Data Wrangling in R

Yunkyu Sohn
September 28, 2017

Research Associate, Department of Politics

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

https://compass-workshops.github.io/info/

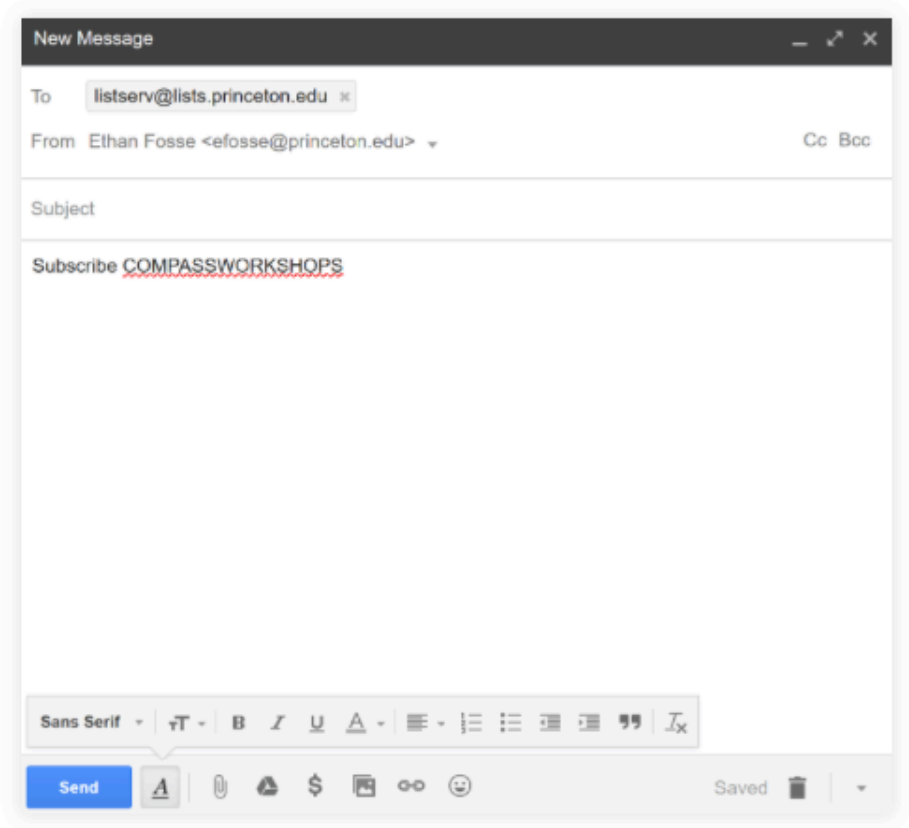
Spring 2017 Schedule

Date	Topic
February 15	Introduction to R and RStudio
February 22	Data Wrangling in R
TBD	Base R Graphics
TBD	Data Visualization in R with ggplot2
TBD	Programming Loops in R
TBD	Probability and Simulations in R
TBD	Monte Carlo Simulations in R
TBD	Text Analysis in R
TBD	Hypothesis Testing in R
TBD	Regression Analysis in R
TBD	Social Network Analysis in R

Connect with Us:

- Visit our website
- Join our mailing list

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)
  - Marc Ratkovic (Assistant 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

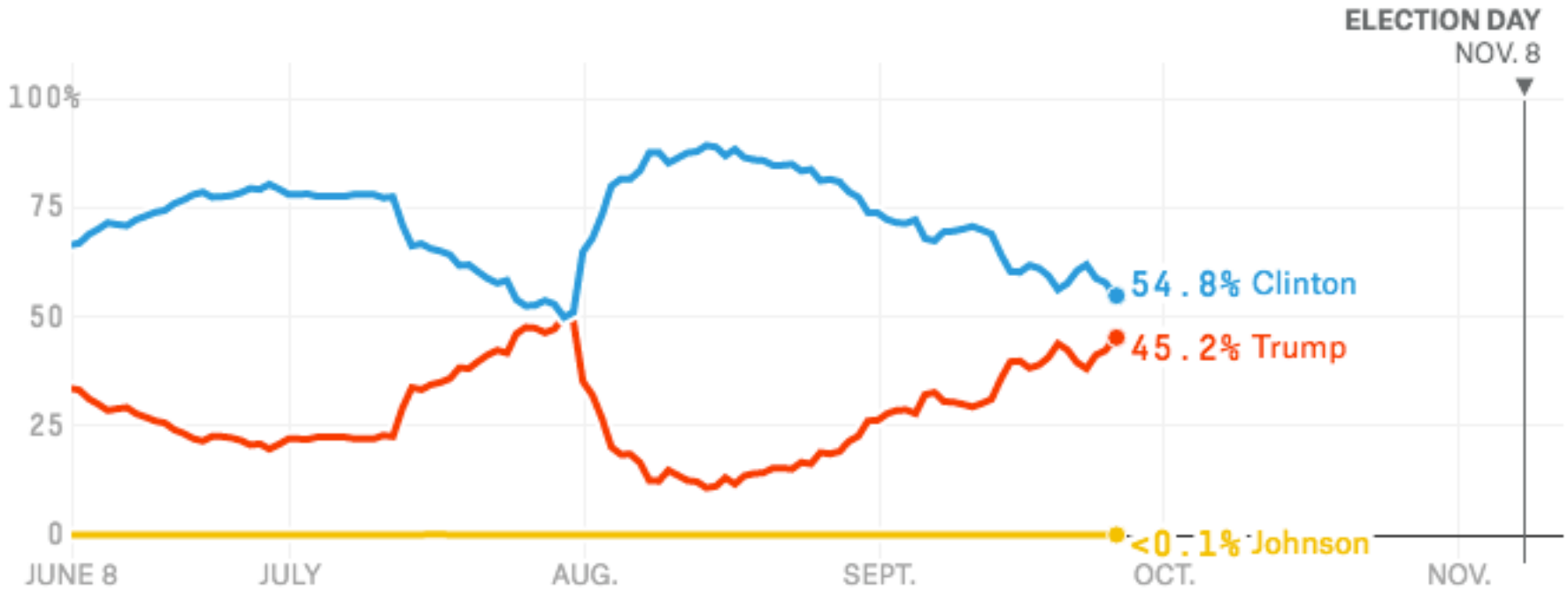
Before You Begin

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

That's all!

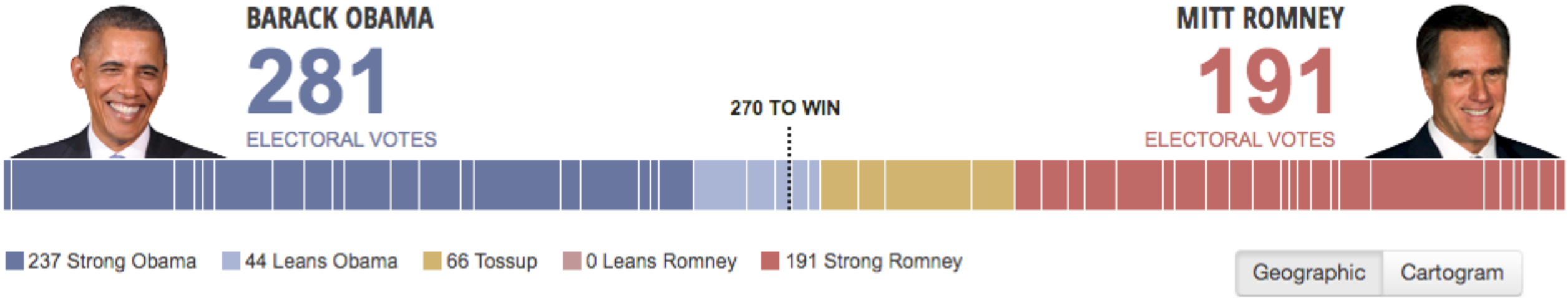
Today’s Project

Election Prediction = Data + Statistics



FiveThirtyEight

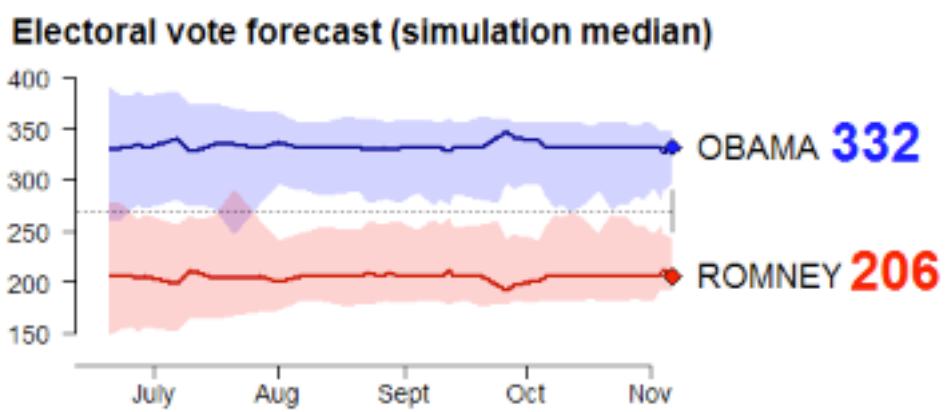
Election Prediction = Data + Statistics



HUFFPOST POLLSTER

Get the latest polls and public opinion updates in your inbox daily

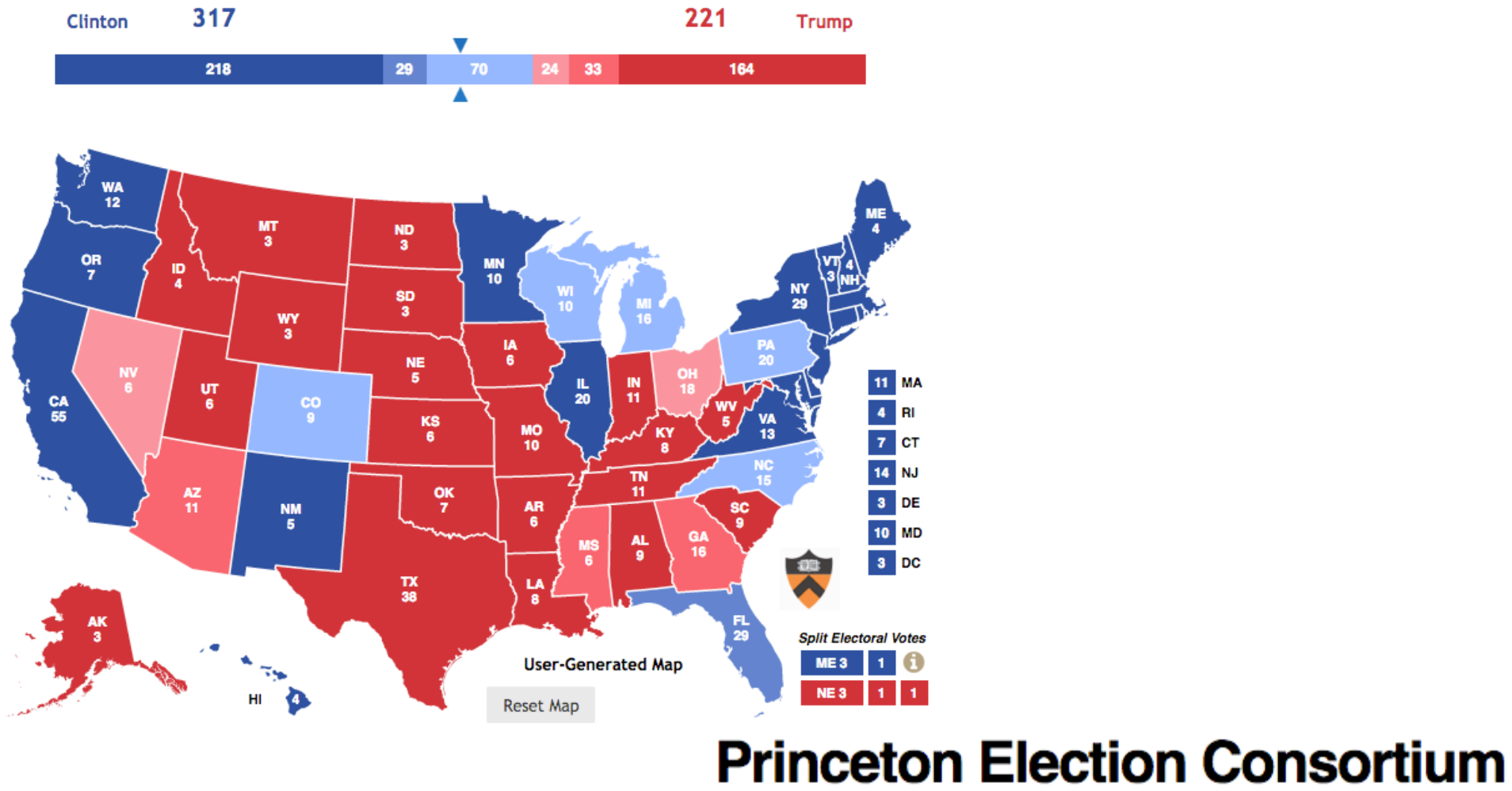
Election Prediction = Data + Statistics



VOTAMATIC

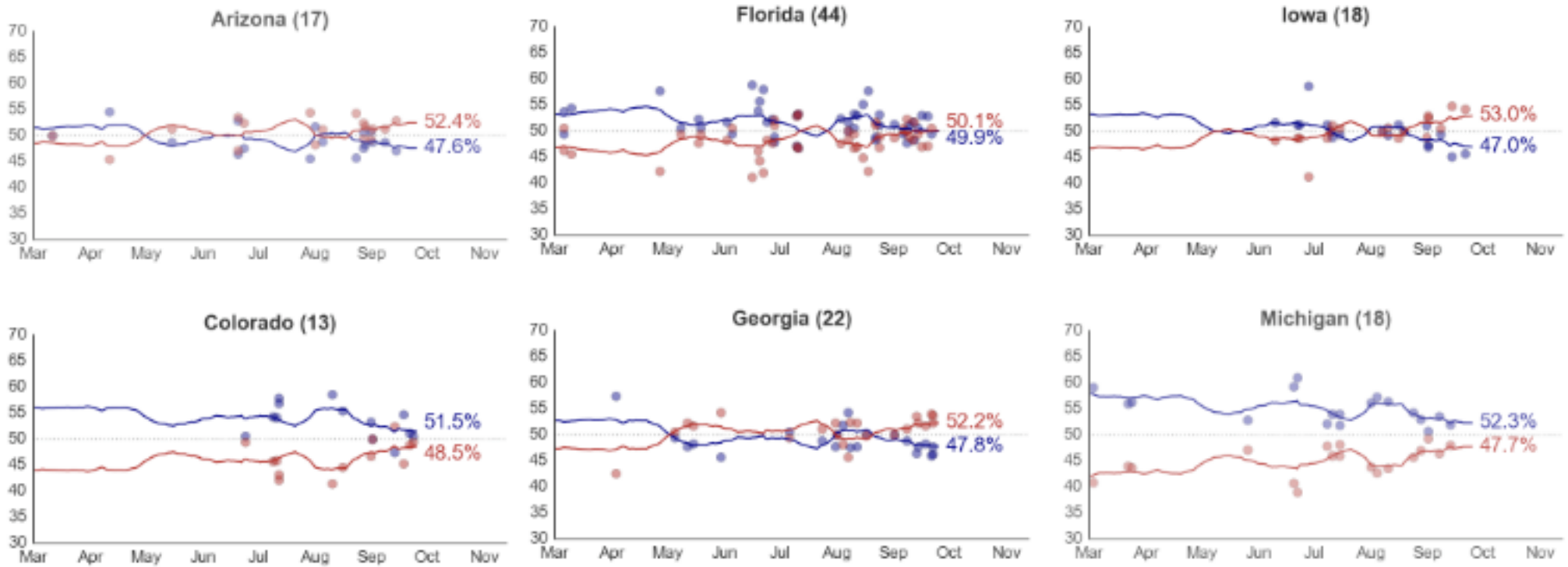
Polling Analysis and Election Forecasting

Election Prediction = Data + Statistics

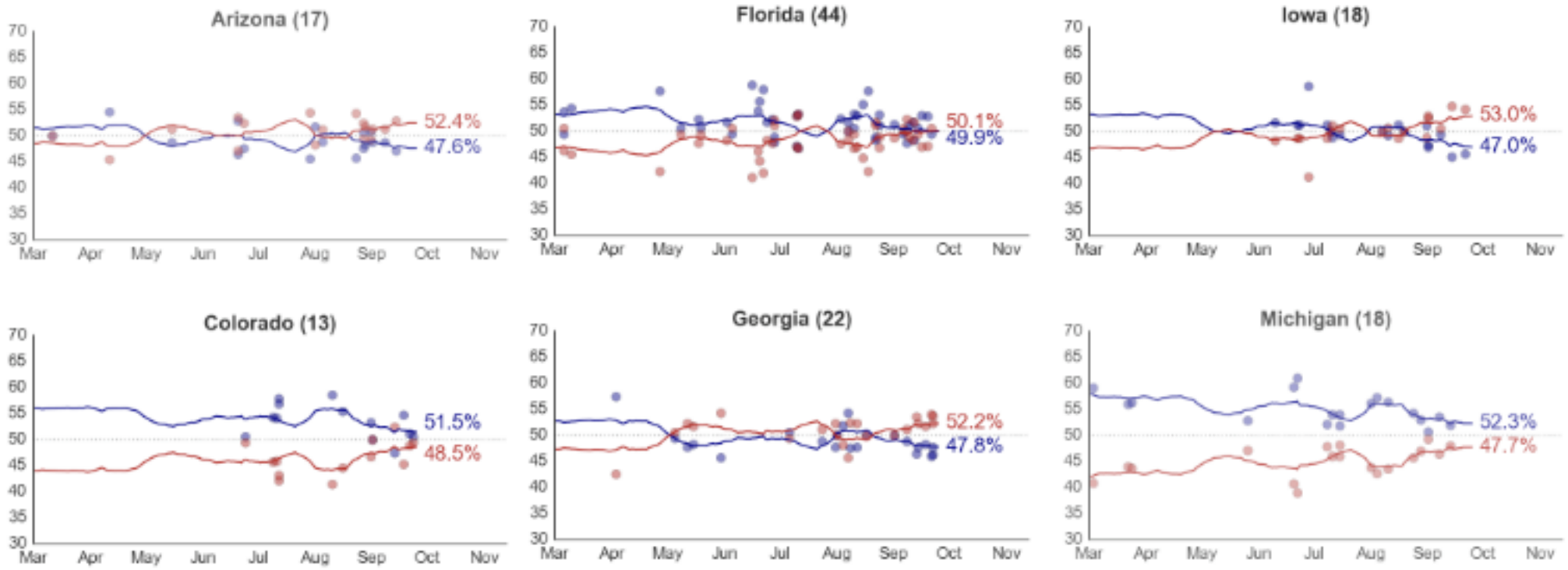


Princeton Election Consortium

Generic Approach to Election Prediction

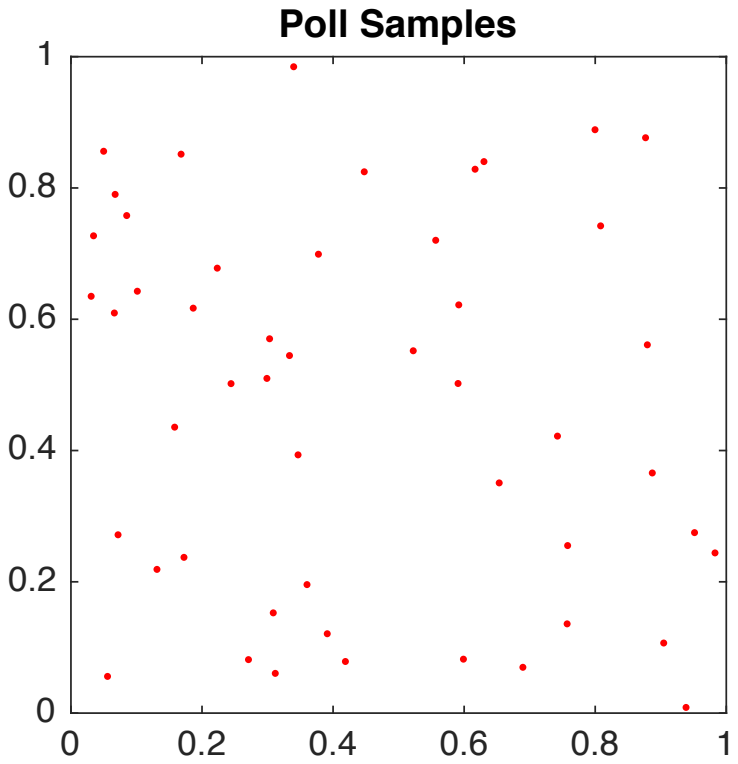


Generic Approach to Election Prediction

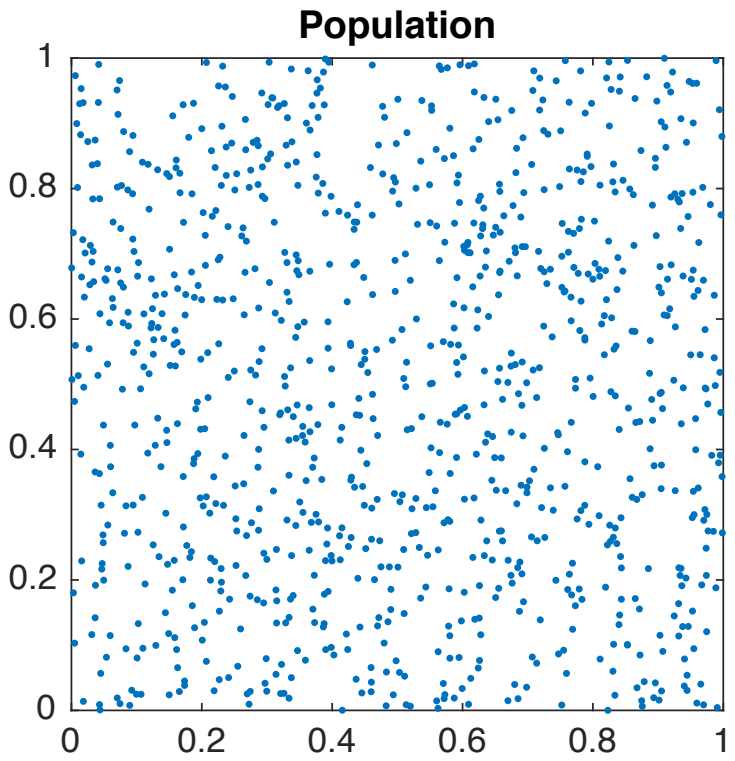


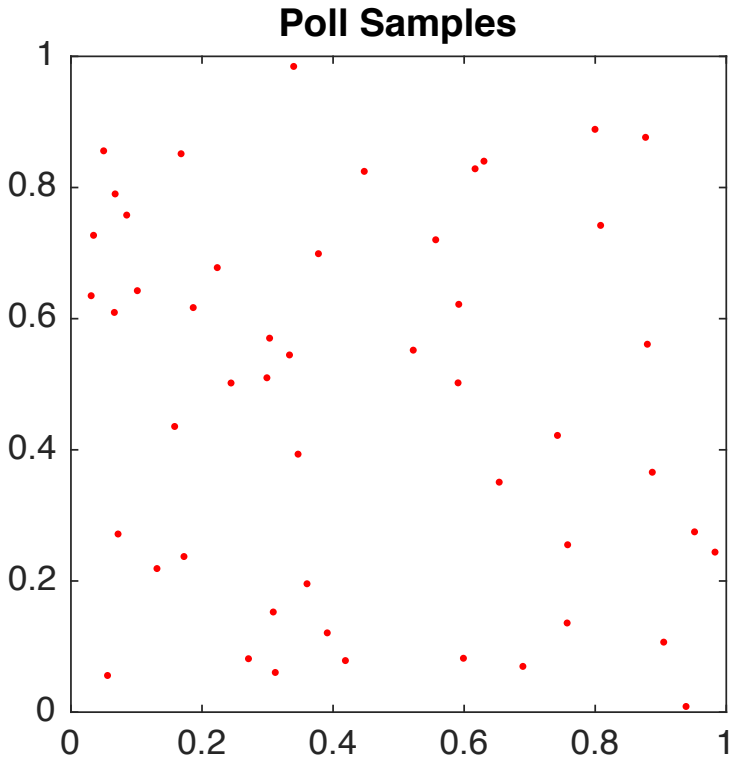
Statistical **aggregation** of state **poll** results over time

Election Prediction = Data + Statistics



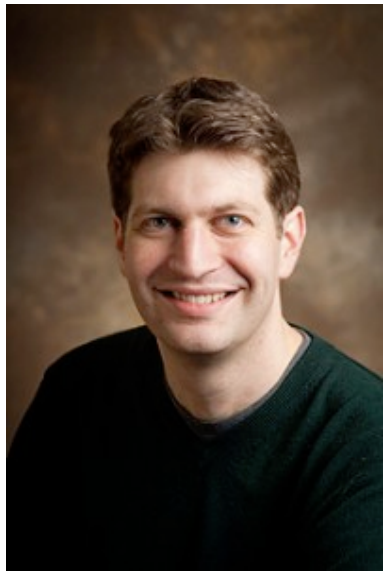
Election Prediction = Data + Statistics





Using Statistics to infer
latent population preference

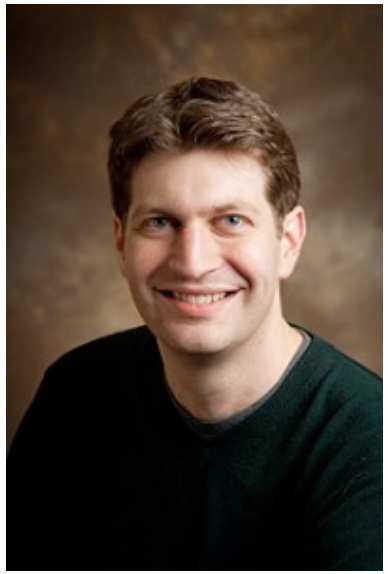
Generic Approach to Election Prediction

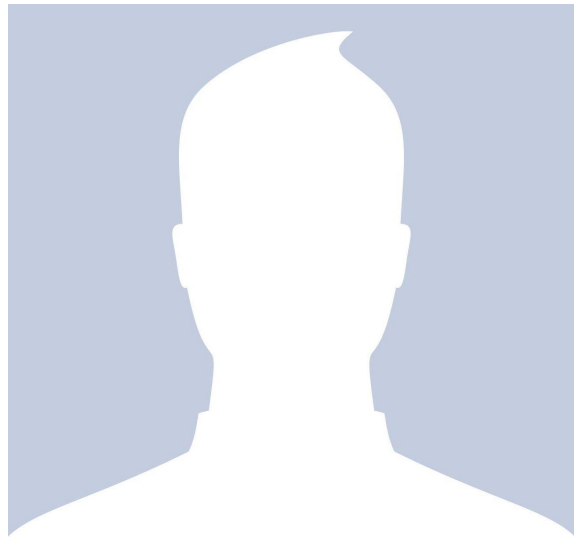



Statistical **aggregation** of state **poll** results over time

After 60 minutes, You will be able to

After 60 minutes, You will be able to





Yourself

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

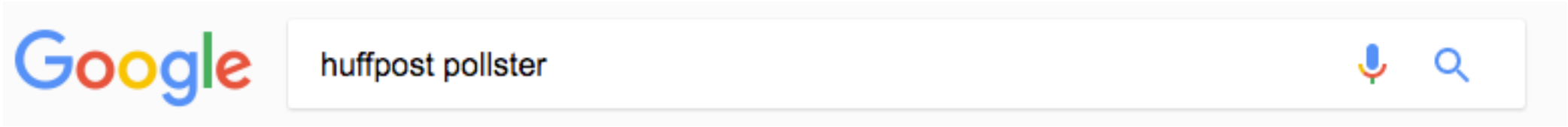
- Data acquisition from the Internet and import to R
- Undersanding R data structures
- Subsetting Observations
- Subsetting Variables
- Subsetting Both Observations and Variables
- Summarizing Data
- Creating and Renaming Variables
- Merging Data Sets

Dataset for Today

- https://compass-workshops.github.io/info/ : Week 2 Data
- Dataset downloaded on September 26th, 2016 from HuffPost Pollster

Not for today

You can Download by Yourself



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 ...

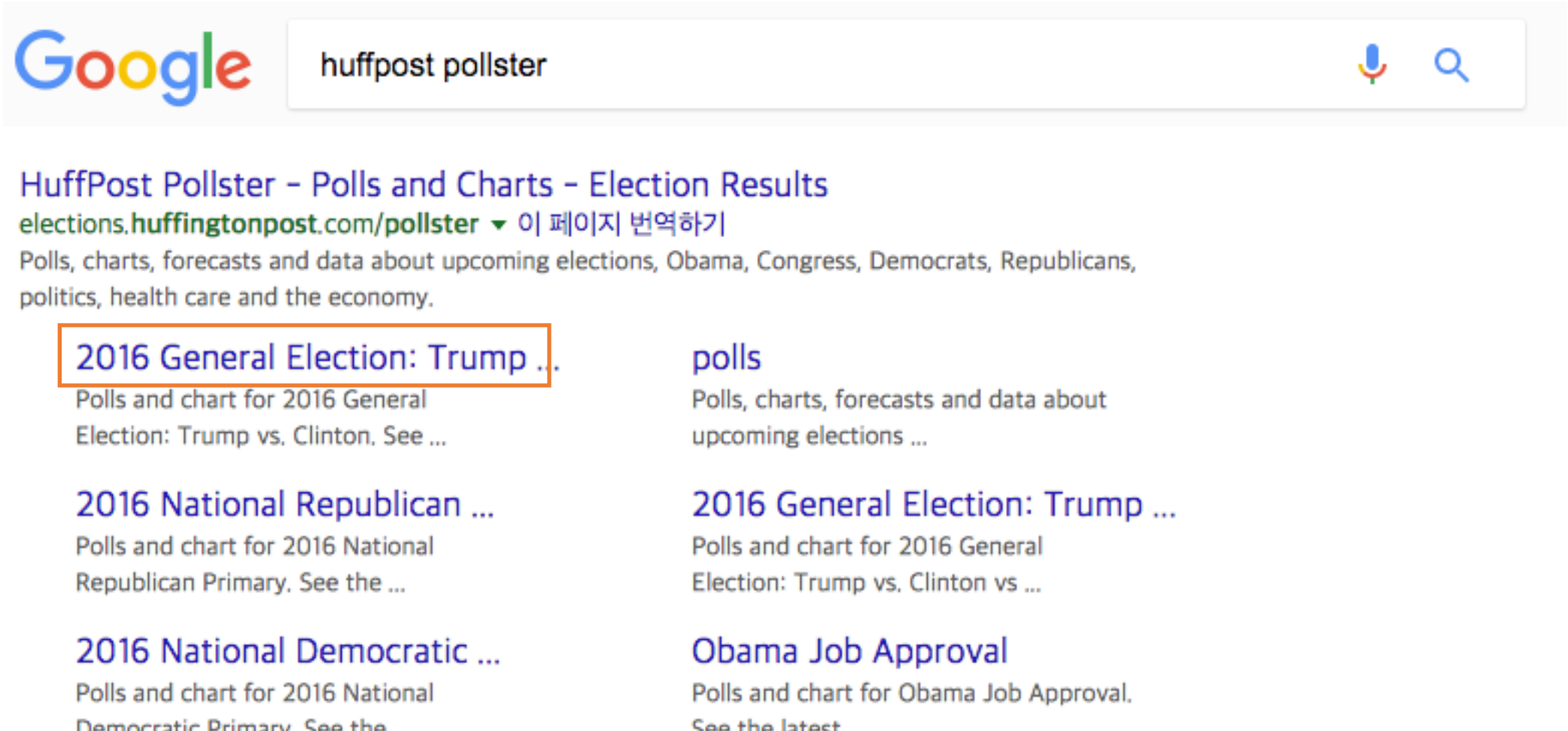
2016 General Election: Trump ...
Polls and chart for 2016 General Election: Trump vs. Clinton vs ...

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 ...

Not for today

You can Download by Yourself



Not for today

You can Download by Yourself

McClatchy/Marist Sep 15 – Sep 20 758 Likely Voters	41	**48**	9	2	Clinton +7

SHOW MORE ▼

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 CSV (Comma Separated Values) File into an R object.

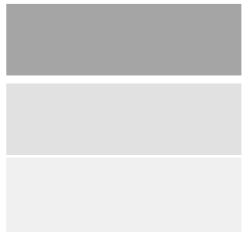
Other datatypes can be imported similarly: e.g. read.dta, read.spss

readr package for large datasets

```
rm(list=ls())
## Delete your workspace
getwd()
## Check your current working directory
setwd("<location of your dataset>")
## Set your working directory
poll <- read.csv("09262016.csv")
## Load data
```

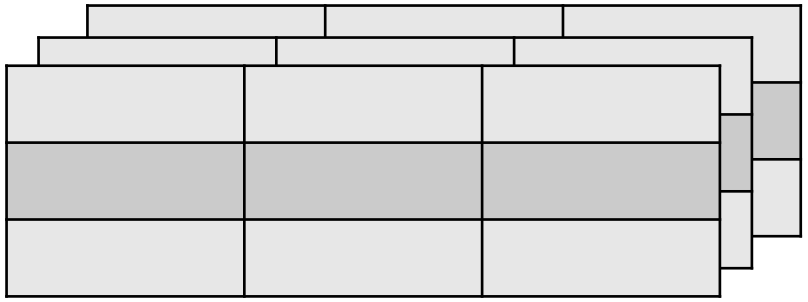
R Data Structures

Vector



Matrix

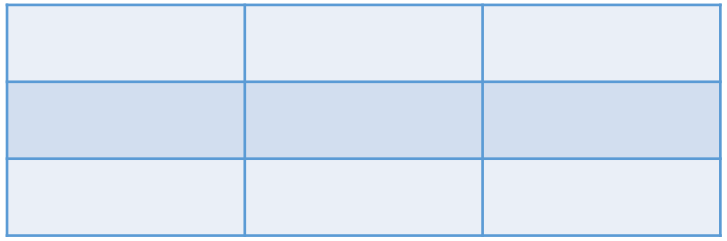
Array



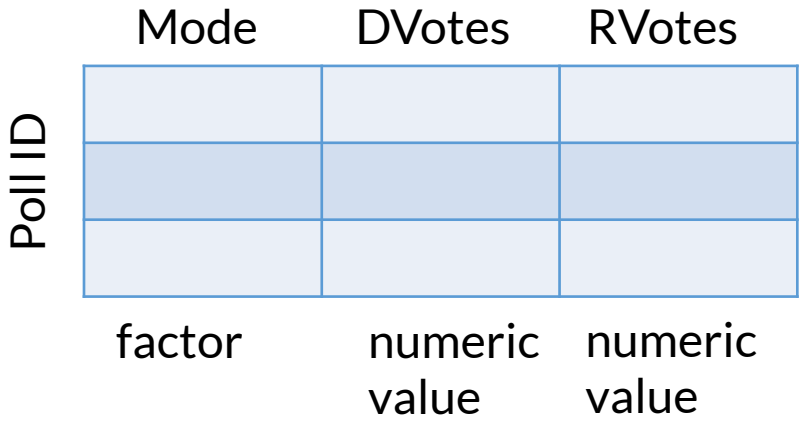
Data frame

Variables with heterogeneous elements

Observations

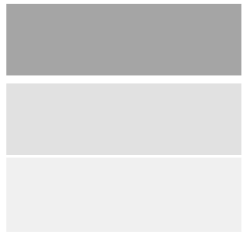


e.g. Data frame for a poll dataset



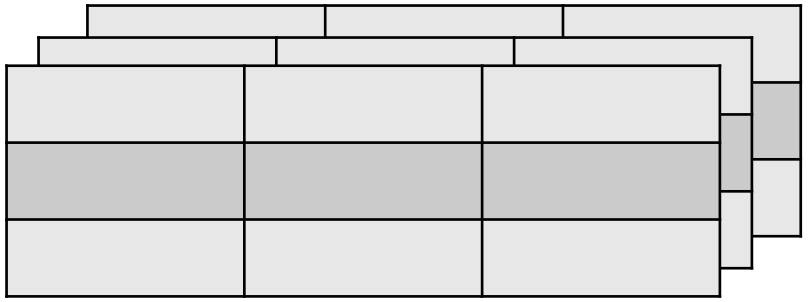
R Data Structures

Vector

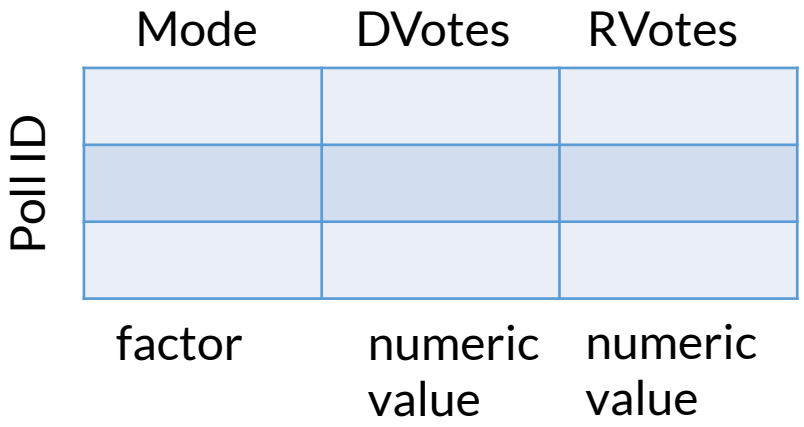


Matrix

Array



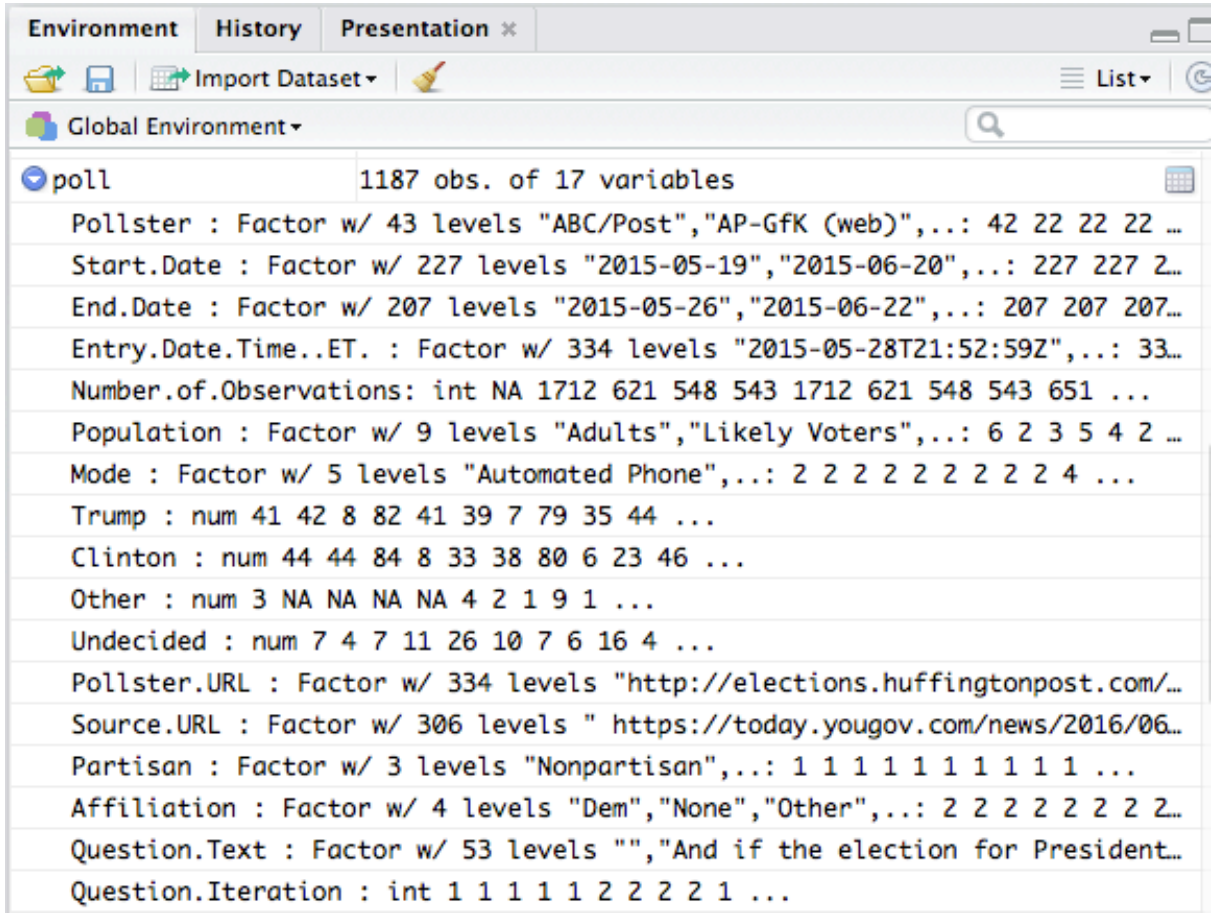
Data frame

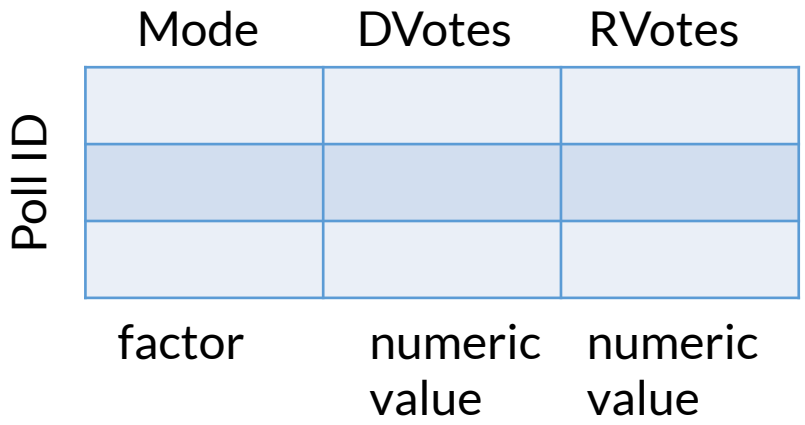


Create a data frame for **toy example**

```
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





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 $

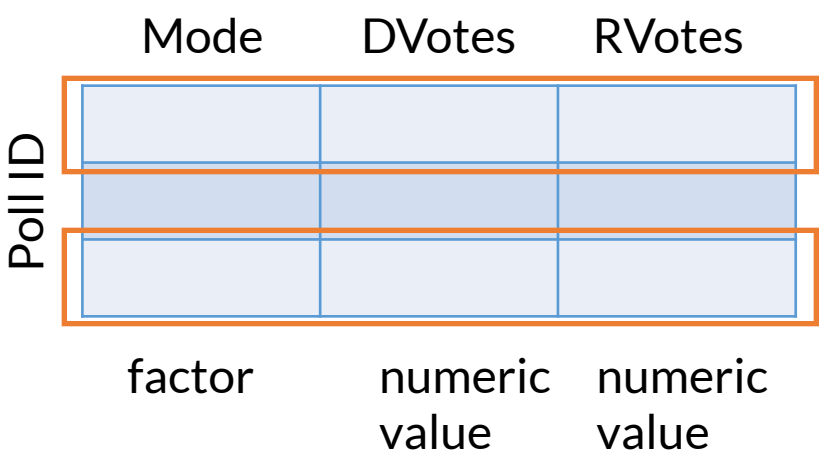
```
poll$VariableName
```

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

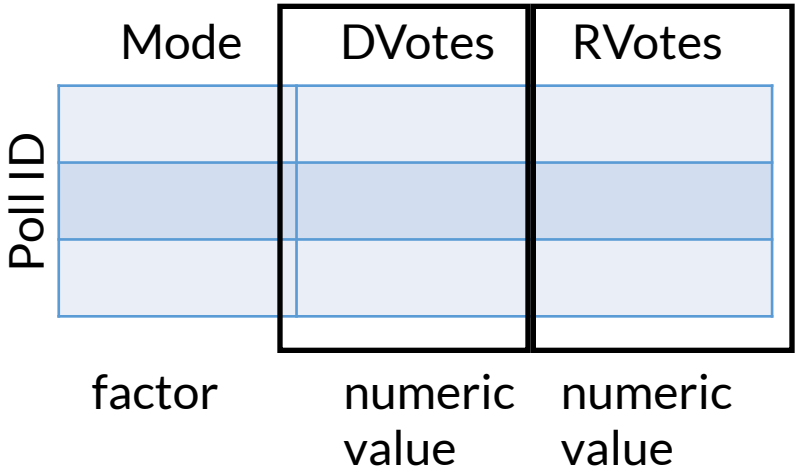
```
poll$Affiliation
```

Subsetting by Direct Indexing

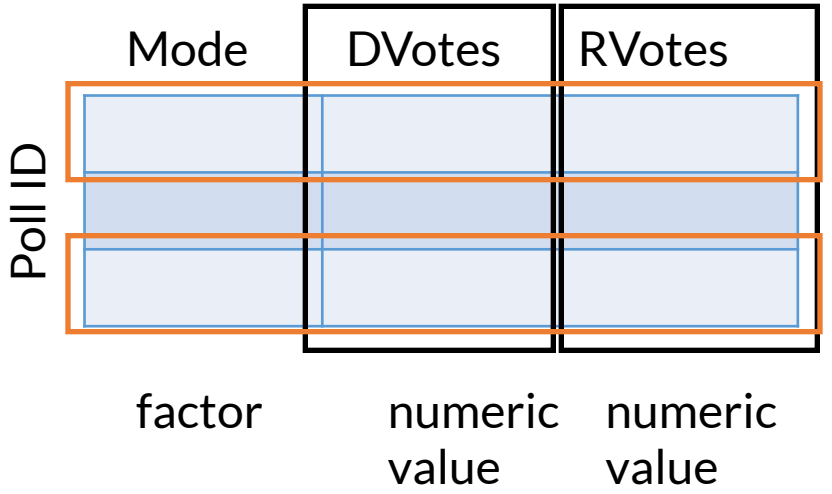
- Subsetting observations



- Subsetting variables

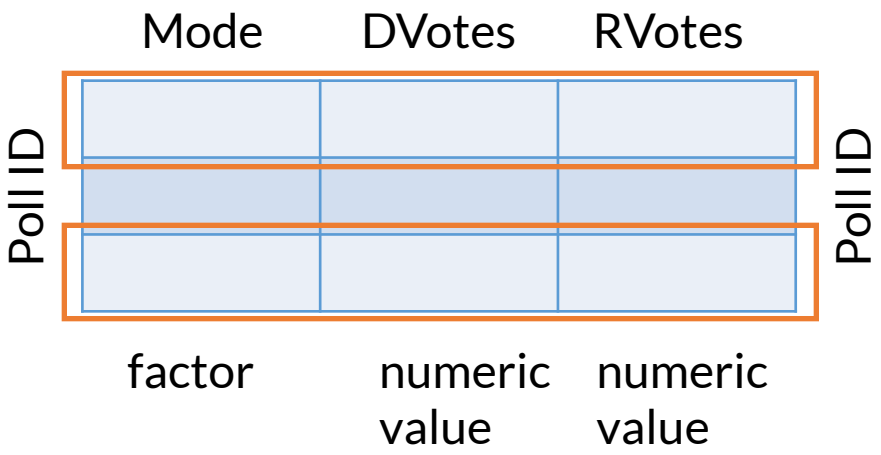


- Subsetting both



Subsetting by Direct Indexing

- Subsetting observations



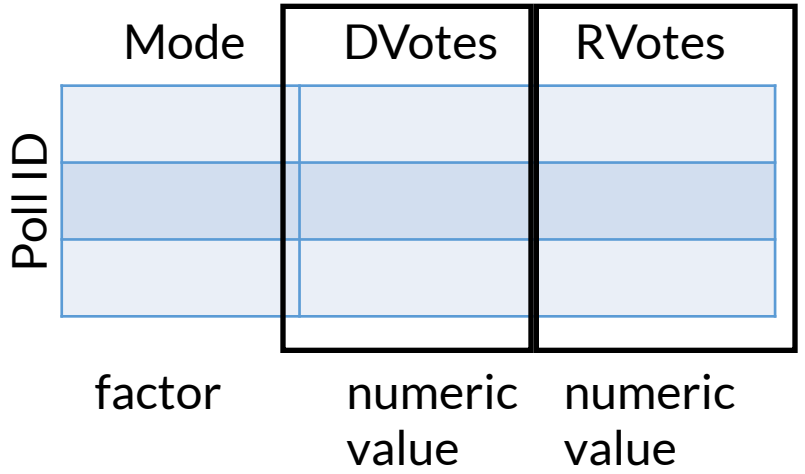

```
dfex[c(1,3),]
```

selecting rows

not selecting columns

Subsetting by Direct Indexing

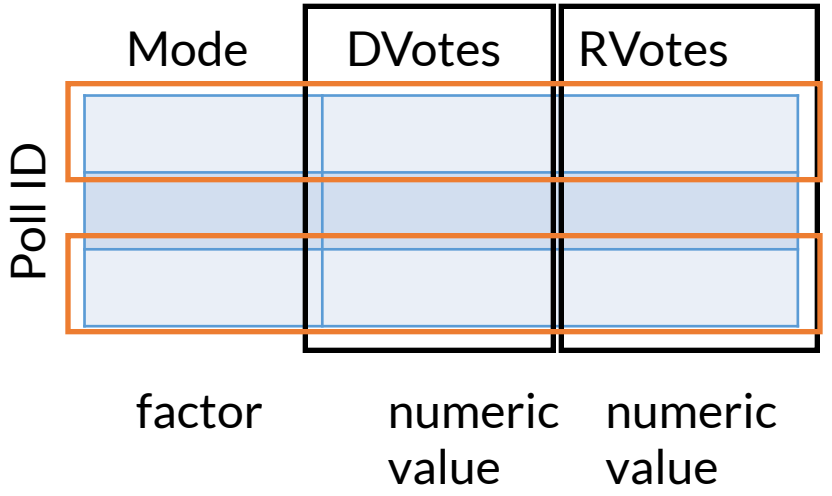
- Subsetting variables




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

Subsetting by Direct Indexing

- Subsetting both




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

Subsetting by Values

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

- **DVotes>55**

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

- **Mode=="Internet"**

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

Logical Operators

- A list of TRUE FALSE indicators

- 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 (I) operations
  - TRUE if and only if (TRUE, TRUE)

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


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

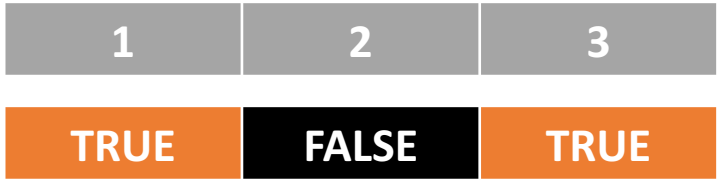
```
poll$Affiliation == "Dem" | poll$Clinton > 50
```

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)`

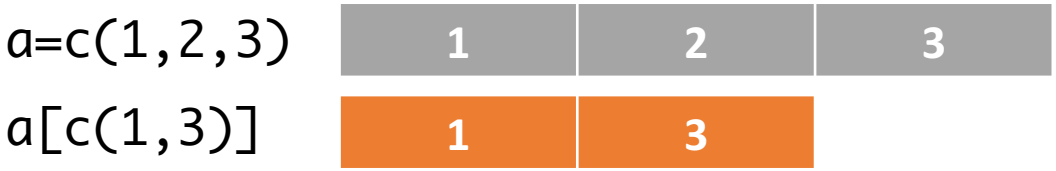
`a[a!=2]`

Universal Routine

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

3 Different Ways of Subsetting Data

0. Direct indexing



3 Different Ways of Subsetting Data

1. Logical indicator


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

3 Different Ways of Subsetting Data

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

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


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


```
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

```
New_Dataframe <- subset(dataframe,                    , select=                    )
```

```
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

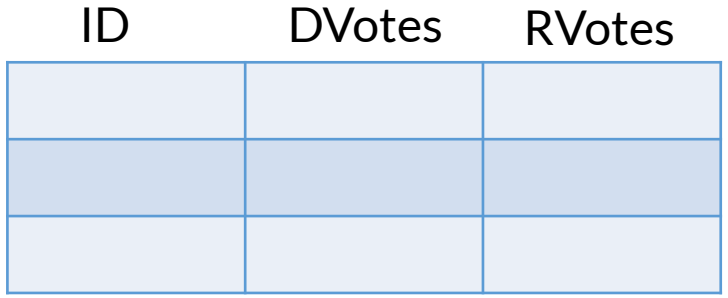
3. subset function

```
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)
```

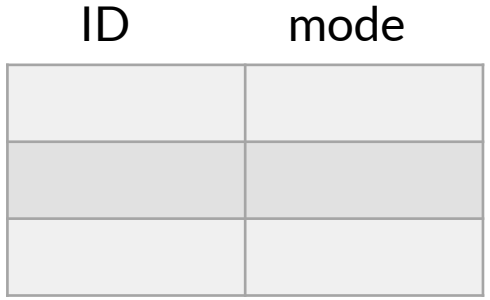
Merging Data frames

Observations



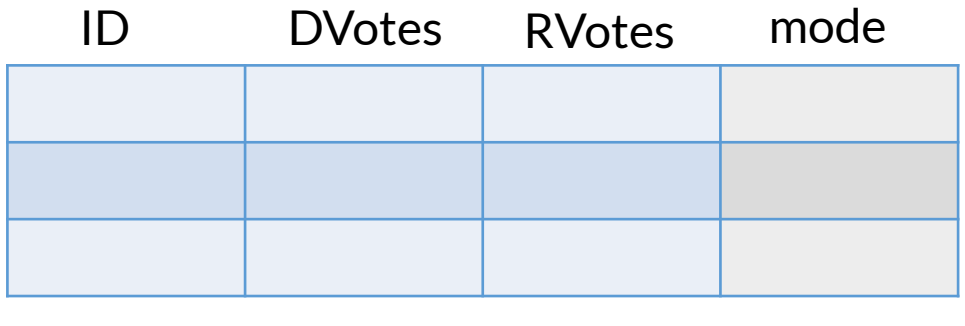

```
dfex1 <- data.frame(ID = c(1,2,3), DVotes = c(40,50,60),
RVotes = c(60,50,40))
dfex1
```

Observations




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

Observations




```
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()
```

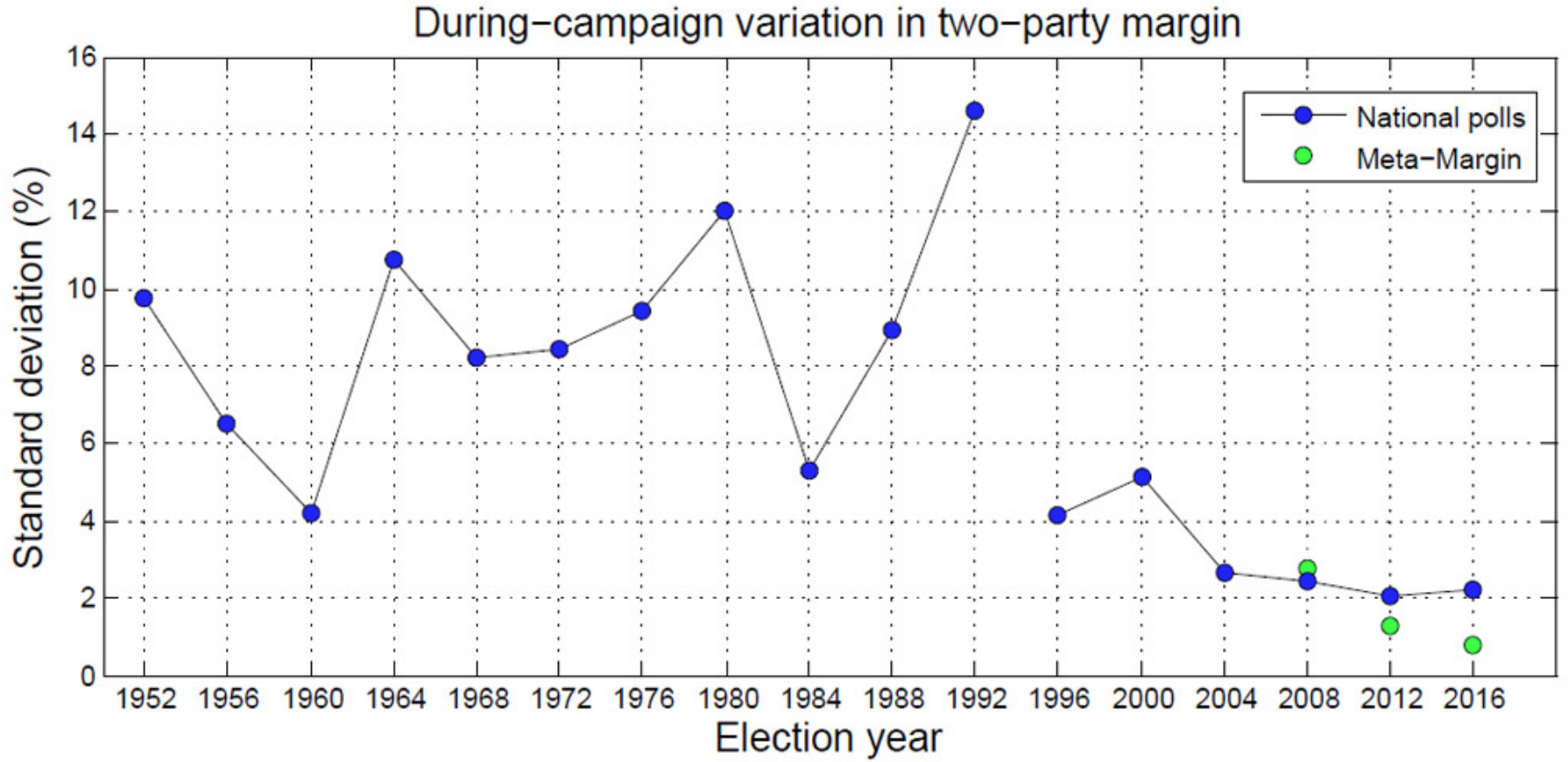
- Save everything in the environment as RData file.

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

- wite.csv, write.dta for exporting to other formats

Now Ready to Answer the Questions!

Question 1



Partisan sorting and electoral volatility

Princeton Election Consortium

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 was the trend afterwards (e.g. the first debate)?

Question 3: Take Home

What was the trend afterwards (e.g. the first debate)?

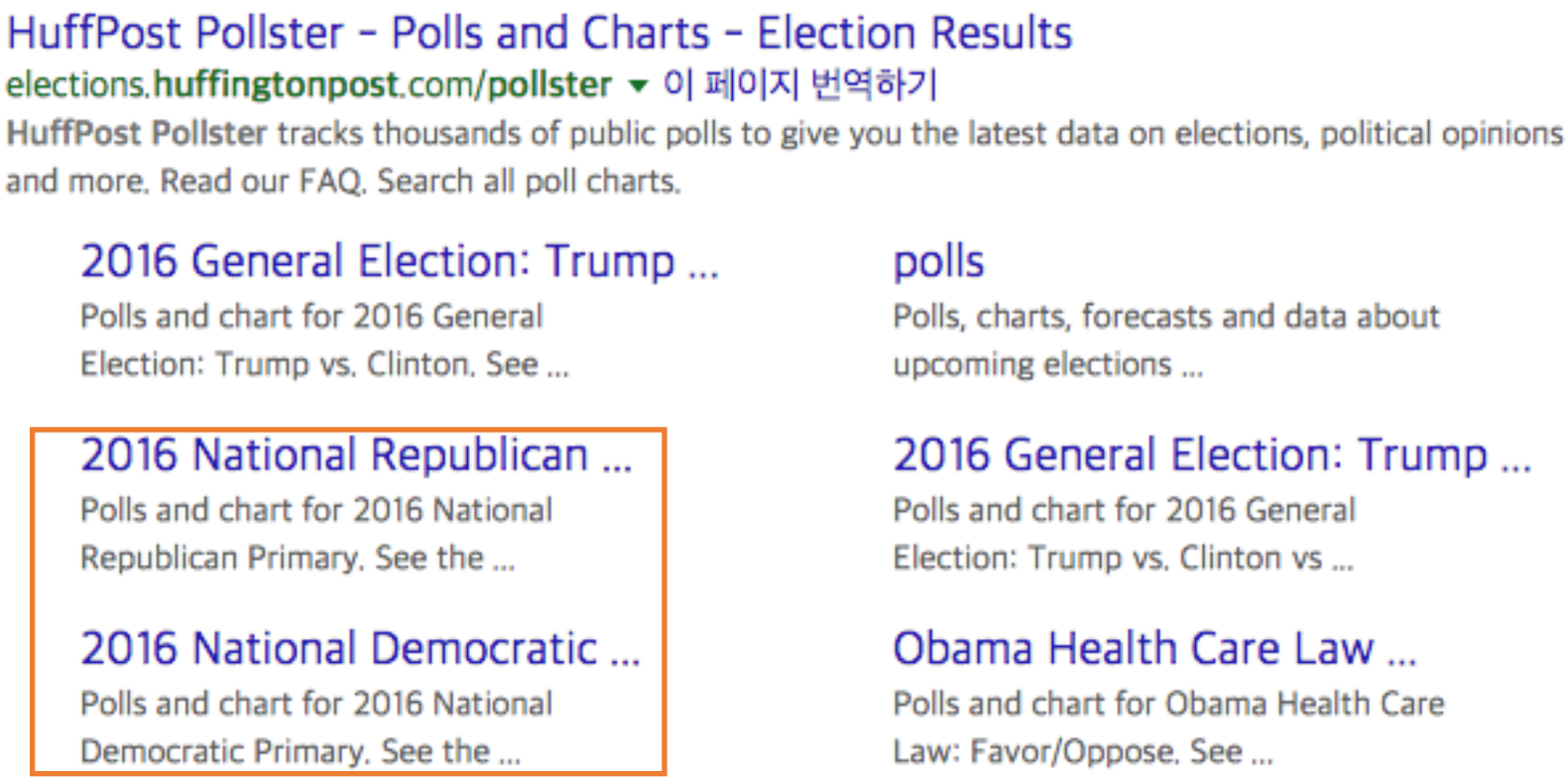
You can see what happened by following the exactly same routine we did today by downloading the complete poll dataset!

1. Right click on CSV
2. Save as to a preferred location

More Interested Participants can check

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

More Interested Participants can go to



Poll records during the primary season

Detailed Poll Results by Demographics

Not provided in CSV format

YouGov US

Feed Back Survey

https://docs.google.com/forms/d/e/1FAIpQLSfyuPoNw7tM_DaJ57sy7NN2tQ52WiF_pr9eZtV0rTo5zU9xvA/viewform?c=0&w=1

2~3 minute survey; We would appreciate if all of you can participate!

Thank you

Date	Topic
February 15	Introduction to R and RStudio
February 22	Data Wrangling in R
March 1	Base R Graphics
TBD	Data Visualization in R with ggplot2
TBD	Programming Loops in R
TBD	Probability and Simulations in R
TBD	Monte Carlo Simulations in R
TBD	Text Analysis in R
TBD	Hypothesis Testing in R
TBD	Regression Analysis in R
TBD	Social Network Analysis in R