Introduction to data wrangling

Ken Mwai

Quote

there are many 1M+1 to do something in R , we teach what we think is the best

Using packages in R

```
install.packages('name')
```

- Downloads the files to your computer
- Do this once per computer

```
# How to use the package
library('name')
```

- Loads the package
- Do this once per session

coherent system of packages for data manipulation

The tidyverse

The tidyverse

Tidyverse?

- The tidyverse is an opinionated collection of R packages designed for data science.
- All packages share an underlying design philosophy, grammar, and data structures.
- The tidyverse makes data science faster, easier and more fun

Installing the tidyverse package

```
install.packages('tidyverse')
```

How do we install and load the `rio` package?

```
install.packages('rio')
```

```
## Error in install.packages : Updating loaded packages
```

```
library(rio)
```

The tidyverse help

- https://ggplot2.tidyverse.org/
- https://www.tidyverse.org/learn/

Data frames and tibbles

- Data frames are the most common kind of data objects; used for rectangular data (like spreadsheets)
- Data frames: R's native data object
- Tibbles (tbl): a fancier enhanced kind of data frame
- (You really won't notice a difference in this class)

Before importing data ensure

1. You have the right project created
2. Have an RScript with the introdcution comments

Import data with rio

```
## Data wrangling code
## Author Ken Mwai
## July 2024
library(rio)
library(tidyverse)
liberia <- import('data/liberia_data.csv')
```

The tidyverse: dplyr

Dataset to use

- We use the `linelist` hostpital data
- It has data on hospital admissions –> *put it in the data folder*
- Shared via email

Import the data

```
## loading the liberia data
liberia <- import('data/liberia_data.csv')
## Load the linelist data
hosp_data <- import("data/line_hospital.xlsx")
```

Data overview

- Check variable names
- Check the number of variables
- Check the number of rows

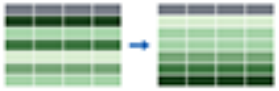
```
names(hosp_data)  # Check variable names
ncol(hosp_data) # Check the number of variables
nrow(hosp_data) # Check the number of rows
```

Checking the overall structure

```
glimpse(hosp_data)
```

```
## Rows: 198
## Columns: 14
## $ case_id         <chr> "5fe599", "8689b7", "11f8ea", "b8812a", "893f25", "be99c8", "07e3e8", "369449", "f393b4",
## $ `date onset`    <dttm> 2014-05-13, 2014-05-13, 2014-05-16, 2014-05-18, 2014-05-21, 2014-05-22, 2014-05-27, 2014
## $ `hosp date`     <dttm> 2014-05-15, 2014-05-14, 2014-05-18, 2014-05-20, 2014-05-22, 2014-05-23, 2014-05-29, 2014
## $ date_of_outcome <dttm> NA, 2014-05-18, 2014-05-30, NA, 2014-05-29, 2014-05-24, 2014-06-01, 2014-06-07, 2014-06-
## $ outcome         <chr> NA, "Recover", "Recover", NA, "Recover", "Recover", "Recover", "Death", "Recover", "Death
## $ gender          <chr> "m", "f", "m", "f", "m", NA, "f", "f", "m", "f", NA, "m", "m", NA, "f", "m", "f", "f", "f
## $ age             <dbl> 2, 3, 56, 18, 3, 16, 16, 0, 61, 27, 12, 42, 19, 7, 7, 13, 35, 17, 11, 11, 19, 54, 14, 28,
## $ wt_kg           <dbl> 27, 25, 91, 41, 36, 56, 47, 0, NA, 69, 67, 84, 68, NA, 34, 66, 78, 47, 53, 47, 71, 86, 53
## $ ht_cm           <dbl> 48, 59, 238, 135, 71, 116, 87, 11, NA, 174, 112, 186, 174, NA, 91, 152, 214, 137, 117, 13
## $ ct_blood        <dbl> 22, 22, 21, 23, 23, 21, 21, 22, 22, 22, 22, 22, 22, 21, 23, 22, 23, 21, 22, 23, 21, 23, 2
## $ chills          <chr> "no", NA, NA, "no", "no", "no", NA, "no", "no", "no", "no", "no", "no", "no", "no", "no",
## $ cough           <chr> "yes", NA, NA, "no", "yes", "yes", NA, "yes", "yes", "yes", "yes", "yes", "yes", "yes", "
## $ aches           <chr> "no", NA, NA, "no", "no", "no", NA, "no", "no", "no", "no", "no", "no", "no", "no", "yes"
## $ vomit           <chr> "yes", NA, NA, "no", "yes", "yes", NA, "yes", "yes", "no", "yes", "no", "no", "no", "yes"
```

dplyr: functions for manipulating data

Extract rows with `filter()`	
Extract columns with `select()`	
Arrange/sort rows with `arrange()`	
Make new columns with `mutate()`	
Make group summaries with `group_by() %>% summarize()`	

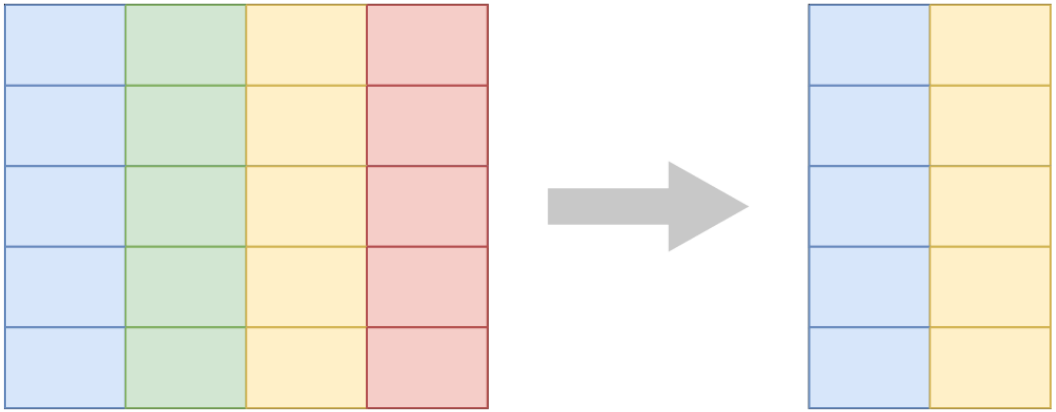
dplyr: select

Select a subset of variables

```
select(.data = DATA, ...)
```

- `DATA` = Data frame to transform
- `...` = variables to select

Select specific columns



Our data

```
glimpse(hosp_data)
```

```
## Rows: 198
## Columns: 14
## $ case_id         <chr> "5fe599", "8689b7", "11f8ea", "b8812a", "893f25", "be99c8", "07e3e8", "369449", "f393b4",
## $ `date onset`    <dttm> 2014-05-13, 2014-05-13, 2014-05-16, 2014-05-18, 2014-05-21, 2014-05-22, 2014-05-27, 2014
## $ `hosp date`     <dttm> 2014-05-15, 2014-05-14, 2014-05-18, 2014-05-20, 2014-05-22, 2014-05-23, 2014-05-29, 2014
## $ date_of_outcome <dttm> NA, 2014-05-18, 2014-05-30, NA, 2014-05-29, 2014-05-24, 2014-06-01, 2014-06-07, 2014-06-
## $ outcome         <chr> NA, "Recover", "Recover", NA, "Recover", "Recover", "Recover", "Death", "Recover", "Death
## $ gender          <chr> "m", "f", "m", "f", "m", NA, "f", "f", "m", "f", NA, "m", "m", NA, "f", "m", "f", "f", "f
## $ age             <dbl> 2, 3, 56, 18, 3, 16, 16, 0, 61, 27, 12, 42, 19, 7, 7, 13, 35, 17, 11, 11, 19, 54, 14, 28,
## $ wt_kg           <dbl> 27, 25, 91, 41, 36, 56, 47, 0, NA, 69, 67, 84, 68, NA, 34, 66, 78, 47, 53, 47, 71, 86, 53
## $ ht_cm           <dbl> 48, 59, 238, 135, 71, 116, 87, 11, NA, 174, 112, 186, 174, NA, 91, 152, 214, 137, 117, 13
## $ ct_blood        <dbl> 22, 22, 21, 23, 23, 21, 21, 22, 22, 22, 22, 22, 22, 21, 23, 22, 23, 21, 22, 23, 21, 23, 2
## $ chills          <chr> "no", NA, NA, "no", "no", "no", NA, "no", "no", "no", "no", "no", "no", "no", "no", "no",
## $ cough           <chr> "yes", NA, NA, "no", "yes", "yes", NA, "yes", "yes", "yes", "yes", "yes", "yes", "yes", "
## $ aches           <chr> "no", NA, NA, "no", "no", "no", NA, "no", "no", "no", "no", "no", "no", "no", "no", "yes"
## $ vomit           <chr> "yes", NA, NA, "no", "yes", "yes", NA, "yes", "yes", "no", "yes", "no", "no", "no", "yes"
```

Task

- Create a data frame `sub1` with case_id, date of onset ,gender and age
- NB: remember how to create objects in R

```
select(.data = DATA, variables , to , select)
```

```
select(hosp_data,case_id,"date onset",gender, age )
```

```
##     case_id date onset gender age
## 1    5fe599 2014-05-13      m   2
## 2    8689b7 2014-05-13      f   3
## 3    11f8ea 2014-05-16      m  56
## 4    b8812a 2014-05-18      f  18
## 5    893f25 2014-05-21      m   3
## 6    be99c8 2014-05-22   <NA>  16
## 7    07e3e8 2014-05-27      f  16
## 8    369449 2014-06-02      f   0
## 9    f393b4 2014-06-05      m  61
## 10   1389ca 2014-06-05      f  27
## 11   2978ac 2014-06-06   <NA>  12
## 12   57a565 2014-06-13      m  42
## 13   fc15ef 2014-06-16      m  19
## 14   2eaa9a 2014-06-17   <NA>   7
## 15   bbfa93 2014-06-18      f   7
## 16   c97dd9 2014-06-19      m  13
## 17   f50e8a 2014-06-22      f  35
## 18   3a7673 2014-06-23      f  17
## 19   7f5a01 2014-06-25      f  11
## 20   ddddee 2014-06-26   <NA>  11
## 21   99e8fa 2014-06-28      m  19
## 22   567136 2014-07-02   <NA>  54
## 23   9371a9 2014-07-08      f  14
## 24   bc2adf 2014-07-09      m  28
## 25   403057 2014-07-09      f   6
## 26   8bd1e8 2014-07-10      m   3
## 27   f327be 2014-07-12      m  31
## 28   42e1a9 2014-07-12      f   6
## 29   90e5fe 2014-07-13      m  67
## 30   959170 2014-07-13      f  14
## 31   8ebf6e 2014-07-14      f  10
## 32   e56412 2014-07-15      f  21
```

Pick rows: rename()

dplyr: rename

rename columns

```
rename(.data = DATA,
     'new_name' = 'old_name')
```

- `DATA` = Data frame to transform
- “new_name”=“old_name”

- Why rename?
- Rename variables names with spaces
- Shorten very long variable names
- have a common naming approach

```
select(hosp_data,case_id,"date onset",gender, age )
```

```
rename(hosp_data,"date_onset"='date onset' )
rename(hosp_data,"hosp_date"='hosp date' )
```

```
hosp_data <- rename(hosp_data,"date_onset"='date onset' )
hosp_data <- rename(hosp_data,"hosp_date"='hosp date' )
glimpse(hosp_data)
```

```
## Rows: 198
## Columns: 14
## $ case_id         <chr> "5fe599", "8689b7", "11f8ea", "b8812a", "893f25", "be99c8", "07e3e8", "369449", "f393b4",
## $ date_onset      <dttm> 2014-05-13, 2014-05-13, 2014-05-16, 2014-05-18, 2014-05-21, 2014-05-22, 2014-05-27, 2014
## $ hosp_date       <dttm> 2014-05-15, 2014-05-14, 2014-05-18, 2014-05-20, 2014-05-22, 2014-05-23, 2014-05-29, 2014
## $ date_of_outcome <dttm> NA, 2014-05-18, 2014-05-30, NA, 2014-05-29, 2014-05-24, 2014-06-01, 2014-06-07, 2014-06-
## $ outcome         <chr> NA, "Recover", "Recover", NA, "Recover", "Recover", "Recover", "Death", "Recover", "Death
## $ gender          <chr> "m", "f", "m", "f", "m", NA, "f", "f", "m", "f", NA, "m", "m", NA, "f", "m", "f", "f", "f
## $ age             <dbl> 2, 3, 56, 18, 3, 16, 16, 0, 61, 27, 12, 42, 19, 7, 7, 13, 35, 17, 11, 11, 19, 54, 14, 28,
## $ wt_kg           <dbl> 27, 25, 91, 41, 36, 56, 47, 0, NA, 69, 67, 84, 68, NA, 34, 66, 78, 47, 53, 47, 71, 86, 53
## $ ht_cm           <dbl> 48, 59, 238, 135, 71, 116, 87, 11, NA, 174, 112, 186, 174, NA, 91, 152, 214, 137, 117, 13
## $ ct_blood        <dbl> 22, 22, 21, 23, 23, 21, 21, 22, 22, 22, 22, 22, 22, 21, 23, 22, 23, 21, 22, 23, 21, 23, 2
## $ chills          <chr> "no", NA, NA, "no", "no", "no", NA, "no", "no", "no", "no", "no", "no", "no", "no", "no",
## $ cough           <chr> "yes", NA, NA, "no", "yes", "yes", NA, "yes", "yes", "yes", "yes", "yes", "yes", "yes", "
## $ aches           <chr> "no", NA, NA, "no", "no", "no", NA, "no", "no", "no", "no", "no", "no", "no", "no", "yes"
## $ vomit           <chr> "yes", NA, NA, "no", "yes", "yes", NA, "yes", "yes", "no", "yes", "no", "no", "no", "yes"
```

Pick rows: filter()

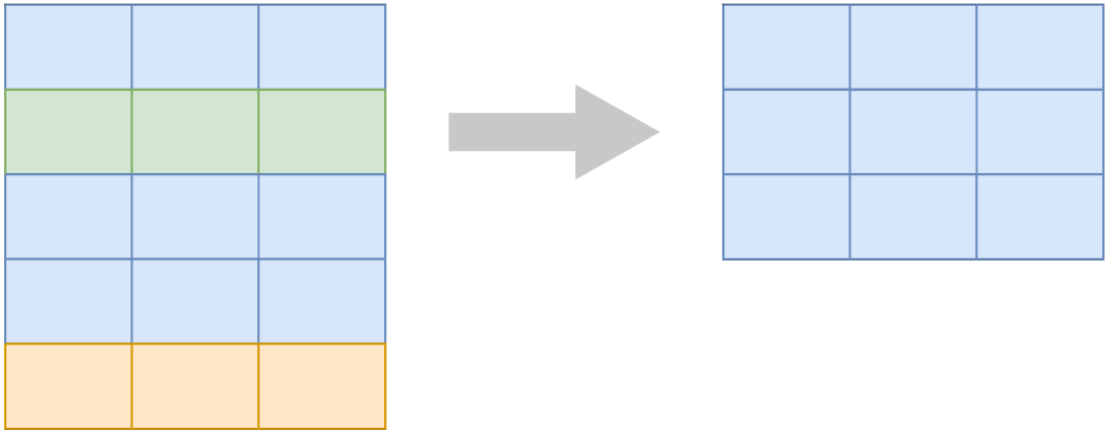
dplyr: filter

Extract rows that meet some sort of condition

```
filter(.data = DATA, ...)
```

- `DATA` = Data frame to transform
- `...` = One or more condition
- `filter()` returns each row for which the test is TRUE

Filter specific rows



```
filter(.data = DATA, ...)
```

- `DATA` = Data frame to transform
- `...` = One or more condition
- `filter()` returns each row for which the condition is TRUE

Filtering only Female data data

```
## use upper case as an example
female <- filter(hosp_data,gender=="f" )
```

- NB: We use == which tests if equal (equality test)

- Create a data of pediatrics

```
## use upper case as an example
paed <- filter(hosp_data,... )
skim(paed)
```

Filter using logical tests

Test	Meaning	Test	Meaning
`x < y`	Less than	`x %in% y`	In (group membership)
`x > y`	Greater than	`is.na(x)`	Is missing
`==`	Equal to	`!is.na(x)`	Is not missing
`x <= y`	Less than or equal to		
`x >= y`	Greater than or equal to		
`x != y`	Not equal to		

Task : Filter using logical

Use filter() and logical tests to create data for female adults

```
female_adult <- filter(hosp_data,gender=="f" , age>17 )
```

Common mistakes

- Using = instead of ==
- Forgeting quote

```
## Wrong
filter(hosp_data,       gender = "f")
filter(hosp_data,       gender = f)


## Correct
filter(hosp_data,       gender == "f")
```

Boolean operators

Operator	Meaning
a & b	and
a | b	or
!a	not

These do the same thing:

```
female_adult <- filter(hosp_data,gender=="f" , age>17 )
```

```
female_adult <- filter(hosp_data,gender=="f" & age>17 )
```

Task : Filtering and boolean

- **Use filter() and Boolean logical tests to show...**
  - Adult females that died

```
## explain with died
female_died <- filter(hosp_data,gender=="f" & outcome=="Death" & age>18 )
```

Filter out rows with missing values

- **Use filter() and Boolean logical tests to show…**
  - Remove patients with missing gender and outcome
  - In R missing is represented by `NA`

```
nonmiss <- drop_na(hosp_data, gender, outcome)
```

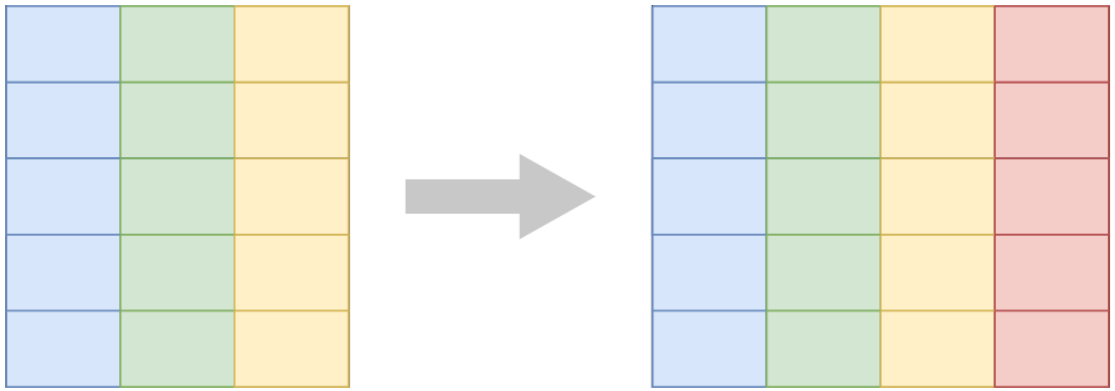
Create new columns: mutate()

dplyr: mutate

```
mutate(.data = DATA, ...)
```

- `DATA` = Data frame to transform
- `...` = Columns to make

Mutate a new column



```
mutate(.data = DATA, ...)
```

- `DATA` = Data frame to transform
- `...` = Columns to make

mutate the age months variable

```
hospt_data_clean <- mutate(hosp_data, age_months = age * 12)
```

mutate the 2 variables

```
## age in years
## height in meters
hospt_data_clean <- mutate(hosp_data, age_months = age * 12,
       height_m=ht_cm/100)
```

Do conditional tests within mutate()

dplyr: ifelse()

```
ifelse(TEST,
       VALUE_IF_TRUE,
       VALUE_IF_FALSE)
```

- TEST = A logical test
- VALUE_IF_TRUE = What happens if test is true
- VALUE_IF_FALSE = What happens if test is false

Create a variable to show Adults and Paed

```
mutate(hospt_data,
       age_group = ifelse(age > 14, "Adults", "Paed"))
```

Piping

- You have realized we have run multiple conditions
- `R pipes` are a way to chain multiple operations together in a concise and expressive way.
- The `%>%` operator (pipe) takes an object on the left
- Then passes it as the first argument of the function on the right

- These two commands do the same thing

```
filter(hosp_data, gender == "f")

hosp_data %>% filter(gender == "f")
```

%>% : Pipe operator

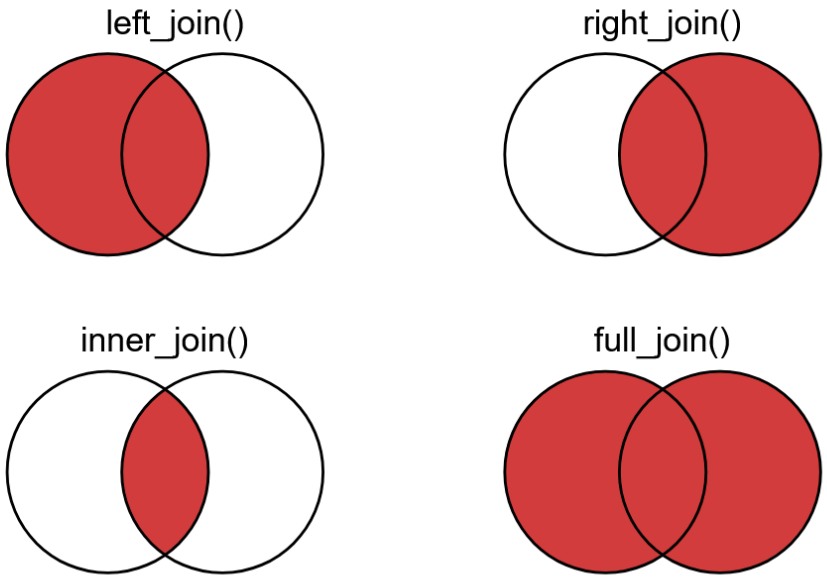
```
leave_house(
  take_breakfast(
    get_dressed(
      wake_up(
        me, ## start here
        time = "8:00"),
      trouser = TRUE, shirt = TRUE , socks=FALSE),
    mayai = TRUE, viazi = TRUE , chai=TRUE),
  nduthi = TRUE, car = FALSE)
```

```
me %>%
  wake_up(time = "8:00") %>%
  get_dressed(trouser = TRUE, shirt = TRUE , socks=FALSE) %>%
  take_breakfast( mayai = TRUE, viazi = TRUE , chai=TRUE, ukwanju=FALSE) %>%
  leave_house( nduthi = TRUE, car = FALSE)
```

- Put together the mutate, filter,select commands we have ran using the pipe operator

left_join() and right_join()

tidyr: join data



- Join the current file with the hospital location data

left_join()

```
left_join(
  x = data1,
  y = data2,
  by = "common id",
  ...
)
```

- Import the data to join first

```
hospital_location <- import('data/line_hospital_sub.csv')
```

```
head(hosp_data)
```

```
##   case_id date_onset  hosp_date date_of_outcome outcome gender age wt_kg ht_cm ct_blood chills cough aches vomit
## 1  5fe599 2014-05-13 2014-05-15           <NA>    <NA>      m   2    27    48       22      no   yes    no   yes
## 2  8689b7 2014-05-13 2014-05-14     2014-05-18 Recover      f   3    25    59       22    <NA> <NA>  <NA>  <NA>
## 3  11f8ea 2014-05-16 2014-05-18     2014-05-30 Recover      m  56    91   238       21    <NA> <NA>  <NA>  <NA>
## 4  b8812a 2014-05-18 2014-05-20           <NA>    <NA>      f  18    41   135       23      no    no    no    no
## 5  893f25 2014-05-21 2014-05-22     2014-05-29 Recover      m   3    36    71       23      no   yes    no   yes
## 6  be99c8 2014-05-22 2014-05-23     2014-05-24 Recover   <NA>  16    56   116       21      no   yes    no   yes
```

```
hosp_left_joined <- left_join(
  x = hosp_data,
  y = hospital_location,
  by = "case_id"
)
head(hosp_left_joined)
```

```
##   case_id date_onset  hosp_date date_of_outcome outcome gender age wt_kg ht_cm ct_blood chills cough aches vomit
## 1  5fe599 2014-05-13 2014-05-15            <NA>    <NA>      m   2    27    48       22      no   yes    no   yes
## 2  8689b7 2014-05-13 2014-05-14      2014-05-18 Recover      f   3    25    59       22    <NA> <NA>  <NA>  <NA>
## 3  11f8ea 2014-05-16 2014-05-18      2014-05-30 Recover      m  56    91   238       21    <NA> <NA>  <NA>  <NA>
## 4  b8812a 2014-05-18 2014-05-20            <NA>    <NA>      f  18    41   135       23      no   no    no    no
## 5  893f25 2014-05-21 2014-05-22      2014-05-29 Recover      m   3    36    71       23      no   yes    no   yes
## 6  be99c8 2014-05-22 2014-05-23      2014-05-24 Recover   <NA>  16    56   116       21      no   yes    no   yes
##                              hospital       lon      lat
## 1                               Other -13.21574 8.468973
## 2                                     -13.21523 8.451719
## 3 St. Mark's Maternity Hospital (SMMH) -13.21291 8.464817
## 4                        Port Hospital -13.23637 8.475476
## 5                    Military Hospital -13.22286 8.460824
## 6                        Port Hospital -13.22263 8.461831
```

```
hosp_left_joined <- left_join(
  x = hosp_data,
  y = hospital_location,
  by = "case_id"
)
head(hosp_left_joined)
```

```
##   case_id date_onset  hosp_date date_of_outcome outcome gender age wt_kg ht_cm ct_blood chills cough aches vomit
## 1  5fe599 2014-05-13 2014-05-15            <NA>    <NA>      m   2    27    48       22      no   yes    no   yes
## 2  8689b7 2014-05-13 2014-05-14      2014-05-18 Recover      f   3    25    59       22    <NA> <NA>  <NA>  <NA>
## 3  11f8ea 2014-05-16 2014-05-18      2014-05-30 Recover      m  56    91   238       21    <NA> <NA>  <NA>  <NA>
## 4  b8812a 2014-05-18 2014-05-20            <NA>    <NA>      f  18    41   135       23      no   no    no    no
## 5  893f25 2014-05-21 2014-05-22      2014-05-29 Recover      m   3    36    71       23      no   yes    no   yes
## 6  be99c8 2014-05-22 2014-05-23      2014-05-24 Recover   <NA>  16    56   116       21      no   yes    no   yes
##                              hospital       lon      lat
## 1                               Other -13.21574 8.468973
## 2                                     -13.21523 8.451719
## 3 St. Mark's Maternity Hospital (SMMH) -13.21291 8.464817
## 4                        Port Hospital -13.23637 8.475476
## 5                    Military Hospital -13.22286 8.460824
## 6                        Port Hospital -13.22263 8.461831
```

right_join()

```
right_join(
  x = data1,
  y = data2,
  by = "common id",
  ...
)
```

```
hosp_right_joined <- right_join(
  x = hosp_data,
  y = hospital_location,
  by = "case_id"
)
head(hosp_right_joined)
```

```
##   case_id date_onset  hosp_date date_of_outcome outcome gender age wt_kg ht_cm ct_blood chills cough aches vomit
## 1  5fe599 2014-05-13 2014-05-15            <NA>    <NA>      m   2    27    48       22      no   yes    no   yes
## 2  8689b7 2014-05-13 2014-05-14      2014-05-18 Recover      f   3    25    59       22    <NA> <NA>  <NA>  <NA>
## 3  11f8ea 2014-05-16 2014-05-18      2014-05-30 Recover      m  56    91   238       21    <NA> <NA>  <NA>  <NA>
## 4  b8812a 2014-05-18 2014-05-20            <NA>    <NA>      f  18    41   135       23      no   no    no    no
## 5  893f25 2014-05-21 2014-05-22      2014-05-29 Recover      m   3    36    71       23      no   yes    no   yes
## 6  be99c8 2014-05-22 2014-05-23      2014-05-24 Recover   <NA>  16    56   116       21      no   yes    no   yes
##                              hospital       lon      lat
## 1                               Other -13.21574 8.468973
## 2                                     -13.21523 8.451719
## 3 St. Mark's Maternity Hospital (SMMH) -13.21291 8.464817
## 4                        Port Hospital -13.23637 8.475476
## 5                    Military Hospital -13.22286 8.460824
## 6                        Port Hospital -13.22263 8.461831
```

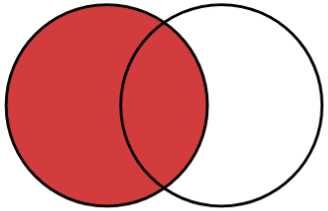
inner_join()

```
inner_join(
  x = data1,
  y = data2,
  by = "common id",
  ...
)
```

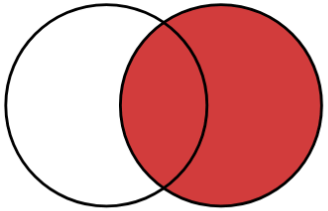
```
hosp_inner_joined <- inner_join(
  x = hosp_data,
  y = hospital_location,
  by = "case_id"
)
head(hosp_inner_joined)
```

```
##   case_id date_onset  hosp_date date_of_outcome outcome gender age wt_kg ht_cm ct_blood chills cough aches vomit
## 1  5fe599 2014-05-13 2014-05-15            <NA>    <NA>      m   2    27    48       22      no   yes    no   yes
## 2  8689b7 2014-05-13 2014-05-14      2014-05-18 Recover      f   3    25    59       22    <NA> <NA>  <NA>  <NA>
## 3  11f8ea 2014-05-16 2014-05-18      2014-05-30 Recover      m  56    91   238       21    <NA> <NA>  <NA>  <NA>
## 4  b8812a 2014-05-18 2014-05-20            <NA>    <NA>      f  18    41   135       23      no   no    no    no
## 5  893f25 2014-05-21 2014-05-22      2014-05-29 Recover      m   3    36    71       23      no   yes    no   yes
## 6  be99c8 2014-05-22 2014-05-23      2014-05-24 Recover   <NA>  16    56   116       21      no   yes    no   yes
##                              hospital       lon      lat
## 1                               Other -13.21574 8.468973
## 2                                     -13.21523 8.451719
## 3 St. Mark's Maternity Hospital (SMMH) -13.21291 8.464817
## 4                        Port Hospital -13.23637 8.475476
## 5                    Military Hospital -13.22286 8.460824
## 6                        Port Hospital -13.22263 8.461831
```

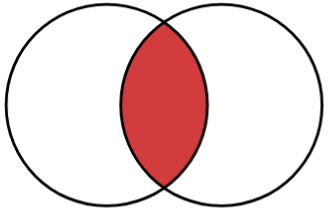
left_join()



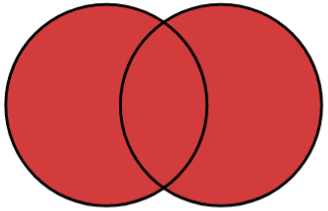
right_join()



inner_join()



full_join()



```
str(hosp_inner_joined)
```

```
## 'data.frame':    25 obs. of  17 variables:
##  $ case_id        : chr  "5fe599" "8689b7" "11f8ea" "b8812a" ...
##  $ date_onset     : POSIXct, format: "2014-05-13" "2014-05-13" "2014-05-16" "2014-05-18" ...
##  $ hosp_date      : POSIXct, format: "2014-05-15" "2014-05-14" "2014-05-18" "2014-05-20" ...
##  $ date_of_outcome: POSIXct, format: NA "2014-05-18" "2014-05-30" NA ...
##  $ outcome        : chr  NA "Recover" "Recover" NA ...
##  $ gender         : chr  "m" "f" "m" "f" ...
##  $ age            : num  2 3 56 18 3 16 16 0 61 27 ...
##  $ wt_kg          : num  27 25 91 41 36 56 47 0 NA 69 ...
##  $ ht_cm          : num  48 59 238 135 71 116 87 11 NA 174 ...
##  $ ct_blood       : num  22 22 21 23 23 21 21 22 22 22 ...
##  $ chills         : chr  "no" NA NA "no" ...
##  $ cough          : chr  "yes" NA NA "no" ...
##  $ aches          : chr  "no" NA NA "no" ...
##  $ vomit          : chr  "yes" NA NA "no" ...
##  $ hospital       : chr  "Other" "" "St. Mark's Maternity Hospital (SMMH)" "Port Hospital" ...
##  $ lon            : num  -13.2 -13.2 -13.2 -13.2 -13.2 ...
##  $ lat            : num  8.47 8.45 8.46 8.48 8.46 ...
```

```
str(hosp_left_joined)
```

```
## 'data.frame':    198 obs. of  17 variables:
##  $ case_id        : chr  "5fe599" "8689b7" "11f8ea" "b8812a" ...
##  $ date_onset     : POSIXct, format: "2014-05-13" "2014-05-13" "2014-05-16" "2014-05-18" ...
##  $ hosp_date      : POSIXct, format: "2014-05-15" "2014-05-14" "2014-05-18" "2014-05-20" ...
##  $ date_of_outcome: POSIXct, format: NA "2014-05-18" "2014-05-30" NA ...
##  $ outcome        : chr  NA "Recover" "Recover" NA ...
##  $ gender         : chr  "m" "f" "m" "f" ...
##  $ age            : num  2 3 56 18 3 16 16 0 61 27 ...
##  $ wt_kg          : num  27 25 91 41 36 56 47 0 NA 69 ...
##  $ ht_cm          : num  48 59 238 135 71 116 87 11 NA 174 ...
```

Exporting data from R

- There are many 1M+1 …but
- rio::export function

```
export(data_to_export, file_name_location_extension,)
```

Exporting the clean version data

```
export(hosp_data, "data/hosp_data_clean.csv")
```