

HUDK 4050: CORE METHODS IN EDM

In the news



GEORGE W. BUSH
INSTITUTE

STATE OF OUR CITIES

Profiles of Education Performance Around the Nation

<http://www.bushcenter.org/stateofourcities/explore/>

Data privacy worries shield thousands of Colorado test scores from public scrutiny



Chalkbeat

Graduation rates at core of potential accreditation overhaul



Education **DIVE**

California AG Harris Issues Educational Data-Sharing Guidelines for Foster Youth

Census Bureau Revamps Statistics Program for K–12 Students and Teachers

Online learning program offers interactive activities and databases to teach data analytics across all grade levels.

Thousands of Nigerian Girls to Benefit from Ed Tech and 3D Printing



Youth for Technology
Foundation

PewResearchCenter

Only 28 Percent of American Adults 'Very' Familiar With Edtech Terms

Education Innovation Technology Summit (ETIS16)

Register ☆ BOOKMARK

Wednesday, September 28
8:00am–7:00pm

McHenry Row Campus
1215 E. Fort Avenue 21230

Events



<http://bit.ly/2cXzRzl>

https://www.eventbrite.com/e/ny-edtech-meetup-edsurge-ny-edtech-jobs-fair-tickets-27396309098?utm_source=blurb&utm_medium=edsurge&utm_campaign=es-blurb-ny-fall16-eventbrite

.....

LA Happy Hour this Thursday

<http://bit.ly/2cX5Zmj>

.....

Monday, October 3rd

5:00PM-6:00PM

Thorndike 157

R Markdown

<http://rmarkdown.rstudio.com/gallery.html>

(Charles, this is to remind you to demonstrate what the interface looks like)

Action Item

- Go to RStudio
- File -> New File -> R Markdown
- Fill the document with some code
- Save
- Knit to HTML

Tidy Data

(#notidynodessert)

Data Frames & Vectors

Why is tidy data?

- Difference between “clean” and “tidy”
- Data comes in a lot different structures, some which are difficult to analyze
- We want to make them manageable
- We want them to be “intuitive” to R (vectorized)
- BUT we want to keep a very precise record of how we did that

What is tidy data?

1. Observations are in rows
2. Variables are in columns
3. In a single data set

But...?

- What is a variable?
- What is an observation?
- What goes where in a data matrix?

Wide Format

- Repeated measures are in a single row

Student	Quiz 2-1-16	Quiz 2-10-16	Quiz 2-20-16
Francis	10	10	11
Alex	14	15	18
Kaji	11	17	14
Miriam	8	10	8

Long (Narrow) Format

- Each row is one time point per subject

Student	Quiz	Date
Francis	10	2-1-16
Francis	10	2-10-16
Francis	11	2-20-16
Alex	14	2-1-16

Generalize

Male	Female
4	10
7	10

How many variables are in the above matrix?

1. Male
2. Female
3. Count

Types of Messiness

- Column headers are values, not variable names
- Multiple variables are stored in one column
- Variables are stored in both rows and columns
- Multiple types of experimental unit stored in the same table
- One type of experimental unit stored in multiple tables

Tidy System

- There are many commands and several packages for doing this in R
- We are going to try to stick to two: tidyr & dplyr (we may end up using more)
- Reshape, Subset, Variable generation, Combine, Summarize

Reshape

- Similar to generating pivot tables
- Long format \longleftrightarrow Wide format

Student	Quiz 1	Quiz 2	Quiz 3
Francis	10	10	11
Alex	14	15	18
Kaji	11	17	14
Miriam	8	10	8

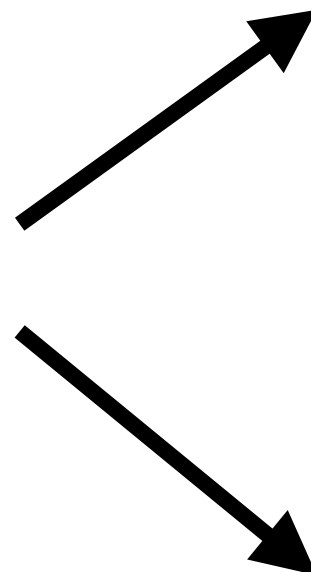


Student	Quiz	Date
Francis	10	2-1-16
Francis	10	2-10-16
Francis	11	2-20-16
Alex	14	2-1-16

Subset

- Splitting data frames

Student	Quiz	Date
Francis	10	2-1-16
Francis	10	2-10-16
Francis	11	2-20-16
Alex	14	2-1-16



Student	Quiz	Date
Francis	10	2-1-16
Francis	10	2-10-16
Francis	11	2-20-16

Student	Quiz	Date
Alex	14	2-1-16

Variable Generation

- Create new variable from current variables

Student	Quiz 2-1-16	Quiz 2-10-16	Quiz 2-20-16		mean
Francis	10	10	11	→	10.3
Alex	14	15	18		15.7
Kaji	11	17	14		14
Miriam	8	10	8		8.7

Combine

- Merge and bind dataframes
- Mutate or Filter

Student	Quiz 2-1-16	Quiz 2-10-16
Francis	10	10
Alex	14	15
Kaji	11	17


+

Student	Quiz 2-1-16	Quiz 2-20-16
Francis	10	9
Suchi	14	5
Kaji	11	10

Summarize

- Collapse data into a limited number of values according to a function

Student	Quiz 2-1-16	Quiz 2-10-16
Francis	10	10
Alex	14	15
Kaji	11	17



Av(Score/ Quiz/ Student)
12.8

Data Wrangling with dplyr and tidyr

Cheat Sheet



Syntax - Helpful conventions for wrangling

dplyr::tbl_df(iris)

Converts data to tbl class. tbl's are easier to examine than data frames. R displays only the data that fits onscreen:

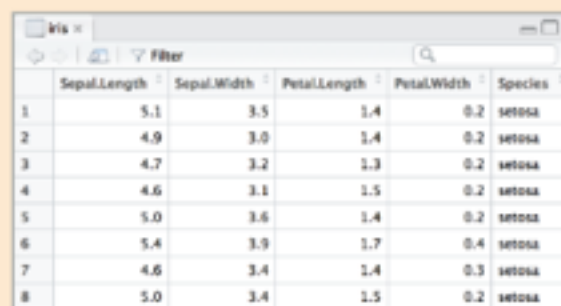
```
Source: local data frame [150 x 5]
   Sepal.Length Sepal.Width Petal.Length
1           5.1         3.5         1.4
2           4.9         3.0         1.4
3           4.7         3.2         1.3
4           4.6         3.1         1.5
5           5.0         3.6         1.4
..          ...          ...          ...
Variables not shown: Petal.Width (dbl),
Species (fctr)
```

dplyr::glimpse(iris)

Information dense summary of tbl data.

utils::View(iris)

View data set in spreadsheet-like display (note capital V).



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

dplyr::%>%

Passes object on left hand side as first argument (or . argument) of function on righthand side.

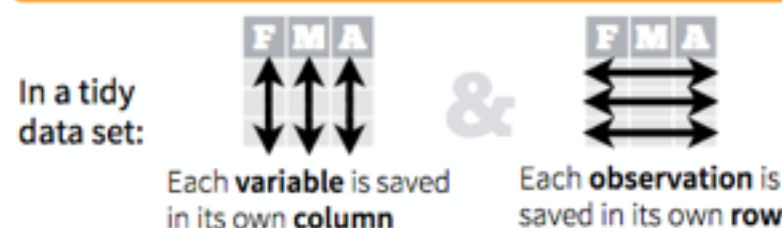
$x \%>\% f(y)$ is the same as $f(x, y)$

$y \%>\% f(x, ., z)$ is the same as $f(x, y, z)$

"Piping" with %>% makes code more readable, e.g.

```
iris %>%
  group_by(Species) %>%
  summarise(avg = mean(Sepal.Width)) %>%
  arrange(avg)
```

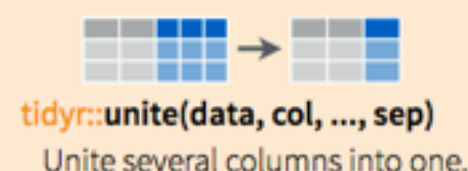
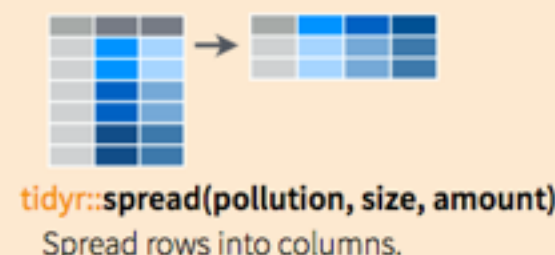
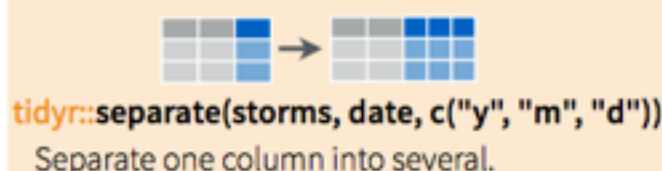
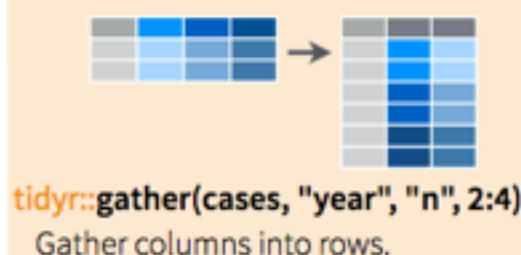
Tidy Data - A foundation for wrangling in R



Tidy data complements R's **vectorized operations**. R will automatically preserve observations as you manipulate variables. No other format works as intuitively with R.



Reshaping Data - Change the layout of a data set



dplyr::data_frame(a = 1:3, b = 4:6)
Combine vectors into data frame (optimized).

dplyr::arrange(mtcars, mpg)
Order rows by values of a column (low to high).

dplyr::arrange(mtcars, desc(mpg))
Order rows by values of a column (high to low).

dplyr::rename(tb, y = year)
Rename the columns of a data frame.

Subset Observations (Rows)



dplyr::filter(iris, Sepal.Length > 7)
Extract rows that meet logical criteria.

dplyr::distinct(iris)
Remove duplicate rows.

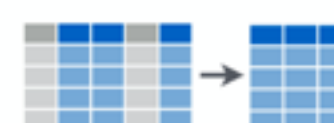
dplyr::sample_frac(iris, 0.5, replace = TRUE)
Randomly select fraction of rows.

dplyr::sample_n(iris, 10, replace = TRUE)
Randomly select n rows.

dplyr::slice(iris, 10:15)
Select rows by position.

dplyr::top_n(storms, 2, date)
Select and order top n entries (by group if grouped data).

Subset Variables (Columns)



dplyr::select(iris, Sepal.Width, Petal.Length, Species)
Select columns by name or helper function.

Helper functions for select - ?select

select(iris, contains(" "))
Select columns whose name contains a character string.

select(iris, ends_with("Length"))
Select columns whose name ends with a character string.

select(iris, everything())
Select every column.

select(iris, matches("t."))
Select columns whose name matches a regular expression.

select(iris, num_range("x", 1:5))
Select columns named x1, x2, x3, x4, x5.

select(iris, one_of(c("Species", "Genus")))
Select columns whose names are in a group of names.

select(iris, starts_with("Sepal"))
Select columns whose name starts with a character string.

select(iris, Sepal.Length:Petal.Width)
Select all columns between Sepal.Length and Petal.Width (inclusive).

select(iris, -Species)
Select all columns except Species.

Logic in R - ?Comparison, ?base::Logic

<	Less than	!=	Not equal to
>	Greater than	%in%	Group membership
==	Equal to	is.na	Is NA
<=	Less than or equal to	!is.na	Is not NA
>=	Greater than or equal to	&, , !, xor, any, all	Boolean operators