Icahn
School of
Medicine at
Mount
Sinai

Data transformation

Part I

Joselyn C. Chávez Fuentes

02/04/2025

Let's recap

- What are the properties of data frames?
- What are the properties of lists?
- How do we access the data of a data frame?
- How do we access the data of a list?

What is Tidyverse?

https://www.tidyverse.org

- A collection of R packages designed for data science.
- The packages share a philosophy of design, grammar and data structure.

Which packages are part of the Tidyverse?

Tidyverse base

- **ggplot2**: creation of graphics
- **dplyr**: data handling
- **tidyr**: tidy data array
- **readr**: reading and writing of tabular data
- purrr: functional programming
- **tibble**: re-design of data frames
- **stringr**: text handling (strings)

Reading data

- **readxl**: reading . xls and . xlsx
- **googlesheets4**: reading Google Sheets
- **DBI**: relational databases
- haven: SPSS, Stata, and SAS data
- httr: web APIs.
- googledrive: reading Google Drive files
- rvest: web scraping.
- jsonlite: JSON
- xml2: XML reading

Handling data

- **lubridate**: handling dates
- hms: time zones.
- blob: binary data storage

Modeling

- **tidymodels**: modeling and machine learning

Programming

- **magrittr**: provides pipe %>% and other specialized operators (%$%, %<>%)
- glue: alternative to paste() to combine data and text

Tidyverse packages for data analysis

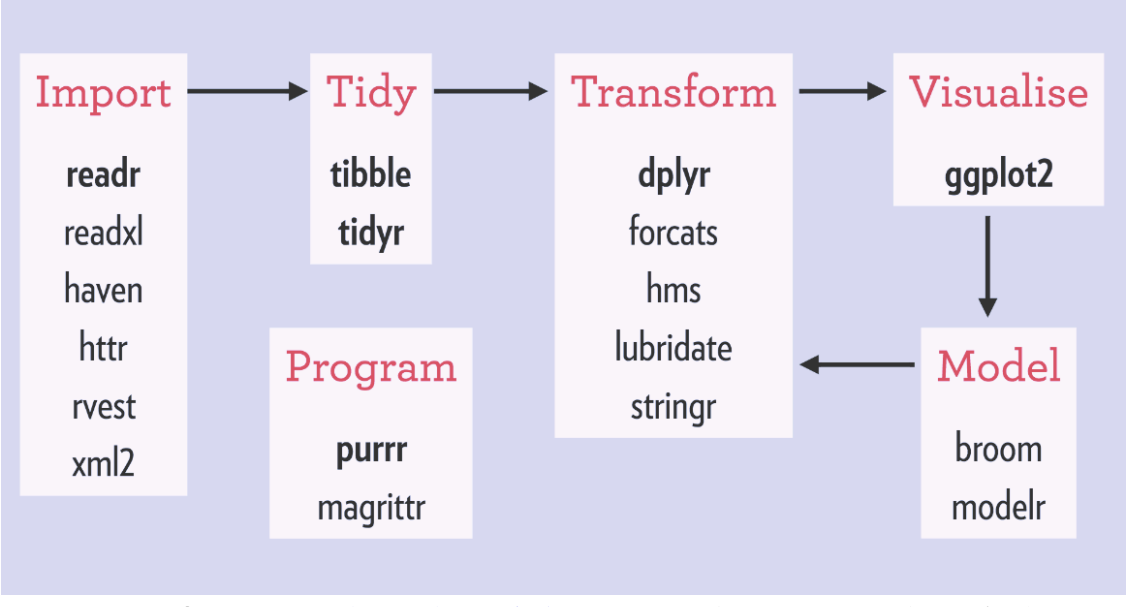


Image from [Why the tidyverse](#) by Joseph Rickert

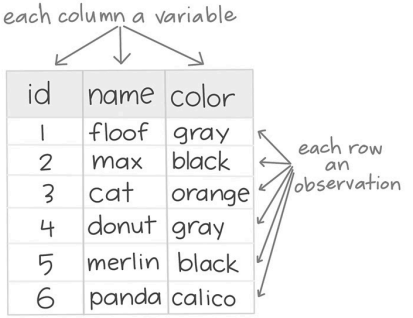
tibbles

"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

Artwork by @allison_horst

tibbles

```
library(dplyr)

df <- data.frame(x = letters,
                 y = 1:26,
                 z = sample(c(TRUE, FALSE), 26,
                            replace = TRUE))
dt <- tibble(x = letters,
                 y = 1:26,
                 z = sample(c(TRUE, FALSE), 26,
                            replace = TRUE))
```

Differences between data frames and tibbles

```
head(df)
```

```
##   x y     z
## 1 a 1 FALSE
## 2 b 2  TRUE
## 3 c 3 FALSE
## 4 d 4  TRUE
## 5 e 5 FALSE
## 6 f 6  TRUE
```

```
head(dt)
```

```
## # A tibble: 6 × 3
##   x         y z
##   <chr> <int> <lgl>
## 1 a         1 FALSE
## 2 b         2 TRUE
## 3 c         3 TRUE
## 4 d         4 TRUE
## 5 e         5 FALSE
## 6 f         6 FALSE
```

Reading and writing of tibbles

```
library(readr)

sinai_covid <- read_csv("Sinai_covid.csv")
```

```
head(sinai_covid)

write_csv(sinai_covid, "printed_sinai_covid.csv")
```

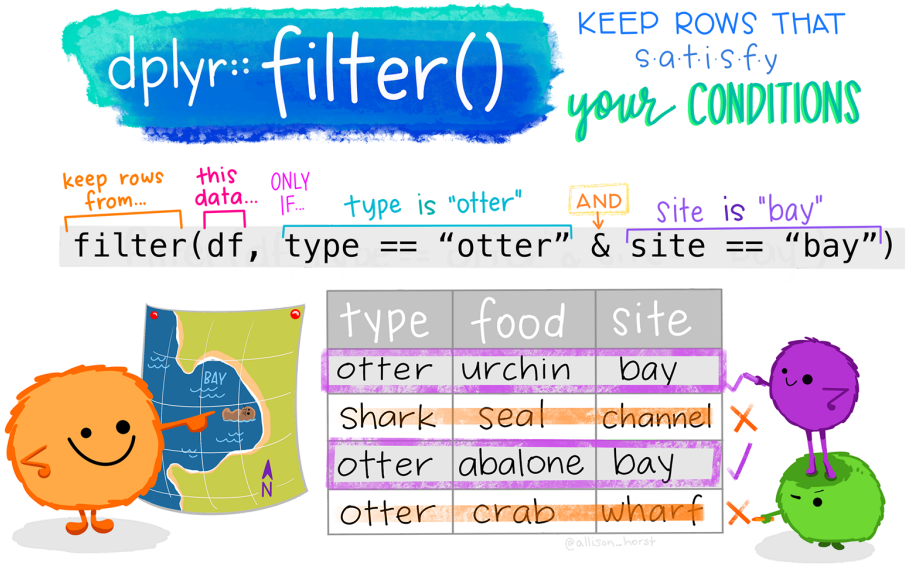
Arrange

```
library(dplyr)

arrange(sinai_covid, by = age)
```

```
## # A tibble: 500 × 18
##       id   age sex    race  ethnicity facility smoking_status asthm
##    <dbl> <dbl> <chr>  <chr> <chr>     <chr>    <chr>           <dbl
##  1  3313     0 MALE   OTHER MEXICAN … THE MOU… NEVER
##  2  1081     9 FEMALE OTHER LATIN AM… THE MOU… NEVER
##  3   940    11 FEMALE ASIA… NON-HISP… THE MOU… NEVER
##  4  4159    15 MALE   OTHER NON-HISP… THE MOU… NEVER
##  5  4995    18 FEMALE OTHER PUERTO R… THE MOU… NEVER
##  6   196    20 MALE   WHITE NON-HISP… THE MOU… NEVER
##  7    21    21 FEMALE WHITE NON-HISP… THE MOU… NEVER
##  8  2774    21 MALE   OTHER NON-HISP… THE MOU… NEVER
##  9  4722    21 FEMALE AFRI… NON-HISP… MOUNT S… NEVER
## 10  3782    22 MALE   OTHER PERUVIAN  THE MOU… NEVER
## # ℹ 490 more rows
## # ℹ 8 more variables: obesity <dbl>, diabetes <dbl>,
## #   chronic_kidney_disease <dbl>, hiv_flag <dbl>, cancer_flag <dbl>
## #   deceased_indicator <dbl>, deceased_days_since_encounter <dbl>
```

Filter


dplyr::filter()
KEEP ROWS THAT satisfy your CONDITIONS
keep rows from...
this data...
ONLY IF...
type is "otter"
AND
site is "bay"
filter(df, type == "otter" & site == "bay")
BAY
N
type food site
otter urchin bay
Shark seal channel
otter abalone bay
otter crab wharf
@allison_horst

Filter

```
filter(sinai_covid, age > 50)
```

```
## # A tibble: 374 × 18
##       id   age sex    race  ethnicity facility smoking_status asthm
##    <dbl> <dbl> <chr>  <chr> <chr>     <chr>    <chr>           <dbl
##  1     6    63 MALE   WHITE NON-HISP… THE MOU… QUIT
##  2    11    64 MALE   AFRI… NON-HISP… THE MOU… NEVER
##  3    46    51 FEMALE OTHER NON-HISP… MOUNT S… NEVER
##  4    47    72 MALE   WHITE NON-HISP… THE MOU… NEVER
##  5    63    67 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  6    67    83 FEMALE WHITE DOMINICAN MOUNT S… NEVER
##  7    86    73 FEMALE AFRI… UNKNOWN   MOUNT S… QUIT
##  8    90    59 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  9    92    65 FEMALE WHITE NON-HISP… THE MOU… NEVER
## 10   103    60 MALE   WHITE NON-HISP… MOUNT S… NEVER
## # ℹ 364 more rows
## # ℹ 8 more variables: obesity <dbl>, diabetes <dbl>,
## #   chronic_kidney_disease <dbl>, hiv_flag <dbl>, cancer_flag <dbl>
## #   deceased_indicator <dbl>, deceased_days_since_encounter <dbl>
```

Filter

```
filter(sinai_covid, age > 50 & sex == "FEMALE")
```

```
## # A tibble: 168 × 18
##       id   age sex    race  ethnicity facility smoking_status asthm
##    <dbl> <dbl> <chr>  <chr> <chr>     <chr>    <chr>           <dbl
##  1    46    51 FEMALE OTHER NON-HISP… MOUNT S… NEVER
##  2    63    67 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  3    67    83 FEMALE WHITE DOMINICAN MOUNT S… NEVER
##  4    86    73 FEMALE AFRI… UNKNOWN   MOUNT S… QUIT
##  5    90    59 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  6    92    65 FEMALE WHITE NON-HISP… THE MOU… NEVER
##  7   138    83 FEMALE OTHER DOMINICAN MOUNT S… NEVER
##  8   186    68 FEMALE CHIN… UNKNOWN   THE MOU… NEVER
##  9   259    72 FEMALE AFRI… NON-HISP… MOUNT S… NEVER
## 10   260    63 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
## # ℹ 158 more rows
## # ℹ 8 more variables: obesity <dbl>, diabetes <dbl>,
## #   chronic_kidney_disease <dbl>, hiv_flag <dbl>, cancer_flag <dbl>
## #   deceased_indicator <dbl>, deceased_days_since_encounter <dbl>
```

Filter

```
filter(sinai_covid, age > 50, sex == "FEMALE")
```

```
## # A tibble: 168 × 18
##       id   age sex    race  ethnicity facility smoking_status asthm
##    <dbl> <dbl> <chr>  <chr> <chr>     <chr>    <chr>           <dbl
##  1    46    51 FEMALE OTHER NON-HISP… MOUNT S… NEVER
##  2    63    67 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  3    67    83 FEMALE WHITE DOMINICAN MOUNT S… NEVER
##  4    86    73 FEMALE AFRI… UNKNOWN   MOUNT S… QUIT
##  5    90    59 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  6    92    65 FEMALE WHITE NON-HISP… THE MOU… NEVER
##  7   138    83 FEMALE OTHER DOMINICAN MOUNT S… NEVER
##  8   186    68 FEMALE CHIN… UNKNOWN   THE MOU… NEVER
##  9   259    72 FEMALE AFRI… NON-HISP… MOUNT S… NEVER
## 10   260    63 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
## # ℹ 158 more rows
## # ℹ 8 more variables: obesity <dbl>, diabetes <dbl>,
## #   chronic_kidney_disease <dbl>, hiv_flag <dbl>, cancer_flag <dbl>
## #   deceased_indicator <dbl>, deceased_days_since_encounter <dbl>
```

Filter

```
filter(sinai_covid, age > 50 | sex == "FEMALE")
```

```
## # A tibble: 428 × 18
##       id   age sex    race  ethnicity facility smoking_status asthm
##    <dbl> <dbl> <chr>  <chr> <chr>     <chr>    <chr>           <dbl
##  1     6    63 MALE   WHITE NON-HISP… THE MOU… QUIT
##  2    11    64 MALE   AFRI… NON-HISP… THE MOU… NEVER
##  3    21    21 FEMALE WHITE NON-HISP… THE MOU… NEVER
##  4    46    51 FEMALE OTHER NON-HISP… MOUNT S… NEVER
##  5    47    72 MALE   WHITE NON-HISP… THE MOU… NEVER
##  6    63    67 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
##  7    67    83 FEMALE WHITE DOMINICAN MOUNT S… NEVER
##  8    86    73 FEMALE AFRI… UNKNOWN   MOUNT S… QUIT
##  9    90    59 FEMALE AFRI… NON-HISP… MOUNT S… QUIT
## 10    92    65 FEMALE WHITE NON-HISP… THE MOU… NEVER
## # ℹ 418 more rows
## # ℹ 8 more variables: obesity <dbl>, diabetes <dbl>,
## #   chronic_kidney_disease <dbl>, hiv_flag <dbl>, cancer_flag <dbl>
## #   deceased_indicator <dbl>, deceased_days_since_encounter <dbl>
```

Select

```
select(sinai_covid, id, age, facility)
```

```
## # A tibble: 500 × 3
##       id   age facility
##    <dbl> <dbl> <chr>
##  1     6    63 THE MOUNT SINAI HOSPITAL
##  2    11    64 THE MOUNT SINAI HOSPITAL
##  3    21    21 THE MOUNT SINAI HOSPITAL
##  4    46    51 MOUNT SINAI QUEENS HOSPITAL
##  5    47    72 THE MOUNT SINAI HOSPITAL
##  6    63    67 MOUNT SINAI QUEENS HOSPITAL
##  7    67    83 MOUNT SINAI WEST
##  8    86    73 MOUNT SINAI BI BROOKLYN
##  9    90    59 MOUNT SINAI QUEENS HOSPITAL
## 10    92    65 THE MOUNT SINAI HOSPITAL
## # ℹ 490 more rows
```

Combine functions

```
x <- filter(sinai_covid, age > 50)

x <- select(x, id, age, facility)

arrange(x, age)
```

```
## # A tibble: 374 × 3
##       id   age facility
##    <dbl> <dbl> <chr>
##  1    46    51 MOUNT SINAI QUEENS HOSPITAL
##  2  2323    51 THE MOUNT SINAI HOSPITAL
##  3  3894    51 THE MOUNT SINAI HOSPITAL
##  4  4230    51 MOUNT SINAI BI BROOKLYN
##  5   439    52 THE MOUNT SINAI HOSPITAL
##  6   450    52 MOUNT SINAI BI BROOKLYN
##  7  3812    52 THE MOUNT SINAI HOSPITAL
##  8  3900    52 MOUNT SINAI WEST
##  9  4104    52 MOUNT SINAI BI BROOKLYN
## 10  4864    52 MOUNT SINAI QUEENS HOSPITAL
## # ℹ 364 more rows
```

Your turn!

Use sinai_covid tibble

- Extract all patients from THE MOUNT SINAI HOSPITAL and MOUNT SINAI BI BROOKLYN
- Select id, sex, ethnicity and bmi
- Order the result by bmi

Mutate

Mutate

```
x <- mutate(sinai_covid, bmi_log = log10(bmi))

select(x, id, bmi, bmi_log)
```

```
## # A tibble: 500 × 3
##       id   bmi bmi_log
##    <dbl> <dbl>   <dbl>
##  1     6  28.7    1.46
##  2    11  36.0    1.56
##  3    21  25.0    1.40
##  4    46  24.6    1.39
##  5    47  25.8    1.41
##  6    63  22.0    1.34
##  7    67  27.6    1.44
##  8    86  24.6    1.39
##  9    90  30.2    1.48
## 10    92  19.4    1.29
## # ℹ 490 more rows
```

Mutate

```
x <- mutate(sinai_covid,
            new_facility = paste("Facility:",facility))

select(x, id, new_facility)
```

```
## # A tibble: 500 × 2
##       id new_facility
##    <dbl> <chr>
##  1     6 Facility: THE MOUNT SINAI HOSPITAL
##  2    11 Facility: THE MOUNT SINAI HOSPITAL
##  3    21 Facility: THE MOUNT SINAI HOSPITAL
##  4    46 Facility: MOUNT SINAI QUEENS HOSPITAL
##  5    47 Facility: THE MOUNT SINAI HOSPITAL
##  6    63 Facility: MOUNT SINAI QUEENS HOSPITAL
##  7    67 Facility: MOUNT SINAI WEST
##  8    86 Facility: MOUNT SINAI BI BROOKLYN
##  9    90 Facility: MOUNT SINAI QUEENS HOSPITAL
## 10    92 Facility: THE MOUNT SINAI HOSPITAL
## # ℹ 490 more rows
```

Mutate

```
x <- mutate(sinai_covid,
            facility = tolower(facility))

select(x, id, facility)
```

```
## # A tibble: 500 × 2
##       id facility
##    <dbl> <chr>
##  1     6 the mount sinai hospital
##  2    11 the mount sinai hospital
##  3    21 the mount sinai hospital
##  4    46 mount sinai queens hospital
##  5    47 the mount sinai hospital
##  6    63 mount sinai queens hospital
##  7    67 mount sinai west
##  8    86 mount sinai bi brooklyn
##  9    90 mount sinai queens hospital
## 10    92 the mount sinai hospital
## # ℹ 490 more rows
```

Your turn!

Use mutate to:

- Add a new column "random_num" adding random numbers (tip: use rnorm)
- Change sex column from uppercase to lowercase
- Add a new column with the result of multiplying bmi by 10

Count

```
count(sinai_covid, facility)
```

```
## # A tibble: 5 × 2
##   facility                        n
##   <chr>                       <int>
## 1 MOUNT SINAI BI BROOKLYN        84
## 2 MOUNT SINAI QUEENS HOSPITAL    89
## 3 MOUNT SINAI ST. LUKE'S         86
## 4 MOUNT SINAI WEST               53
## 5 THE MOUNT SINAI HOSPITAL      188
```

Summarise

```
summarise(sinai_covid, mean_age = mean(age))
```

```
## # A tibble: 1 × 1
##   mean_age
##      <dbl>
## 1     61.1
```

Group

```
x <- group_by(sinai_covid, facility)

summarise(x, mean_age = mean(age))
```

```
## # A tibble: 5 × 2
##   facility                      mean_age
##   <chr>                            <dbl>
## 1 MOUNT SINAI BI BROOKLYN           64.2
## 2 MOUNT SINAI QUEENS HOSPITAL       64.5
## 3 MOUNT SINAI ST. LUKE'S            65.5
## 4 MOUNT SINAI WEST                  67.6
## 5 THE MOUNT SINAI HOSPITAL          54.3
```

Combining functions using the pipe

```
group_by(sinai_covid, facility) %>%
  summarise(mean_age = mean(age))
```

```
## # A tibble: 5 × 2
##   facility                      mean_age
##   <chr>                            <dbl>
## 1 MOUNT SINAI BI BROOKLYN           64.2
## 2 MOUNT SINAI QUEENS HOSPITAL       64.5
## 3 MOUNT SINAI ST. LUKE'S            65.5
## 4 MOUNT SINAI WEST                  67.6
## 5 THE MOUNT SINAI HOSPITAL          54.3
```

Combining functions using the pipe

```
sinai_covid %>%
  group_by(facility) %>%
  summarise(mean_age = mean(age))
```

```
## # A tibble: 5 × 2
##   facility                     mean_age
##   <chr>                           <dbl>
## 1 MOUNT SINAI BI BROOKLYN          64.2
## 2 MOUNT SINAI QUEENS HOSPITAL      64.5
## 3 MOUNT SINAI ST. LUKE'S           65.5
## 4 MOUNT SINAI WEST                 67.6
## 5 THE MOUNT SINAI HOSPITAL         54.3
```

Combining functions using the pipe

```
sinai_covid %>%
  filter(age > 50) %>%
  count(facility, name = "patients_older_than_50")
```

```
## # A tibble: 5 × 2
##   facility                    patients_older_than_50
##   <chr>                                        <int>
## 1 MOUNT SINAI BI BROOKLYN                         70
## 2 MOUNT SINAI QUEENS HOSPITAL                     71
## 3 MOUNT SINAI ST. LUKE'S                          72
## 4 MOUNT SINAI WEST                                46
## 5 THE MOUNT SINAI HOSPITAL                       115
```

Your turn!

Exercise 1

Use the pipe to:

- Extract all patients from THE MOUNT SINAI HOSPITAL and MOUNT SINAI BI BROOKLYN
- Select id, sex, ethnicity and bmi
- Order by bmi

Your turn!

Exercise 2

Use the pipe to:

- Extract all patients from THE MOUNT SINAI HOSPITAL and MOUNT SINAI BI BROOKLYN
- Select id, sex, ethnicity and bmi
- Count ethnicity by facility

Your turn!

Exercise 3

Use the pipe to:

- Extract all patients from THE MOUNT SINAI HOSPITAL and MOUNT SINAI BI BROOKLYN
- Select id, age, ethnicity and bmi
- Calculate mean of age by facility

Thanks!