sample_PRJEB6070-DE-077_at
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   ncbi_taxon_id = col_double(),
##   relative_abundance = col_character(),
##   taxon_rank_level = col_double(),
##   scientific_name = col_character()
## )

## Warning: 76 parsing failures.
```

```
## row col  expected    actual
##   1  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   2  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   3  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   4  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
##   5  -- 4 columns 5 columns 'data/talk06/relative_abundance_for_curated_sample_PRJEB
## ... ... ......... ......... ..................................................
## See problems(...) for more details.
```

```
txt1_a<-txt1 %>% summarise(scientific_name=scientific_name,
                           relative_abundance=taxon_rank_level)
txt2_a<-txt2 %>% summarise(scientific_name=scientific_name,
                           relative_abundance=taxon_rank_level)
txt3_a<-txt3 %>% summarise(scientific_name=scientific_name,
                           relative_abundance=taxon_rank_level)
txt4_a<-txt4 %>% summarise(scientific_name=scientific_name,
                           relative_abundance=taxon_rank_level)
txt5_a<-txt5 %>% summarise(scientific_name=scientific_name,
                           relative_abundance=taxon_rank_level)

txt1_a<-add_column(txt1_a,sample_name = "PRJEB6070-DE-073")
txt2_a<-add_column(txt2_a,sample_name = "PRJEB6070-DE-074")
txt3_a<-add_column(txt3_a,sample_name = "PRJEB6070-DE-075")
txt4_a<-add_column(txt4_a,sample_name = "PRJEB6070-DE-076")
txt5_a<-add_column(txt5_a,sample_name = "PRJEB6070-DE-077")

a12<-full_join(txt1_a,txt2_a)
```

```
## Joining, by = c("scientific_name", "relative_abundance", "sample_name")
```

```
a123<-full_join(a12,txt3_a)
```

```
## Joining, by = c("scientific_name", "relative_abundance", "sample_name")
```

```
a1234<-full_join(a123,txt4_a)
```

```
## Joining, by = c("scientific_name", "relative_abundance", "sample_name")
```

```
txt_a<-full_join(a1234,txt5_a)
```

```
## Joining, by = c("scientific_name", "relative_abundance", "sample_name")
```

```
txt_a
```

```
## # A tibble: 418 x 3
##    scientific_name  relative_abundance sample_name
##    <chr>                        <dbl> <chr>
##  1 Faecalibacterium              19.9  PRJEB6070-DE-073
##  2 [Eubacterium]                  9.49 PRJEB6070-DE-073
##  3 Bacteroides                    7.15 PRJEB6070-DE-073
##  4 Coprococcus                    5.02 PRJEB6070-DE-073
##  5 Roseburia                      4.69 PRJEB6070-DE-073
##  6 Bacteroides                    4.57 PRJEB6070-DE-073
##  7 Bacteroides                    4.36 PRJEB6070-DE-073
##  8 Ruminococcus                   4.24 PRJEB6070-DE-073
##  9 Alistipes                      2.59 PRJEB6070-DE-073
## 10 Bacteroides                    2.59 PRJEB6070-DE-073
## # ... with 408 more rows
```

```
txt_a[is.na(txt_a)]=0
txt_a
```

```
## # A tibble: 418 x 3
##    scientific_name  relative_abundance sample_name
##    <chr>                        <dbl> <chr>
##  1 Faecalibacterium              19.9  PRJEB6070-DE-073
##  2 [Eubacterium]                  9.49 PRJEB6070-DE-073
```

```
##  3 Bacteroides                  7.15 PRJEB6070-DE-073
##  4 Coprococcus                  5.02 PRJEB6070-DE-073
##  5 Roseburia                    4.69 PRJEB6070-DE-073
##  6 Bacteroides                  4.57 PRJEB6070-DE-073
##  7 Bacteroides                  4.36 PRJEB6070-DE-073
##  8 Ruminococcus                 4.24 PRJEB6070-DE-073
##  9 Alistipes                    2.59 PRJEB6070-DE-073
## 10 Bacteroides                  2.59 PRJEB6070-DE-073
## # ... with 408 more rows
```