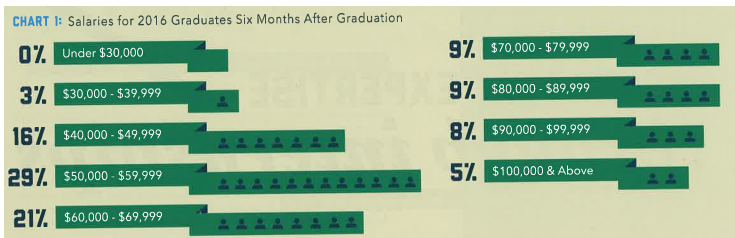


# Lecture 1: Welcome to Data Visualization Using R

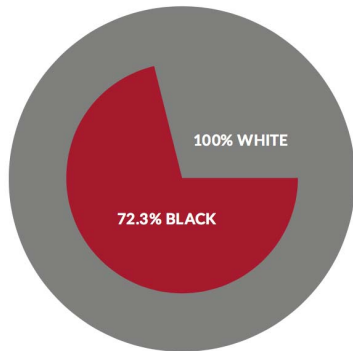
January 28, 2019

# Take This Class So You Won't Make This Graphic



From Trachtenberg's 2018 magazine.

## Or This One

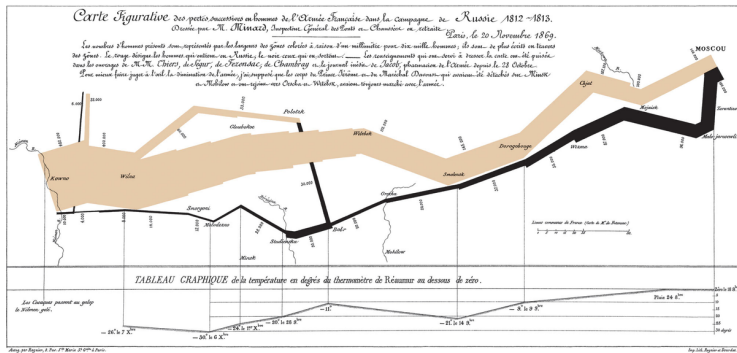


**EQUALITY INDEX OF BLACK AMERICA, 2016-2017**

	REVISED 2016	2017
<b>EQUALITY INDEX</b>	<b>72.2%</b>	<b>72.3%</b>
Economics	56.2%	56.5%
Health	79.4%	80.0%
Education	77.4%	78.2%
Social Justice	60.9%	57.4%
Civic Engagement	100.6%	100.6%

“U.S. Metros Ranked on Black-White Income Inequality,” *Next City*, May 2, 2017

# Instead, Aspire to This



See Tuftte for citation.



# Overview

Course Administration

Some R Examples

Tufte, Grandfather of Visualization

Getting Started with R

# Course Administration

## 1. Syllabus

- Good/bad/ugly assignments online
- Propose and trade and notify me if you can't do your assigned date

## 2. Bring a name tent to class

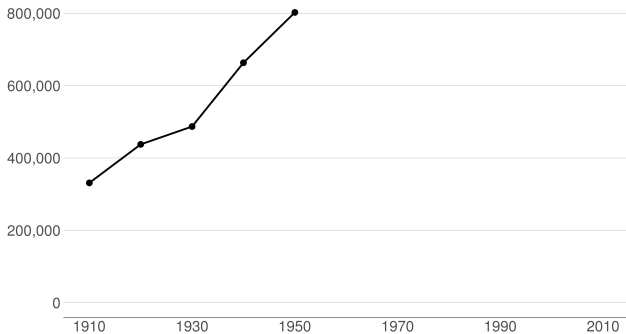
## 3. Questions/issues with readings?

## 4. Make sure you're signed up for Piazza

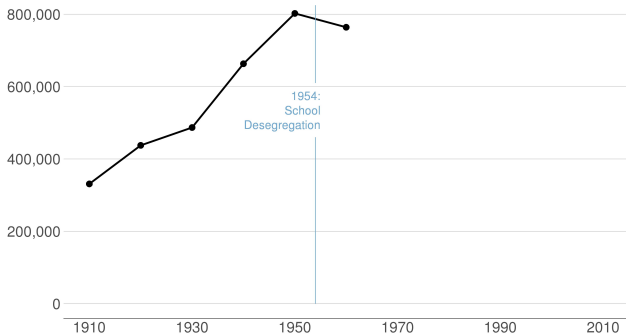
## 5. Introductions

- name and degree
- why this course?
- what you do now
- what you'd like to do when you're done

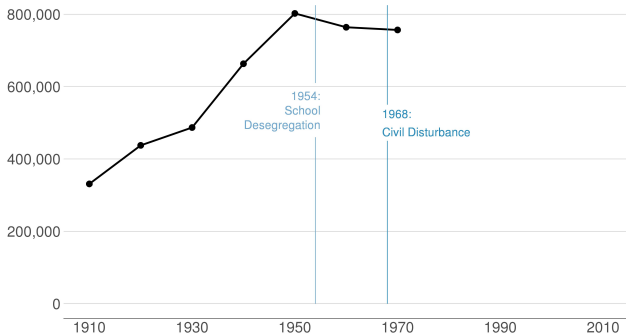
## DC Gains Population Through 1950



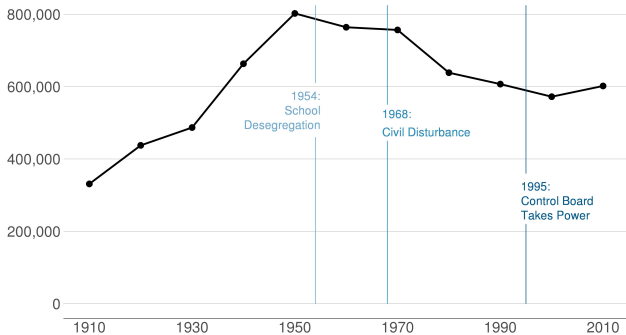
## Population Losses Start with Desegregation



## Continue After Civil Disturbance



## Population Turns Up After 2000



## Profound Changes: Share African American by Neighborhood

**1930**

1940

1950

1960

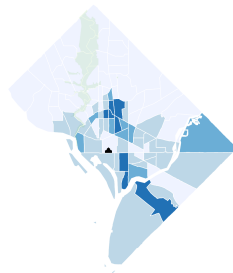
1970

1980

1990

2000

2010



0.15

0.5

0.85

## Profound Changes: Share African American by Neighborhood

1930

**1940**

1950

1960

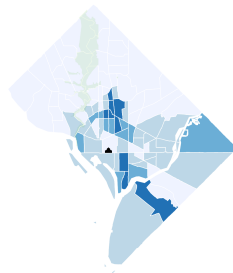
1970

1980

1990

2000

2010



0.15

0.5

0.85



## Profound Changes: Share African American by Neighborhood

1930

1940

**1950**

1960

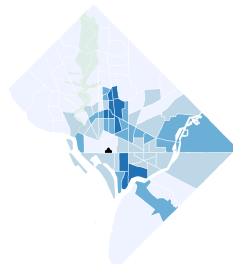
1970

1980

1990

2000

2010



0.15

0.5

0.85

## Profound Changes: Share African American by Neighborhood

1930

1940

1950

**1960**

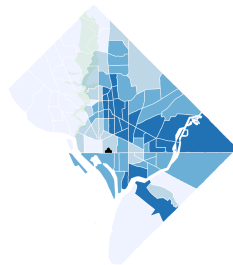
1970

1980

1990

2000

2010



0.15

0.5

0.85

## Profound Changes: Share African American by Neighborhood

1930

1940

1950

1960

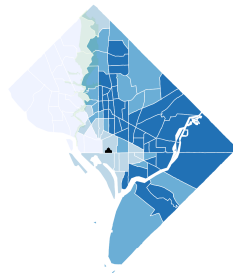
**1970**

1980

1990

2000

2010



0.15

0.5

0.85

## Profound Changes: Share African American by Neighborhood

1930

1940

1950

1960

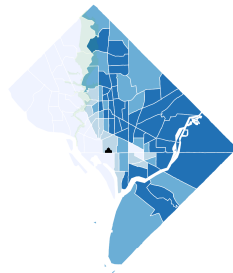
1970

**1980**

1990

2000

2010



0.15

0.5

0.85

## Profound Changes: Share African American by Neighborhood

1930

1940

1950

1960

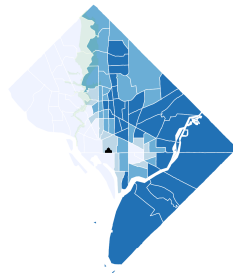
1970

1980

**1990**

2000

2010



0.15

0.5

0.85

## Profound Changes: Share African American by Neighborhood

1930

1940

1950

1960

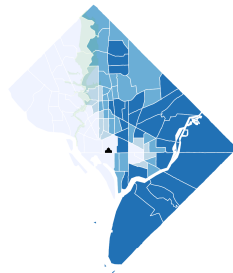
1970

1980

1990

**2000**

2010



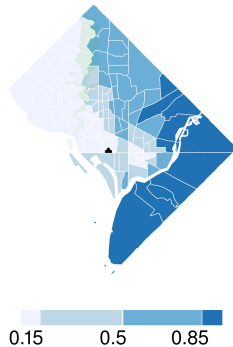
0.15

0.5

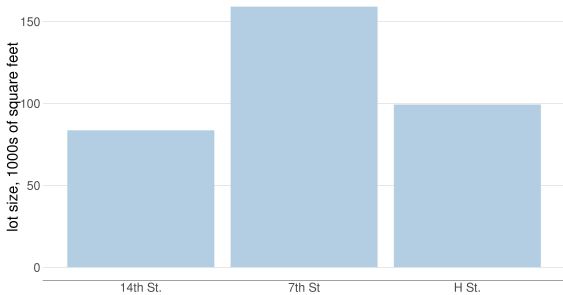
0.85

## Profound Changes: Share African American by Neighborhood

1930  
1940  
1950  
1960  
1970  
1980  
1990  
2000  
**2010**

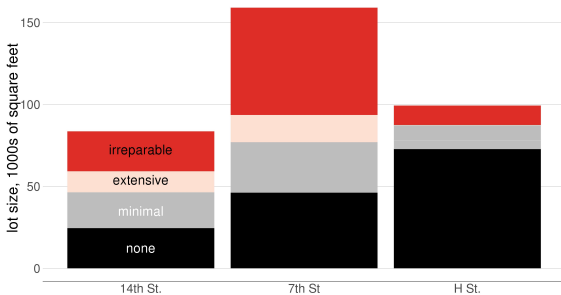


## By Square Footage, 7th Street is Most Impacted

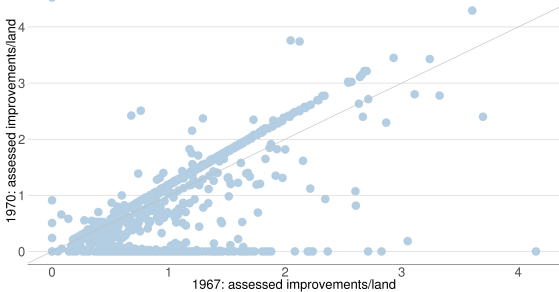




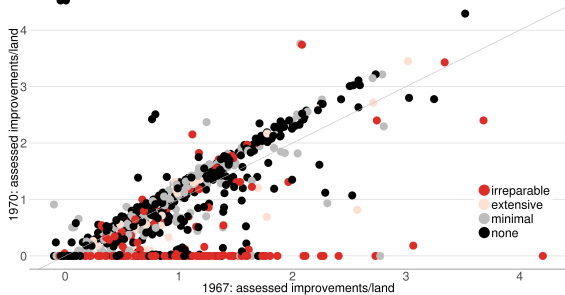
## Roughly Half of Square Footage Damaged to Some Degree



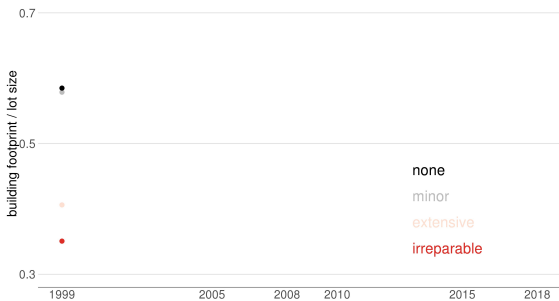
## Assessed Value of Most Improvements Drops, 1967 to 1970



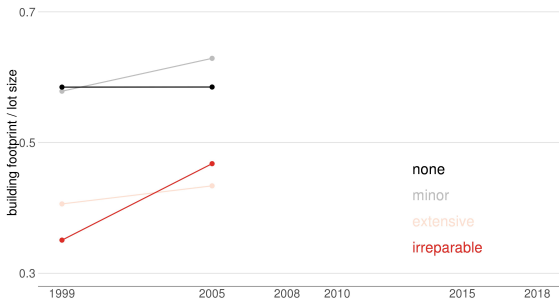
## Damaged Properties Lose Improvements, A Few Rebuild



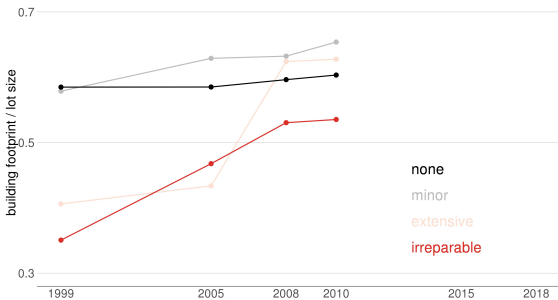
## 1999: Damaged Properties Have Smaller Structures



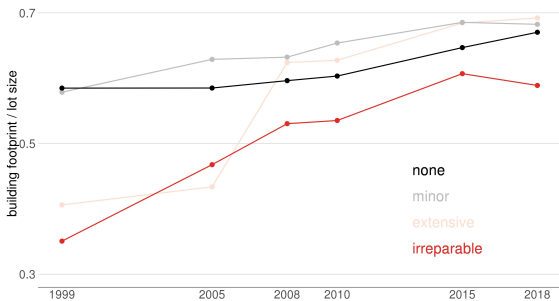
## 2005: Damaged Properties Show Some Catch-up



## 2010: Damaged Properties Approaching Undamaged Ones



## 2018: Near Convergence of Damaged Properties



# Tufte



# Edward Tufte

- A quantitative political scientist
- Writing in the mid-1970s
- Became interested in visualization by working with pioneering statistician John Tukey
- Remember that this is the pre-Excel era, in which data graphics are difficult to make

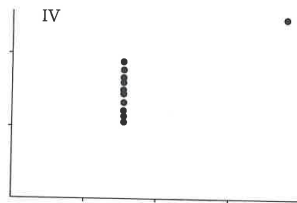
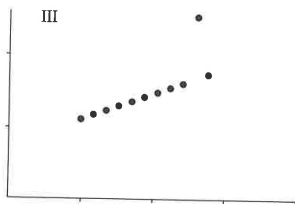
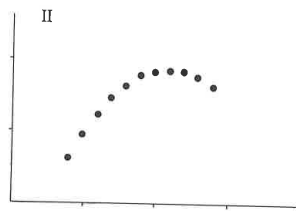
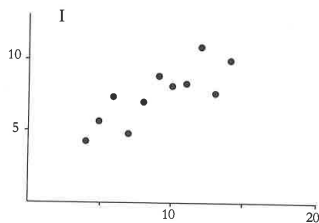
# An Argument for Better Visualization

Because good visualizations tell the most compelling story

I		II		III		IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

# An Argument for Better Visualization

Because good visualizations tell the most compelling story



# Tufte's Types of Graphs

1. Data maps
2. Time series
3. Space-time narrative designs
4. Relational graphs – the holy grail

# Data Maps

- Describe the location of numbers
- This can be revealing or obfuscating
- We will make these in this class

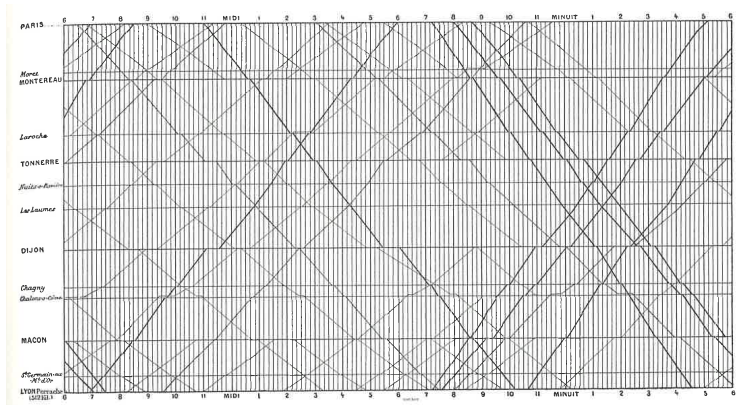
# Data Map Example



# Time Series

- Time on the horizontal axis
- Something else on the vertical axis
- One of the first types of data graphics

# Train, Paris to Lyon



See Tufte for citation.

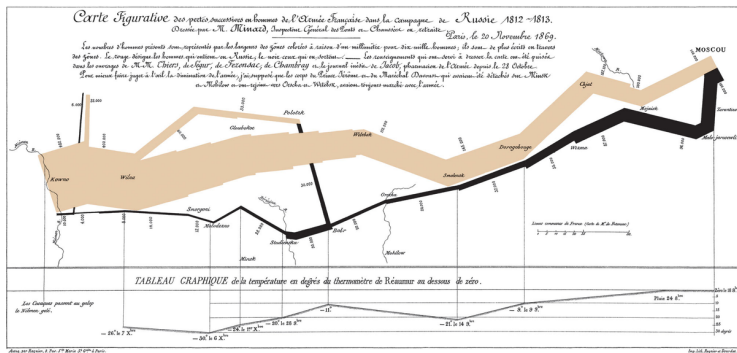


# Space-Time Narrative Designs

- Move over space and time at the same time
- A time series plus



# Space-Time Narrative Example



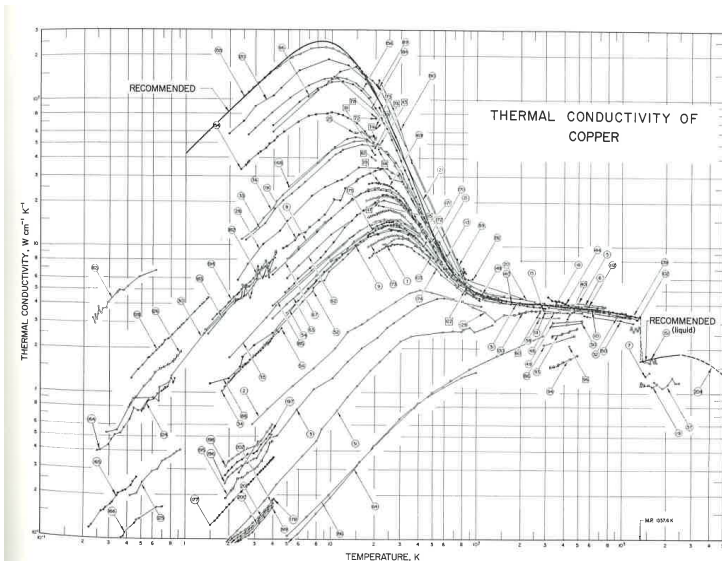
Which dimensions?

1. army size
2. army location, N/S
3. army location, E/W
4. direction of movement
5. temperature
6. by date

# Relational Graphics

- One variable on the vertical, another on the horizontal
- A conceptual advance in graphics
- A more sophisticated way of thinking

# Relational Graphics Example

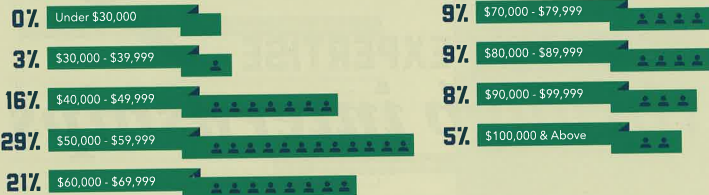


# Tufte's Main Causes of Distortion in Graphics

1. Data are bad
  - should be per capita and are not
  - data are not consistent over time
  - don't adjust for inflation
2. Graphics are rotten
  - size doesn't match the numbers
  - colors and styles are misleading
  - graphic fails to highlight key point
3. Graphics are irrelevant
  - too much extraneous stuff

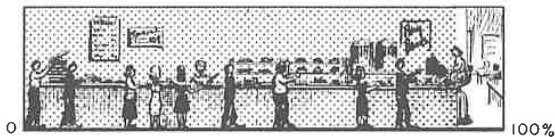
# Size and Number Don't Match

**CHART 1:** Salaries for 2016 Graduates Six Months After Graduation



# A Continuing Problem: Graphics are Irrelevant

The Company Cafeteria was used by 9 Out of 10  
Employees during the Fiscal Year 1949



Source: COMPANY REPORTS



## Tufte's Six Rules of Graphic Integrity, 1 to 3 of 6

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
2. Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data.
3. Show data variation, not design variation.

## Tufte's Six Rules of Graphic Integrity, 4 to 6

4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units.
5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data.
6. Graphics must not quote data out of context.

R

# What is R?

- A programming language
- Developed by statisticians from New Zealand
- Open source, and therefore free
- Based on “S,” developed by Bell Labs

# Strengths of R

- Free
- Open-source, so packages by all kinds of users are available
- There are frequently many ways to do the same task
- Very good graphics
- Very flexible
- Can have many datasets in memory at once

# Weaknesses of R

- Not always enterprise-ready: packages break and there is no central help
- There are frequently many ways to do the same task
- Syntax can be challenging

# Today's Goals

- When you leave today, you will be able to
  - run a R script
  - create a R dataframe
  - do basic operations with a R dataframe
- Download the R tutorial for this class now.
- You'll continue work at home on your own and turn in a problem set next lecture

# Today

- A. A R dataframe
- B. Packages
- C. Subsetting
- D. Summarizing



## A. A R dataframe

- ▶ a dataframe is the basic building block of data analysis in R
- ▶ R has other types of data structures, but this will be the most useful to you
- ▶ dataframe consists of columns
- ▶ each column can be
  - ▶ numeric: 1,2,3.556,-2.6
  - ▶ or
  - ▶ character: "hello", "dogs are good", ""
- ▶ refer to rows and columns

## Sample dataframe

```
new.dataframe <- data.frame(class = c(1,2,3),  
                             subject = c("basics","merging",  
                                         students = c(19,19,18))  
new.dataframe
```

```
##   class subject students  
## 1     1  basics      19  
## 2     2 merging      19  
## 3     3  graphs      18
```

## Referring to parts of the dataframe

```
new.dataframe[ROWS,COLUMNS]
```

## Referring to parts of the dataframe

```
new.dataframe[ROWS,COLUMNS]
```

Just one column, all rows

```
new.dataframe[,c("students")]
```

```
## [1] 19 19 18
```

## Referring to parts of the dataframe

```
new.dataframe[ROWS,COLUMNS]
```

Just one column, all rows

```
new.dataframe[,c("students")]
```

```
## [1] 19 19 18
```

Just two rows, all columns

```
new.dataframe[1:2,]
```

```
##   class subject students
## 1     1   basics      19
## 2     2 merging      19
```

## B. Packages

- ▶ there is “Base R,” which is a set of basic commands
- ▶ and user-written packages that add functionality
- ▶ some packages are maintained by teams, frequently updated, and do many things
- ▶ some are one-function add-ins
- ▶ most famous are those by Hadley Wickham
- ▶ today we’ll use his “plyr” package

## Installing packages

- ▶ install a package once

```
install.packages("plyr", dependencies = TRUE)
```

- ▶ call a package at the beginning of any program in which you'd like to use the package

```
library(plyr)
```

- ▶ this makes commands accessible

## C. Subsetting

- ▶ sometimes you want to work with something smaller than the whole dataframe
- ▶ create a new dataframe that has only part of the big one
- ▶ here we keep just students 1 and 2

```
df.smaller <- new.dataframe[1:2,]  
df.smaller
```

```
##   class subject students  
## 1     1  basics      19  
## 2     2 merging      19
```



## Subset by attributes

- ▶ take only classes with 19 students

```
df19 <-  
  new.dataframe[which(new.dataframe$students == 19),]  
df19
```

```
##   class subject students  
## 1     1  basics      19  
## 2     2 merging      19
```

## D. Summarizing

- ▶ frequently, you'd like to know something at a level of aggregation not in your dataset
- ▶ in our case, maybe average attendance
- ▶ make a new dataframe with this information
- ▶ use `plyr` library

## Summarize data: load dplyr package

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:plyr':  
##  
##   arrange, count, desc, failwith, id, mutate, rename,  
##   summarize  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

## Making a new dataset that is a function of the old one

```
av.attendance <- summarize(new.dataframe,  
                           av.at=mean(students, na.rm = TRUE))  
av.attendance
```

```
##      av.at  
## 1 18.66667
```

- ▶ more complicated example in tutorial

## Next Lecture

- Turn in PS 1, which is at the end of the tutorial
- Read Few Chapters 3 and 5
- Look at “Smart Graph Chart”