

Base R Graphics

Yunkyu Sohn
October 4, 2016



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



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.

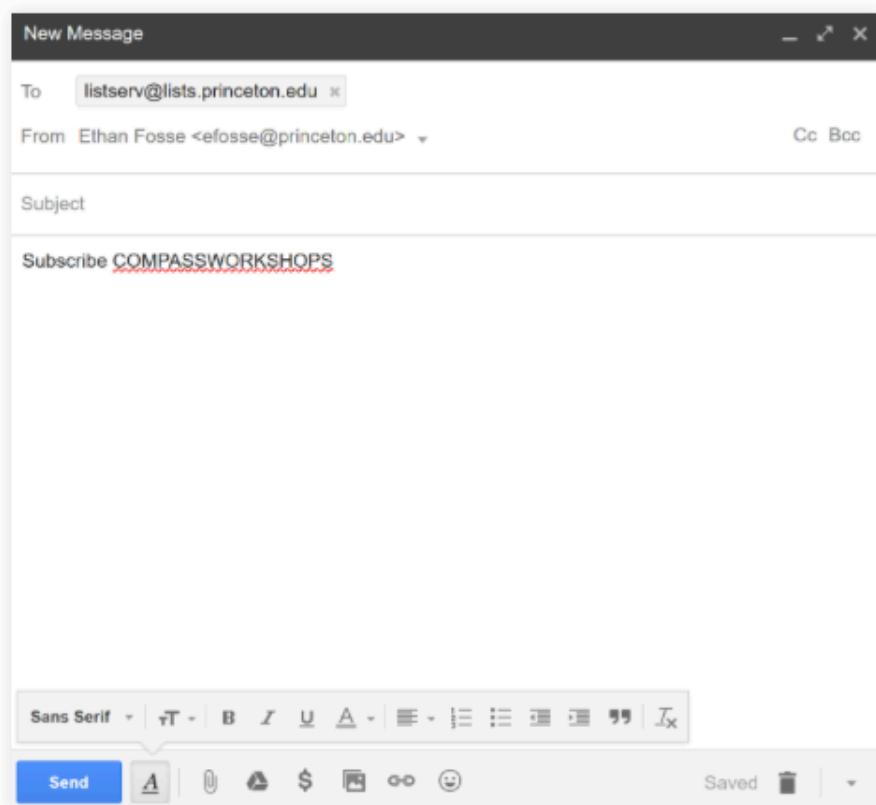
To find out more about the workshops, join the COMPASS Workshops mailing list by sending an email to listserv@lists.princeton.edu with "Subscribe COMPASSWORKSHOPS" in the body and all other lines blank including the subject.



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 3 (right click -> save as) at
<https://compass-workshops.github.io/info/>
4. Start Rstudio

That's all!

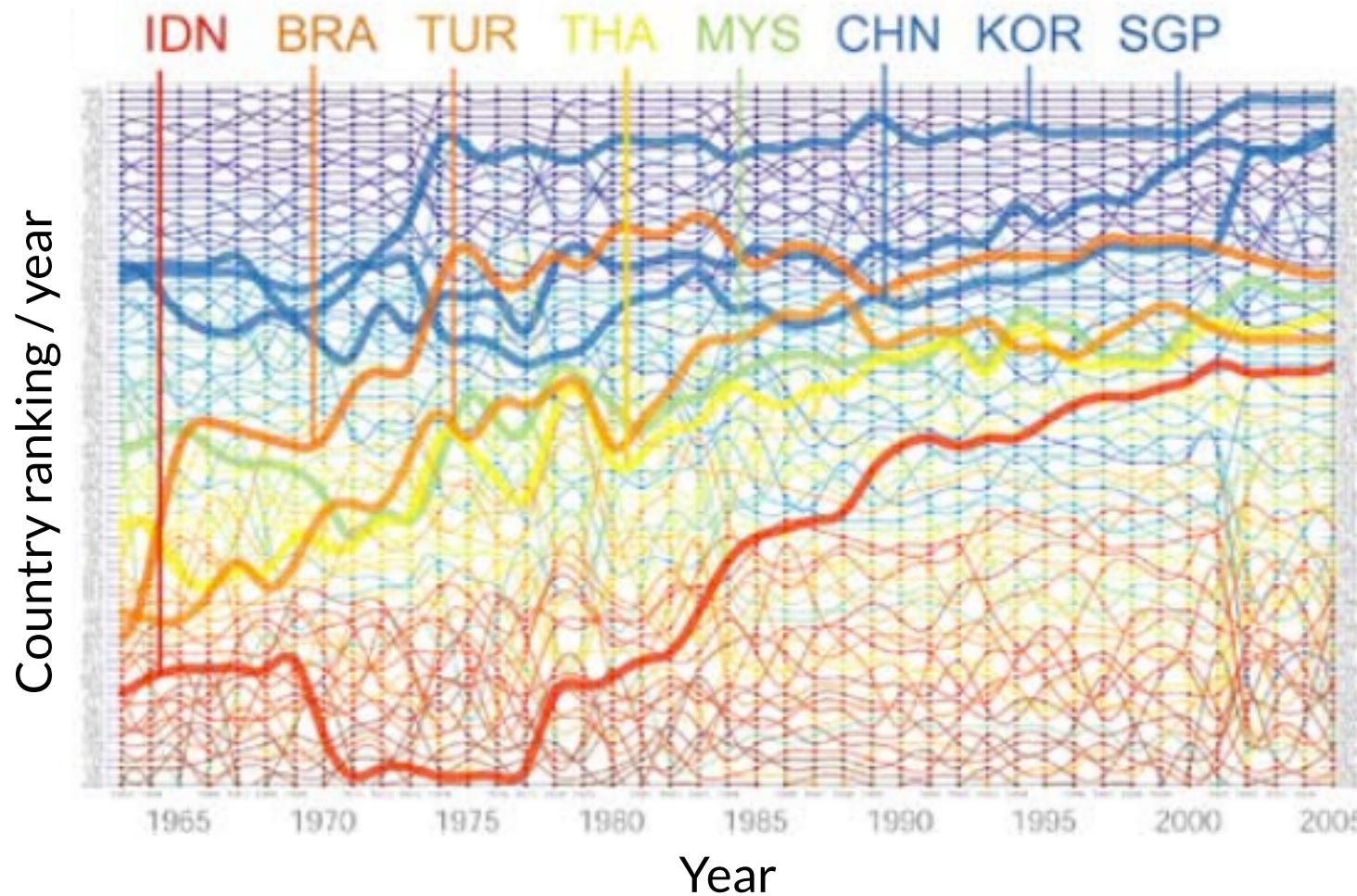
Motivation

A dataset is just a collection of numbers and strings (very complex; not understandable in its naïve format). In order to understand the systemic patterns behind a dataset, we use statistics and graphs (simpler; understandable).

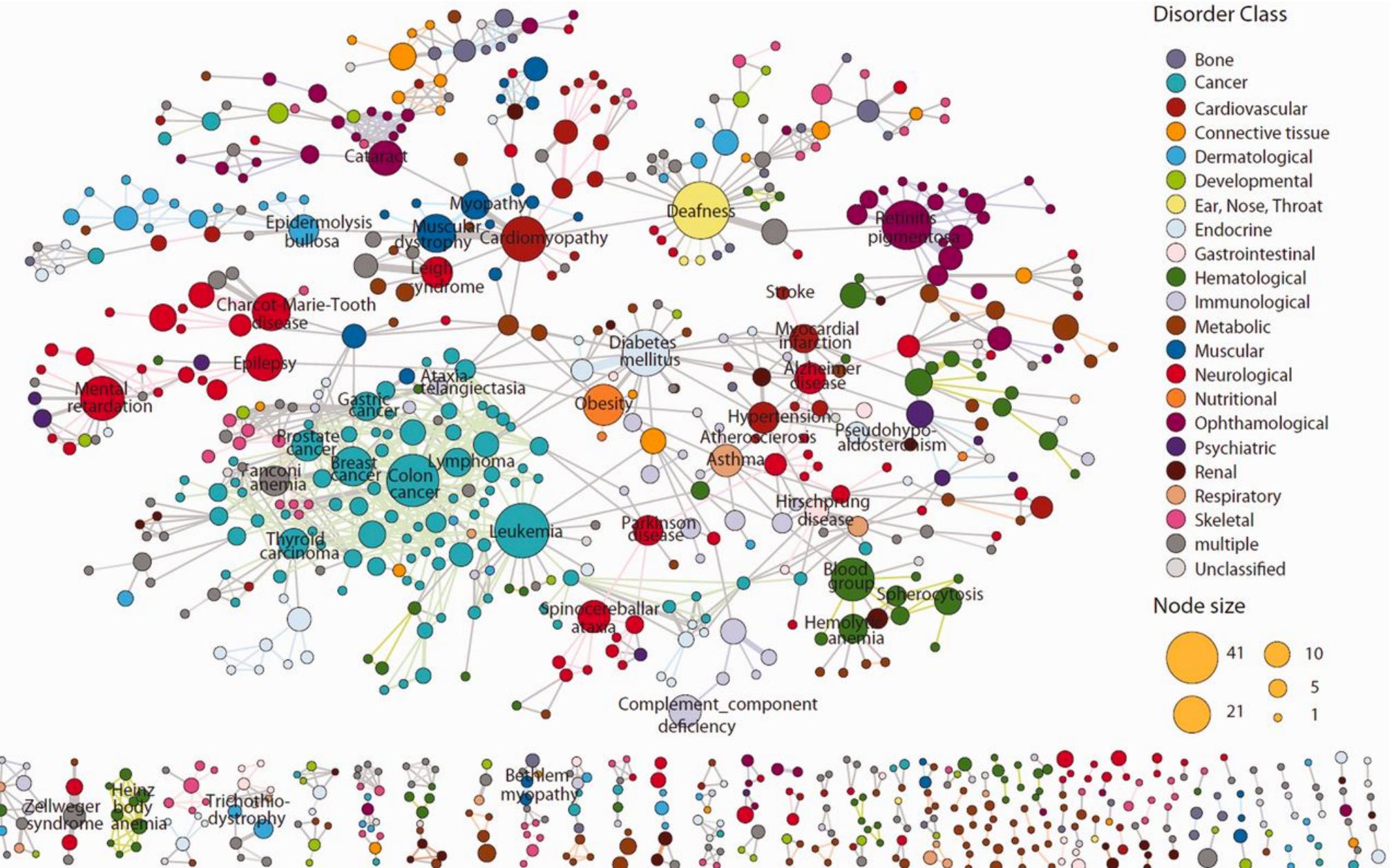
Power of Graphs

Often, graphs convey more information with less complexity in a very effective way.

Evolution (Ranking) of Industrial Complexity



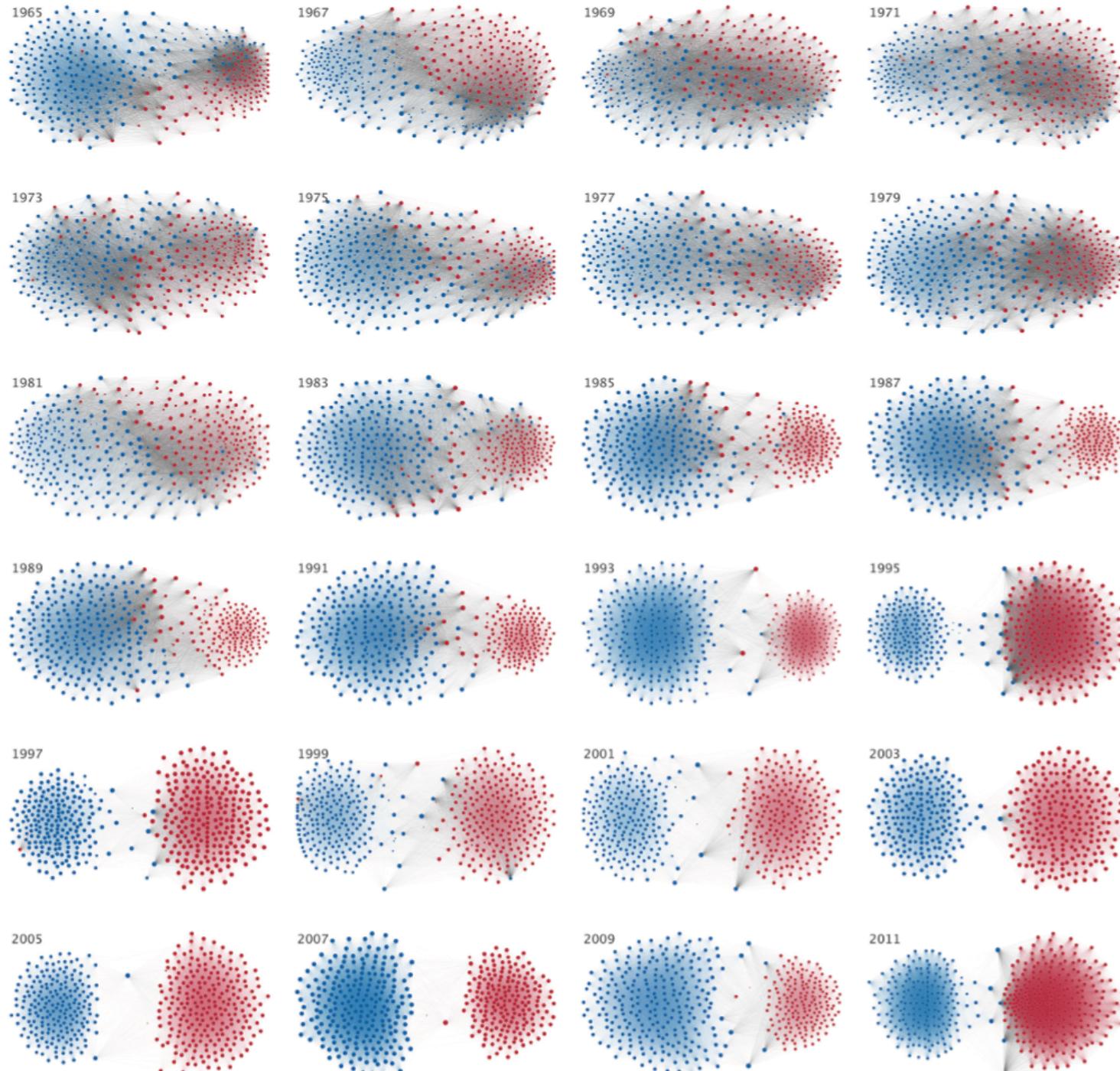
Hidalgo and Hausmann (2009)



Two disease nodes (circles) are connected if they share a common genetic component according to disorder disease-gene associations list

Koh et al (2007)

Rise of Polarization in US House



Blue: Democrat

Red: Republican

Tie weight: roll call vote similarity

Layout: High weight pairs more
likely to be located closely.

Things to Cover :: Base R Graphics

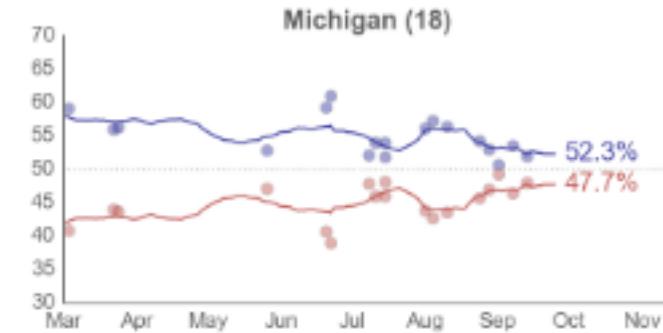
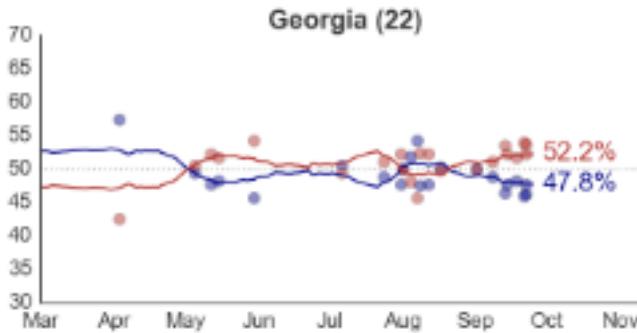
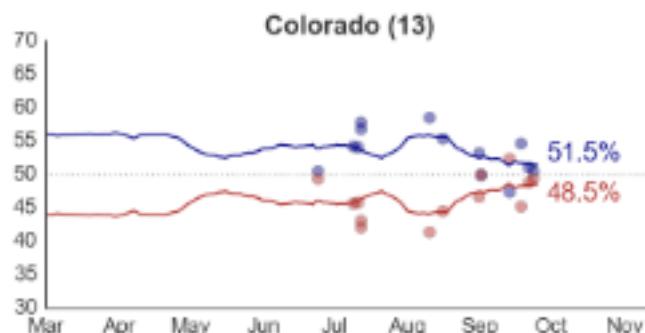
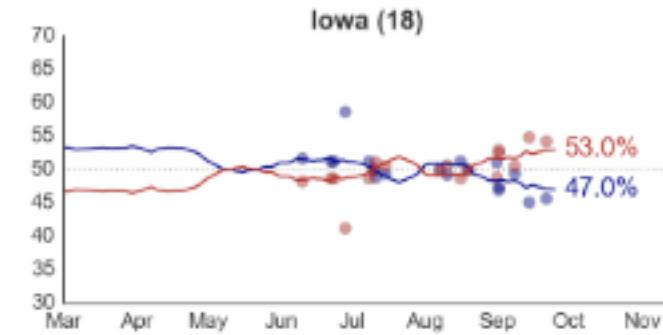
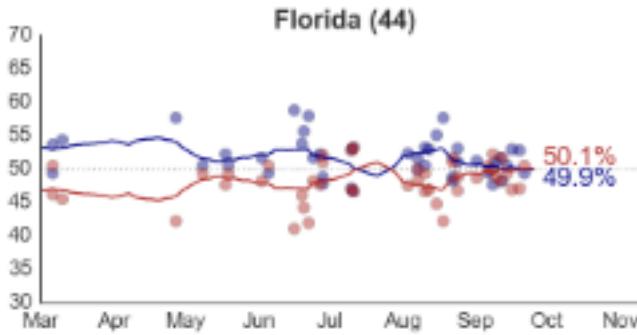
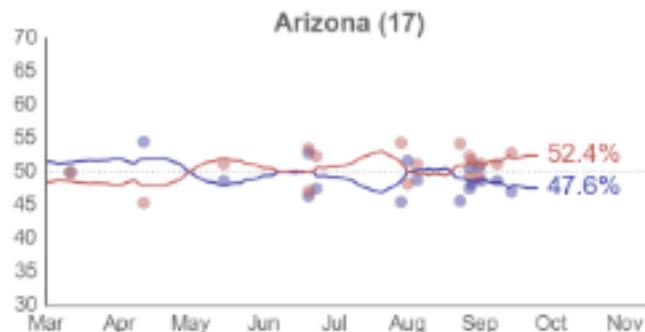
- Bar plot
- Box plot
- Scatter plot
- Histogram
- Kernel density plot

Project 1



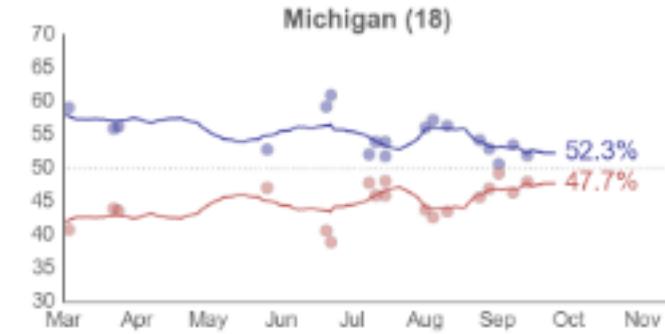
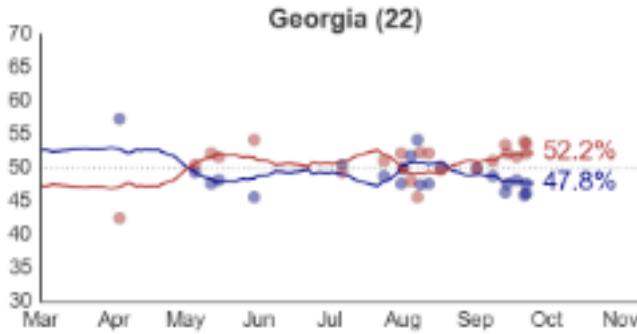
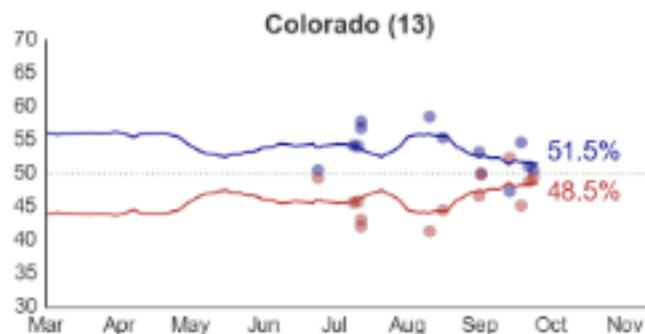
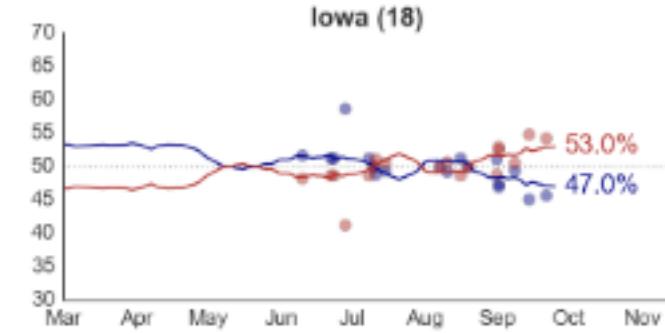
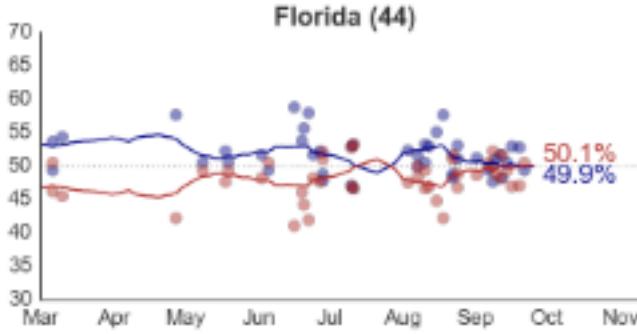
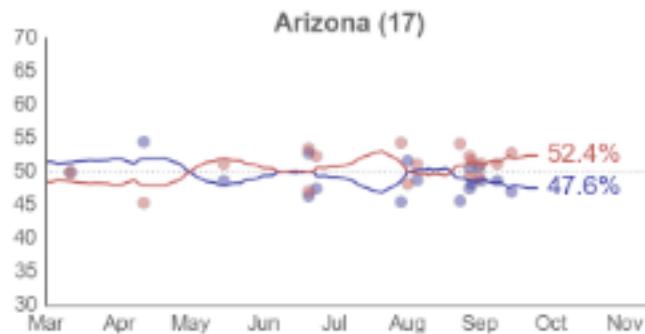
State-level Polls

Generic Approach to Election Prediction



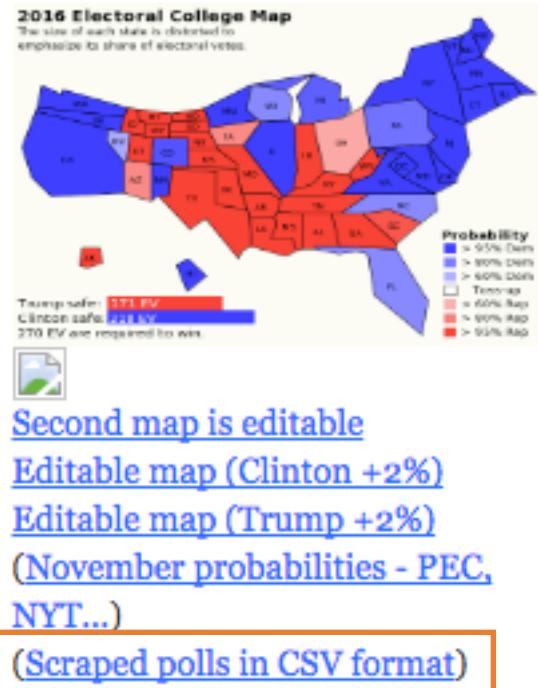
VOTAMATIC
Polling Analysis and Election Forecasting

Generic Approach to Election Prediction



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

Download the Dataset/Slides/Codes



<http://election.princeton.edu/>

1. Right click on [\(Scraped polls in CSV format\)](#)
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()); ← ; is added since copy/paste sometimes  
does not recognize line breaks  
## Delete your workspace;  
getwd();  
## Check your current working directory  
setwd("<location of your dataset>");  
## Set your working directory;  
polls = read.csv("2016_StatePolls.csv");  
## Load data;
```

Quick Inspection of Poll Data Frame

```
polls      849 obs. of 19 variables
  State : Factor w/ 50 levels "AK", "AL", "AR", ...: 6 6 6 6 6 6 ...
  pollster : Factor w/ 155 levels "Abt SRBI Inc/Atlanta Jour...
  pop : int 622 400 694 694 538 534 784 784 991 644 ...
  vtype : Factor w/ 3 levels "Adults", "Likely Voters", ...: 2 ...
  method : Factor w/ 7 levels "Automated Phone", ...: 2 5 4 4 ...
  begmm : int 9 9 9 9 9 9 9 9 9 9 ...
  begdd : int 9 29 27 27 12 2 20 20 21 13 ...
  begyy : int 2016 2016 2016 2016 2016 2016 2016 2016 2016 2...
  endmm : int 9 10 9 9 9 9 9 9 9 9 ...
  enddd : int 29 2 28 28 25 22 25 25 23 21 ...
  endyy : int 2016 2016 2016 2016 2016 2016 2016 2016 2016 2...
  trump : num 45 38 40 44 45 44 42 47 39 42 ...
  clinton : num 43 49 46 51 49 42 41 49 40 44 ...
  other : num 12 1 NA NA 6 14 NA NA NA NA ...
  undecided: num NA 3 6 NA NA NA NA NA NA ...
  Begdate : Factor w/ 167 levels "1/13/2016", "1/15/2016", ...
  Enddate : Factor w/ 166 levels "1/18/2016", "1/19/2016", ...
  Middate : Factor w/ 163 levels "1/15/2016", "1/16/2016", ...
  etc : num 12 13 14 5 6 14 17 4 21 14 ...
```

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

In order to avoid missing values (NA)

```
summary(polls$other);  
summary(polls$undecided);  
polls$etc<-100-polls$trump-polls$clinton
```

- Now `polls$etc` contains rest of the responses other than Clinton and Trump

plot command structure

```
plot_command(command-specific, main="title", xlab="xlabel", ylab="ylabel")
```

e.g.

- plot
- barplot
- pie
- hist

e.g.

- single variable, x
- multiple variables: x,y
- variables, **parameters**

generic parameters

- symbol
- color
- line style
- line width
-

par() :: subplots

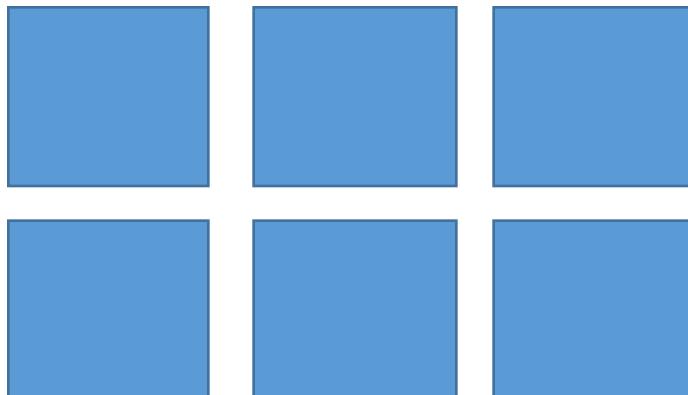
```
par(mfrow=c(#rows, #cols))
```

e.g.

```
par(mfrow=c(1,2))
```



```
par(mfrow=c(2,3))
```



dev.off() :: reset/complete graphic device

```
dev.off()
```

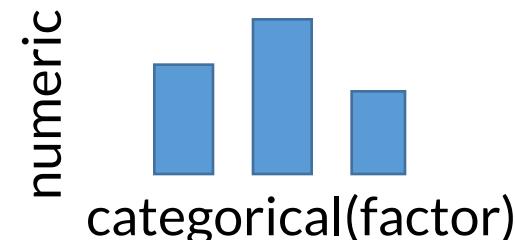
plot.new() :: starting a new plot

```
plot.new()
```

barplot() :: values by factor

```
boxplot(y~x,main="title",xlab=" xlabel",ylab=" ylabel")
```

↑
numeric categorical(factor)
 numeric



```
boxplot(table,main="title",xlab=" xlabel",ylab=" ylabel")
```

↑
N by 2 table:
[categorical, numeric]

e.g.

Mode	DVotes
factor	numeric value

Question 1

- Are contested states more likely to be polled?

(with the assumption that our dataset contains almost all of the poll records w/o bias)

Goal: First compute the frequency of polling by state

```
table(polls$State)
```

```
barplot(table(polls$State),main="Unordered")
```

```
## simple bar plot (Q: what is the order in the x values?)
```

- Goal: Reorder x values depending on support rate

Question 1

- Are contested states more likely to be polled?

(with the assumption that our dataset contains almost all of the poll records w/o bias)

```
barplot(table(polls$State),main="Unordered");
## simple bar plot (Q: what is the order in the x values?)
polls_r <- transform(polls,State = reorder(State, trump));
## reorder states by Trump support rate
barplot(table(polls_r$State),main="Ordered by %Trump");
## ordered plot!
```

Question 1

- Are contested states more likely to be polled?

(with the assumption that our dataset contains almost all of the poll records w/o bias)

```
barplot(table(polls$State),main="Unordered");
## simple bar plot (Q: what is the order in the x values?)
polls_r <- transform(polls,State = reorder(State, trump));
barplot(table(polls_r$State),main="Ordered by %Trump");
## ordered plot!
par(mfrow=c(2,1))
barplot(table(polls$State),main="Simple Bar Plot");
barplot(table(polls_r$State),main="Ordered by %Trump");
## with subplot function
```

Question 1

- Are contested states more likely to be polled?

(with the assumption that our dataset contains almost all of the poll records w/o bias)

```
barplot(table(polls$State),main="Unordered")
## simple bar plot (Q: what is the order in the x values?)
```

- Goal: Reorder x values by support rate

Question 1

- Are contested states more likely to be polled?

(with the assumption that our dataset contains almost all of the poll records w/o bias)

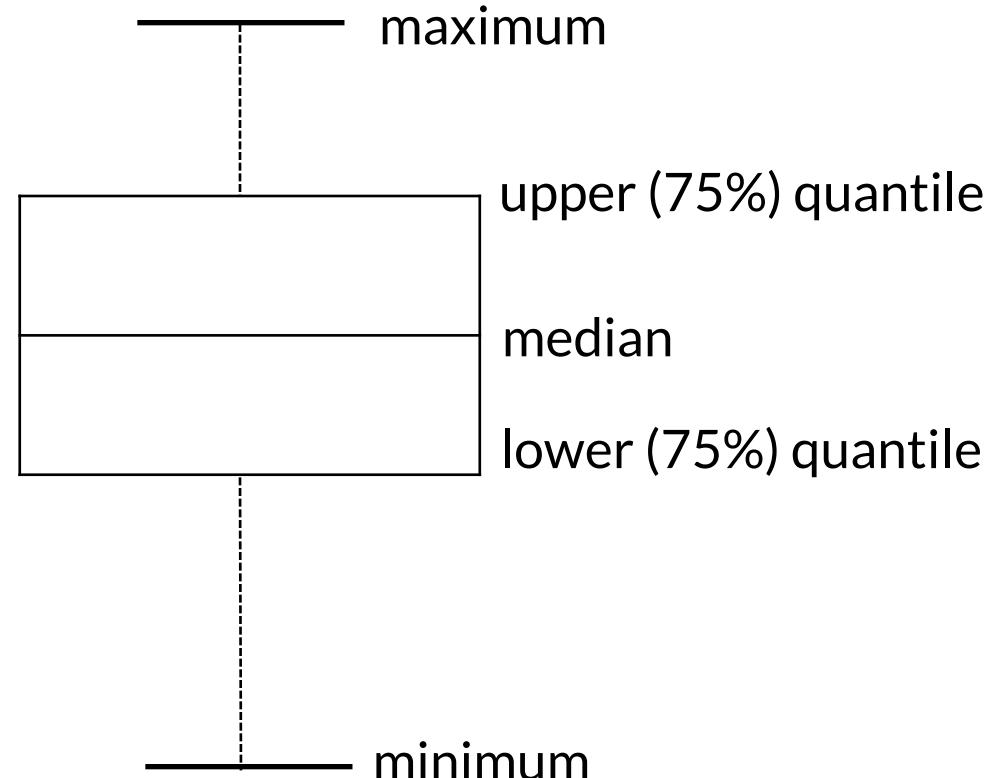
```
barplot(table(polls$State),main="Unordered");
## simple bar plot (Q: what is the order in the x values?)
polls_r <- transform(polls,State = reorder(State, trump));
barplot(table(polls_r$State),main="Ordered by %Trump");
## ordered plot!
par(mfrow=c(2,1));
barplot(table(polls_r$State),main="Ordered by %Trump");
barplot(table(polls_r$State,polls$vtype),beside=T);
## with subplot function
```

boxplot() :: include distributional information

```
boxplot(y~x,main="title",xlab=" xlabel",ylab=" ylabel")
```

↑
numeric categorical(factor)

e.g.



Question 1

- Are contested states more likely to be polled?

(with the assumption that our dataset contains almost all of the poll records w/o bias)

```
boxplot(polls_r$clinton~polls_r$State,main="title",xlab="xlabel",ylab  
="ylabel");
```

- More robust inference by looking at the dispersion of each value set

Project 2



Information for 136 movies released from Hollywood in 2011

Loading your Dataset

Import Hollywood movie dataset using a package **Lock5Data**

```
install.packages("Lock5Data");
## Install package Lock5Data which contains the Hollywood dataset;
data(HollywoodMovies2011);
## Load data;
movies<- na.omit(HollywoodMovies2011);
## drop all observations with at least one NA
```

Information for 136 movies released from Hollywood in 2011

Quick Inspection of Poll Data Frame

```
movies      111 obs. of 14 variables
Movie : Factor w/ 136 levels "30 Minutes or Less",...: 50 73
LeadStudio : Factor w/ 34 levels "20th Century Fox",...: 24 1
RottenTomatoes : int 67 68 44 96 90 93 75 35 69 69 ...
AudienceScore : int 65 58 38 92 77 84 91 58 73 72 ...
Story : Factor w/ 22 levels "", "Comedy", "Discovery", ...: 10 1
Genre : Factor w/ 9 levels "Action", "Adventure", ...: 7 7 4 6
TheatersOpenWeek : int 2408 3321 3049 4375 2918 944 2534 361
BOAverageOpenWeek: int 5511 15829 10365 38672 8995 6177 1027
DomesticGross : num 54 104 100 381 169 ...
ForeignGross : num 43 98.2 115.9 947.1 119.3 ...
WorldGross : num 97 202 216 1328 288 ...
Budget : num 1.5 5 20 125 32.5 17 25 80 27 35 ...
Profitability : num 64.67 40.38 10.81 10.62 8.87 ...
OpeningWeekend : num 13.3 52.6 31.6 169.2 26.2 ...
attr(*, "na.action")=Class 'omit' Named int [1:25] 9 21 22 2
.. ..- attr(*, "names")= chr [1:25] "9" "21" "22" "25" ...
```

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

hist () :: distribution of values

```
hist(x, breaks=bins, main="title", xlab="xlabel", ylab="ylabel")
```

x: numeric

breaks: number of bins

```
hist(movies$RottenTomatoes, breaks=10, col="red", xlab="Rating",  
main="Colored histogram with 10 bins");
```



TOP BOX OFFICE			Get Tickets
🍅 65%	Miss Peregrine's Home for Peculi...	\$28.9M	
🟡 83%	Deepwater Horizon	\$20.2M	
🍅 63%	The Magnificent Seven	\$15.6M	
🍅 62%	Storks	\$13.5M	
🟡 82%	Sully	\$8.3M	
✳️ 36%	Masterminds	\$6.5M	

Question 2

- What is more correlated with gross income, budget or critics rating?

Question 2

- What is more correlated with gross income, budget or critics rating?
- Check bivariate correlations of each pair on different graphs

Question 2

- What is more correlated with gross income, budget or critics rating?
- Check bivariate correlations of each pair on different graphs

```
par(mfrow=c(1,2));
## 1 by 2 subplots;
plot(movies$RottenTomatoes, log10(movies$WorldGross));
plot(log10(movies$Budget), log10(movies$WorldGross));
## log10: logarithm function with base 10;
```

Question 2

- What is more correlated with gross income, budget or critics rating?
- Check bivariate correlations of each pair on different graphs

```
par(mfrow=c(1,2));
## 1 by 2 subplots;
plot(movies$RottenTomatoes,log10(movies$WorldGross),col=movies$Genre);
plot(log10(movies$Budget),log10(movies$WorldGross),col=movies$Genre);
## log10: logarithm function with base 10;
```

plot() :: add linear trend plot

```
lines(x,y_predicted,main="title",xlab="xlabel",ylab="ylabel")
```



x, y_predicted: numeric

plot() :: add linear trend plot

```
lines(x,y_predicted,main="title",xlab="xlabel",ylab="ylabel")
```

x, y_predicted: numeric

```
mod1 <- lm(log10(movies$WorldGross) ~ movies$RottenTomatoes);  
## Linear regression  
preds1 <- predict(mod1);  
## predicted value obtained by linear regression  
plot(movies$RottenTomatoes,log10(movies$WorldGross));  
lines(movies$RottenTomatoes, preds1);
```

plot() :: add linear trend plot

```
lines(x,y_predicted,main="title",xlab="xlabel",ylab="ylabel")
```

x, y_predicted: numeric

```
mod2 <- lm(log10(movies$WorldGross) ~ log10(movies$Budget));  
## Linear regression  
preds2 <- predict(mod2);  
## predicted value obtained by linear regression  
plot(log10(movies$Budget),log10(movies$WorldGross));  
lines(log10(movies$Budget), preds2);
```

Question 3

- Is there significant difference in rating/budget by genre?

Question 3

- Is there significant difference in rating/budget by genre?

```
par(mfrow=c(1,2));
## 1 by 2 subplots;
plot(movies$RottenTomatoes,log10(movies$WorldGross),col=movies$Genre);
plot(log10(movies$Budget),log10(movies$WorldGross),col=movies$Genre);
## log10: logarithm function with base 10;
```

Question 3

- Is there significant difference in rating/budget by genre?
- Check bivariate correlations of each pair on different graphs

Question 3

- Is there significant difference in rating/budget by genre?
- Check bivariate correlations of each pair on different graphs.

```
par(las=2);
## horizontal text;
par(mfrow=c(1,2));
boxplot(movies$RottenTomatoes~movies$Genre,xlab="Genre",ylab="Rating");
## Genre VS Rating;
boxplot(movies$Budget~movies$Genre,xlab="Genre",ylab="Budget");
## Genre VS Budget;
```

Question 3

- Is there significant difference in rating/budget by genre?
- e.g. Compare smoothed distributions of Drama and Animation budgets.

```
R.Drama <- density(movies$Budget[which(movies$Genre=="Drama")]);  
R.Ani <- density(movies$Budget[which(movies$Genre=="Animation")]);  
plot(R.Drama, col="green4", lwd=2, main="Drama VS  
Animation", xlab="Budget", xlim=c(-10,250));  
lines(R.Ani, col="orange", lwd=2, lty=2);  
## overlay lines over plot
```

text() :: add location-specific text labels

```
pdf("file_name.pdf")
```

e.g.

```
plot.new();
R.Drama <- density(movies$Budget[which(movies$Genre=="Drama")]);
R.Ani <- density(movies$Budget[which(movies$Genre=="Animation")]);
plot(R.Drama, col="blue",lwd=2,main="Drama VS Animation",xlab="Budget",xlim=c(-10,250));
lines(R.Ani,col="red",lwd=2);legend("topright",c("Drama","Animation"));
###???
```

Question 3

- Is there significant difference in rating/budget by genre?
- e.g. Compare smoothed distributions of Drama and Animation budgets.

```
dev.off();

R.Drama <- density(movies$Budget[which(movies$Genre=="Drama")]);
R.Ani <- density(movies$Budget[which(movies$Genre=="Animation")]);
plot(R.Drama, col="green4", lwd=2, main="Drama VS Animation", xlab="Budget", xlim=c(-10,250));
lines(R.Ani, col="orange", lwd=2, lty=2);
## overlay lines over plot
text(c(10,200),c(0.025,0.007),c("Drama", "Animation"));
## add line labels
```

pdf() :: save graph as pdf

```
pdf("file_name.pdf",width=width_length,height=height_length)
```

e.g.

```
pdf("file_name.pdf");
plot(R.Drama, col="green4",lwd=2,main="Drama VS Animation",
xlab="Budget",xlim=c(-10,250));
lines(R.Ani,col="orange",lwd=2,lty=2);
text(c(10,200),c(0.025,0.007),c("Drama","Animation"));
dev.off()
```

Feed Back Survey

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

Thank you

FALL 2016 SCHEDULE

September 20	Introduction to R and RStudio
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