Introduction to

Programming with R

Tidying Data

Packages

CRAN

tidyverse

dplyr

ggplot2

stringr

tidyr

...

```
install.packages
library
```

dplyr

select

filter

arrange

distinct

group_by

summarize

Tibbles

select

```
contains
ends_with
starts_with
…
```

filter

either will equivalently pipe value on LHS as the 1st argument into a function on the RHS

|>

%>%

```
select(storms, …)
is equivalent to
storms |> select(...)
```

```
storms |>
  select(…) |>
  filter(…)
```

from data, select(columns), filter(rows)

arrange

used for sorting by column
non-decreasing is default
rows having same values, arrange can sort ties by 2nd argument

```
storms |>
  select(…) |>
  filter(…) |>
  arrange(…)
```

from data, select(columns), filter(rows), sorted by(column)

arrange

used for sorting by column
non-decreasing is default
rows having same values, arrange can sort ties by 2nd argument

```
storms |>
  select(…) |>
  filter(…) |>
  arrange(desc(…))
```

from data, select(columns), filter(rows), sorted by(column)

distinct

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct()
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct()
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct()
```

keeps only 1st row of duplicate rows
duplicates means having all columns matching by default if named columns aren't specified

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1985	60
Ana	1985	55

```
storms |> distinct()
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1985	60
Ana	1985	55

```
storms |> distinct()
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name)
```

name	year	wind
Ana	1979	50

```
storms |> distinct(name)
```

name	year	wind
Ana	1979	50

```
storms |> distinct(name)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name, year)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name, year)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name, year)
```

name	year	wind
Ana	1979	50
Ana	1979	40
Ana	1979	40
Ana	1985	60
Ana	1985	55
Ana	1985	55

```
storms |> distinct(name, year)
```

name	year	wind
Ana	1979	50
Ana	1985	60

```
storms |> distinct(name, year)
```

name	year	wind
Ana	1979	50
Ana	1985	60

```
storms |> distinct(name, year)
```

year	name	wind
1975	Gladys	120
1976	Belle	105
1977	Anita	150
1978	Ella	120
1979	David	150
...	...	...

`hurricanes`

Groups

group_by

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

```
hurricanes |> group_by(year)
```

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

```
hurricanes |> group_by(year)
```

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

```
hurricanes |> group_by(year)
```

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

```
hurricanes |> group_by(year) |> arrange(desc(wind))
```

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

```
hurricanes |> group_by(year) |> arrange(desc(wind))
```

year	name	wind
1975	Gladys	120
1975	Eloise	110
1975	Caroline	100
1976	Belle	105
1976	Frances	100
1976	Emmy	90

```
hurricanes |> group_by(year) |> arrange(desc(wind))
```

year	name	wind
1975	Gladys	120
1975	Eloise	110
1975	Caroline	100
1976	Belle	105
1976	Frances	100
1976	Emmy	90

```
hurricanes |> … |> … |> slice_head()
```

year	name	wind
1975	Gladys	120
1975	Eloise	110
1975	Caroline	100
1976	Belle	105
1976	Frances	100
1976	Emmy	90

```
hurricanes |> … |> … |> slice_head()
```

year	name	wind
1975	Gladys	120
1976	Belle	105

```
hurricanes |> … |> … |> slice_head()
```

year	name	wind
1975	Gladys	120
1976	Belle	105

```
hurricanes |> … |> … |> slice_head()
```

slice_head

slice_tail

slice_max

slice_min

…

summarize

Tidy Data

1. Each observation is a row; each row is an observation.

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

1. Each observation is a row; each row is an observation.
2. Each variable is a column; each column is a variable.

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

1. Each observation is a row; each row is an observation.
2. Each variable is a column; each column is a variable.
3. Each value is a cell; each cell is a single value.

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

year	name	wind
1975	Eloise	110
1976	Belle	105
1975	Gladys	120
1975	Caroline	100
1976	Emmy	90
1976	Frances	100

`hurricanes`

Normalizing

student	attribute	value
Mario	major	Statistics
Mario	GPA	3.5
Peach	major	Computer Science
Peach	GPA	4.0
Bowser	major	Data Science
Bowser	GPA	3.7

student	attribute	value
Mario	major	Statistics
Mario	GPA	3.5
Peach	major	Computer Science
Peach	GPA	4.0
Bowser	major	Data Science
Bowser	GPA	3.7

student	concentration	GPA
Mario	Statistics	3.5
Peach	Computer Science	4.0
Bowser	Data Science	3.7

student	concentration	GPA
Mario	Statistics	3.5
Peach	Computer Science	4.0
Bowser	Data Science	3.7

tidyr

pivot_wider

student	attribute	value
Mario	major	Statistics
Mario	GPA	3.5
Peach	major	Computer Science
Peach	GPA	4.0
Bowser	major	Data Science
Bowser	GPA	3.7

student	major	GPA
Mario	Statistics	3.5
Peach	Computer Science	4.0
Bowser	Data Science	3.7

pivot_longer

stringr

str_trim

str_squish

```
str_to_lower
str_to_upper
str_to_title
…
```

`str_detect`

…

Introduction to

Programming with R

Tidying Data