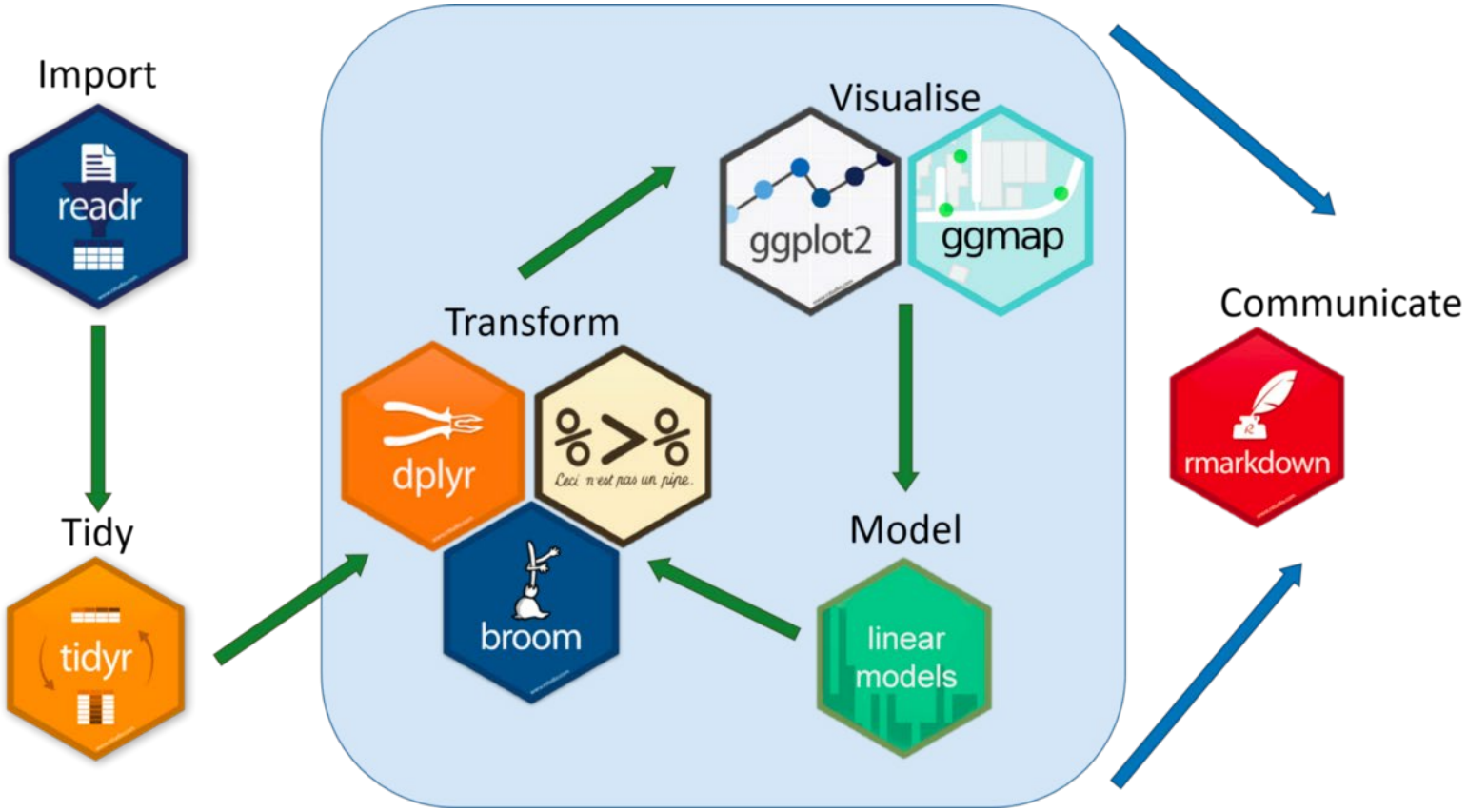
Tidyverse and DataFrames

School of Information Studies
Syracuse University

The Tidyverse (Partial View)



Tibbles vs. DataFrames

A “tibble” updates the behavior of a dataframe to make the structure more useful.

Tibbles are part of the “tidyverse.”

Try ‘glimpse’ (vs ‘str’)

Tibbles:

- Print nicely on the console
- Provide type information
- Tibbles avoid unnecessary type conversions and support improved column naming

```
#> # A tibble: 1,000 x 5
#>   a                   b              c     d e
#>   <dttm>              <date>     <int> <dbl> <chr
#> 1 2019-01-08 17:33:11 2019-01-15     1 0.368 h
#> 2 2019-01-09 11:38:20 2019-01-20     2 0.612 n
#> 3 2019-01-09 06:02:00 2019-01-30     3 0.415 l
#> 4 2019-01-08 19:23:17 2019-01-29     4 0.212 x
#> 5 2019-01-08 15:47:33 2019-01-26     5 0.733 a
#> 6 2019-01-09 02:48:30 2019-01-22     6 0.460 v
#> # … with 994 more rows
```

Exploring a DataFrame

How to explore a dataframe

#Base R

str(mtcars)

'data.frame': 32 obs. of 11 variables:
$ mpg : num 21 21 22.8…
$ cyl : num 6 6 4 6 8 6…

#The tidyverse way

library(tidyverse)

glimpse(mtcars)

Rows: 32
Columns: 11
$ mpg <dbl> 21.0, 21.0, 22.8, 21.4…
$ cyl <dbl> 6, 6, 4, 6, 8, 6, 8, 4, 4…

Selecting a Column

How to select one column

```
#Base R
mtcars[ , "mpg"]
mtcars[ , 1]
mtcars$mpg
```

[1] 21.0 21.0 22.8 21.4 18.7…

```
#The tidyverse way
select(mtcars, "mpg")
select(mtcars, mpg)
```

	mpg
Mazda RX4	21.0
Mazda RX4 Wag	21.0
Datsun 710	22.8

Selecting Columns

How to select more than one column

#Base R

```
mtcars[ , c("mpg", "hp")]
mtcars[ ,c( 1,4)]
```

	mpg	hp
Mazda RX4	21.0	110
Mazda RX4 Wag	21.0	110
Datsun 710	22.8	93

#The tidyverse way

```
select(mtcars, c("mpg", "hp"))
select(mtcars, mpg, hp)
```

	mpg	hp
Mazda RX4	21.0	110
Mazda RX4 Wag	21.0	110
Datsun 710	22.8	93

Selecting Rows

How to select one or more rows

#Base R

mtcars[1:3,]

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160	110
Mazda RX4 Wag	21.0	6	160	110

#The tidyverse way

slice(mtcars,1:3)

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160	110
Mazda RX4 Wag	21.0	6	160	110

Tidyverse Pipes

Passes data from the command before to the one after - Data flows left to right

Makes the code read more like a sentence

"%>%" is the "symbol" for a pipe

Example: min(x) is the same as:

x %>% min #(x data get passed to "min" function)

#Simple example

```
x <- c(4,1, 10,5)
x %>% min
```

[1] 1

#Dataframes example

```
mtcars %>% slice(1:3)
```

	mpg	cyl	disp	hp
Mazda RX4	21.0	6	160	110
Mazda RX4 Wag	21.0	6	160	110

Filtering DataFrames (i.e., Subset Rows Based on a Condition)

How to create a subset of rows based on a condition for a column

#Base R

```
mtcars[mtcars$mpg > 28,  ]
```

	mpg	cyl	disp	hp	drat
Fiat 128	32.4	4	78.7	66	4.08
Honda Civic	30.4	4	75.7	52	4.93

#The tidyverse way

```
mtcars %>% filter(mpg > 28)
```

	mpg	cyl	disp	hp	drat
Fiat 128	32.4	4	78.7	66	4.08
Honda Civic	30.4	4	75.7	52	4.93

Adding a Column

How to create a subset of rows based on a condition for a column

```
#Base R
myCars <- mtcars
carName=rownames(myCars)
myCars$carName <- carName
myCars [1,
     c("mpg", "cyl", "carName", "hp")]
```

```
                  mpg  cyl   carName      hp
Mazda RX4         21     6   Mazda RX4  110
```

```
#The tidyverse way
myCars <- mtcars
myCars <- rownames_to_column(myCars, var =
"carName")
myCars %>%
   slice(1) %>%
   select(mpg, cyl, carName, hp)
```

```
     mpg   cyl   carName      hp
1    21     6   Mazda RX4   110
```

Grouping Information

Do a summary-type report

```
#Base R
mean(myCars[myCars$cyl==4,"mpg"])
[1] 26.66364

mean(myCars[myCars$cyl==6,"mpg"])
[1] 19.74286

mean(myCars[myCars$cyl==8,"mpg"])
[1] 15.1
```

```
#The tidyverse way
myCars %>%
    group_by(cyl) %>%
    summarize(mpg=mean(mpg),
              hp=mean(hp), .groups = 'drop')
# A tibble: 3 x 3
    cyl     mpg     hp
  <dbl>   <dbl>  <dbl>
1     4    26.7   82.6
2     6    19.7  122
3     8    15.1  209
```

Sorting a DataFrame

```
#Base R
indexes <- order(mtcars[,"mpg"])
sortedDF <- mtcars[indexes,]
str(sortedDF)
```

"data.frame": 32 obs. of 11 variables:
$ mpg : num 10.4 10.4 13.3…
$ cyl : num 8 8 8 8 8 8…

```
#The tidyverse way
sortedDF <- mtcars %>% arrange(mpg)
glimpse(sortedDF)
```

Rows: 32
Columns: 11
$ mpg <dbl> 10.4, 10.4, 13.3,…
$ cyl <dbl> 8, 8, 8, 8, 8, 8, 8,…

Sample Questions

#Given the following lines of code were executed

```
names <- c("Jeff", "Pat", "Joe")
height <- c(100,103,120)
myFamily <- data.frame(names, height)
```

#What is returned from the following commands?

```
myFamily[1,1]
myFamily[1,]
select(myFamily, height)
myFamily %>% slice( c(1,3) )
```

Tidyverse and DataFrames (cont.)

School of Information Studies
Syracuse University

Sample Answers

myFamily[1,1]

[1] "Jeff"

myFamily[1,]

	names	height
1	Jeff	100

select(myFamily, height)

	height
1	100
2	103
3	120

myFamily %>% slice(c(1,3))

	names	height
1	Jeff	100
2	Joe	120