

Data Wrangling in R

Yunkyu Sohn

September 27, 2016

Research Associate, Department of Politics



PRINCETON
UNIVERSITY





COMPASS Workshops

Computing for Data Analysis in the Social Sciences

- Free, open-source statistical programming and data analysis workshops using R and RStudio
- Open to everyone with a Princeton ID
- No programming experience is necessary or expected
- Attendees should bring a laptop computer to fully participate in the workshops



COMPASS Workshops

Computing for Data Analysis in the Social Sciences

[News](#) [Events](#) [People](#)

OVERVIEW

The COMPASS Workshops are a series of free, open-source statistical programming and data science workshops held weekly during the academic year at Princeton University. Supported by the [Department of Sociology](#), [Department of Politics](#), and the [Office of the President](#), the workshops focus on using R with RStudio in a variety of real-world applications relevant to social scientists. Topics covered include data wrangling, data visualization, Monte Carlo simulations, linear regression, social network analysis, and text mining.

The COMPASS Workshops are held on **Tuesdays from 7:30 to 9:00pm at Friend Center Auditorium 101** at Princeton University during the academic year. The workshops are open to all students, staff, and affiliates of Princeton as well as the larger research community. To fully participate in the workshops attendees should bring a laptop.

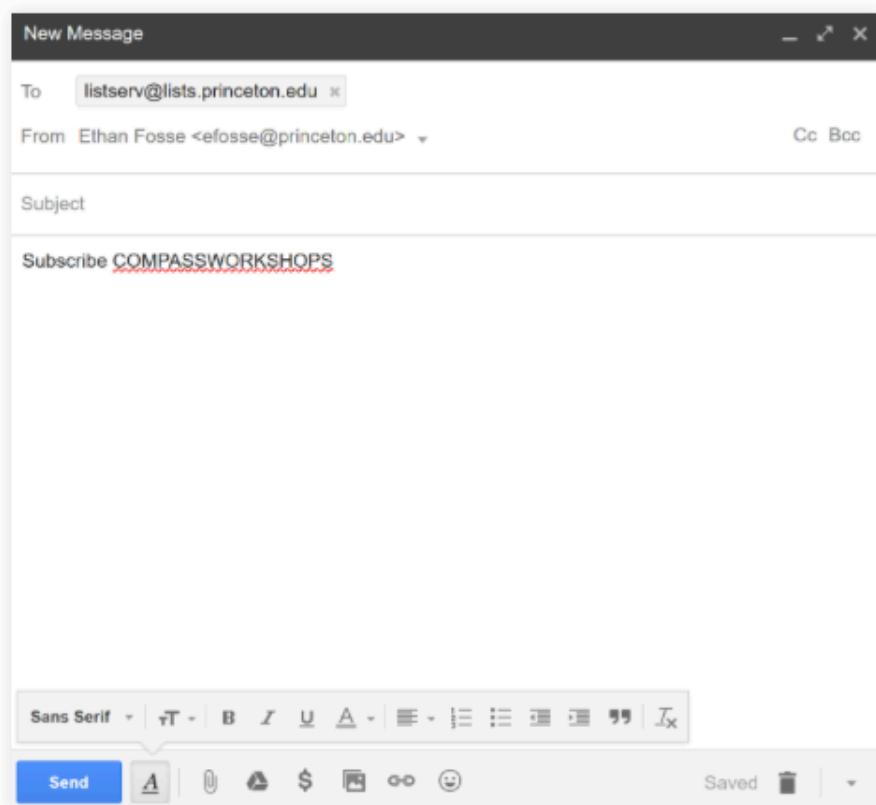
To find out more about the workshops, join the COMPASS Workshops mailing list by sending an email to listserv@lists.princeton.edu with "Subscribe COMPASSWORKSHOPS" in the body and all other lines blank including the subject.



FALL 2016 SCHEDULE

September 20	Introduction to R and RStudio	Slides	Data	Code
September 27	Data Wrangling in R			
October 4	Base R Graphics			
October 11	Data Visualization in R with ggplot2			
October 18	Programming Loops in R			
November 8	Probability and Simulations in R			
November 15	Monte Carlo Simulations in R			
November 29	Text Analysis in R			
December 6	Hypothesis Testing in R			
December 13	Regression Analysis in R			

Our Mailing List



Send an email to
`listserv@lists.princeton.edu`
with "Subscribe
COMPASSWORKSHOPS" in
the body and all other lines
blank, *including the subject*.

People

- **Teaching Staff**

- Ethan Fosse (Research Associate, Department of Sociology)
- Yunkyu Sohn (Research Associate, Department of Politics)

- **Faculty Sponsors**

- Margaret Frye (Assistant Professor, Department of Sociology)
- Kosuke Imai (Professor, Department of Politics)
- Matthew Salganik (Professor, Department of Sociology)

Todays' Contents

1. Before You Begin
2. Today's Project
3. Things to Cover
4. Learning by Doing
5. Research Questions

Before You Begin

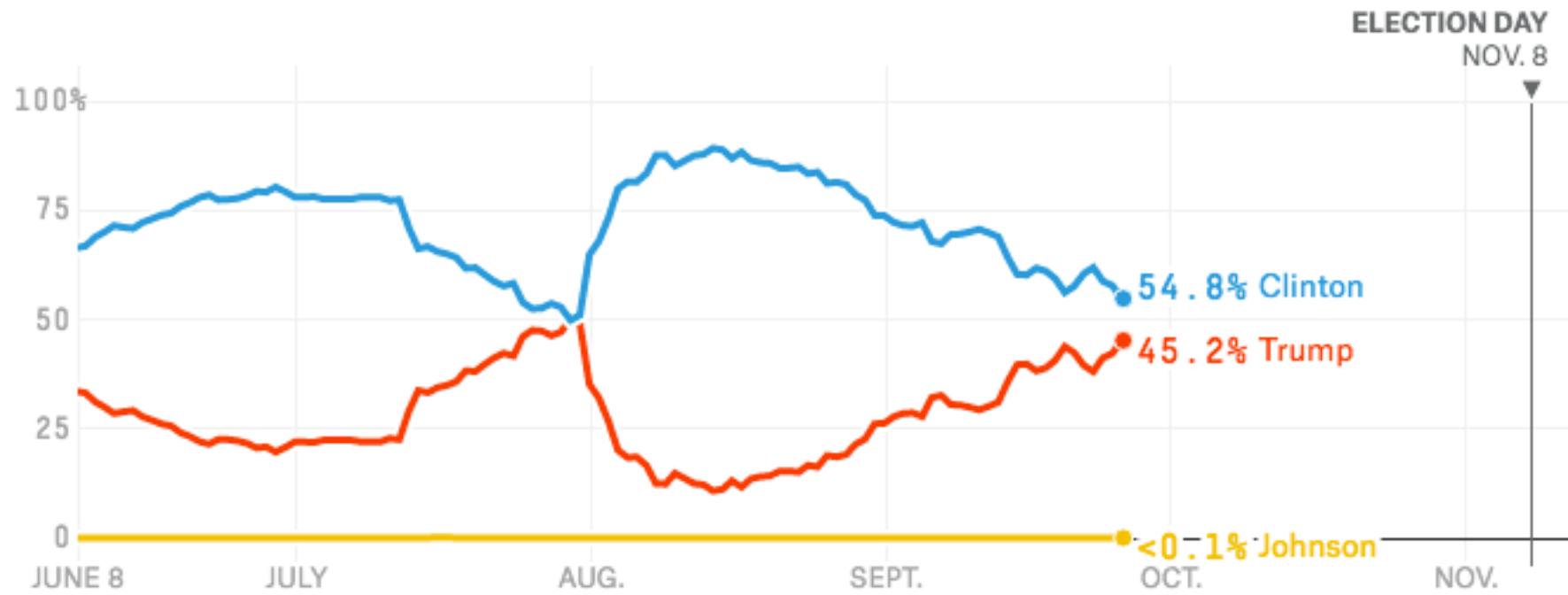
1. You should have a computer with Internet connection.
2. You should have R and RStudio (latest version preferred) installed.
3. Download Slides and Codes for Week 2 (right click -> save as) at
<https://compass-workshops.github.io/info/>
4. Start Rstudio

That's all!

Today's Project

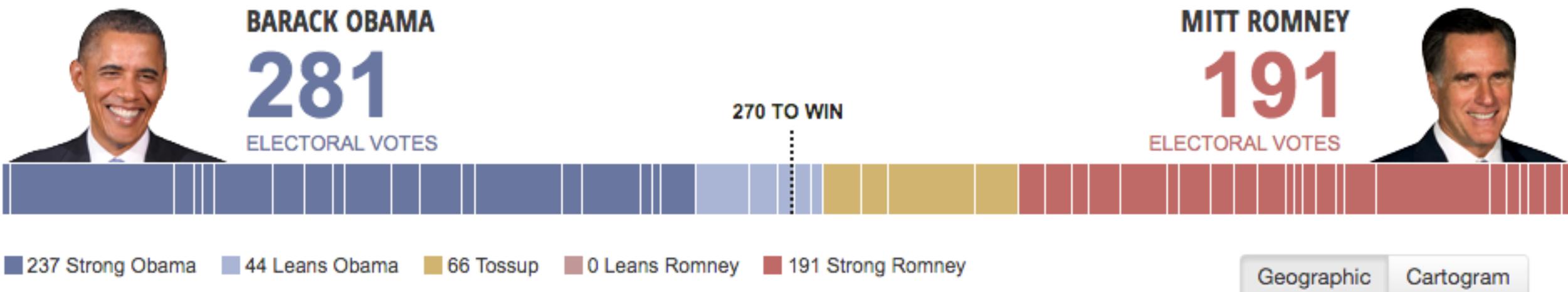


Election Prediction = Data + Statistics

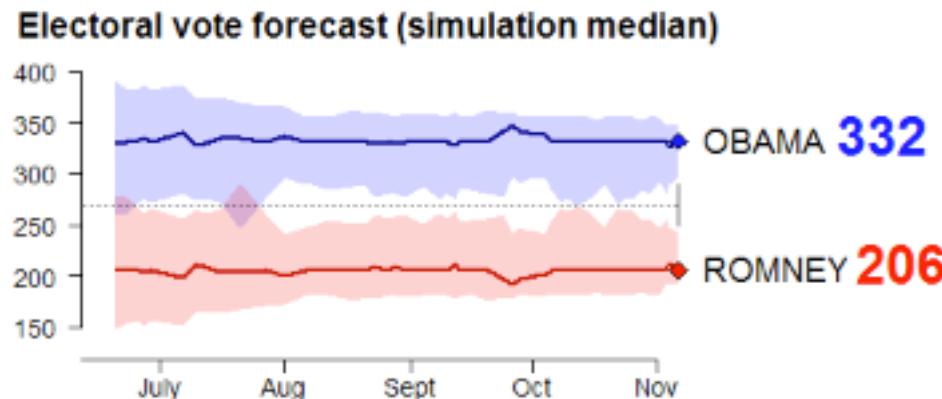


FiveThirtyEight

Election Prediction = Data + Statistics

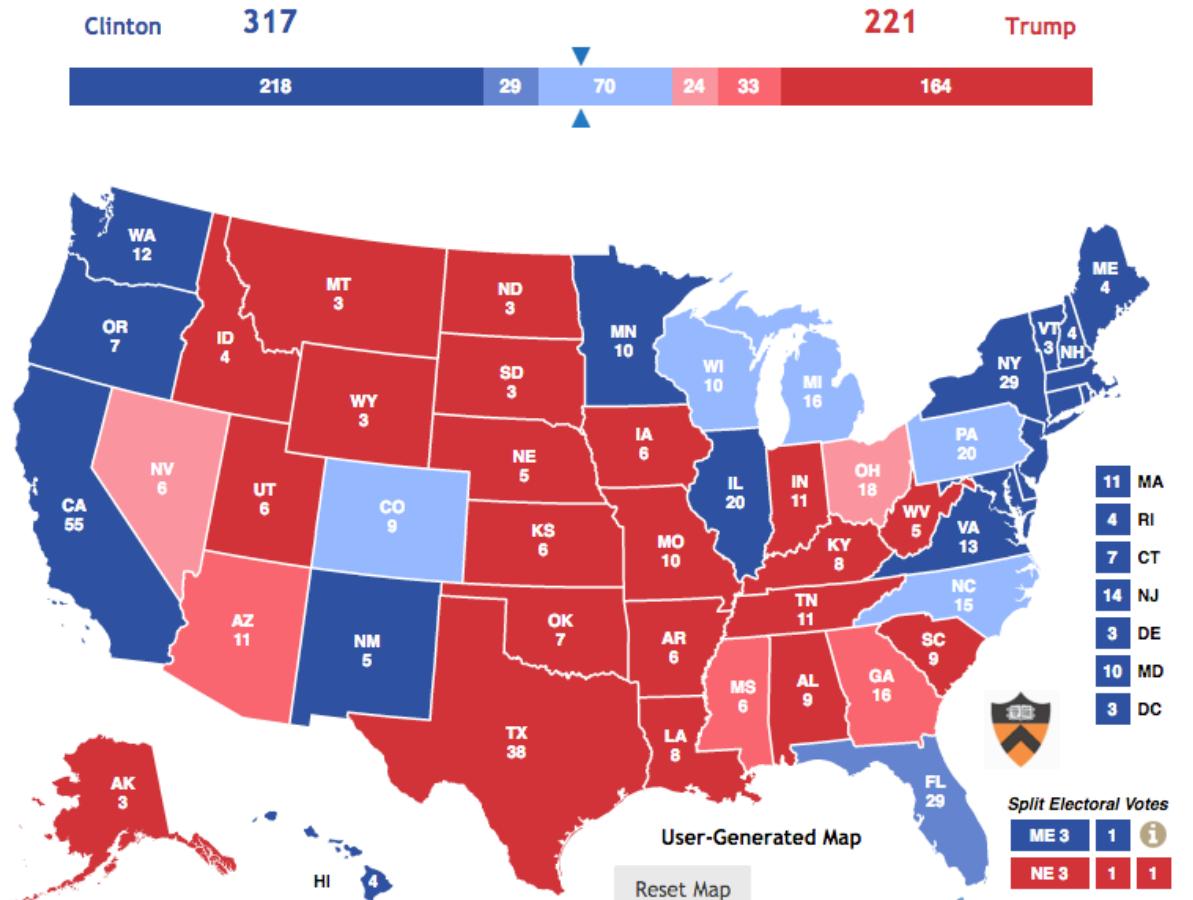


Election Prediction = Data + Statistics



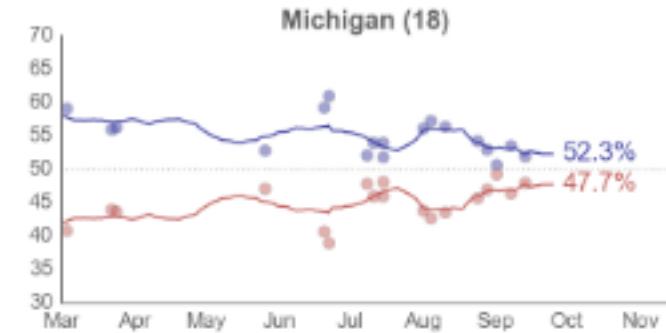
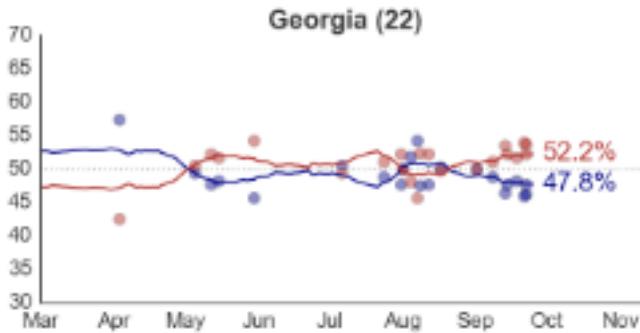
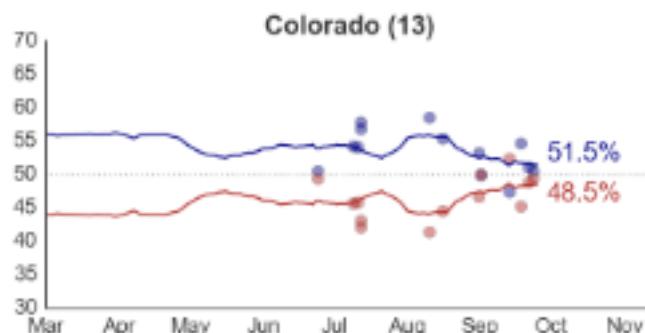
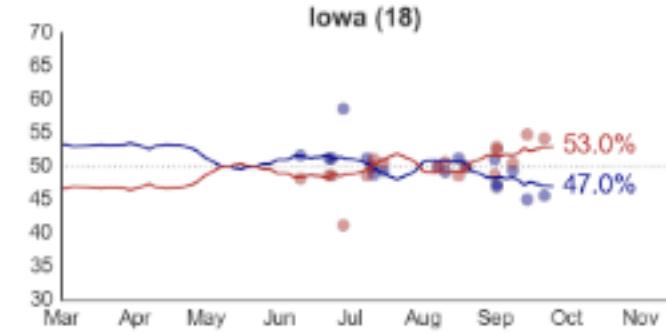
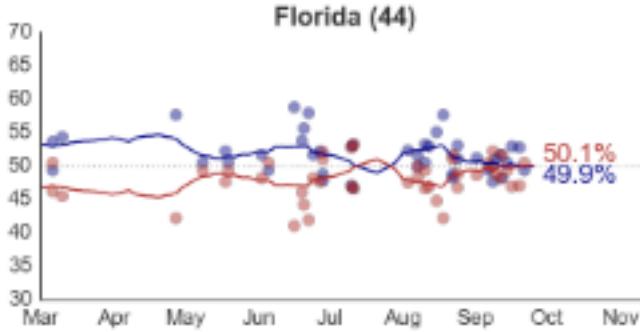
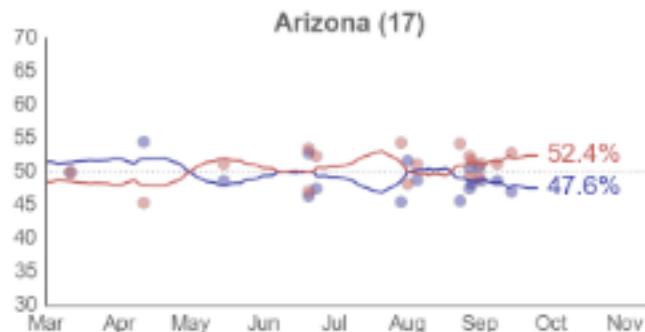
VOTAMATIC
Polling Analysis and Election Forecasting

Election Prediction = Data + Statistics

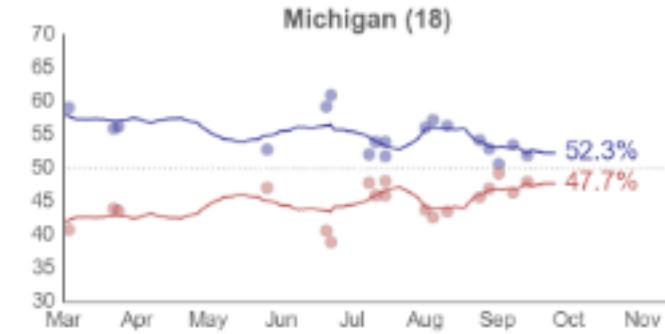
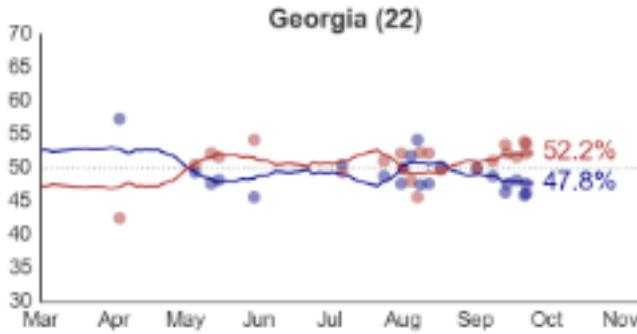
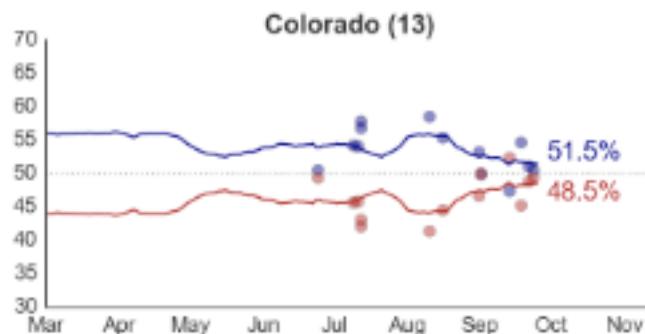
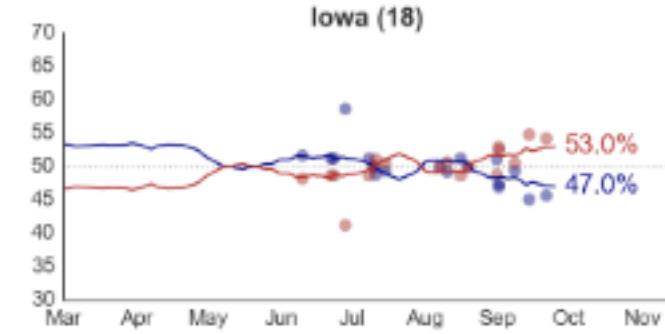
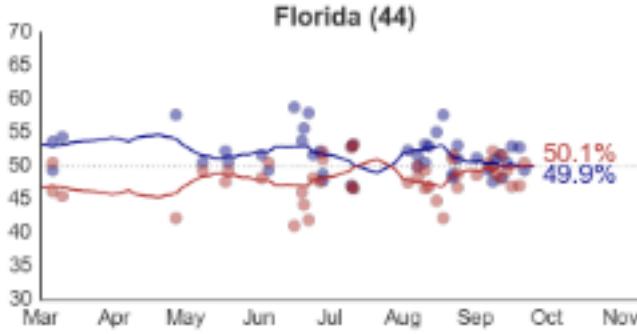
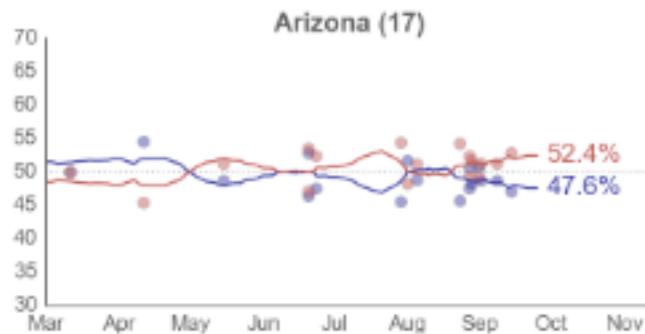


Princeton Election Consortium

Generic Approach to Election Prediction

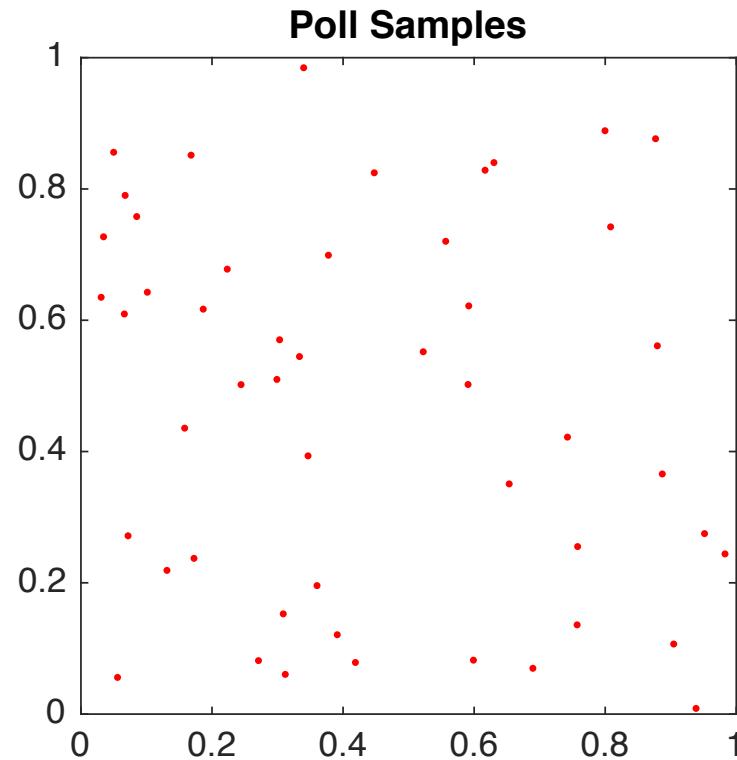


Generic Approach to Election Prediction

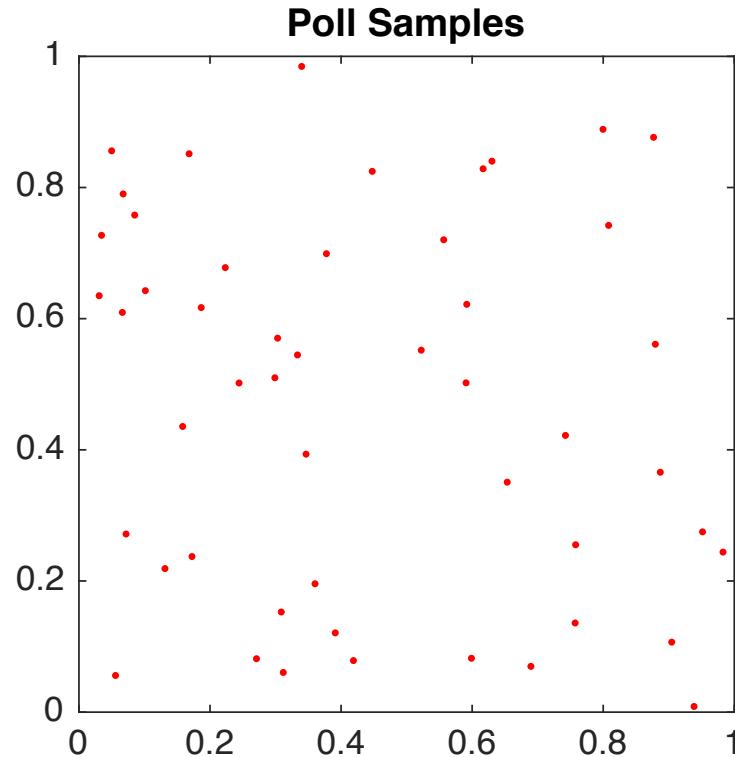
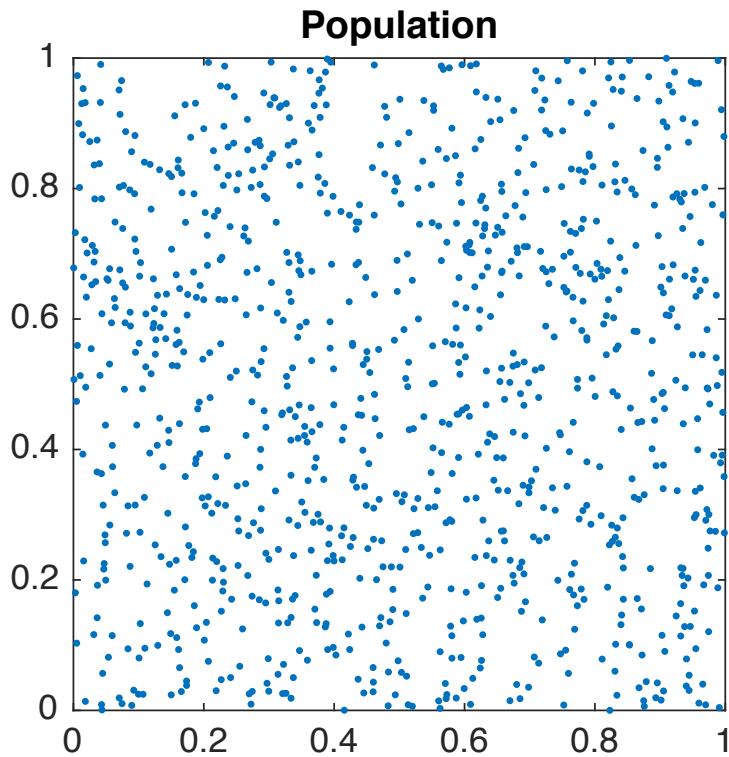


Statistical **aggregation** of state **poll** results over time

Election Prediction = Data + Statistics



Election Prediction = Data + Statistics

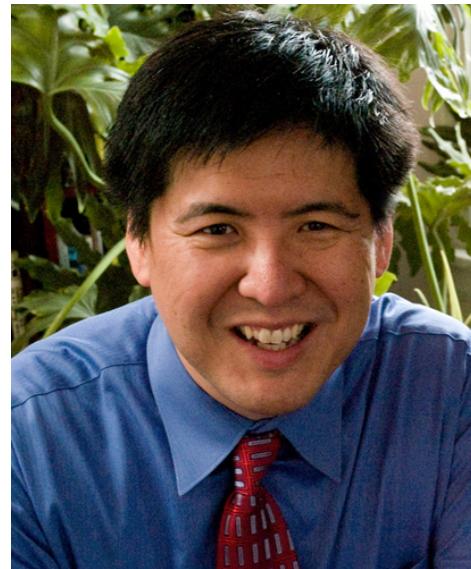
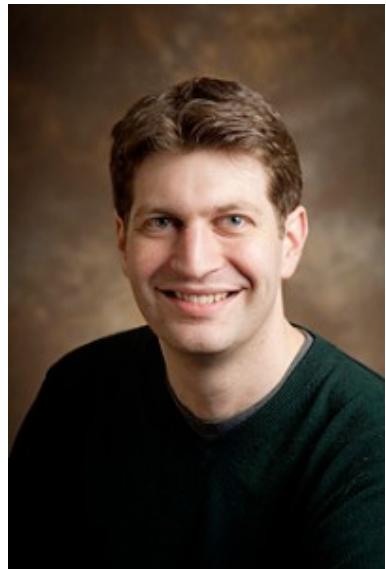


Using Statistics to infer
latent population preference



$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

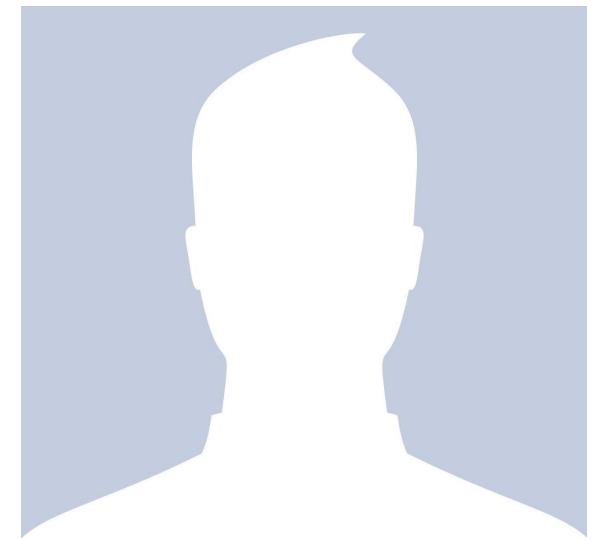
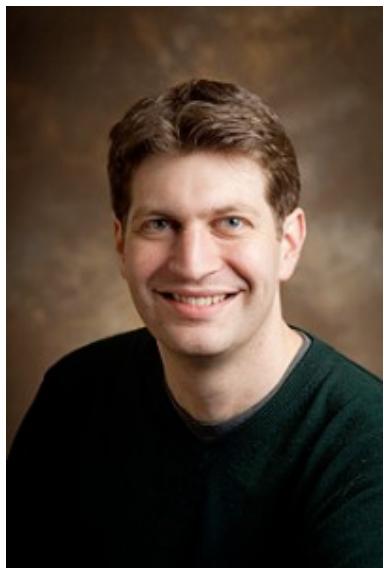
Generic Approach to Election Prediction



Statistical **aggregation** of state **poll** results over time

After 60 minutes, You will be able to

After 60 minutes, You will be able to



Youself

After 60 minutes, You will be able to
conduct preliminary poll analysis from scratch!

After 60 minutes, You will be able to
conduct preliminary poll analysis from scratch!

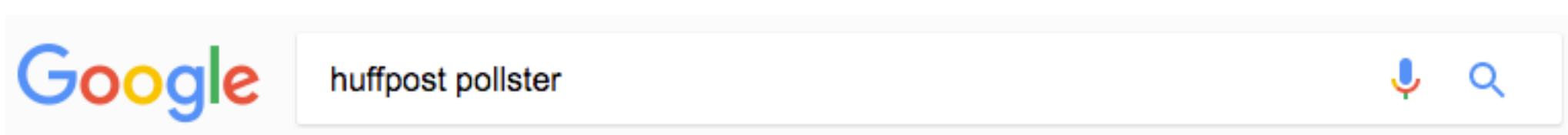
- Obtain a very similar dataset used in popular prediction sites
- Learn how to do basic data import/manipulation using this dataset
- Do elementary data analysis

Things to Cover

- Data acquisition from the Internet and import to R
- Understanding R data structures
- Subsetting Observations
- Subsetting Variables
- Subsetting Both Observations and Variables
- Summarizing Data
- Creating and Renaming Variables
- Merging Data Sets

Learning by Doing

Download the Dataset



HuffPost Pollster - Polls and Charts - Election Results

elections.huffingtonpost.com/pollster ▾ 이 페이지 번역하기

Polls, charts, forecasts and data about upcoming elections, Obama, Congress, Democrats, Republicans, politics, health care and the economy.

[2016 General Election: Trump ...](#)

Polls and chart for 2016 General
Election: Trump vs. Clinton. See ...

[polls](#)

Polls, charts, forecasts and data about
upcoming elections ...

[2016 National Republican ...](#)

Polls and chart for 2016 National
Republican Primary. See the ...

[2016 General Election: Trump ...](#)

Polls and chart for 2016 General
Election: Trump vs. Clinton vs ...

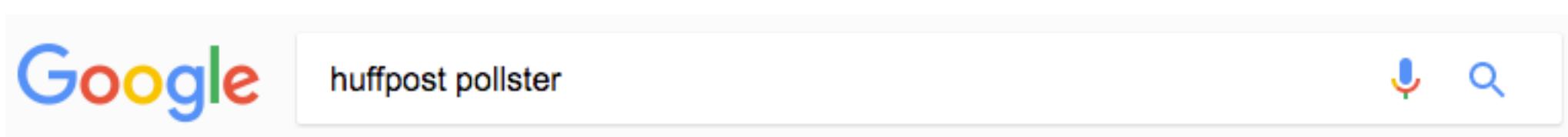
[2016 National Democratic ...](#)

Polls and chart for 2016 National
Democratic Primary. See the ...

[Obama Job Approval](#)

Polls and chart for Obama Job Approval.
See the [latest](#) ...

Download the Dataset



HuffPost Pollster - Polls and Charts - Election Results

elections.huffingtonpost.com/pollster ▾ 이 페이지 번역하기

Polls, charts, forecasts and data about upcoming elections, Obama, Congress, Democrats, Republicans, politics, health care and the economy.

2016 General Election: Trump ...

Polls and chart for 2016 General Election: Trump vs. Clinton. See ...

2016 National Republican ...

Polls and chart for 2016 National Republican Primary. See the ...

2016 National Democratic ...

Polls and chart for 2016 National Democratic Primary. See the ...

polls

Polls, charts, forecasts and data about upcoming elections ...

2016 General Election: Trump ...

Polls and chart for 2016 General Election: Trump vs. Clinton vs ...

Obama Job Approval

Polls and chart for Obama Job Approval. See the latest ...

Download the Dataset



RSS | CSV | CSV (Slim) | JSON | API Docs
House Effects: CSV

1. Right click on CSV
 2. Save as to a preferred location
- Make sure you know the location!**

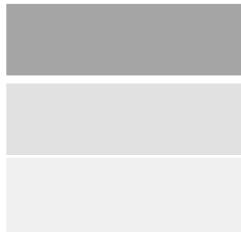
Get back to Rstudio: Loading your Dataset

Task 1: Convert the downloaded CSV (Comma Separated Values) File into an R object.

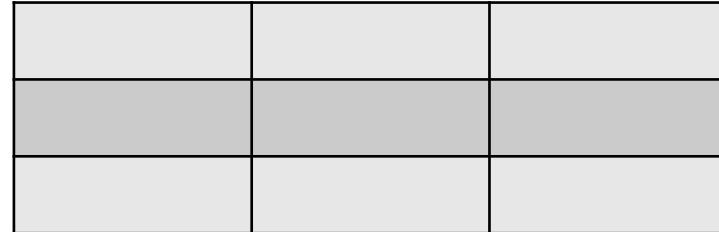
```
rm(list=ls());
## Delete your workspace;
getwd();
## Check your current working directory
setwd("<location of your dataset>");
## Set your working directory;
poll = read.csv("2016-general-election-trump-vs-clinton.csv");
## Load data;
```

R Data Structures

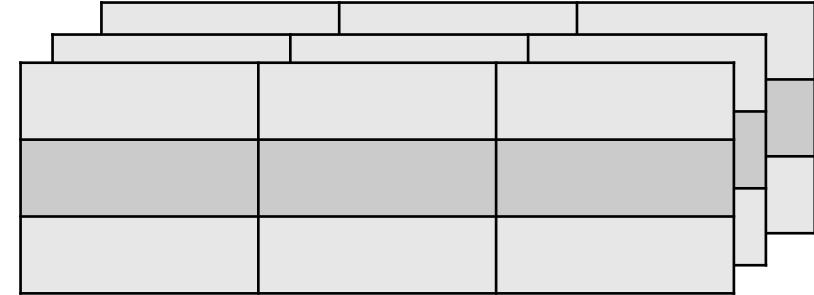
Vector



Matrix



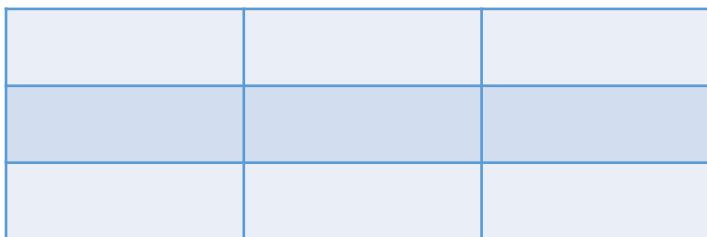
Array



Data frame

Observations

Variables with heterogeneous elements



e.g. Data frame for a poll dataset

Poll ID	Mode	DVotes	RVotes

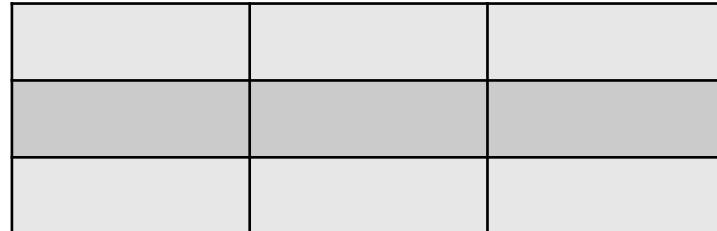
factor numeric numeric
 value value

R Data Structures

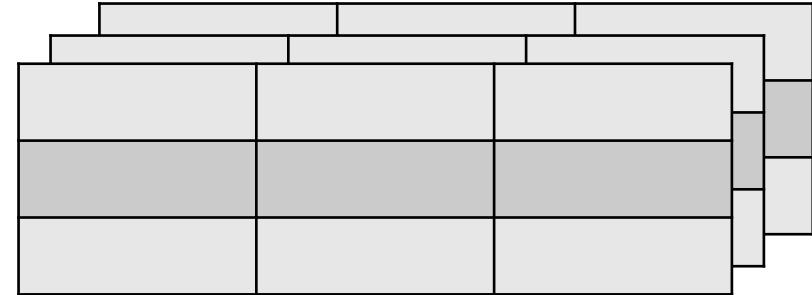
Vector



Matrix



Array



Data frame

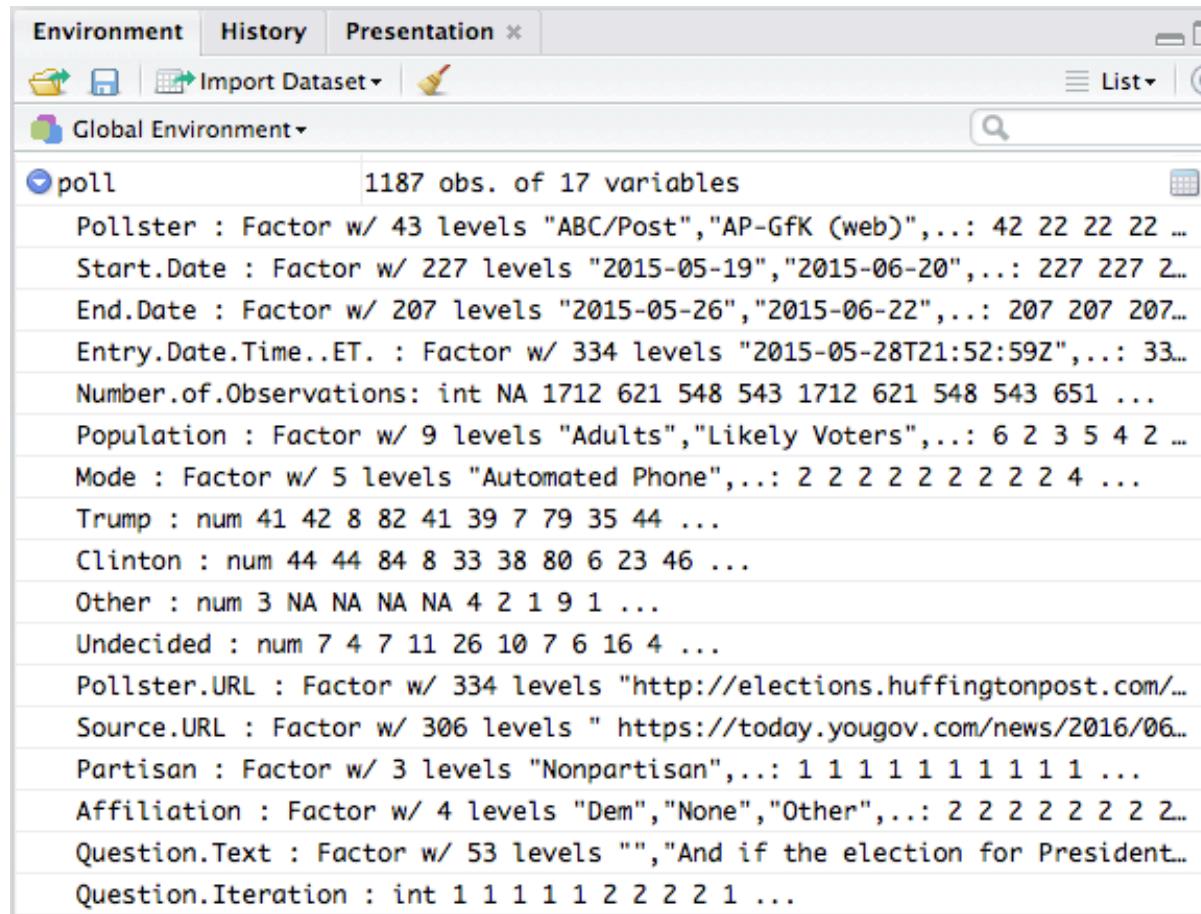
Poll ID	Mode	DVotes	RVotes

factor numeric numeric
 value value

Create a data frame

```
dfex <- data.frame(mode = c("phone", "Internet", "Internet"),
DVotes = c(40, 50, 60), RVotes = c(60, 50, 40));
dfex;
```

Quick Inspection of Poll Data Frame



The screenshot shows the RStudio interface with the 'Global Environment' tab selected. A single object named 'poll' is listed, which contains 1187 observations across 17 variables. The variables include categorical factors like 'Pollster', 'Start.Date', 'End.Date', 'Entry.Date.Time..ET.', 'Mode', and various partisan and affiliation categories, along with numerical variables for 'DVotes' and 'RVotes'. The 'Mode' variable is explicitly labeled as having 5 levels.

```
Environment History Presentation x
Import Dataset | 
Global Environment | 
poll 1187 obs. of 17 variables
  Pollster : Factor w/ 43 levels "ABC/Post","AP-GfK (web)",...
  Start.Date : Factor w/ 227 levels "2015-05-19","2015-06-20",...
  End.Date : Factor w/ 207 levels "2015-05-26","2015-06-22",...
  Entry.Date.Time..ET. : Factor w/ 334 levels "2015-05-28T21:52:59Z",...
  Number.of.Observations: int NA 1712 621 548 543 1712 621 548 543 651 ...
  Population : Factor w/ 9 levels "Adults","Likely Voters",...
  Mode : Factor w/ 5 levels "Automated Phone",...
  Trump : num 41 42 8 82 41 39 7 79 35 44 ...
  Clinton : num 44 44 84 8 33 38 80 6 23 46 ...
  Other : num 3 NA NA NA NA 4 2 1 9 1 ...
  Undecided : num 7 4 7 11 26 10 7 6 16 4 ...
  Pollster.URL : Factor w/ 334 levels "http://elections.huffingtonpost.com/",...
  Source.URL : Factor w/ 306 levels " https://today.yougov.com/news/2016/06/...
  Partisan : Factor w/ 3 levels "Nonpartisan",...
  Affiliation : Factor w/ 4 levels "Dem","None","Other",...
  Question.Text : Factor w/ 53 levels "", "And if the election for President...
  Question.Iteration : int 1 1 1 1 1 2 2 2 2 1 ...
```

Poll ID	Mode	DVotes	RVotes

factor numeric numeric
value value

Quick Inspection of Poll Data Frame

```
View(poll);  
## Spreadsheet-style data viewer;  
summary(poll);  
## Summarize variables on your console;  
names(poll);  
## Names of all variables;  
dim(poll);  
nrow(poll);  
ncol(poll);  
## Dimensional information;  
head(poll);  
tail(poll);
```

Communicating with Your Data

- How to select a specific variable of interest?: Use \$

```
poll$VariableName
```

- e.g. If you want to select the Affiliation variable

```
poll$Affiliation;
```

Subsetting by Direct Indexing

- Subsetting observations

	Mode	DVotes	RVotes
Poll ID	factor	numeric value	numeric value
1	blue	100	200
2	orange	150	250

- Subsetting variables

	Mode	DVotes	RVotes
Poll ID	factor	numeric value	numeric value
1	blue	100	200
2	orange	150	250

- Subsetting both

	Mode	DVotes	RVotes
Poll ID	factor	numeric value	numeric value
1	blue	100	200
2	orange	150	250

Subsetting by Direct Indexing

- Subsetting observations

	Mode	DVotes	RVotes
Poll ID			

factor numeric numeric
 value value

```
dfex[c(1,3),];
```

selecting rows
not selecting columns

Subsetting by Direct Indexing

- Subsetting variables

Mode	DVotes	RVotes
Poll ID		
factor	numeric value	numeric value

```
dfex[,c(2,3)];  
dfex[,c("DVotes","RVotes")];
```

Subsetting by Direct Indexing

- Subsetting both

Poll ID	Mode	DVotes	RVotes
1	factor	100	100
2	factor	100	100
3	factor	100	100

```
dfex[c(1,3),c(2,3)];  
dfex[c(1,3),c("DVotes", "RVotes")];
```

Subsetting by Values

Poll ID	Mode	DVotes	RVotes
factor	Phone	40	60
	Internet	50	50
	Internet	60	40

- DVotes>55

Poll ID	Mode	DVotes	RVotes
numeric value	Phone	40	60
	Internet	50	50
	Internet	60	40

- Mode=="Internet"

Poll ID	Mode	DVotes	RVotes
factor	Phone	40	60
	Internet	50	50
	Internet	60	40

Logical Operators

- A list of TRUE FALSE indicators

TRUE	FALSE	TRUE
------	-------	------

- Q) How to produce indicators under certain criteria?

Logical Operators

- Select a set of observations with a certain value: ==

```
poll$Affiliation == "Dem";
```

- Select a set of observations different from a certain value: !=

```
poll$Affiliation != "Dem";
```

- Select a set of observations with values larger/smaller than a certain value

```
poll$Clinton > 50;
```

Logical Operators

- AND (&) and OR (|) operations

- TRUE if and only if (TRUE, TRUE)

TRUE	FALSE	TRUE	&	TRUE	FALSE	FALSE	=	TRUE	FALSE	FALSE
------	-------	------	---	------	-------	-------	---	------	-------	-------

- TRUE if either one is TRUE: (TRUE, TRUE), (TRUE, FALSE), (FALSE, TRUE)

TRUE	FALSE	TRUE		TRUE	FALSE	FALSE	=	TRUE	FALSE	TRUE
------	-------	------	--	------	-------	-------	---	------	-------	------

```
poll$Clinton >= 40 & poll$Clinton <= 60
```

```
poll$Affiliation == "Dem" | poll$Clinton > 50
```

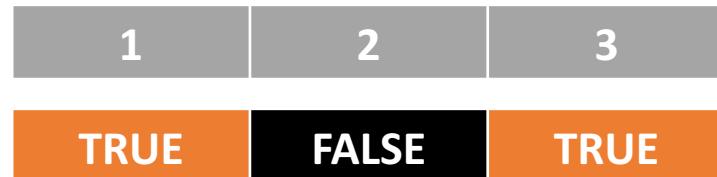
Logical Operators

- By using the TRUE and FALSE indicators, you can choose a subset

e.g. `a=c(1,2,3)`



`a[c(TRUE, FALSE, TRUE)]`



Index vector

e.g. `a=c(1,2,3)`

`a[a>2 | a<2]`

e.g. `a=c(1,2,3)`

`a[a!=2]`

Universal Routine

1. Select a subset with a particular trait
2. Drop/Replace the subset

3 Different Ways of Subsetting Data

0. Direct indexing

```
a=c(1,2,3)
```



```
a[c(1,3)]
```



3 Different Ways of Subsetting Data

1. Logical indicator

```
a=c(1,2,3)
```

1	2	3
---	---	---

```
a[a!=2]
```

TRUE	FALSE	TRUE
------	-------	------

1	3
---	---

```
test1 <- poll[poll$Clinton >= 40 & poll$Clinton <= 60,];
dim(test1);
summary(test1$Clinton);
```

3 Different Ways of Subsetting Data

- which function (fundamentally same as logical operator, but works convenient for high-dimensional objects)

```
a=c(1,2,3)
```

1	2	3
---	---	---

```
a[which(a!=2)]
```

1	3
---	---

```
which(poll$Clinton >= 40 & poll$Clinton <= 60);
test1 <- poll[which(poll$Clinton >= 40 & poll$Clinton <= 60),];
dim(test1);
summary(test1$Clinton);
```

3 Different Ways of Subsetting Data

3. subset function (the easiest; recommended)

```
New_Dataframe <- subset(dataframe, [redacted], select=[redacted]);
```

```
test1 <- subset(poll, Clinton >= 40 & Clinton <= 60, select=c(Pollster, Clinton));
```

Subset observations

Subset variables

```
summary(test1$Clinton);  
dim(test1);  
names(test1);
```

3 Different Ways of Subsetting Data

3. subset function (the easiest; recommended)

```
test1 <- subset(poll,Clinton >= 40 & poll$Clinton <= 60, select=c(Pollster,  
Clinton));  
summary(test1$Clinton);  
dim(test1);  
names(test1);
```

```
test2 <- subset(poll,Clinton >= 40 & poll$Clinton <= 60,names(poll)!="Trump");  
summary(test2$Clinton);  
dim(test2);  
names(test2);
```

Merging Data frames

Observations

ID	DVotes	RVotes

```
dfex1 <- data.frame(ID = c(1,2,3), DVotes = c(40,50,60),
RVotes = c(60,50,40));
dfex1;
```

Observations

ID	mode

```
dfex2 <- data.frame(ID = c(1,3,2), mode =
c("phone", "Internet", "Internet"));
dfex2;
```

Observations

ID	DVotes	RVotes	mode

```
dfex.total <- merge(dfex1,dfex2,by="ID")
dfex.total;
```

Save Data Set

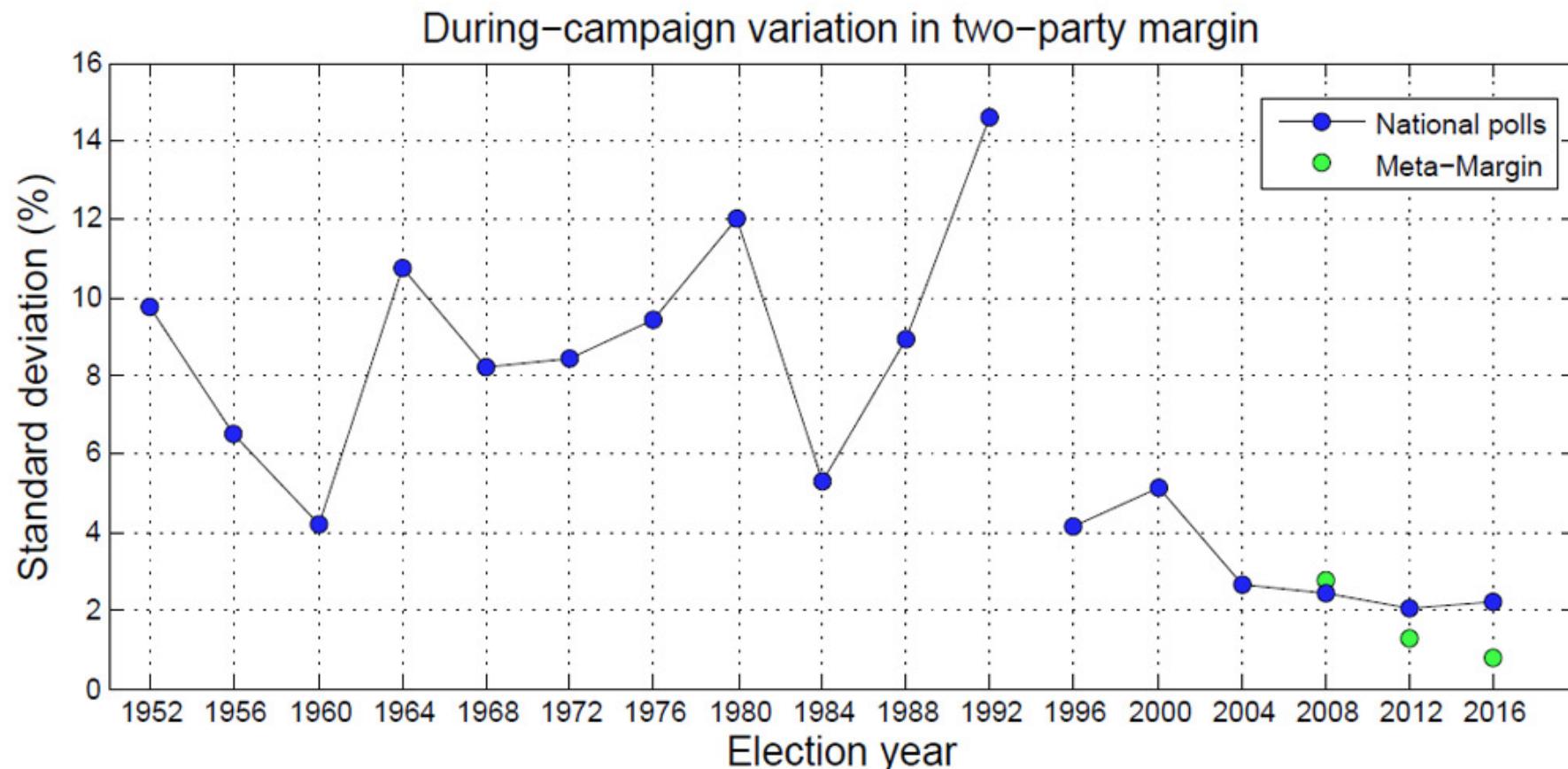
- Save a specific data structure as RData file.

```
save(poll, file="pollonly.Rdata");
dir();
```

- Save everything in the environment as RData file.

```
save.image("everything.RData");
```

Question 1



Partisan sorting and electoral volatility

Question 1

- How has Clinton support rate evolved by respondent party affiliation?

poll\$Population: Respondent type variable

```
summary(poll$Population);
poll_rep <- subset(poll, Population=="Likely Voters - Republican");
poll_dem <- subset(poll, Population=="Likely Voters - Democrat");
par(mfrow=c(1,2));
## 1 by 2 subplots
plot(as.Date(poll_rep$End.Date),poll_rep$Clinton, col = "red");
plot(as.Date(poll_dem$End.Date),poll_dem$Clinton, col = "blue");
## as.Date: Date operator for date variables
```

Create a Variable, Merge into Data Frame

- TrumpWin: Whether Trump won Clinton in each poll

```
TrumpWin <- (poll$Clinton < poll$Trump);
## Create an indicator variable for win/lose status;
poll$TrumpWin <- TrumpWin;
## Add the created variable to poll data frame;
names(poll)[names(poll) == "TrumpWin"] <- "TW"
## Rename variable
```

Question 1.5

- How has Trump support evolved by respondent party affiliation?

```
summary(poll$Population);
poll_rep <- subset(poll, Population=="Likely Voters - Republican");
poll_dem <- subset(poll, Population=="Likely Voters - Democrat");
par(mfrow=c(1,2));
## 1 by 2 subplots
plot(as.Date(poll_rep$End.Date),poll_rep$TW, col = "red");
## as.Date: Date operator for date variables
plot(as.Date(poll_dem$End.Date),poll_dem$TW, col = "blue");
```

Question 2

- Did partisan medias release different results from those by nonpartisan?

```
summary(poll$Affiliation);
summary(poll$Population);
poll.rep <- subset(poll, Affiliation=="Rep" & Population=="Likely Voters");
poll.dem <- subset(poll, Affiliation=="Dem" & Population=="Likely Voters");
poll.none <- subset(poll, Affiliation!="Rep" & Affiliation!="Dem" &
Population=="Likely Voters");
par(mfrow=c(1,3));
## 1 by 3 subplots
plot(as.Date(poll.rep$End.Date),poll.rep$Trump, col = "red");
plot(as.Date(poll.dem$End.Date),poll.dem$Trump, col = "blue");
plot(as.Date(poll.none$End.Date),poll.none$Trump, col = "green");
```

Question 3

What is the trend after the first debate?



Question 3

What is the trend after the first debate?

Poll	Trump	Clinton	Other	Undecided	Spread
Monmouth University					
Sep 22 – Sep 25	46	49	4	2	Clinton +3
729 Likely Voters					
Quinnipiac					
Sep 22 – Sep 25	46	47	1	6	Clinton +1
1,115 Likely Voters					
NBC/SurveyMonkey <small>NEW!</small>					
Sep 19 – Sep 25	44	51	-	6	Clinton +7
13,598 Likely Voters					

Not updated yet!

Question 3: Take Home

What is the trend after the first debate?

You can see what happened by following the exactly same routine we did today by downloading the newly updated dataset over the weekend !

More Interested Participants can check



<https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf>

More Interested Participants can go to

[HuffPost Pollster - Polls and Charts - Election Results](#)
[elections.huffingtonpost.com/pollster](#) ▾ 이 페이지 번역하기

HuffPost Pollster tracks thousands of public polls to give you the latest data on elections, political opinions and more. Read our FAQ. Search all poll charts.

[2016 General Election: Trump ...](#)

Polls and chart for 2016 General
Election: Trump vs. Clinton. See ...

[polls](#)

Polls, charts, forecasts and data about
upcoming elections ...

[2016 National Republican ...](#)

Polls and chart for 2016 National
Republican Primary. See the ...

[2016 General Election: Trump ...](#)

Polls and chart for 2016 General
Election: Trump vs. Clinton vs ...

[2016 National Democratic ...](#)

Polls and chart for 2016 National
Democratic Primary. See the ...

[Obama Health Care Law ...](#)

Polls and chart for Obama Health Care
Law: Favor/Oppose. See ...

Poll records during the primary season

Detailed Poll Results by Demographics

YouGov | Economist/YouGov Poll

<https://today.yougov.com/news/categories/economist/> ▾ YouGov ▾

This is a summary of a YouGov/Economist Poll conducted September 22-24, 2016. The sample is 1300 general population respondents with a Margin of Error .

Not provided in CSV format



Feed Back Survey

<https://goo.gl/forms/GbNI7fhCx70IRVpi1>

Thank you

FALL 2016 SCHEDULE

September 20	Introduction to R and RStudio	Slides	Data	Code
September 27	Data Wrangling in R			
October 4	Base R Graphics			
October 11	Data Visualization in R with ggplot2			
October 18	Programming Loops in R			
November 8	Probability and Simulations in R			
November 15	Monte Carlo Simulations in R			
November 29	Text Analysis in R			
December 6	Hypothesis Testing in R			
December 13	Regression Analysis in R			