

# R Programming for Data Science

EMILY JONES

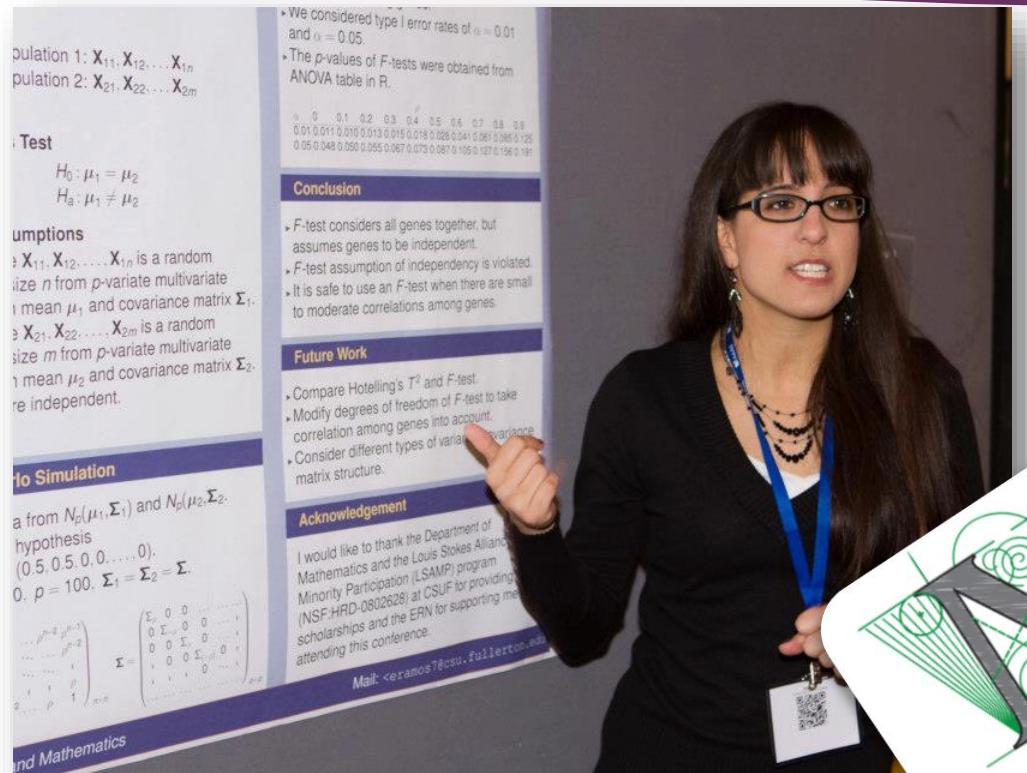
DATA SCIENTIST

11/07/18

# Agenda

- WHO AM I
- WHAT IS A DATA SCIENTIST
- WHAT IS R
- WHY SHOULD YOU LEARN R
- OVERVIEW OF RSTUDIO
- OTHER COOL THINGS IN R
- DEMO (IF TIME)

# Who am I - Education



Population 1:  $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n}$   
Population 2:  $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2m}$

**Hypothesis Test**

$H_0: \mu_1 = \mu_2$   
 $H_a: \mu_1 \neq \mu_2$

**Assumptions**

- $\mathbf{X}_{11}, \mathbf{X}_{12}, \dots, \mathbf{X}_{1n}$  is a random size  $n$  from  $p$ -variate multivariate mean  $\mu_1$  and covariance matrix  $\Sigma_1$ .
- $\mathbf{X}_{21}, \mathbf{X}_{22}, \dots, \mathbf{X}_{2m}$  is a random size  $m$  from  $p$ -variate multivariate mean  $\mu_2$  and covariance matrix  $\Sigma_2$ .  
are independent.

**To Simulation**

a from  $N_p(\mu_1, \Sigma_1)$  and  $N_p(\mu_2, \Sigma_2)$ .  
hypothesis  
(0.5, 0.5, 0, 0, ..., 0).  
 $0, p = 100, \Sigma_1 = \Sigma_2 = \Sigma$

$$\Sigma = \begin{pmatrix} \Sigma_{11} & 0 & \cdots & 0 \\ 0 & \Sigma_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_{pp} \end{pmatrix}$$

and Mathematics

We considered type I error rates of  $\alpha = 0.01$  and  $\alpha = 0.05$ .  
The  $p$ -values of  $F$ -tests were obtained from ANOVA table in R.

$\alpha$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	0.011	0.010	0.013	0.015	0.018	0.028	0.041	0.061	0.085	0.126
0.05	0.048	0.050	0.055	0.067	0.073	0.087	0.105	0.127	0.156	0.191

**Conclusion**

- $F$ -test considers all genes together, but assumes genes to be independent.
- $F$ -test assumption of independency is violated.
- It is safe to use an  $F$ -test when there are small to moderate correlations among genes.

**Future Work**

- Compare Hotelling's  $T^2$  and  $F$ -test.
- Modify degrees of freedom of  $F$ -test to take correlation among genes into account.
- Consider different types of variance-covariance matrix structure.

**Acknowledgement**

I would like to thank the Department of Mathematics and the Louis Stokes Alliance Minority Participation (LSAMP) program (NSF-HRD-0802628) at CSUF for providing scholarships and the ERN for supporting me attending this conference.

Mail: <eramos7@csuf.fullerton.edu>



# Who am I – Education...



**UMASS**  
**AMHERST**



# Who am I – Career Path



# What is a Data Scientist – Venn Diagram

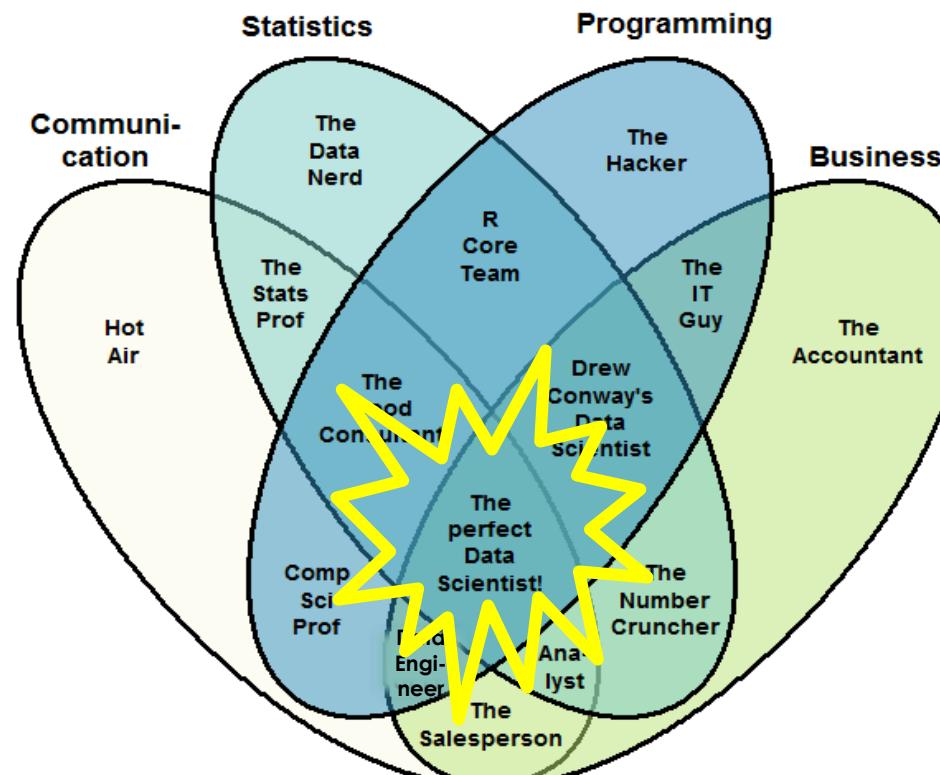
## ► Statistics/Math

- ▶ Basic Stats
- ▶ Modeling
- ▶ Machine Learning

## ► Communication

- ▶ Explain complex concepts
- ▶ Visualization
- ▶ Presentation

The Data Scientist Venn Diagram



## ► Programming

- ▶ Querying
- ▶ Statistical Language
- ▶ Technical Logic

## ► Business Acumen

- ▶ Know your business
- ▶ Fill Business Gaps
- ▶ Passionate

# What is a Data Scientist – Process

## Data Science Process

The data science process involves:

- ▶ Data acquisition, collection, and storage
- ▶ Access, ingest, and integrate data
- ▶ Processing and cleaning data (munging/wrangling)
- ▶ Initial data investigation and exploratory data analysis
- ▶ Apply data science methods and techniques (e.g., machine learning, statistical modeling, ...)
- ▶ Measuring and improving results (validation and tuning)
- ▶ Delivering, communicating, and presenting final results
- ▶ Business decisions and/or changes are made based on the results

# What Do I Do

- ▶ Data Scientist on the SD Analytics Team
  - ▶ Obtain, scrub, explore, model and interpret game play and spending data by blending analytics, statistics and machine learning techniques to provide insights and create data driven narratives to the San Diego Studio.

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21th century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

### MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants



### PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing packages, e.g., R
- ★ Databases: SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

### DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

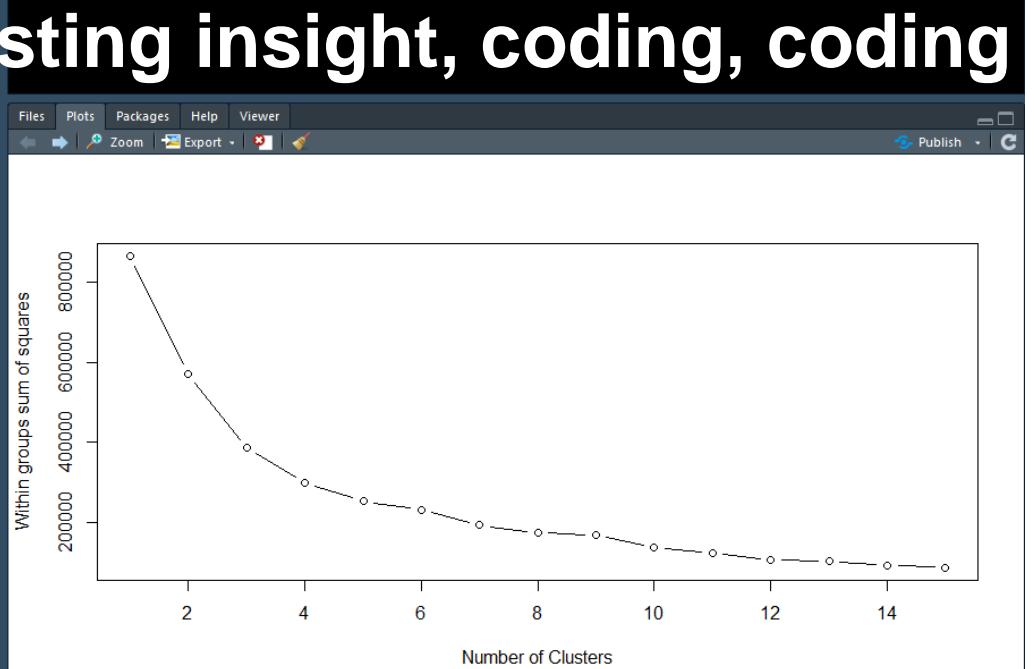
### COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

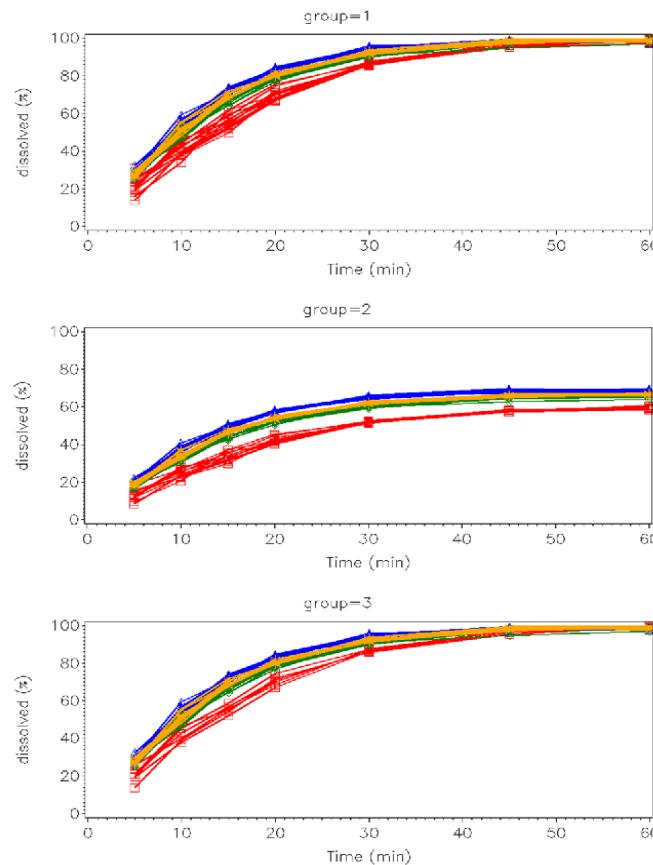
```
132 dat$avg_spent_scale <- as.numeric(scale(dat$avg_spent))
133
134 ## ##### #####
135 ## Running the Clusters #####
136 ## ##### #####
137
138 # Winner winner chicken dinner (this is the clustering approach chosen)
139 # look at purchases, revenue and average
140
141 # subsetting the data
142 mydata <- as.data.frame(cbind(dat$totalpurchases_scale, dat$totalrevenue_scale, dat$avg_spent_scale))
143 colnames(mydata) <- c('totalpurchases_scale', 'totalrevenue_scale', 'avg_spent_scale')
144
145 # Checking the Elbow Chart
146 wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
147 for (i in 2:15) wss[i] <- sum(kmeans(mydata,
148                                     centers=i)$withinss)
149 plot(1:15, wss, type="b", xlab="Number of Clusters",
150       ylab="Within groups sum of squares")
151
152 ### 3 clusters look good on the elbow plot, 4 looks better
153
154 # Create Matrix to run cluster
155 m <- as.matrix(cbind(dat$totalpurchases_scale,dat$totalrevenue_scale,dat$avg_spent_scale
156 ))
157 # running kmeans cluster
158 cl <- (kmeans(m,4))
159
160 # Loading this info back into our data set
161 dat$cluster <- factor(cl$cluster)
162 centers <- as.data.frame(cl$centers)
163
164 #plotting the clusters
165 ggplot(data=dat, aes(x=totalpurchases_scale, y=totalrevenue_scale, color=cluster)) +
166   theme_bw() +
167   geom_point() +
168   geom_point(data=centers, aes(x=V1,y=V2, color='Center')) +
169   geom_point(data=centers, aes(x=V1,y=V2, color='Center'), size=15, alpha=.3, show.legend =
170   FALSE)
171
161:13 # (Untitled) R Script
```

# Coding, coding, make an interesting presentation

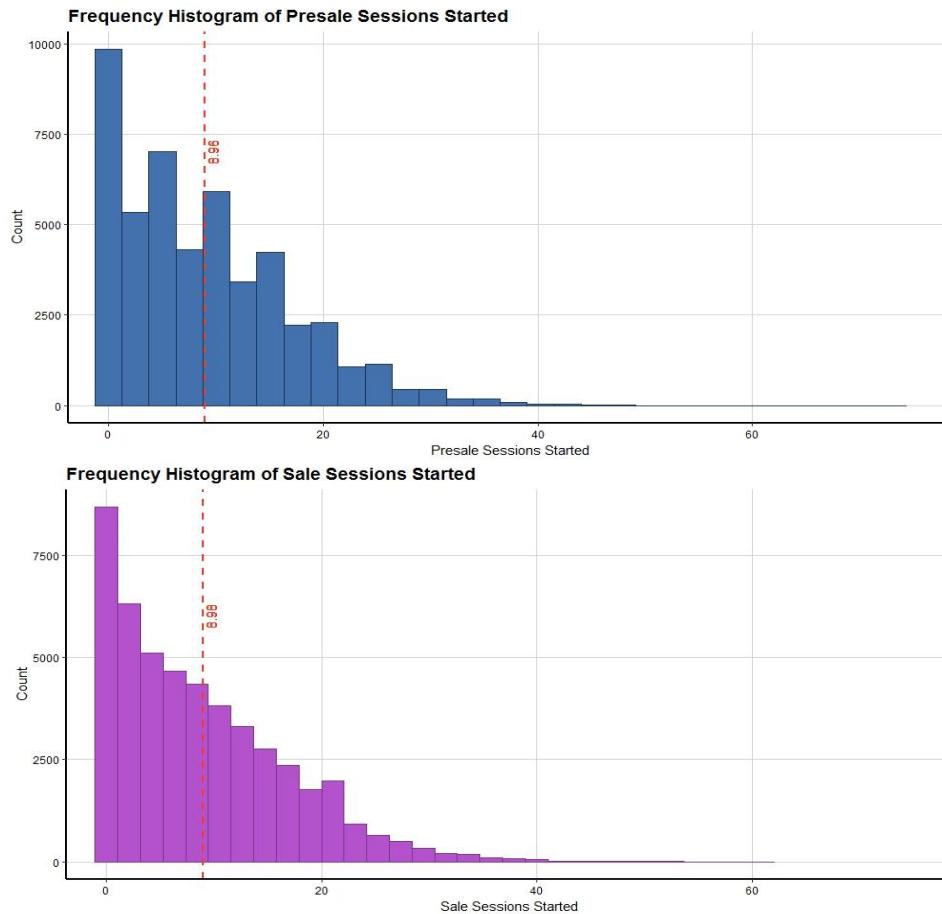
Data	
centers	4 obs. of 3 variables
cl	Large kmeans (9 elements, 1.1 Mb)
dat	287786 obs. of 10 variables
m	Large matrix (863358 elements, 6.6 Mb)
mydata	287786 obs. of 3 variables
Values	
connection	Class 'RODBC' atomic [1:1] 1
i	15L
wss	num [1:15] 863355 570587 387114 299205 254228 ...



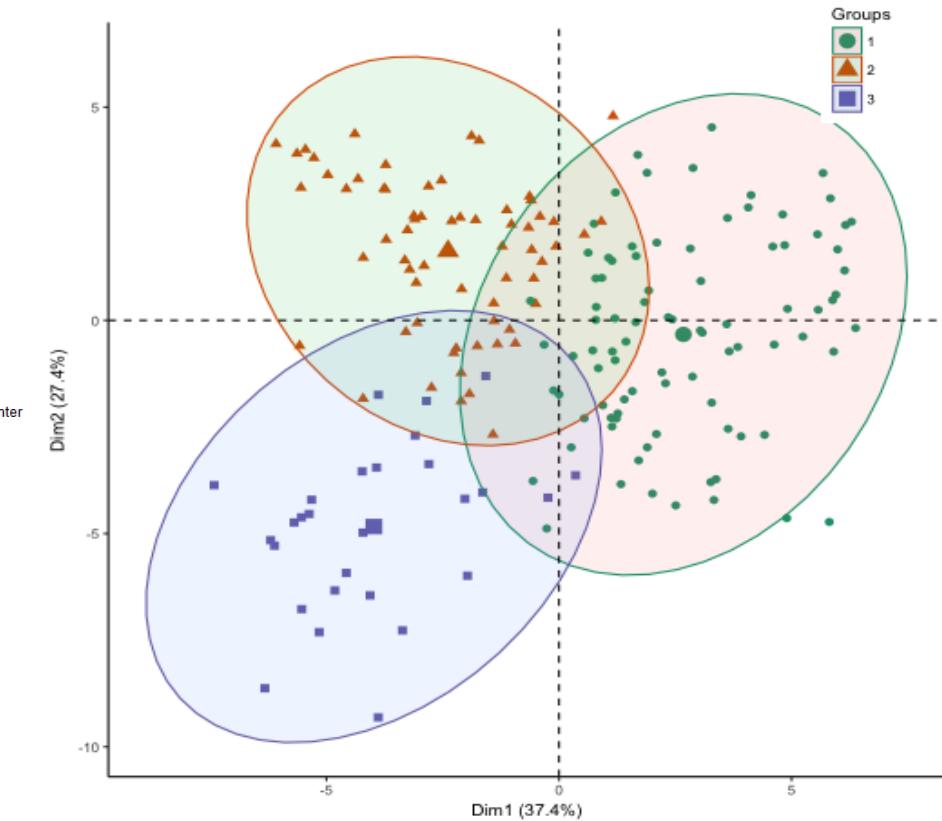
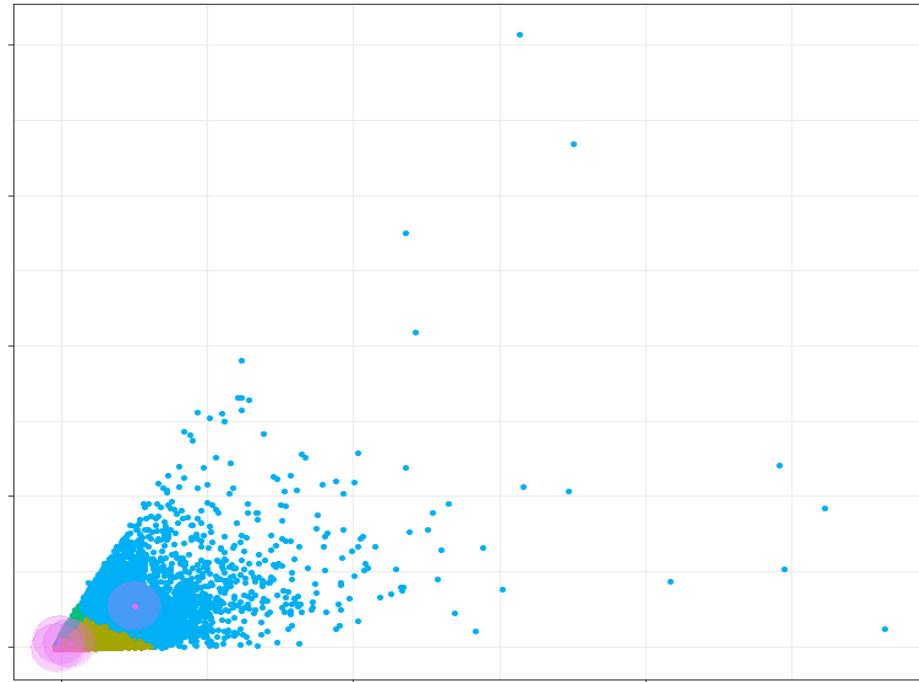
# What Do I Do - Significance Testing



**Fig. 1.** *In vitro* dissolution outside  $F_0$  prerequisite for variability of the

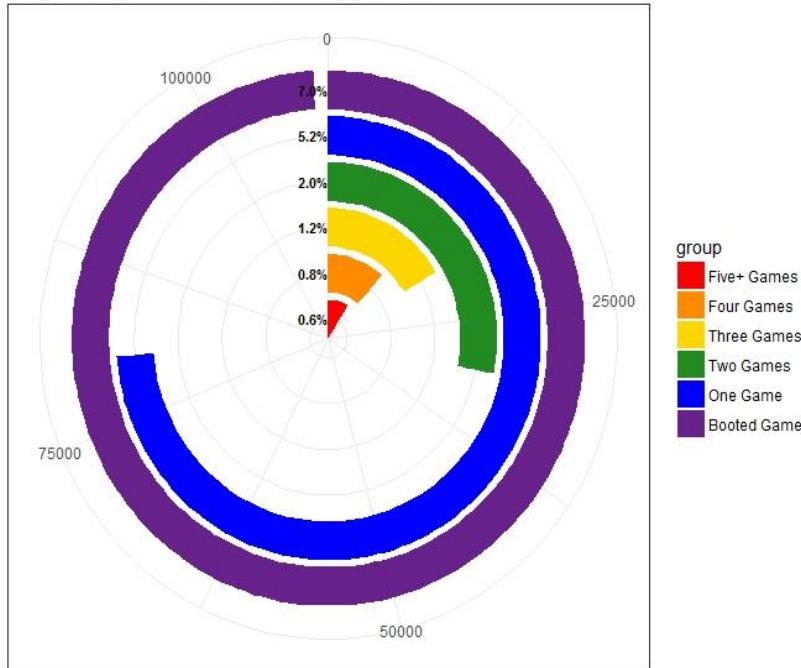


# What Do I Do - Create Player Clusters



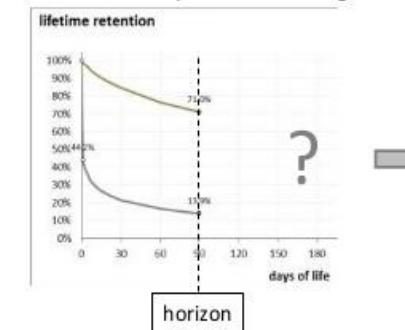
# What Do I Do – Build Retention Models

Player Progression as a Percentage of Entitlement

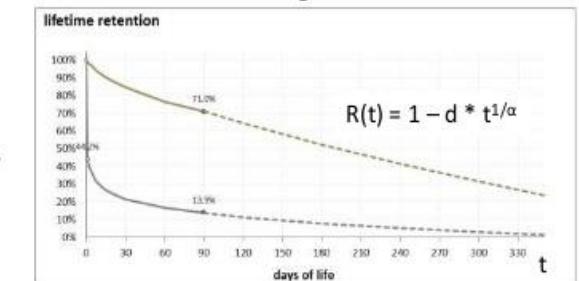


## Lifetime Retention model

Life to date operation of the game



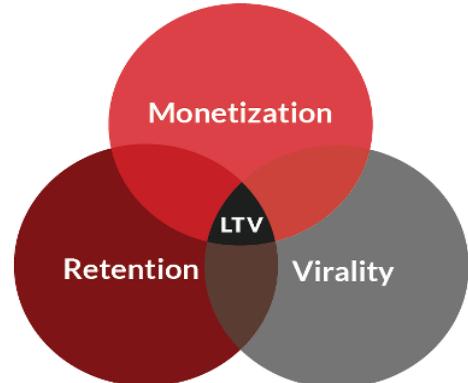
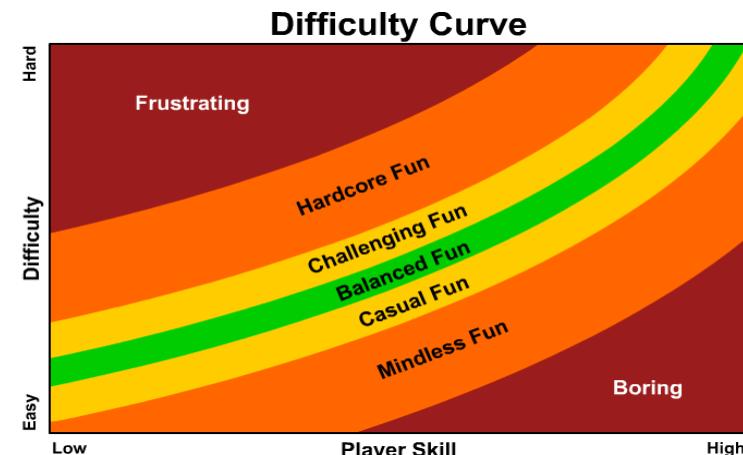
modeling retention curves



parameters  $d$  and  $\alpha$  are found with estimation techniques

- The area under the retention curve is the average lifetime
- KPI : quality of retention  $Q^* = \log(\text{area})$

# What Do I Do - Also...



Discovering people opinions, emotions and feelings about a product or service

---

**DAY IN THE LIFE  
OF A GAMER**

---

# What Do I Do – My Tools



**amazon**  
REDSHIFT



**Studio**<sup>®</sup>



# What is R

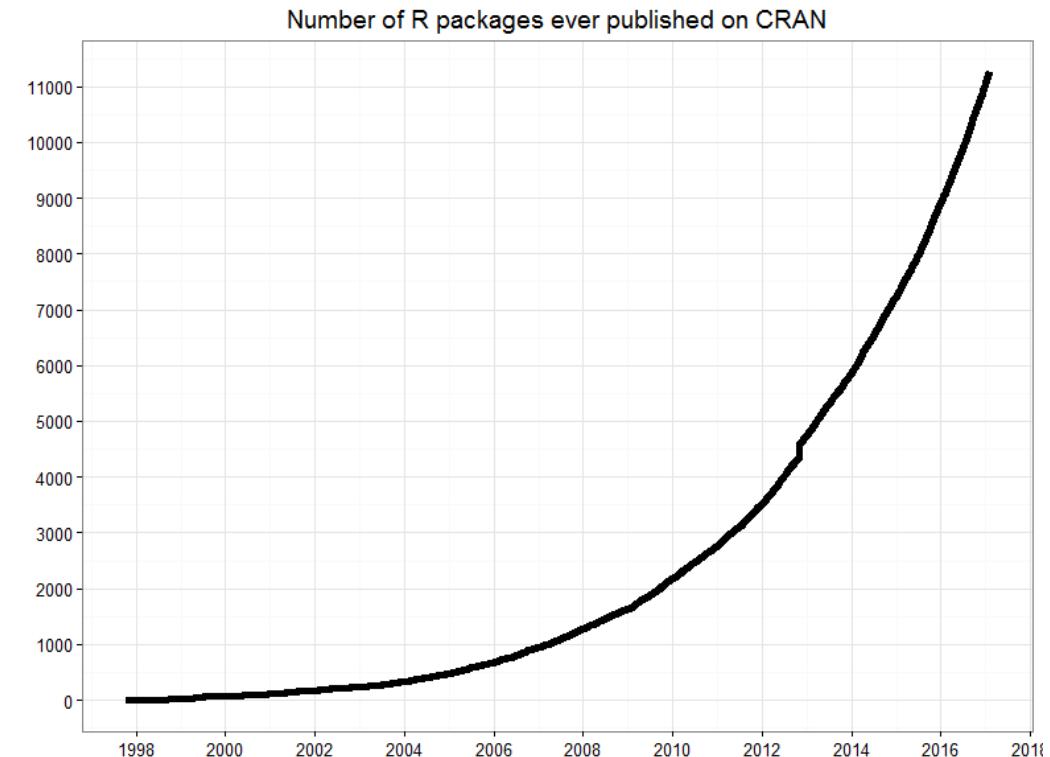
- R IS A DATA ANALYSIS SOFTWARE
- R IS A PROGRAMMING LANGUAGE
- R IS AN ENVIRONMENT FOR STATISTICAL ANALYSIS
- R IS AN OPEN-SOURCE SOFTWARE PROJECT
- R HAS A GROWING AND ACTIVE COMMUNITY
- R IS FREE



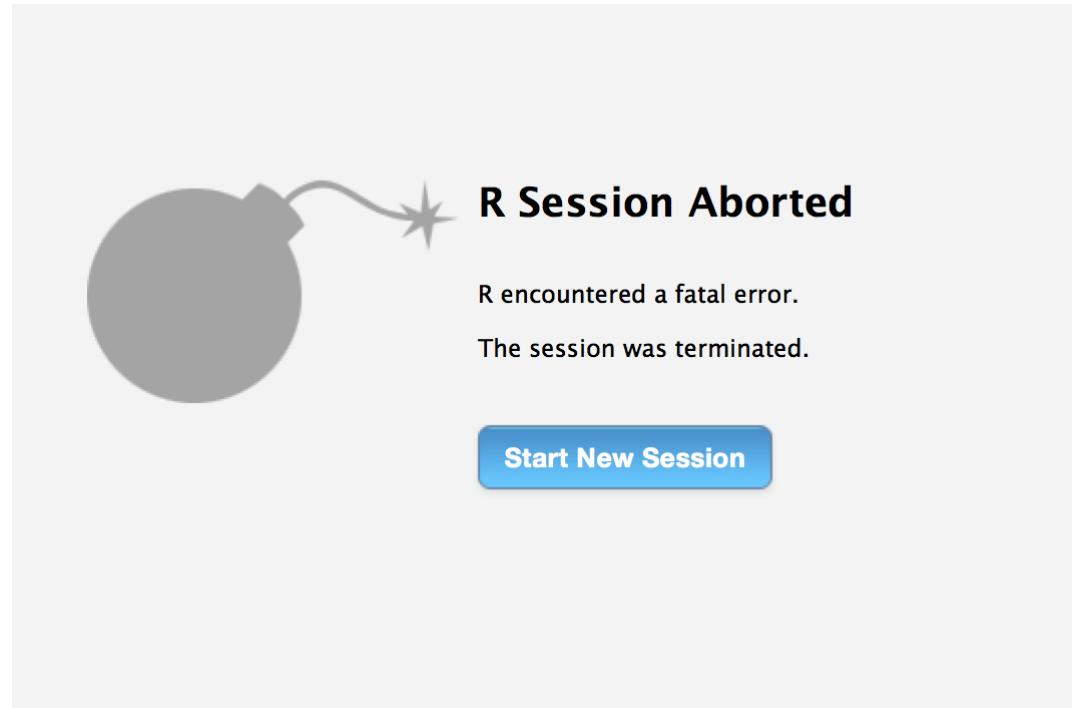
# What is R ...

## Advantages

- ▶ Fast, free and open-source
- ▶ Over 10,000 R packages
- ▶ Outstanding graphical abilities
- ▶ Compatible with most programs/languages
- ▶ Growing & active user community
- ▶ Excellent for simulation, statistics, programming, computer intensive analyses, etc.
- ▶ Great with reporting and documentation



# What is R ...



## Disadvantages

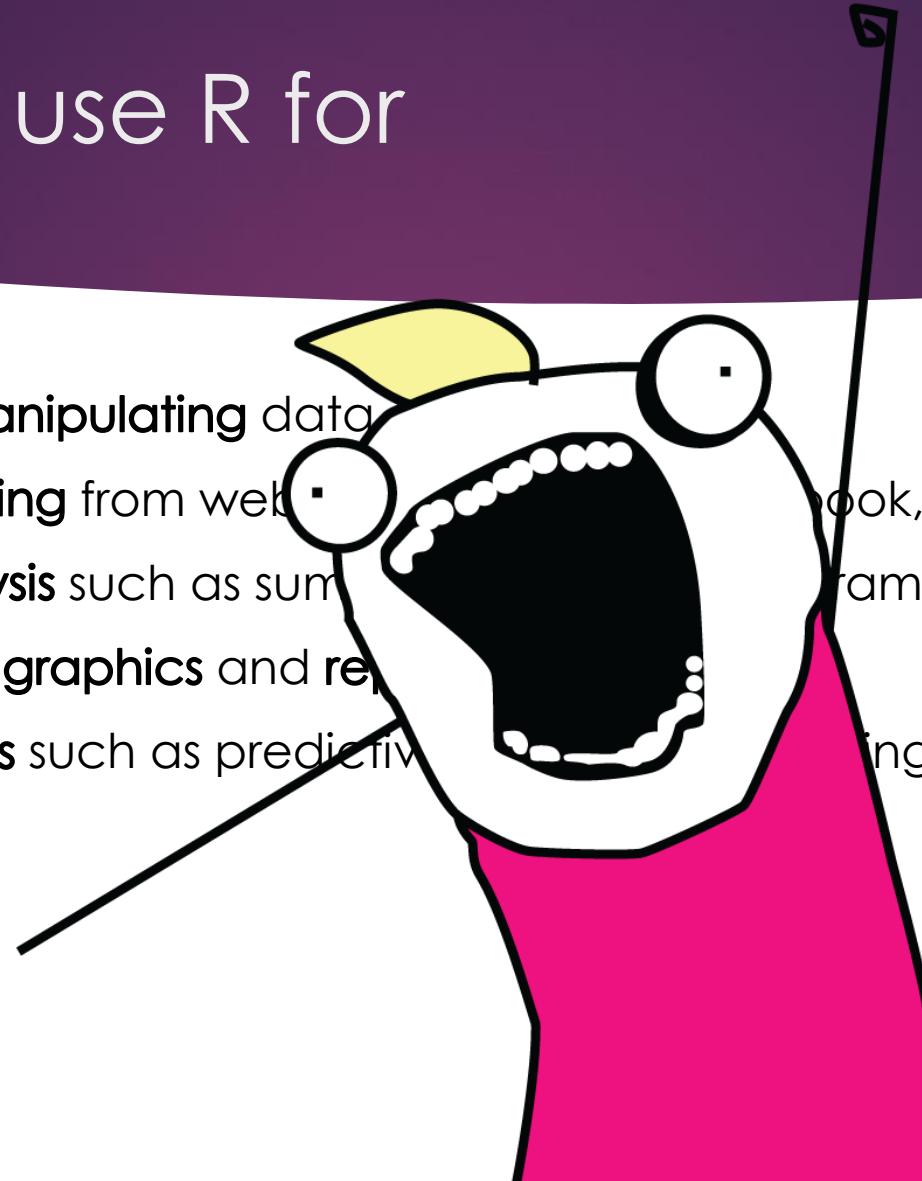
- ▶ Not user friendly, has a steep learning curve, **if you don't understand coding logic!**
- ▶ No professional / commercial support, **but you have an active community for support!**
- ▶ It is not a database, **but connects to DBMS!**
- ▶ Working with large datasets is limited by RAM, **sorry no “buts” here...** ☹
- ▶ Errors can be vague and difficult to debug, **but that's what the internet is for!**
- ▶ There are many ways to do the same thing, which can be confusing, **but also freeing!**

# What do I use R for

- ▶ **Cleaning** and manipulating data.
- ▶ **Data/ Text Scraping** from websites like Twitter, Facebook, Twitch, etc.
- ▶ **Exploratory Analysis** such as summary statistics, histograms, frequency tables, etc.
- ▶ Create beautiful **graphics** and **reports**.
- ▶ **Statistical Analysis** such as predictive modeling, clustering, machine learning, etc.
- ▶ All the **things!**

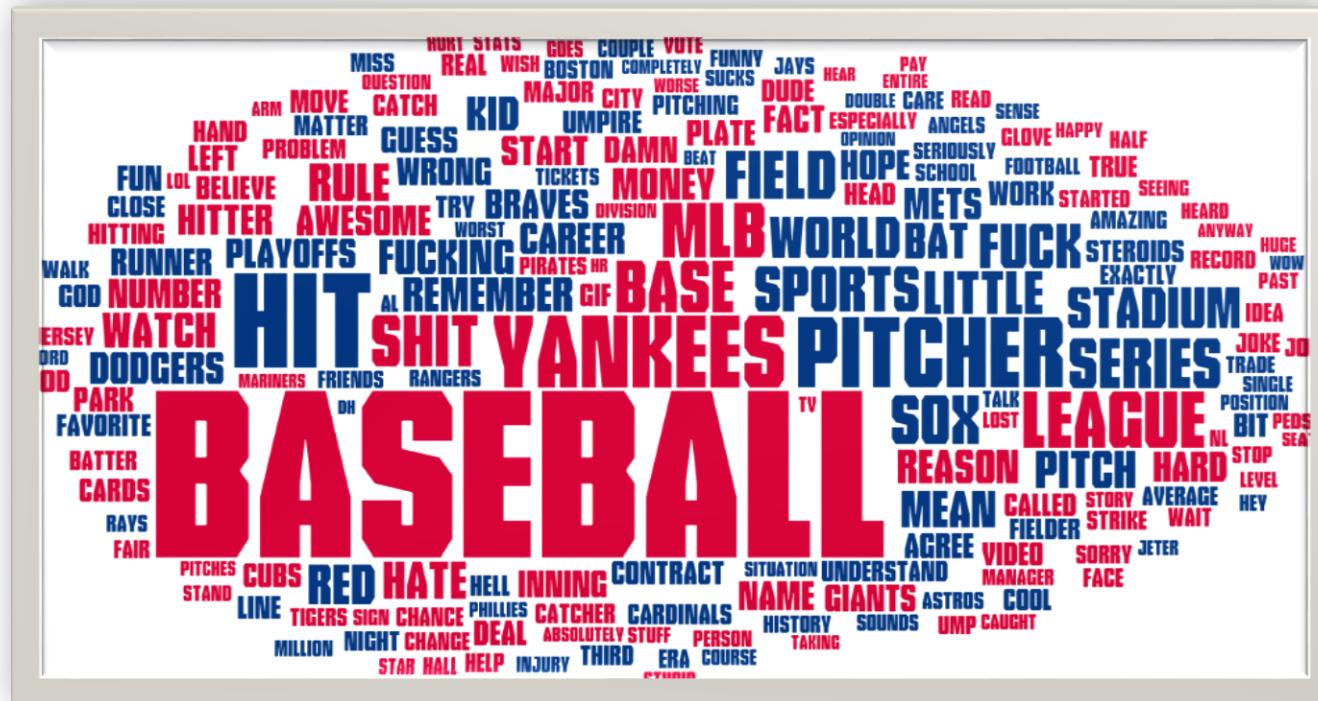
# What do I use R for

- ▶ **Cleaning** and manipulating data
- ▶ Data/ Text Scraping from websites such as YouTube, Instagram, Twitch, etc.
- ▶ Exploratory Analysis such as summaries, plots, histograms, frequency tables, etc.
- ▶ Create beautiful **graphics** and **reports**
- ▶ Statistical Analysis such as predictive modeling, machine learning, etc.
- ▶ All the **things!**



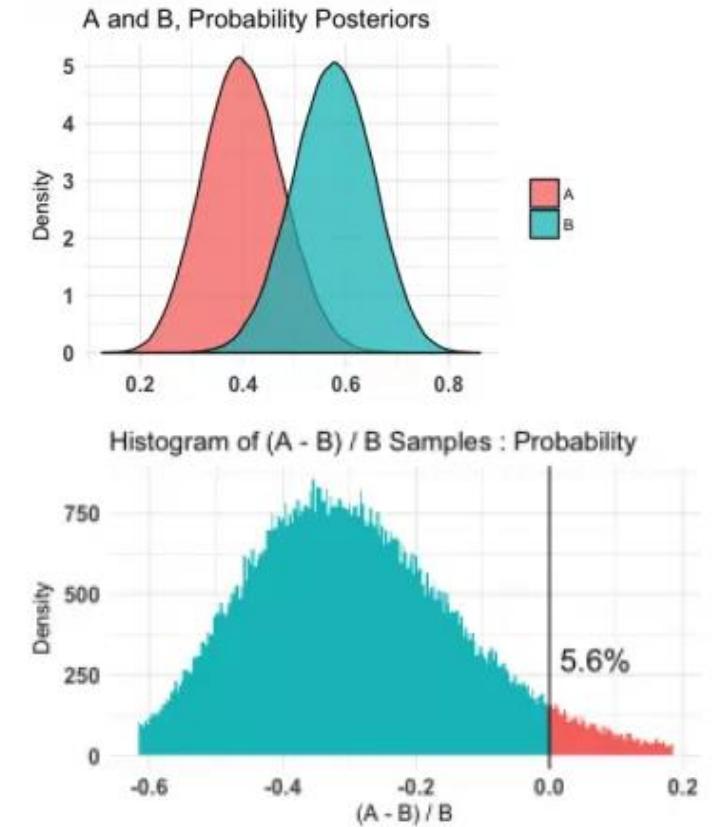
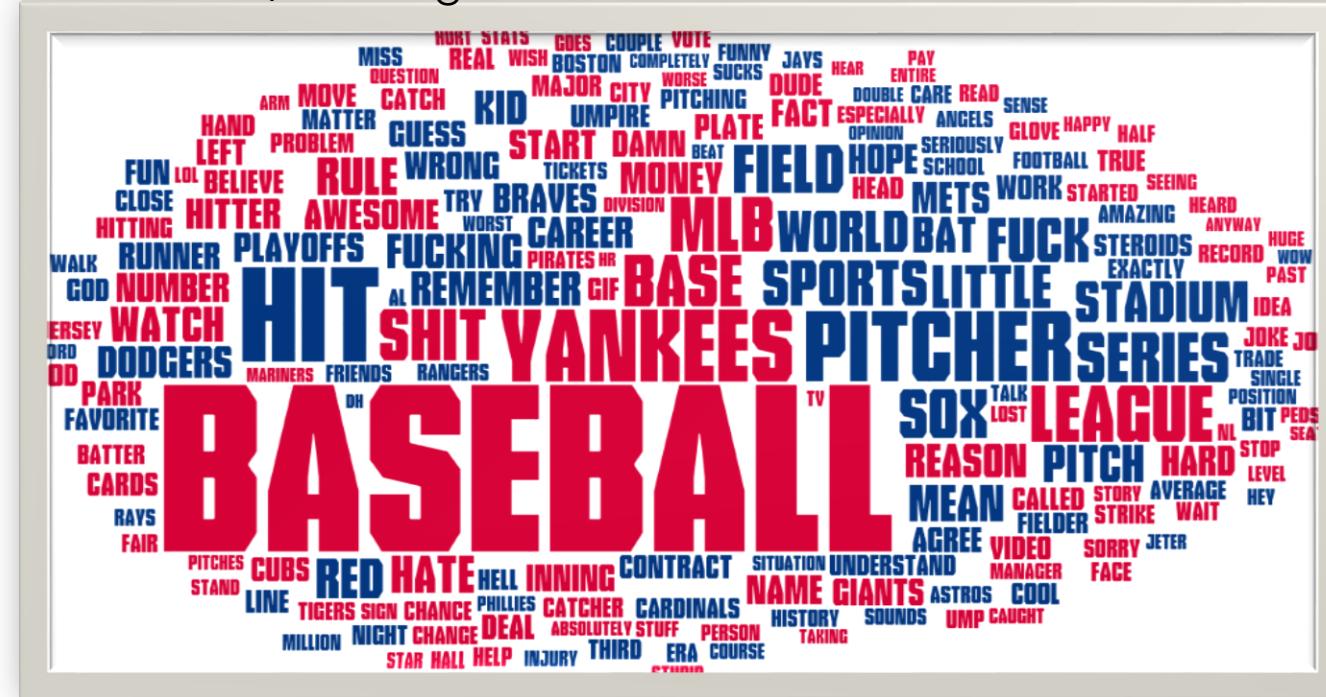
# Why should you learn R

- ▶ Easy to implement techniques such as:
    - ▶ Data/ Text Mining



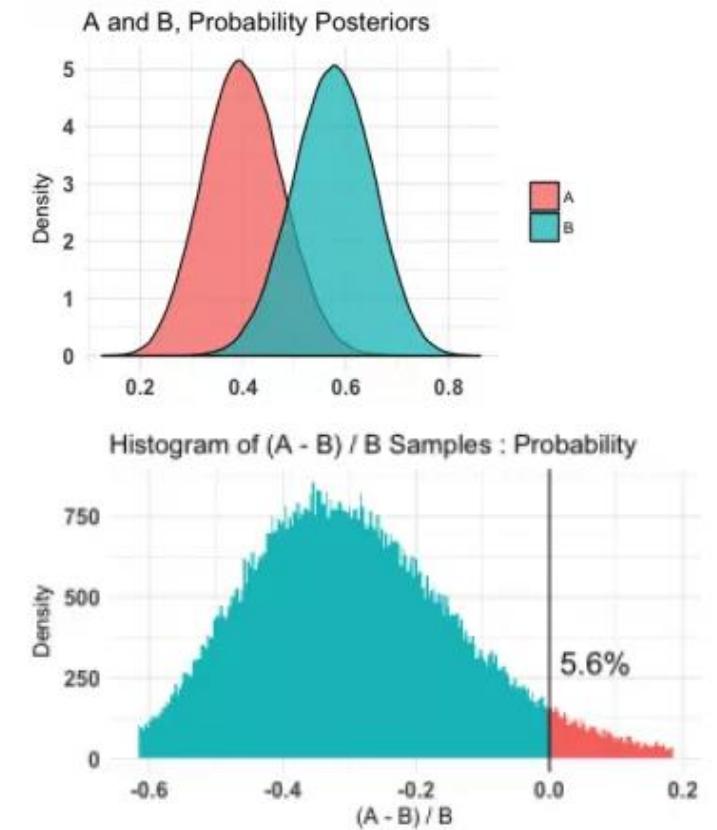
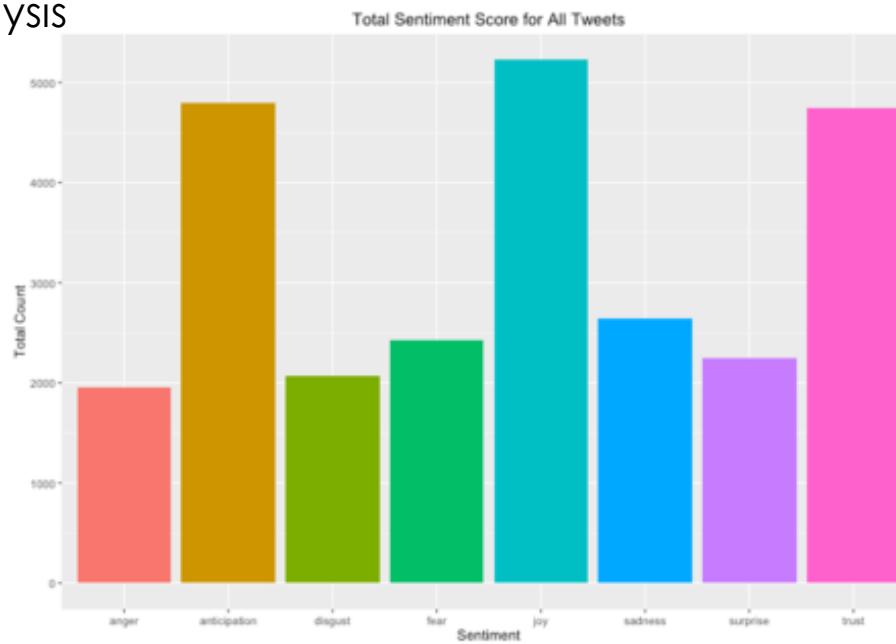
# Why should you learn R

- ▶ Easy to implement techniques such as:
    - ▶ Data/ Text Mining
    - ▶ A/B Testing



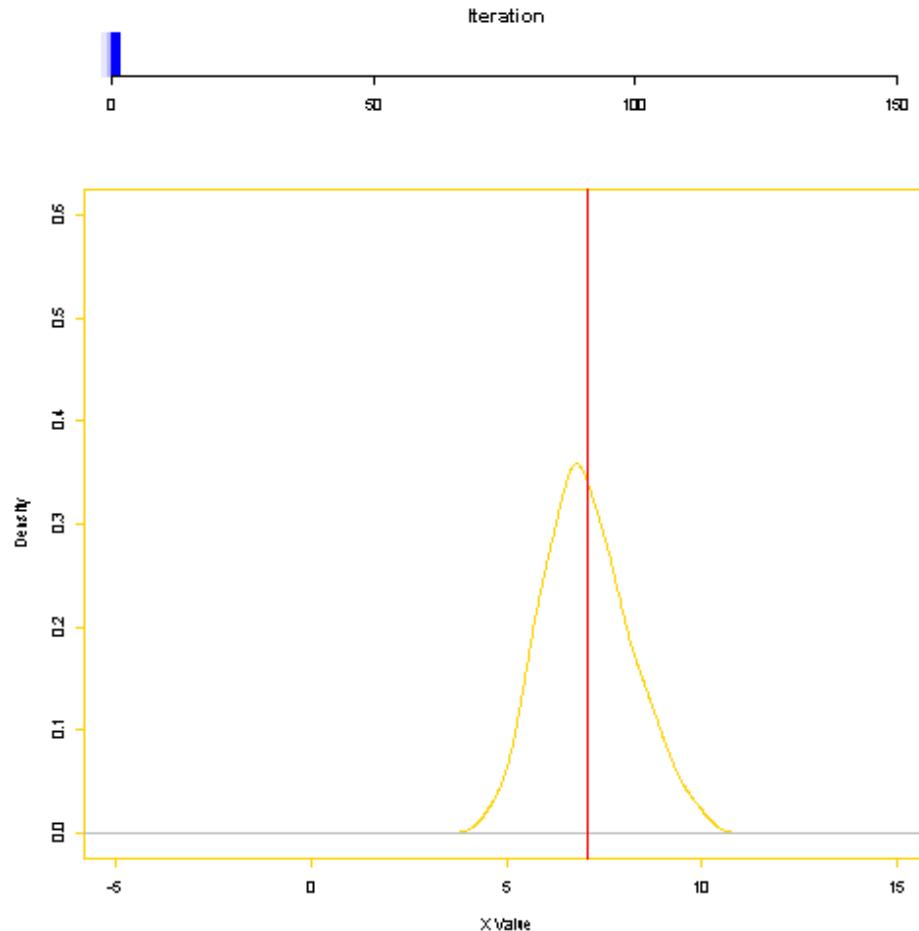
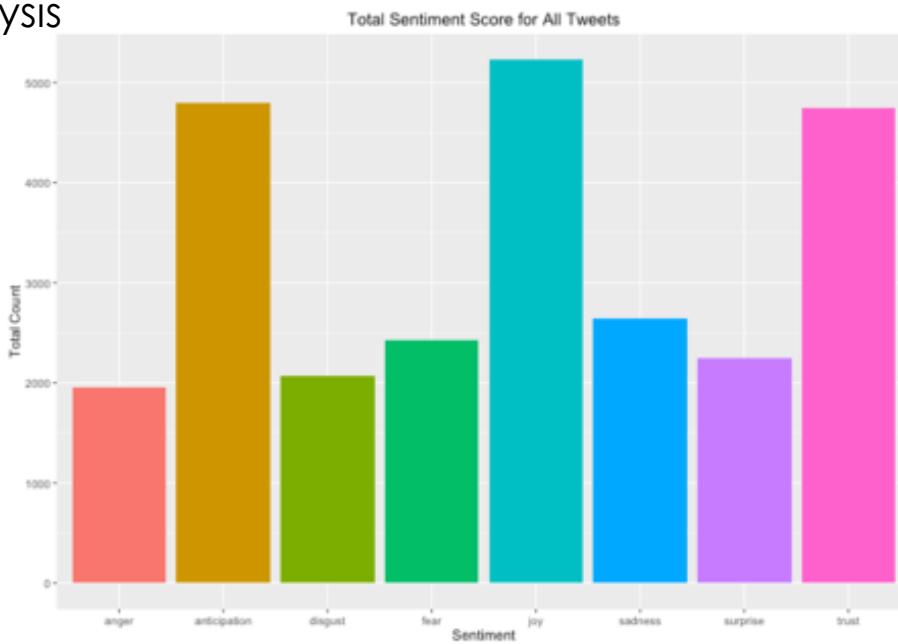
# Why should you learn R

- ▶ Easy to implement techniques such as:
  - ▶ Data/ Text Mining
  - ▶ A/B Testing
  - ▶ Sentiment Analysis



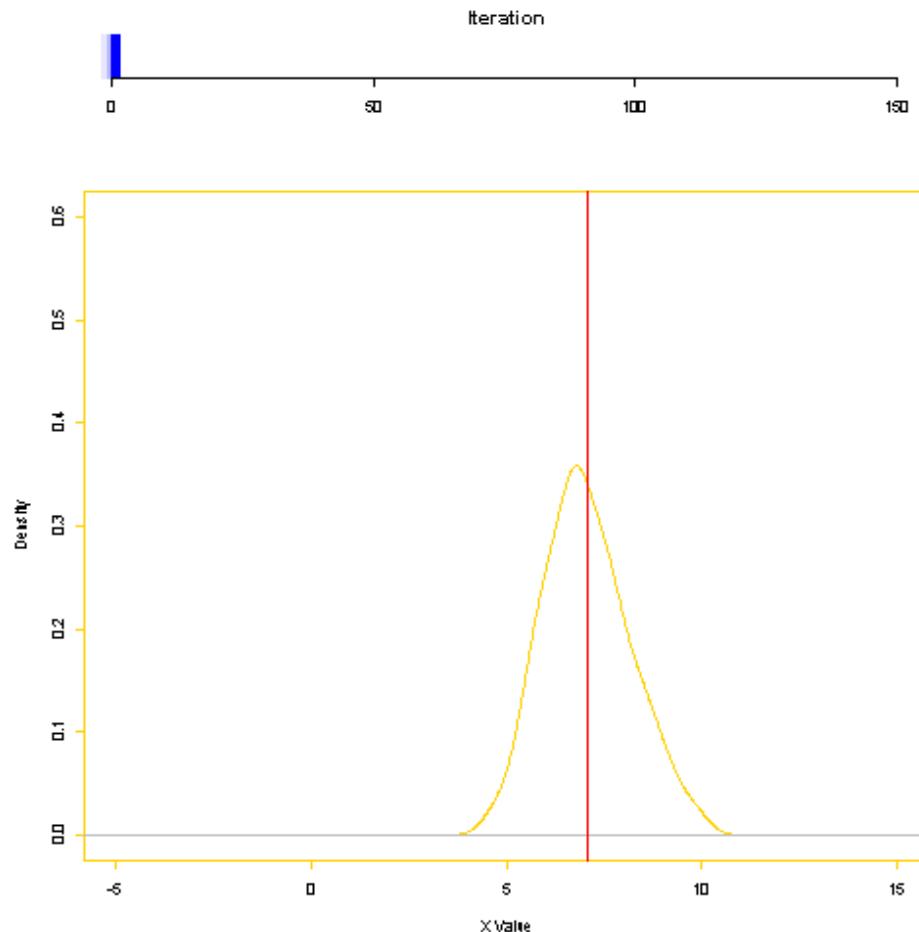
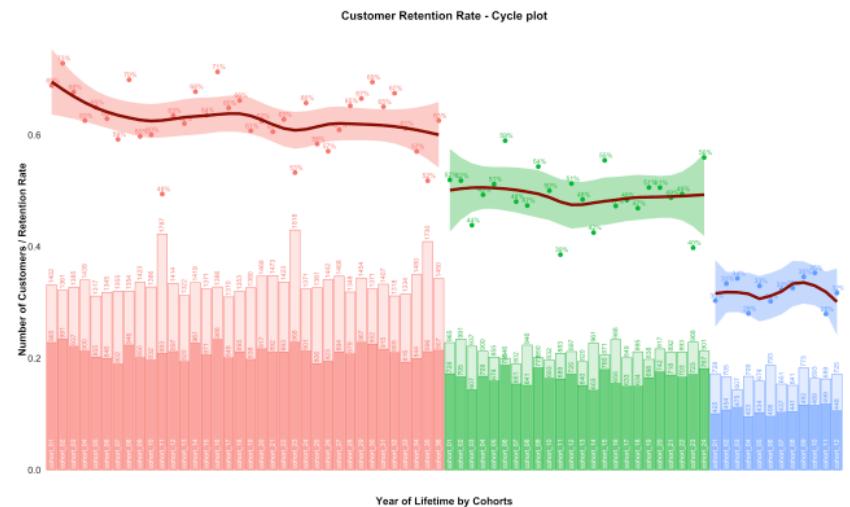
# Why should you learn R

- ▶ Easy to implement techniques such as:
  - ▶ Data/ Text Mining
  - ▶ A/B Testing
  - ▶ Sentiment Analysis
  - ▶ Simulate Data



# Why should you learn R

- ▶ Easy to implement techniques such as:
  - ▶ Data/ Text Mining
  - ▶ A/B Testing
  - ▶ Sentiment Analysis
  - ▶ Simulate Data
  - ▶ Calculate ROI and other financial metrics



# Why should you learn R ...

- ▶ Create beautiful and unique data visualizations
- ▶ Explore and process data from all data sources
- ▶ Add another language/ program to your 'tool kit'
- ▶ Relates to other languages
- ▶ Cross platform compatibility



# Why should you learn R ...

- ▶ Create beautiful and unique data visualizations



# Why should you learn R ...

- ▶ Create beautiful and unique data visualizations
- ▶ Explore and process data from all data sources



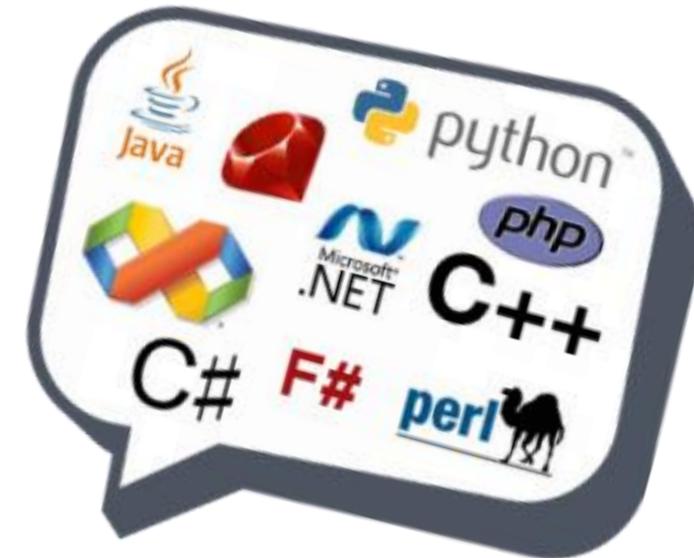
# Why should you learn R ...

- ▶ Create beautiful and unique data visualizations
- ▶ Explore and process data from all data sources
- ▶ Add another language/ program to your 'tool kit'



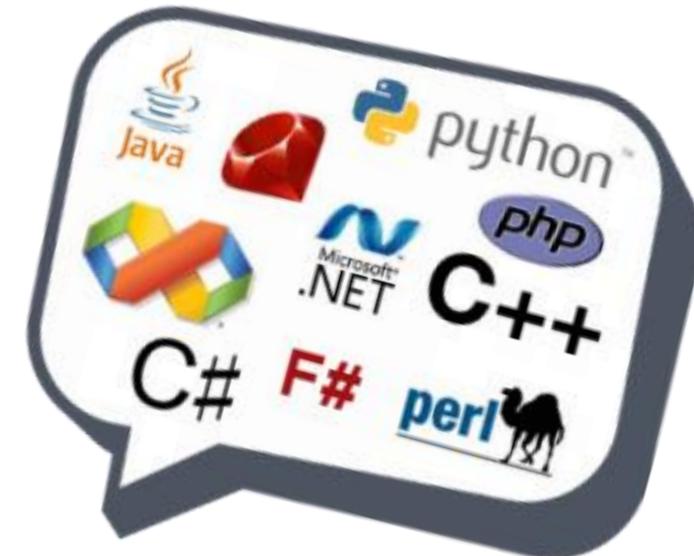
# Why should you learn R ...

- ▶ Create beautiful and unique data visualizations
- ▶ Explore and process data from all data sources
- ▶ Add another language/ program to your 'tool kit'
- ▶ Relates to other languages



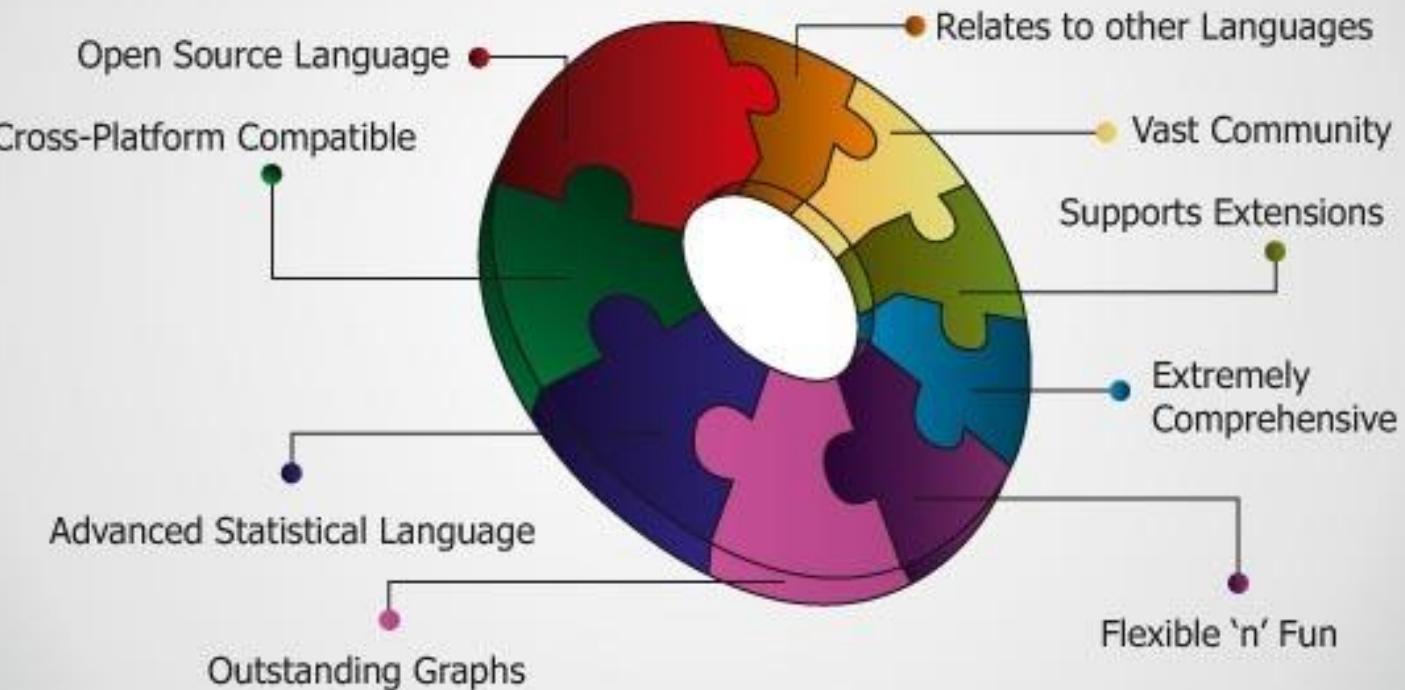
# Why should you learn R ...

- ▶ Create beautiful and unique data visualizations
- ▶ Explore and process data from all data sources
- ▶ Add another language/ program to your 'tool kit'
- ▶ Relates to other languages
- ▶ Cross platform compatibility



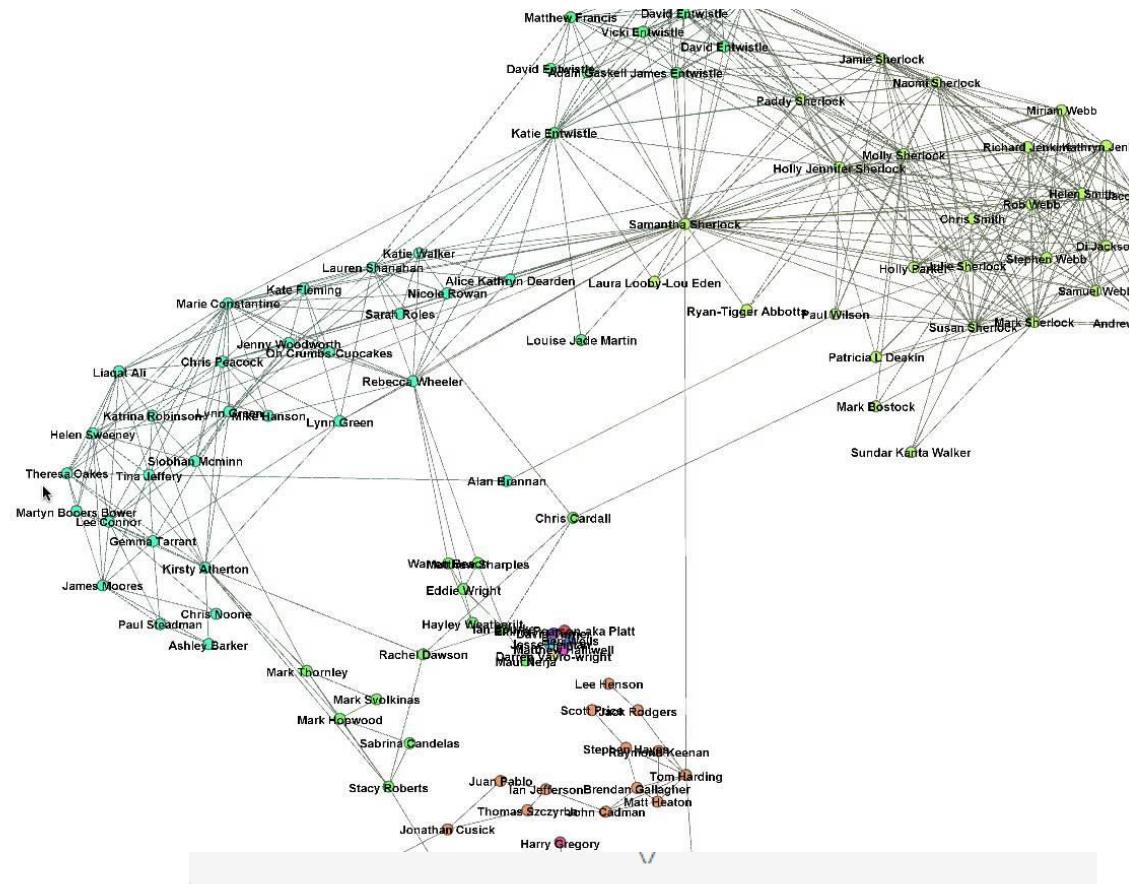
# Why should you learn R ...

## Why Learn R?

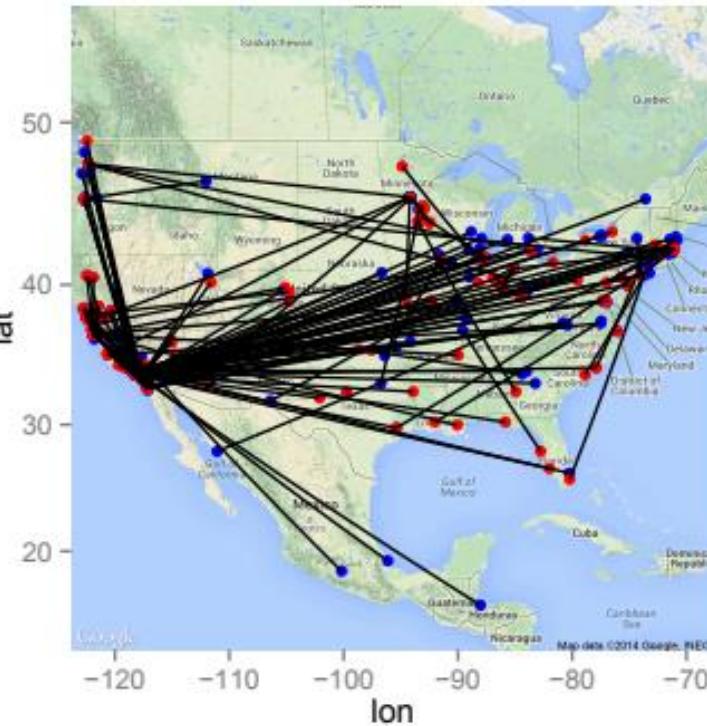


► R is fun!

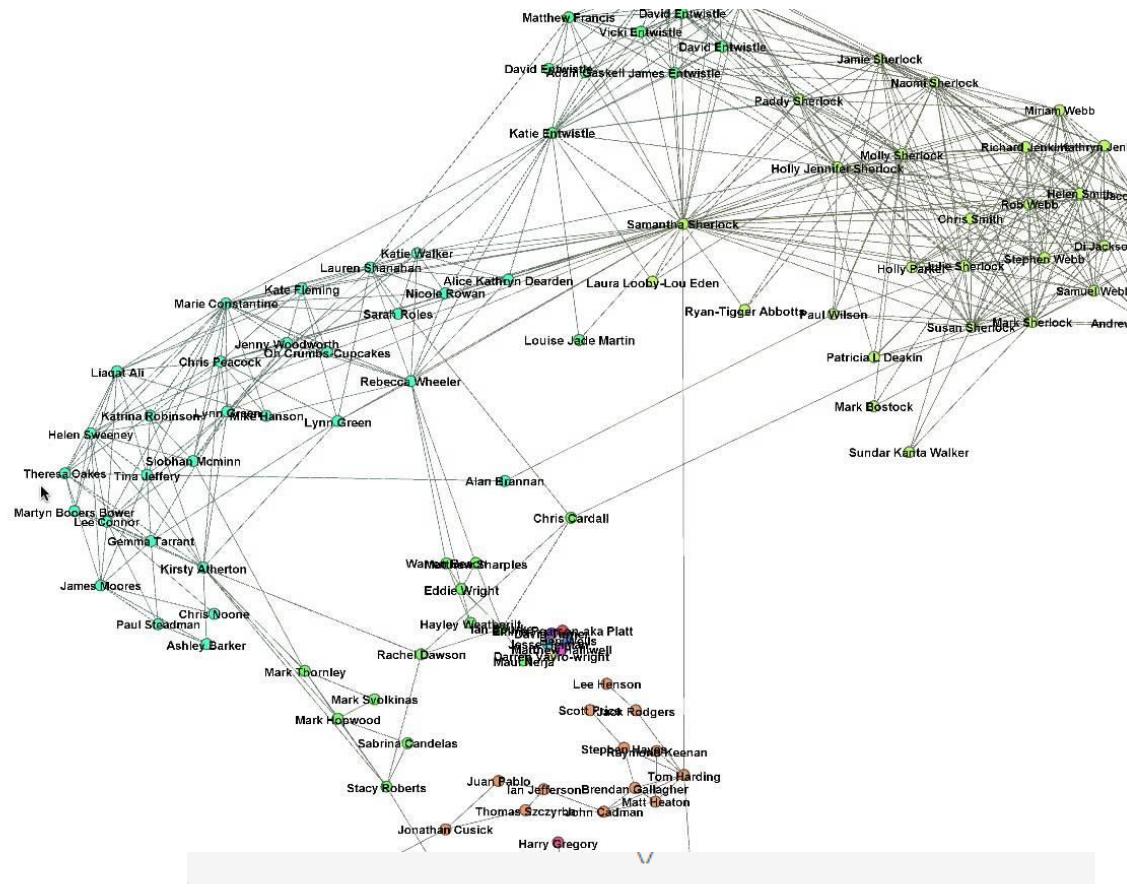
# Why should you learn R ...



- ▶ R is fun!
    - ▶ Integrate with Social Media



# Why should you learn R ...



- ▶ R is fun!
    - ▶ Integrate with Social Media
    - ▶ Play Poker via simulation



# Why should you learn R ...

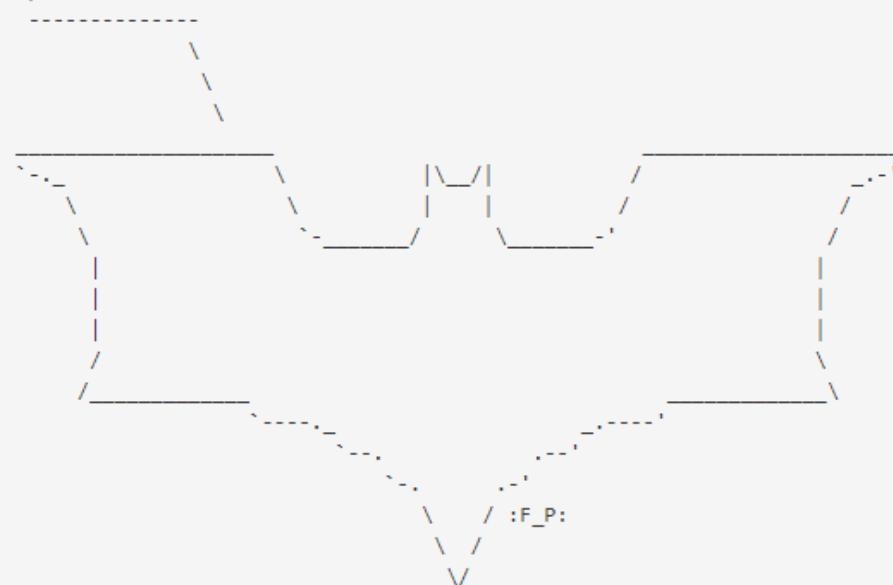
-----  
Only with a very high signal to noise ratio (e.g., high true R^2) can  
torturing data lead to a confession to something other than what the  
analyst wants to hear.

Frank Harrell

NA

R-help

April 2010



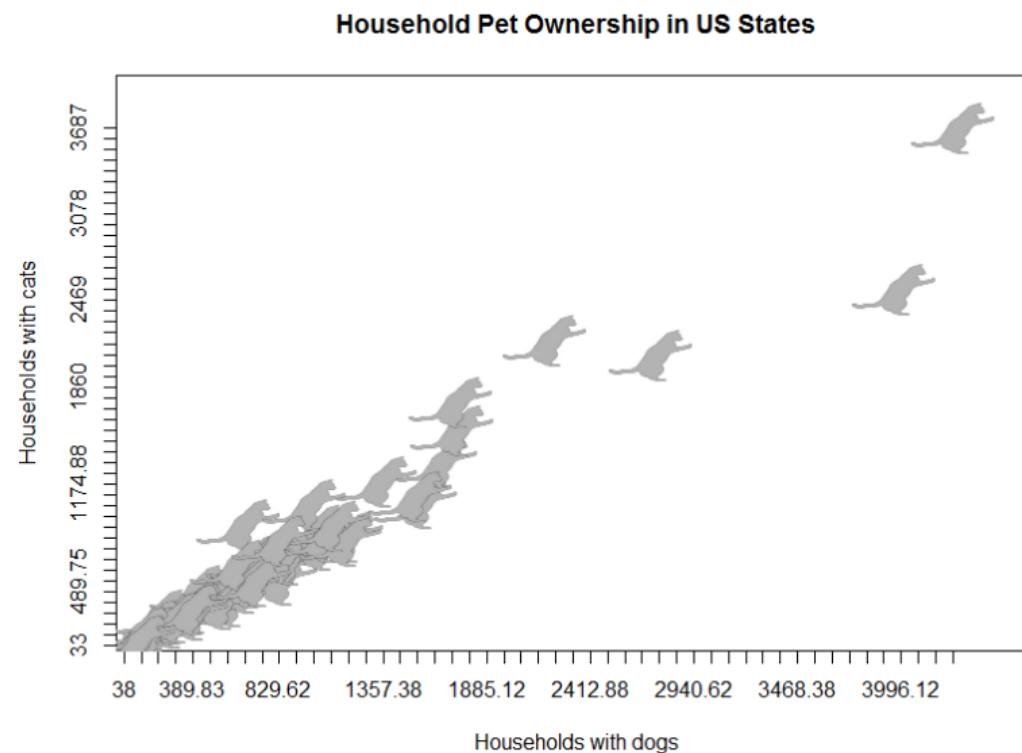
## ► R is fun!

- Integrate with Social Media
- Play Poker via simulation
- Get your fortune from Batman



# Why should you learn R ...

**Even print graphics with cats as data points!**



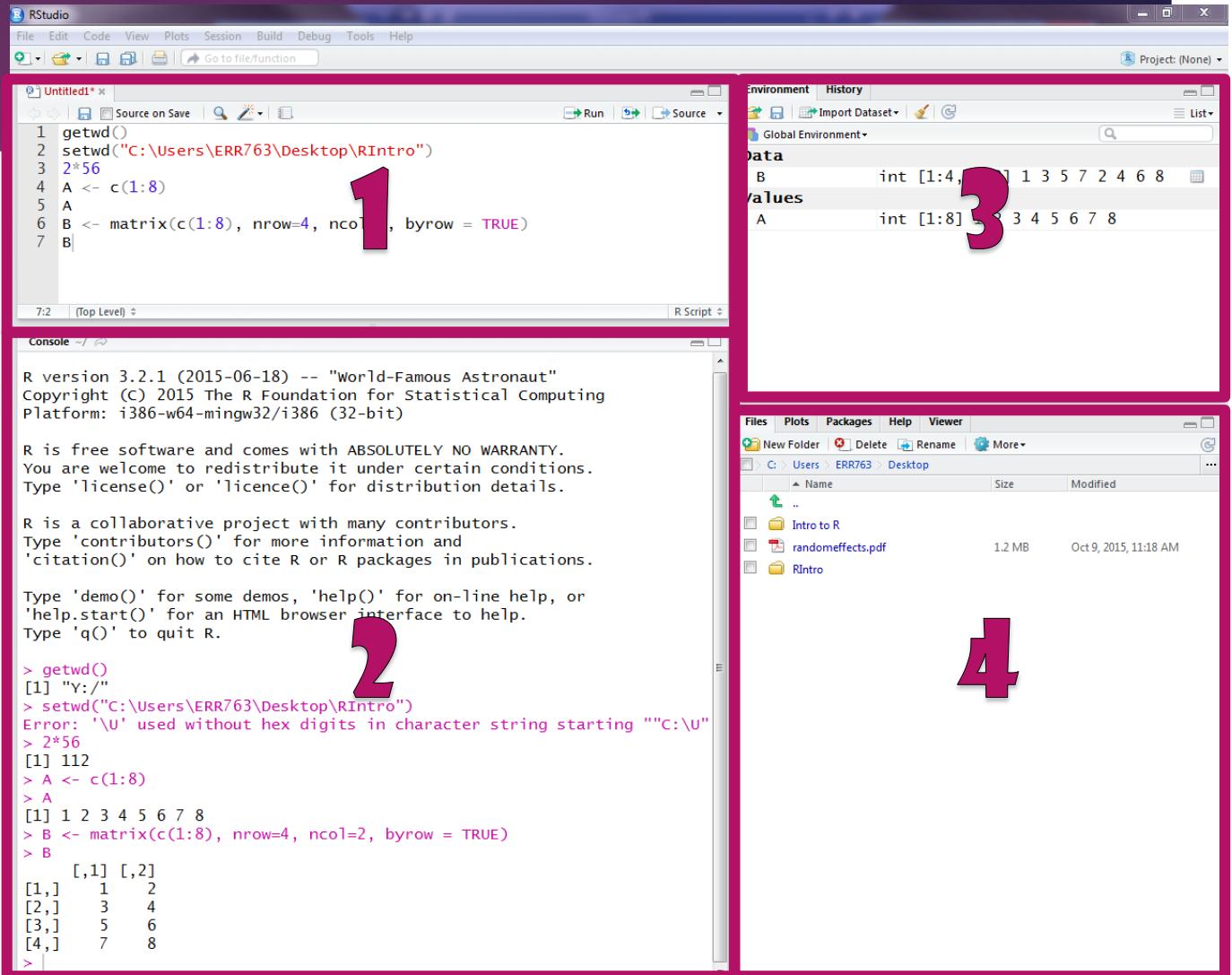
# Overview of RStudio

- ▶ RStudio is a **user interface**
  - ▶ Various tabs to organize graphics, data, file explorer, etc.
  - ▶ 4 windows for easy navigation
- ▶ RStudio runs R in the **background**
  - ▶ Note that you need to install R before installing RStudio
- ▶ RStudio works with Shiny and RMarkdown to create interactive web application and documentation
- ▶ RStudio is a prettier and more user-friendly environment to run R.



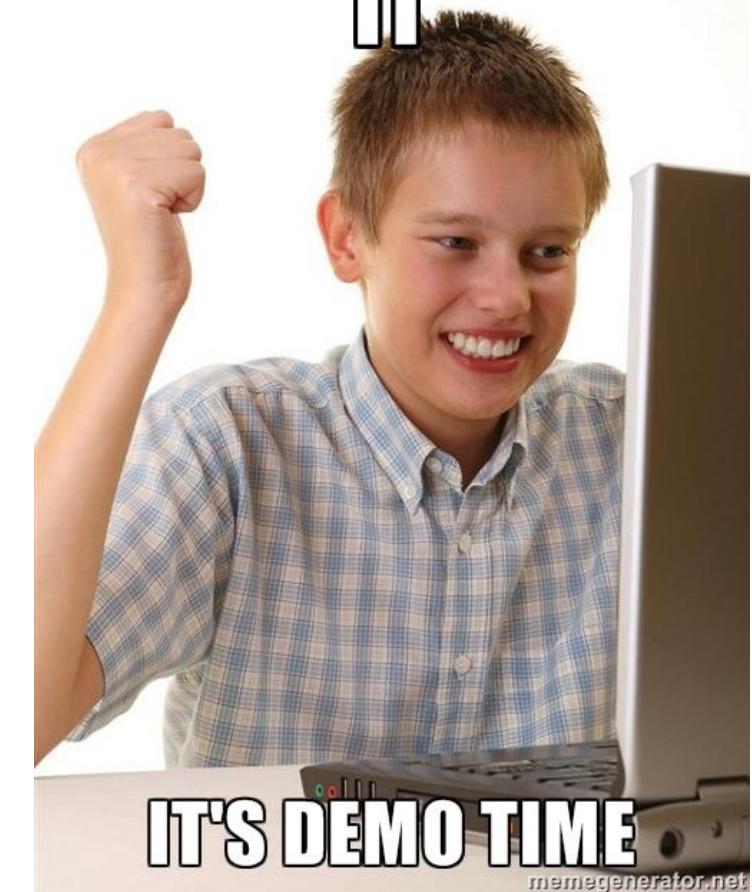
# Overview of RStudio ...

1. Source
2. Console
3. Environment and History
4. Files, Plots, Packages, Help, Viewer



Demo  
(If Time)

I CAN'T BELIEVE  
IT



Thank You!

EMILY.JONES@SONY.COM



# Useful Resources

- [WWW.RSTUDIO.COM/RESOURCES/CHEATSHEETS/](http://www.rstudio.com/resources/cheatsheets/)
- [WWW.COURSERA.COM](http://www.coursera.com)
- [WWW.DATACAMP.COM](http://www.datacamp.com)
- R FOR DATA SCIENCE & GGPLOT2  
(BY HADLEY WICKHAM)
- AN INTRODUCTION TO STATISTICAL LEARNING (BY GARETH JAMES)
- [EMILY.JONES@SONY.COM](mailto:EMILY.JONES@SONY.COM)
- THE INTERNET!