R Programming for Data Science

EMILY JONES

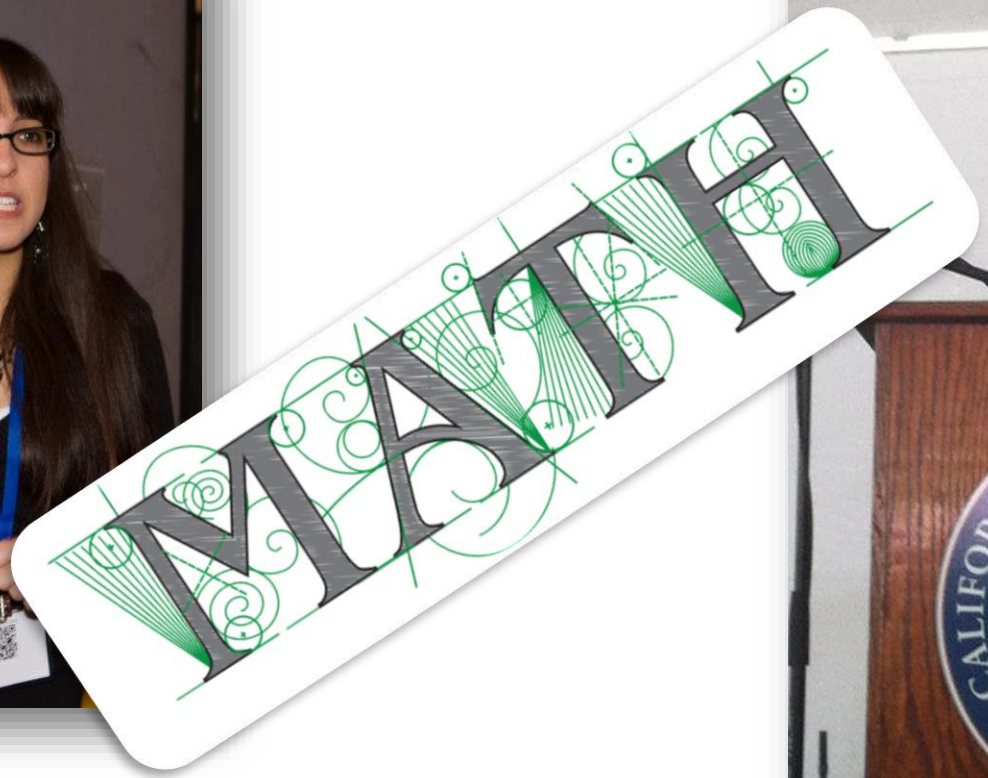
DATA SCIENTIST

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



Who am I – Education…

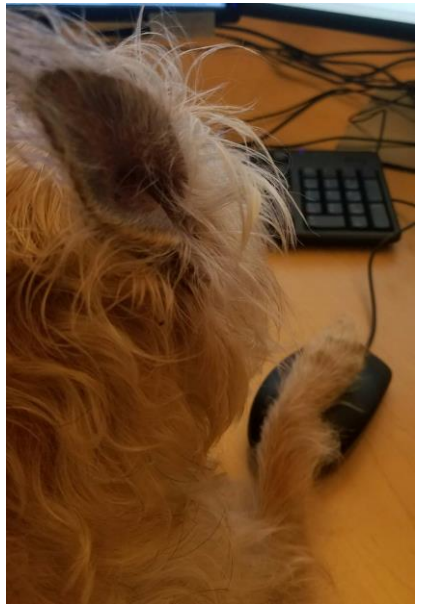
UNIVERSITY OF MASSACHUSETTS AMHERST 1863

UMASS AMHERST

Bridging the sciences

BIOSTATISTICS

Who am I – Career Path

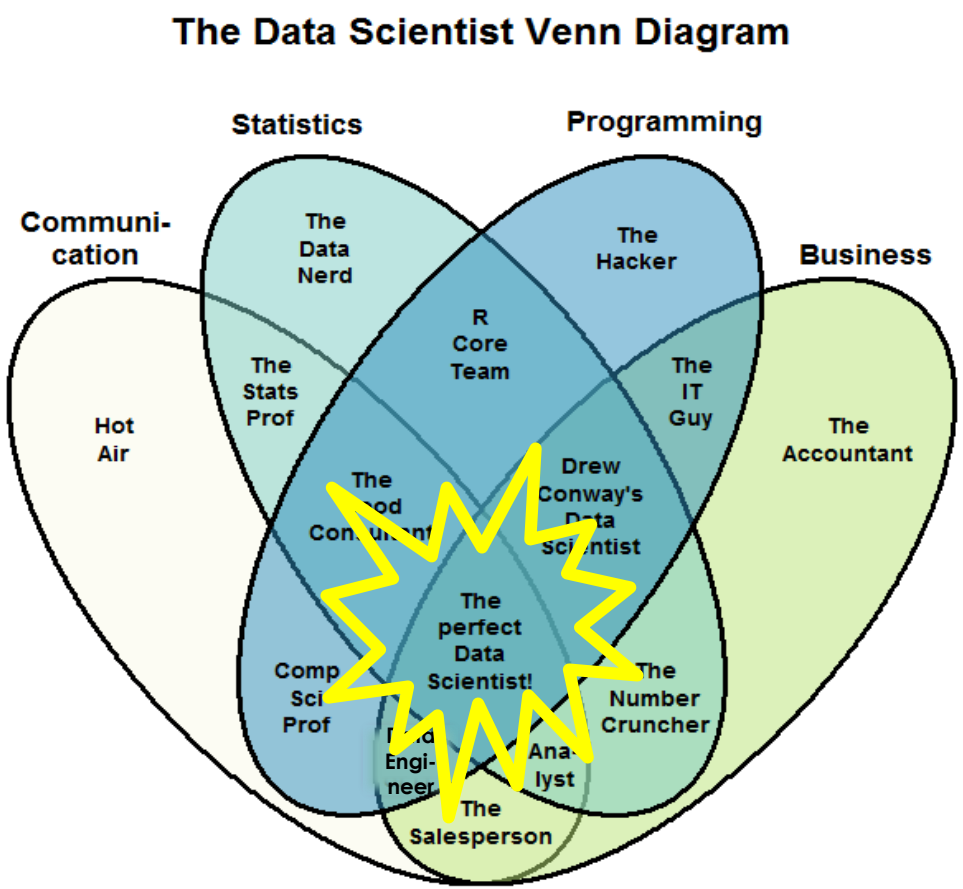


The Hanover Insurance Group®

Rockstar Games

What is a Data Scientist – Venn Diagram

- **Statistics/Math**
  - Basic Stats
  - Modeling
  - Machine Learning
- **Communication**
  - Explain complex concepts
  - Visualization
  - Presentation



- **Programming**
  - Querying
  - Statistical Language
  - Technical Logic
- **Business Acumen**
  - Know your business
  - Fill Business Gaps
  - Passionate

What is a Data Scientist – Process

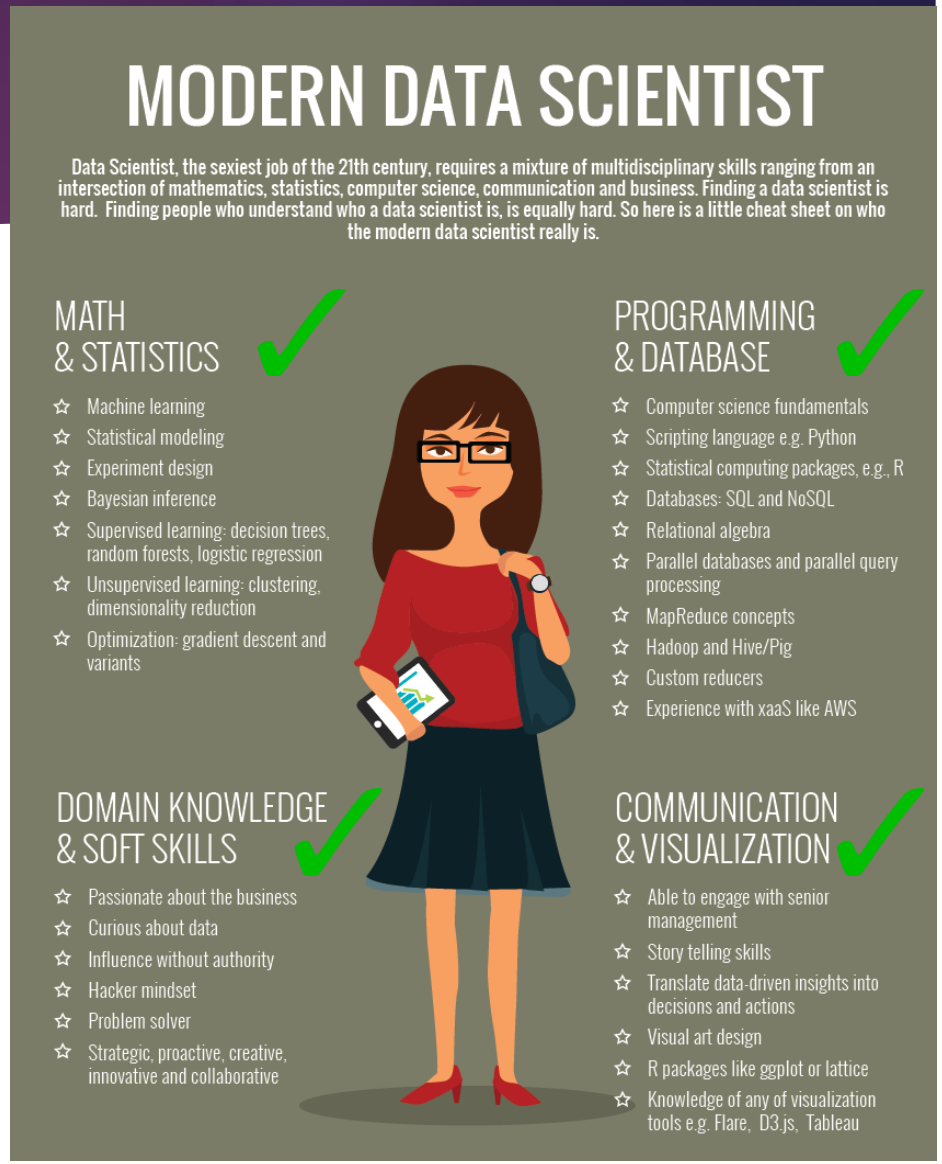
Data Science Process

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.

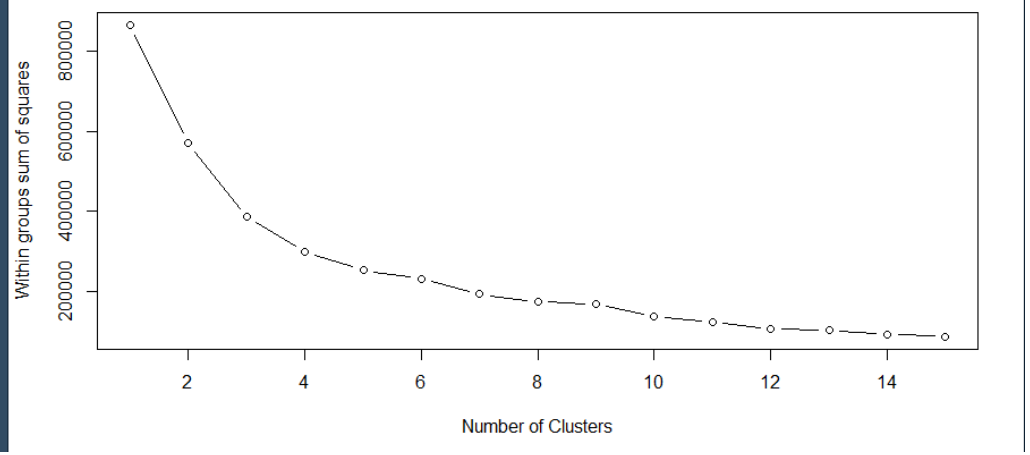


Coding, coding, make an interesting insight, coding, coding

```
dat$avg_spent_scale <- as.numeric(scale(dat$avg_spent))

####################################
## Running the Clusters            ##
####################################

# Winner winner chicken dinner (this is the clustering approach chosen)
# look at purchases, revenue and average

# subsetting the data
mydata <- as.data.frame(cbind(dat$totalpurchases_scale, dat$totalrevenue_scale, dat$avg_spent_scale))
colnames(mydata) <- c('totalpurchases_scale', 'totalrevenue_scale', 'avg_spent_scale')

# Checking the Elbow Chart
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(mydata,
                                     centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",
     ylab="Within groups sum of squares")

### 3 clusters look good on the elbow plot, 4 looks better

# Create Matrix to run cluster
m <- as.matrix(cbind(dat$totalpurchases_scale,dat$totalrevenue_scale,dat$avg_spent_scale
# running kmeans cluster
cl <- (kmeans(m,4))

# Loading this info back into our data set
dat$cluster <- factor(cl$cluster)
centers <- as.data.frame(cl$centers)

#plotting the clusters
ggplot(data=dat, aes(x=totalpurchases_scale, y=totalrevenue_scale, color=cluster)) +
  theme_bw() +
  geom_point() +
  geom_point(data=centers, aes(x=V1,y=V2, color='Center')) +
  geom_point(data=centers, aes(x=V1,y=V2, color='Center'), size=15, alpha=.3, show.legend = FALSE)
```

```
> dat$cluster <- factor(cl$cluster)
> centers <- as.data.frame(cl$centers)
>
> #plotting the clusters
> ggplot(data=dat, aes(x=totalpurchases_scale, y=totalrevenue_scale, color=cluster)) +
+   theme_bw() +
+   geom_point() +
+   geom_point(data=centers, aes(x=V1,y=V2, color='Center')) +
+   geom_point(data=centers, aes(x=V1,y=V2, color='Center'), size=15, alpha=.3, show.legend = FALSE)
>
```

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

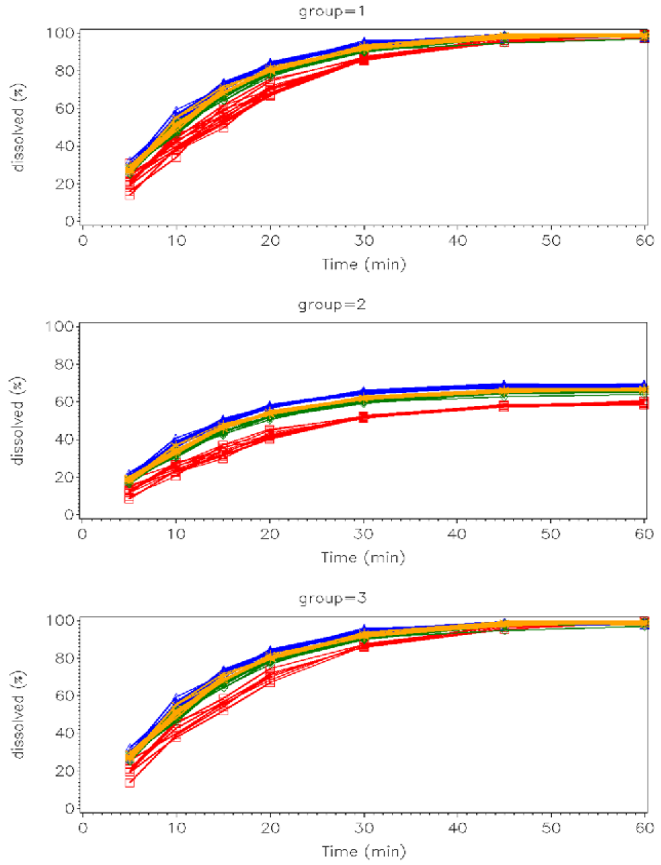
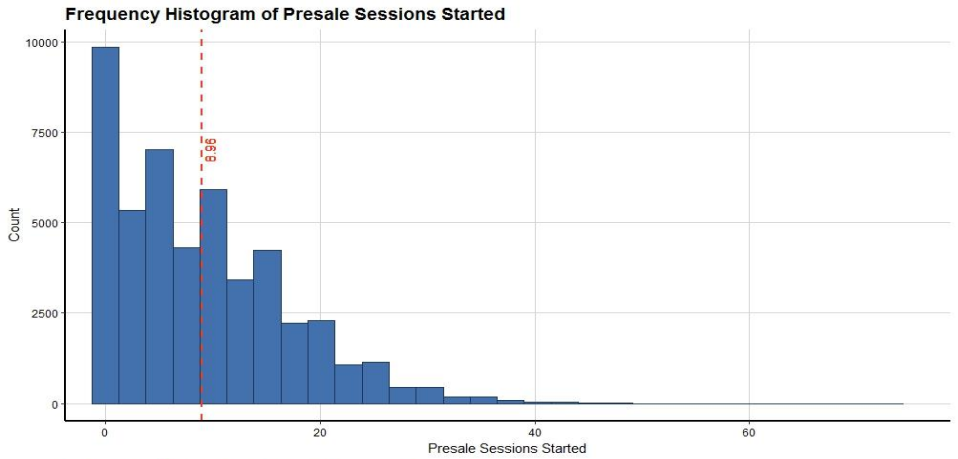
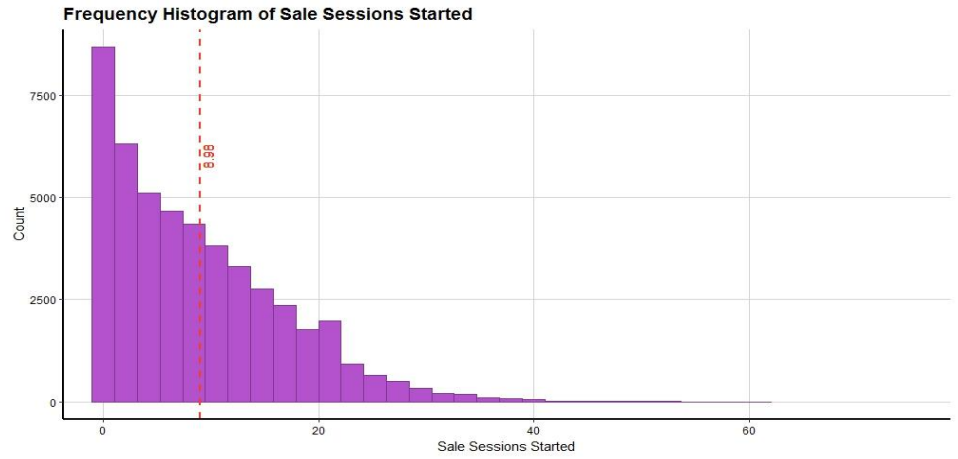
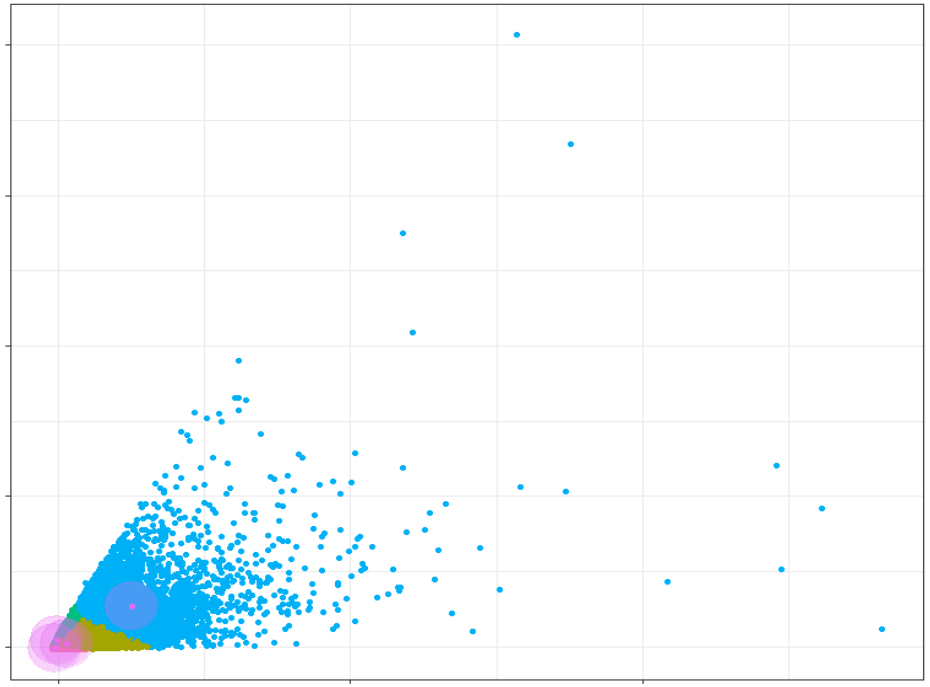


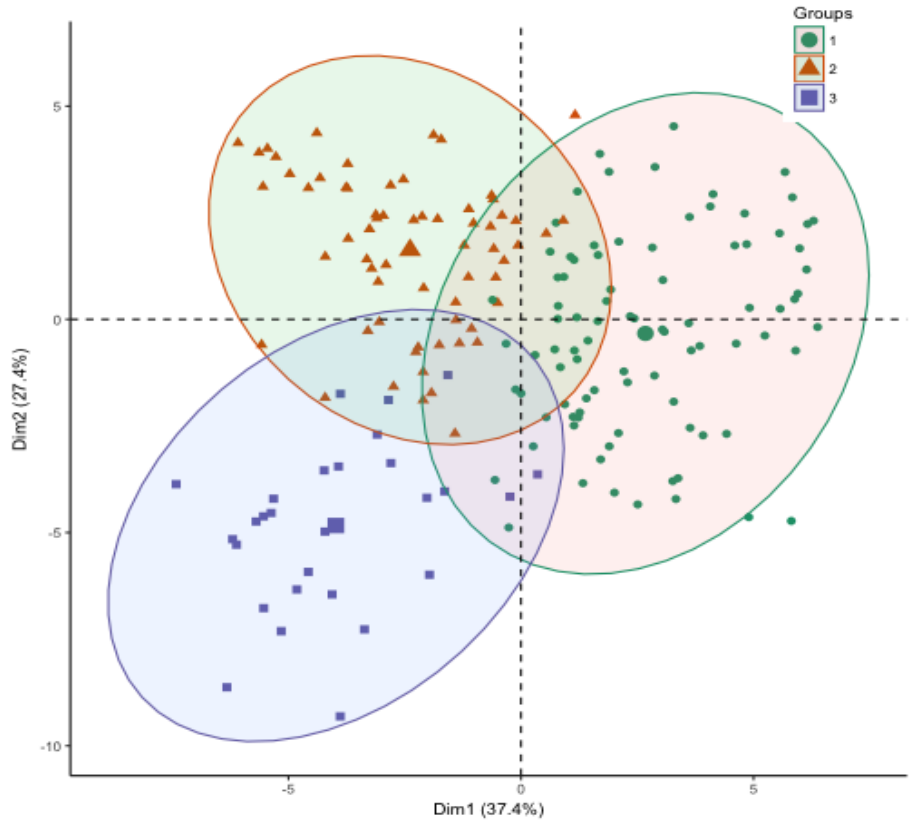
Fig. 1. *In vitro* dissolution outside f_2 prerequisite for variability of the



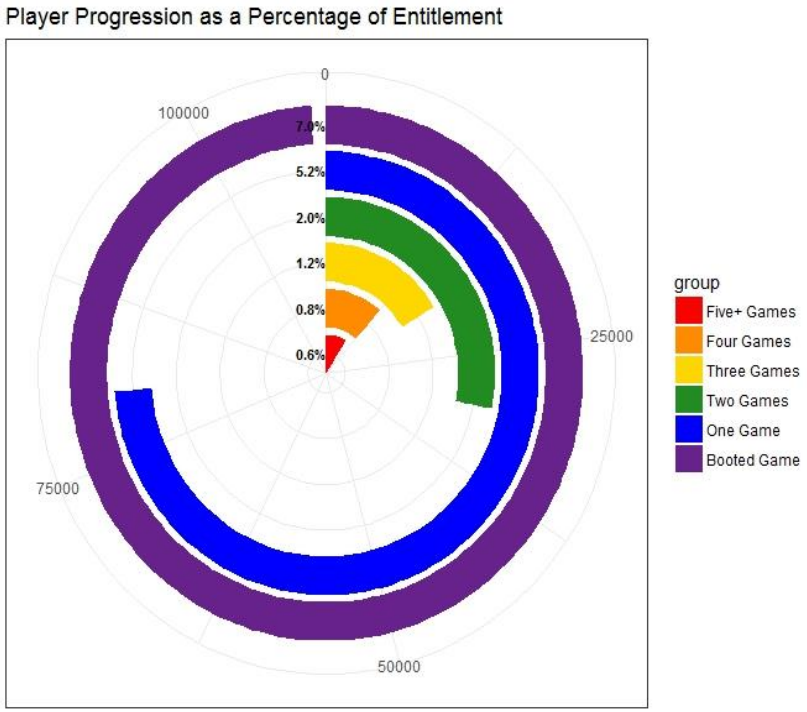


What Do I Do - Create Player Clusters



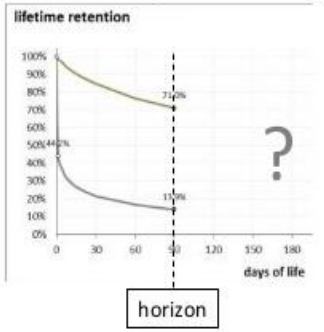


What Do I Do – Build Retention Models

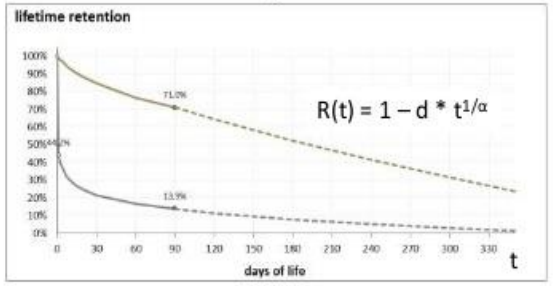


Lifetime Retention model

Life to date operation of the game



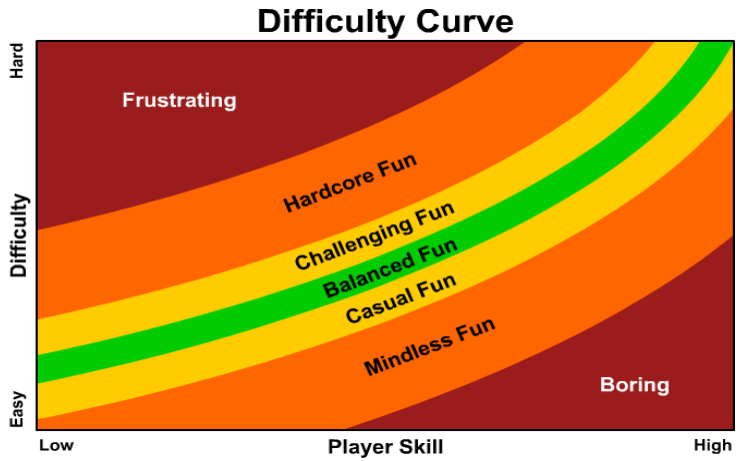
modeling retention curves

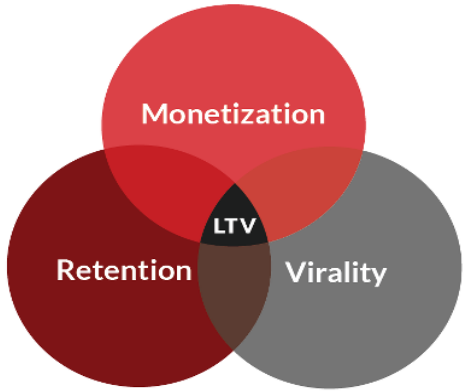


parameters **d** and **α** are found with estimation techniques

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

What Do I Do - Also…





What Do I Do – My Tools

amazon
REDSHIFT

python

RStudio®

+ableau®

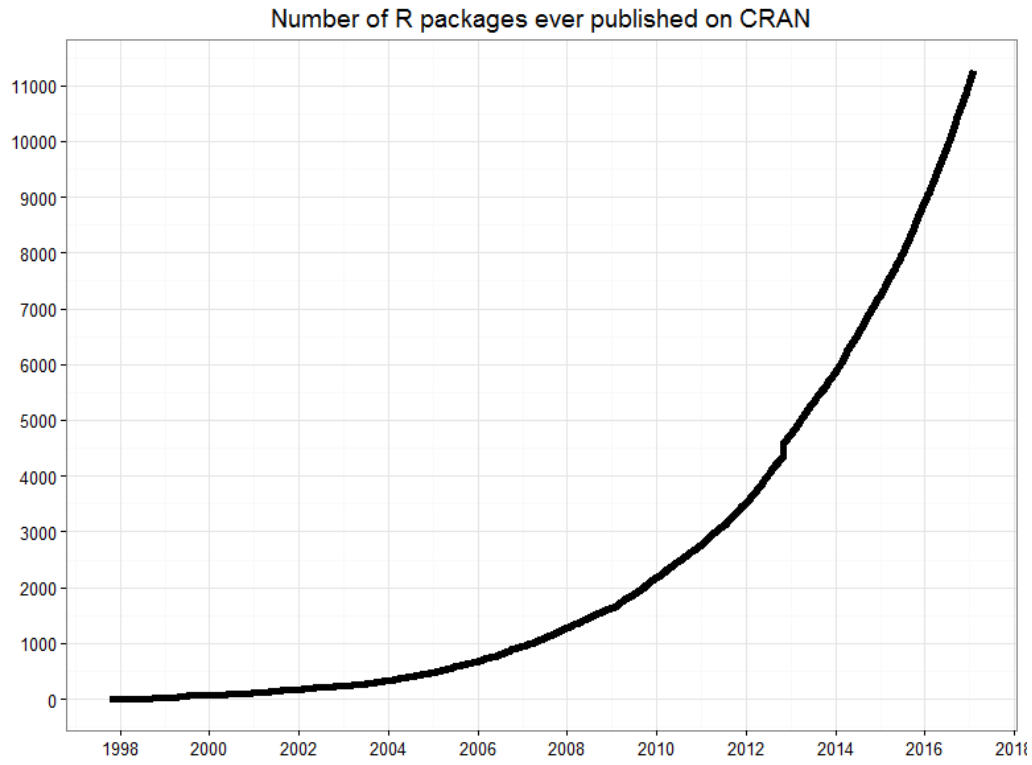
What is R

- 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

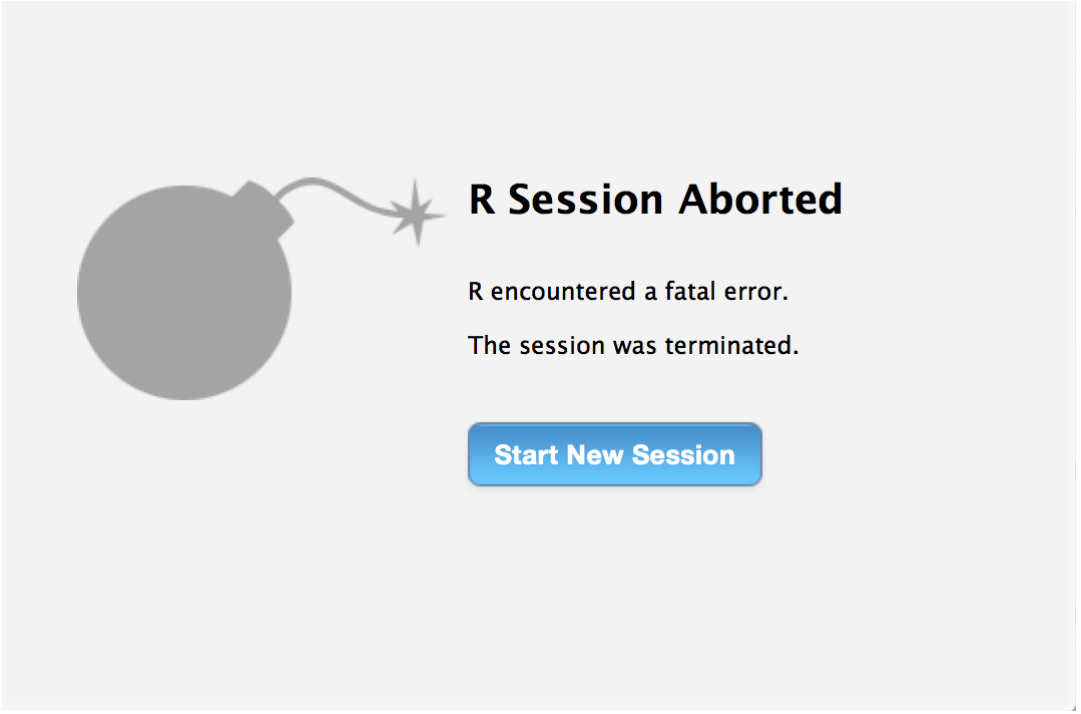
What is R …

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

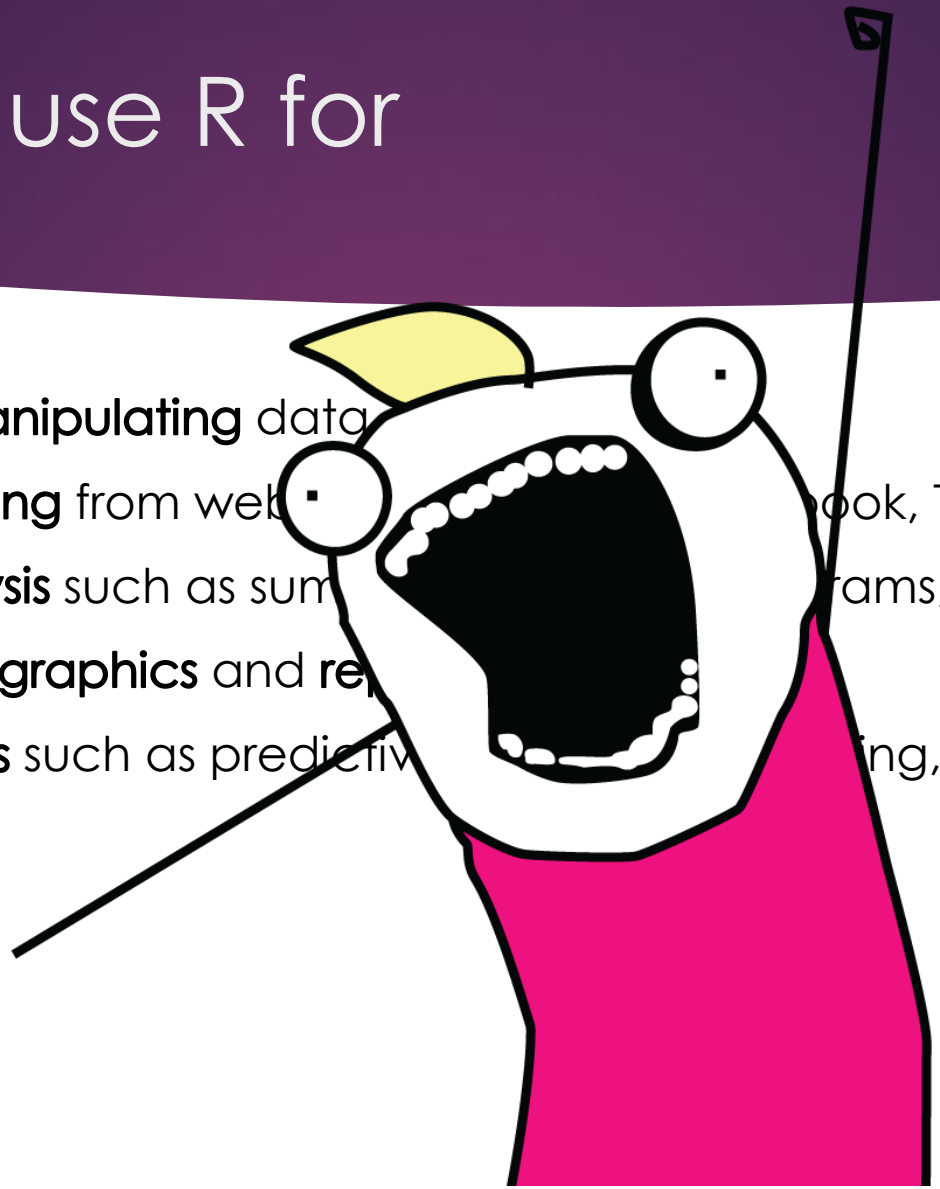
- Not user friendly, has a steep learning curve, **if you don't understand coding logic!**
- No professional / commercial support, **but you have an active community for support!**
- It is not a database, **but connects to DBMS!**
- Working with large datasets is limited by RAM, **sorry no "buts" here**… ☹
- 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**!

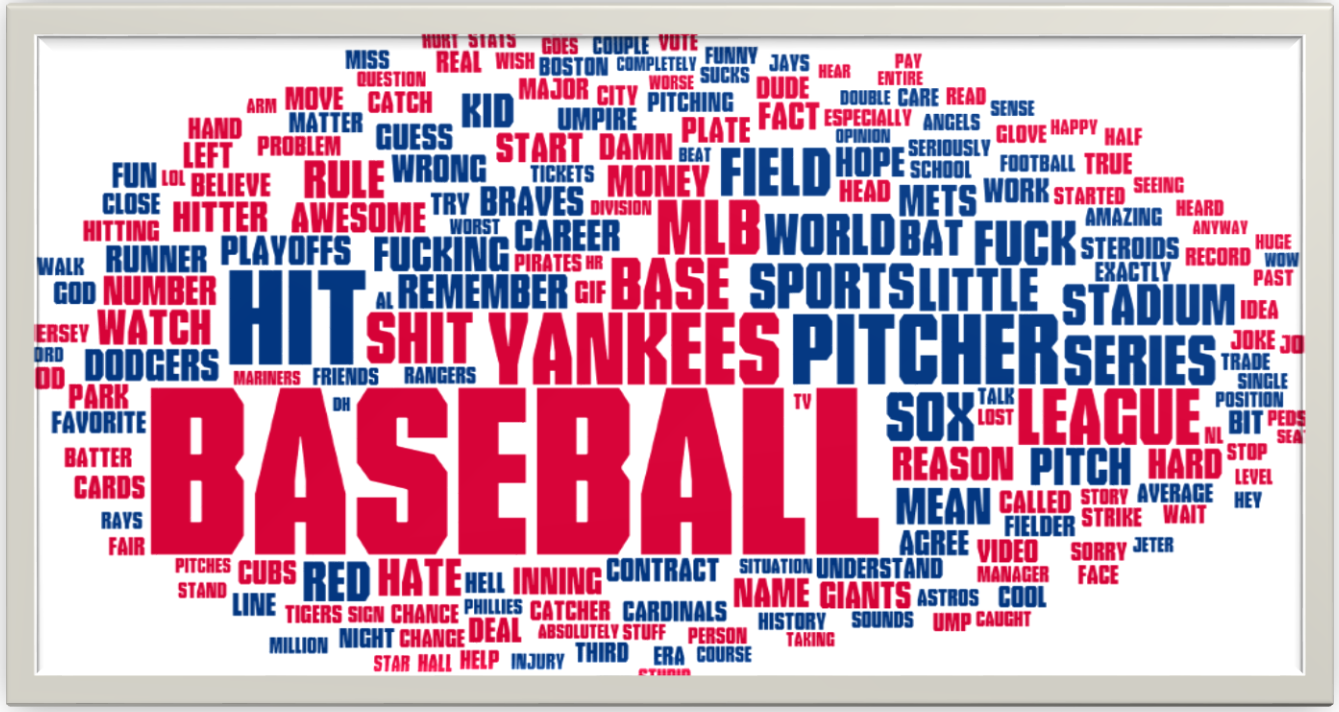
What do I use R for

- **Cleaning** and **manipulating** data
- **Data/ Text Scraping** from web ook, Twitch, etc.
- **Exploratory Analysis** such as sum rams, frequency tables, etc.
- Create beautiful **graphics** and **re**
- **Statistical Analysis** such as predictiv ing, machine learning, etc.
- All the **things**!



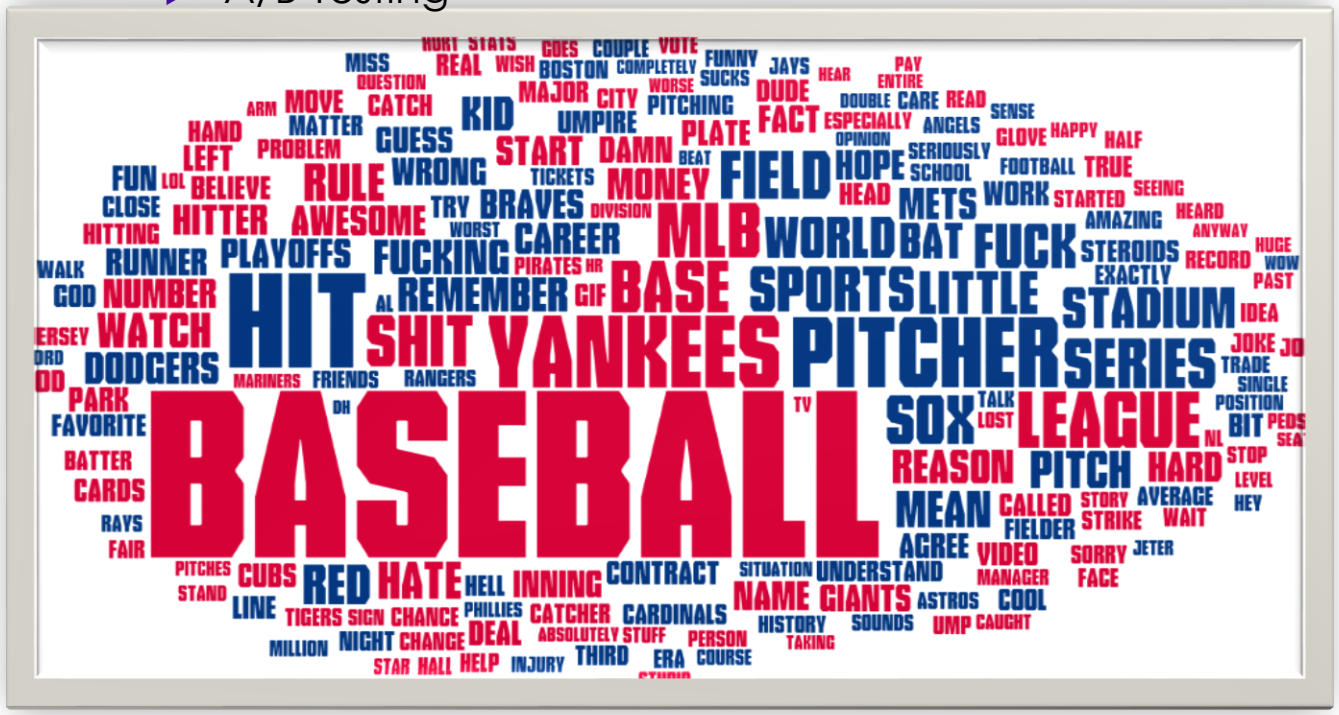
Why should you learn R

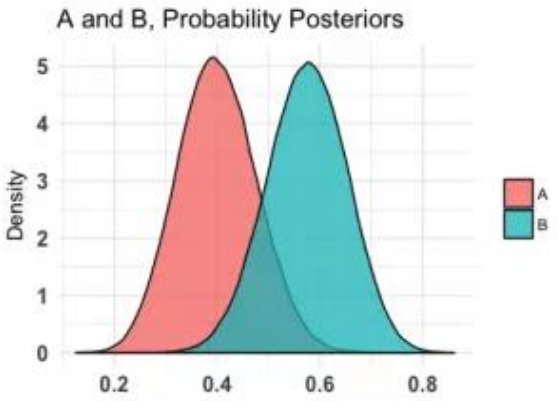
- Easy to implement techniques such as:
  - Data/ Text Mining

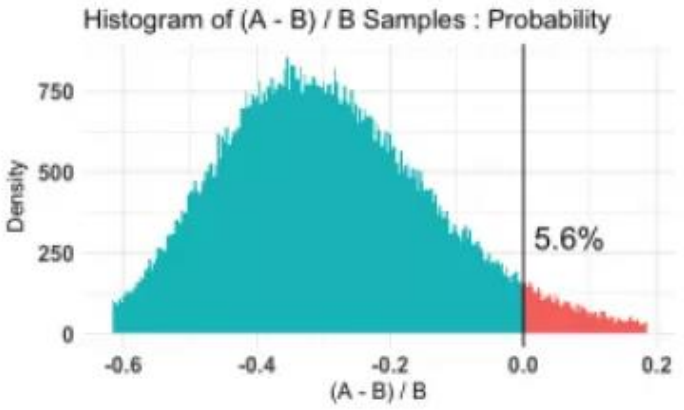


Why should you learn R

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

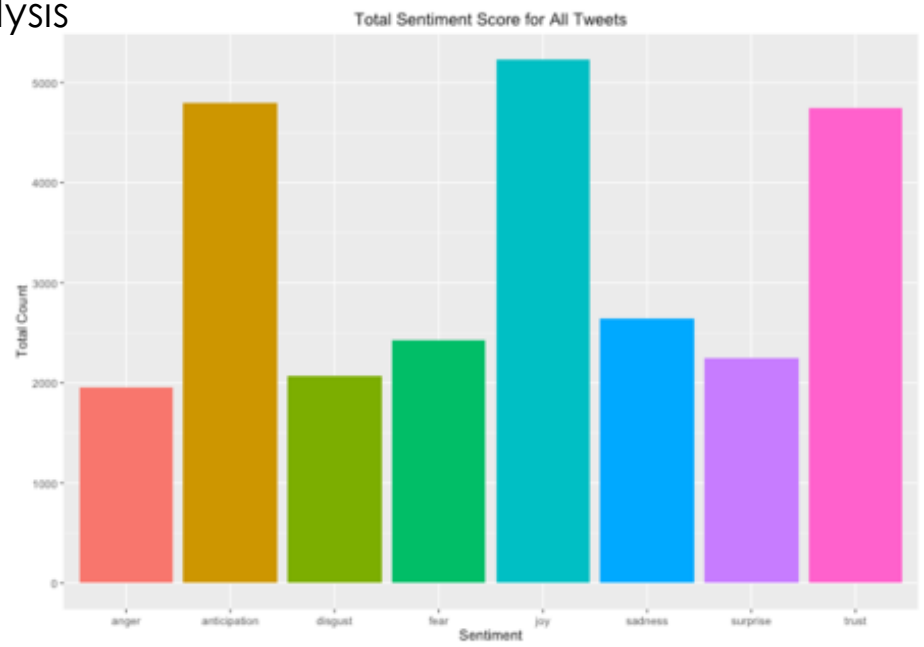


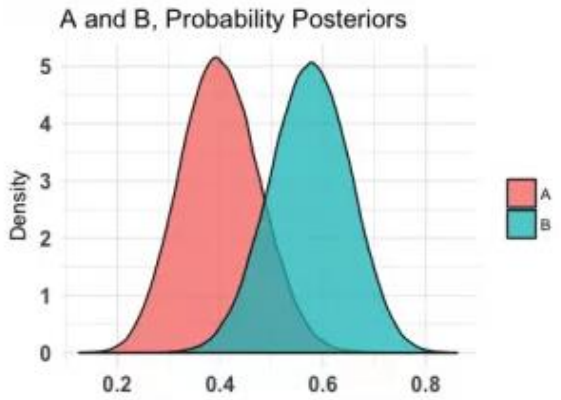


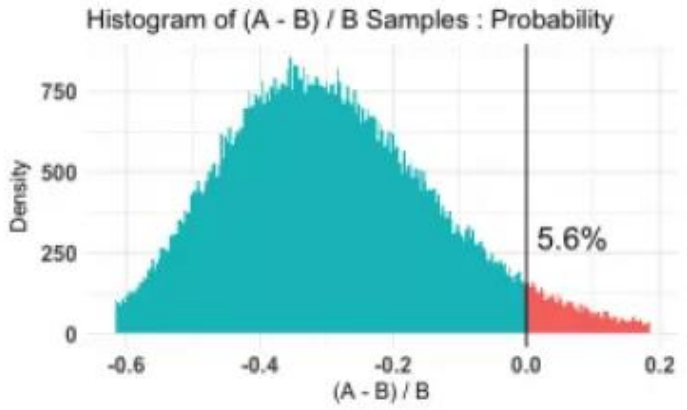


Why should you learn R

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

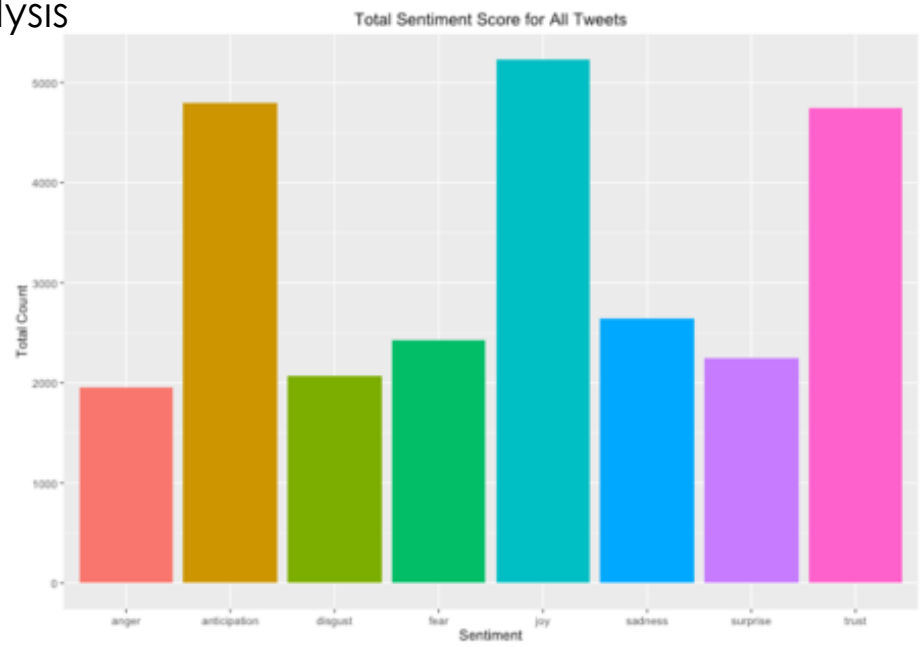


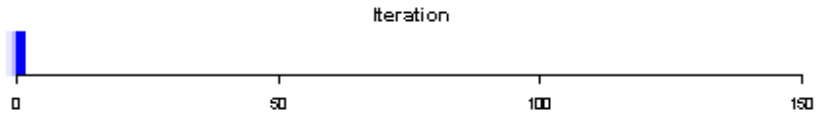


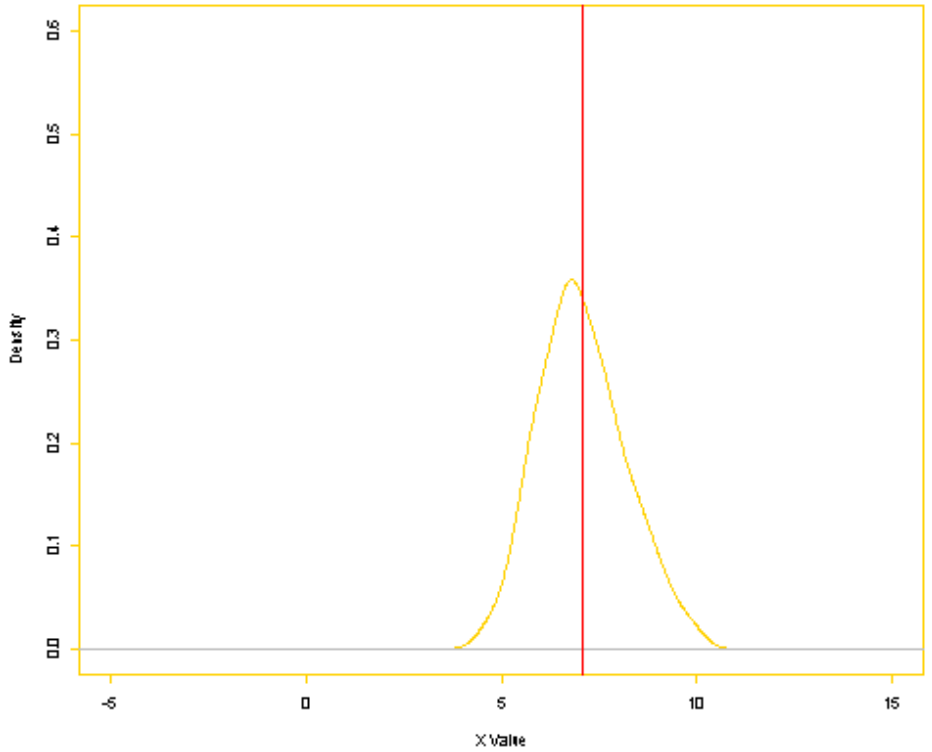


Why should you learn R

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

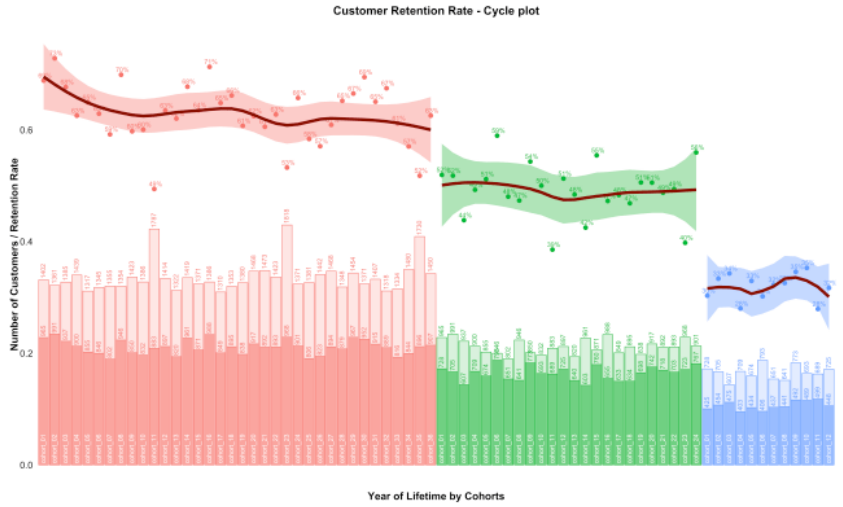


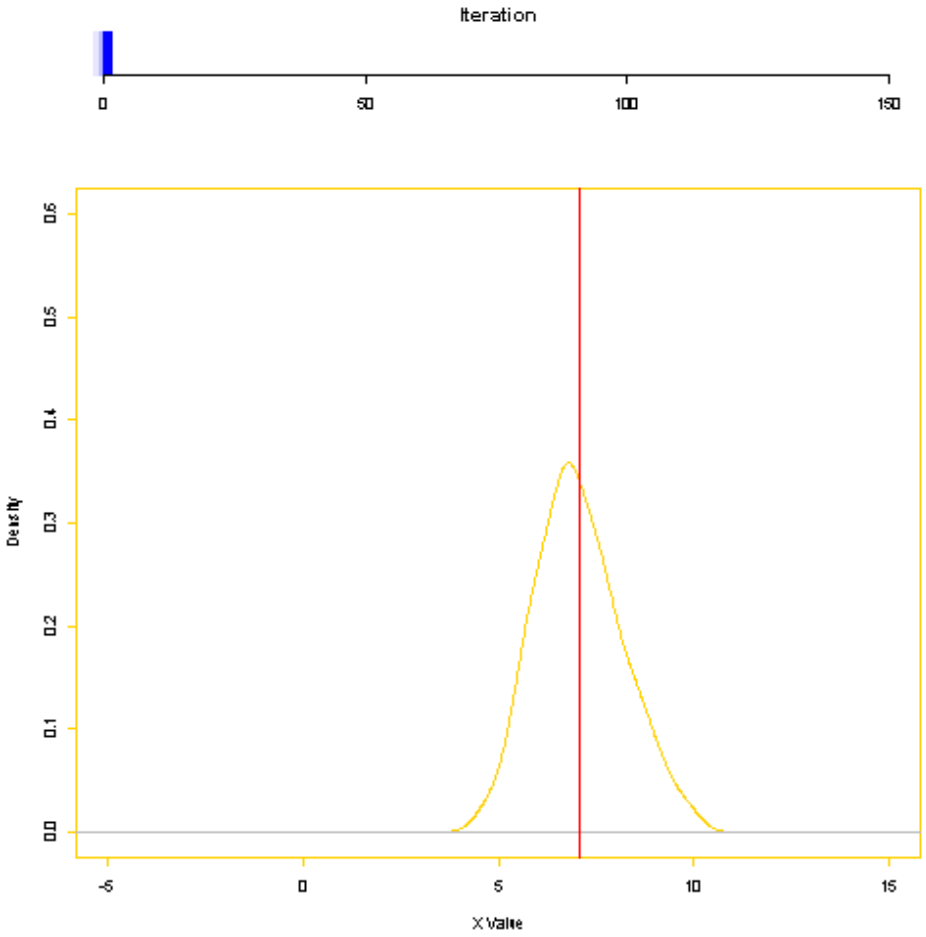




Why should you learn R

- 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’

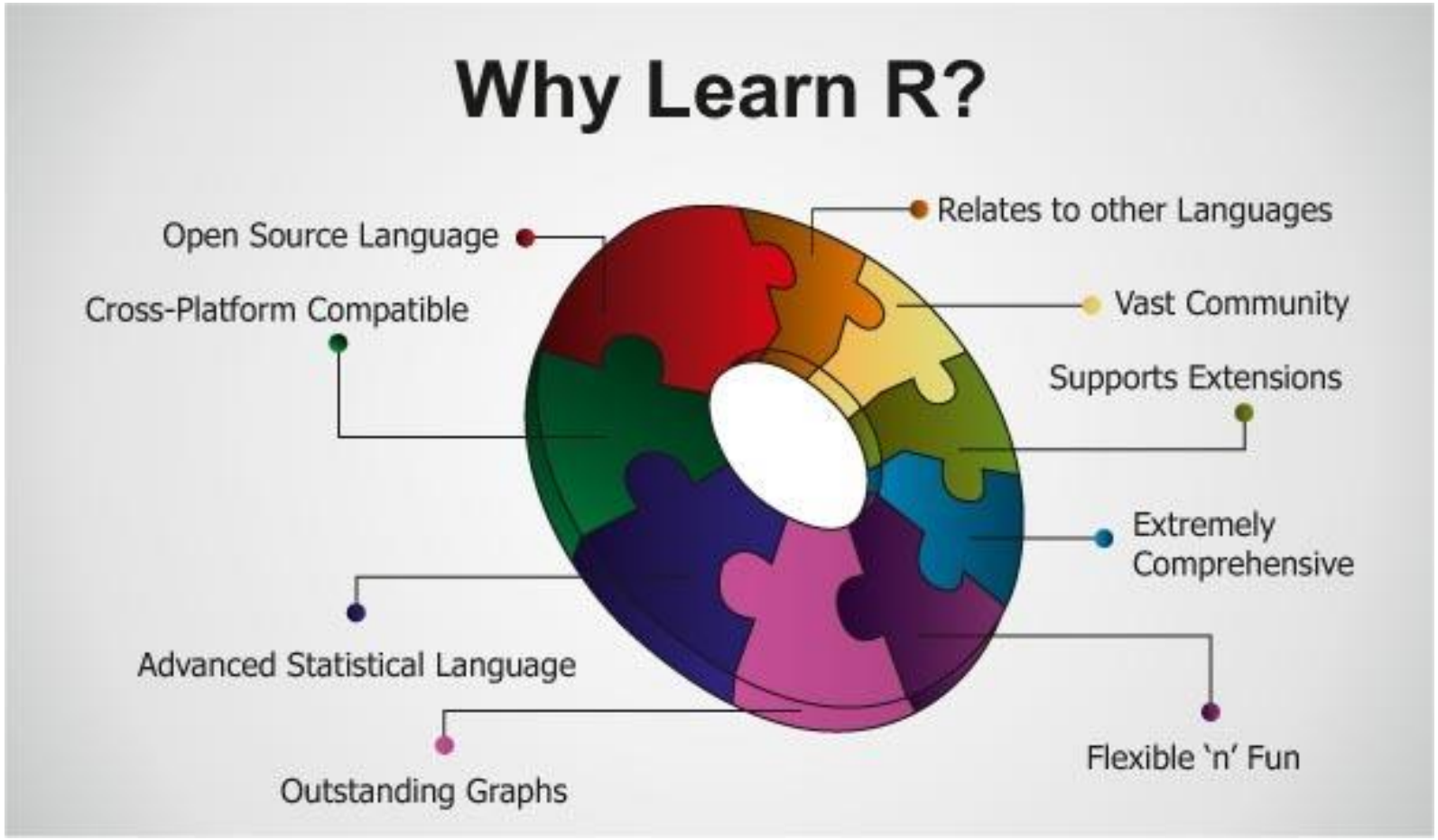
Why should you learn R …

- 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 …

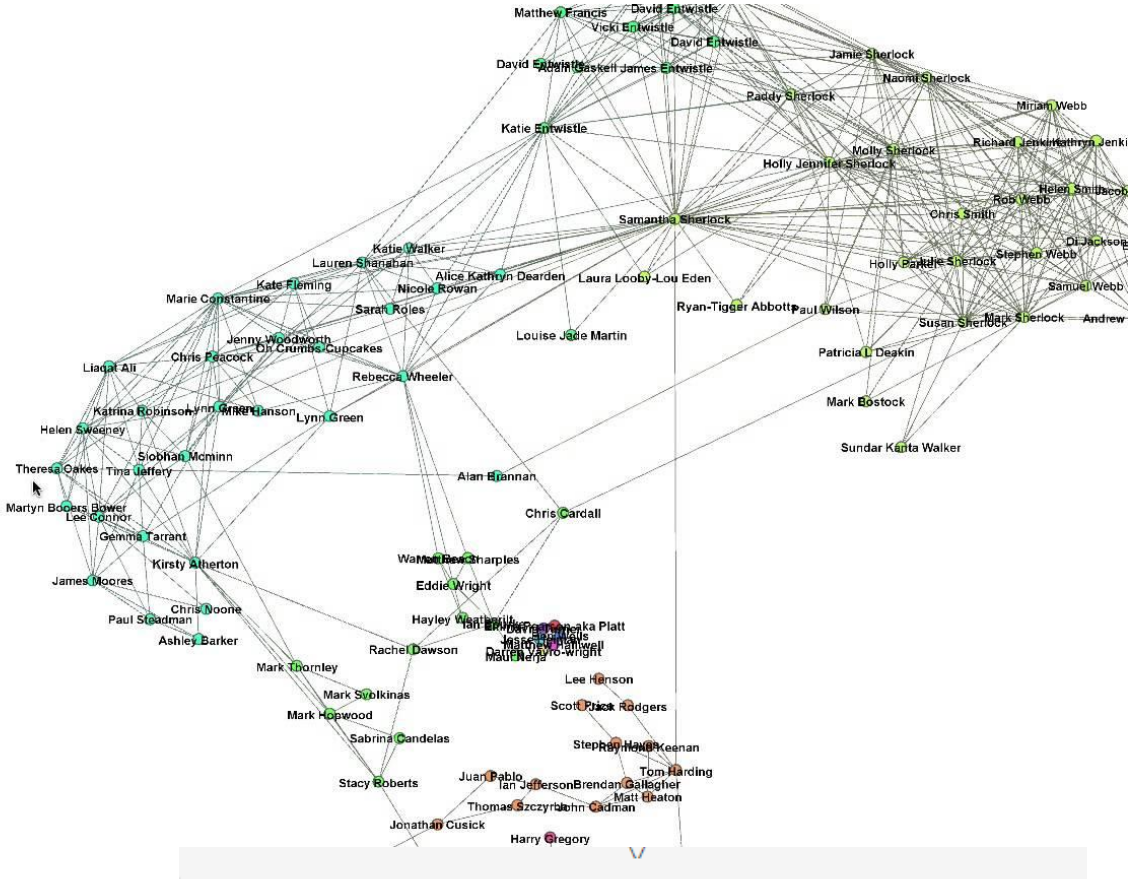
- Create beautiful and unique data visualizations
- Explore and process data from all data sources
- Add another language/ program to your ‘tool kit’
- Relates to other languages
- Cross platform compatibility

Why should you learn R …

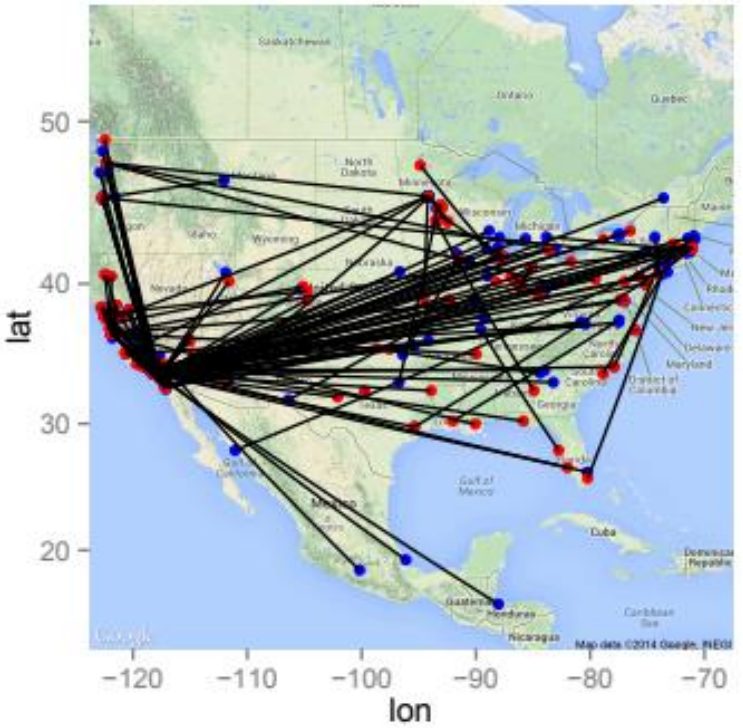


- R is fun!

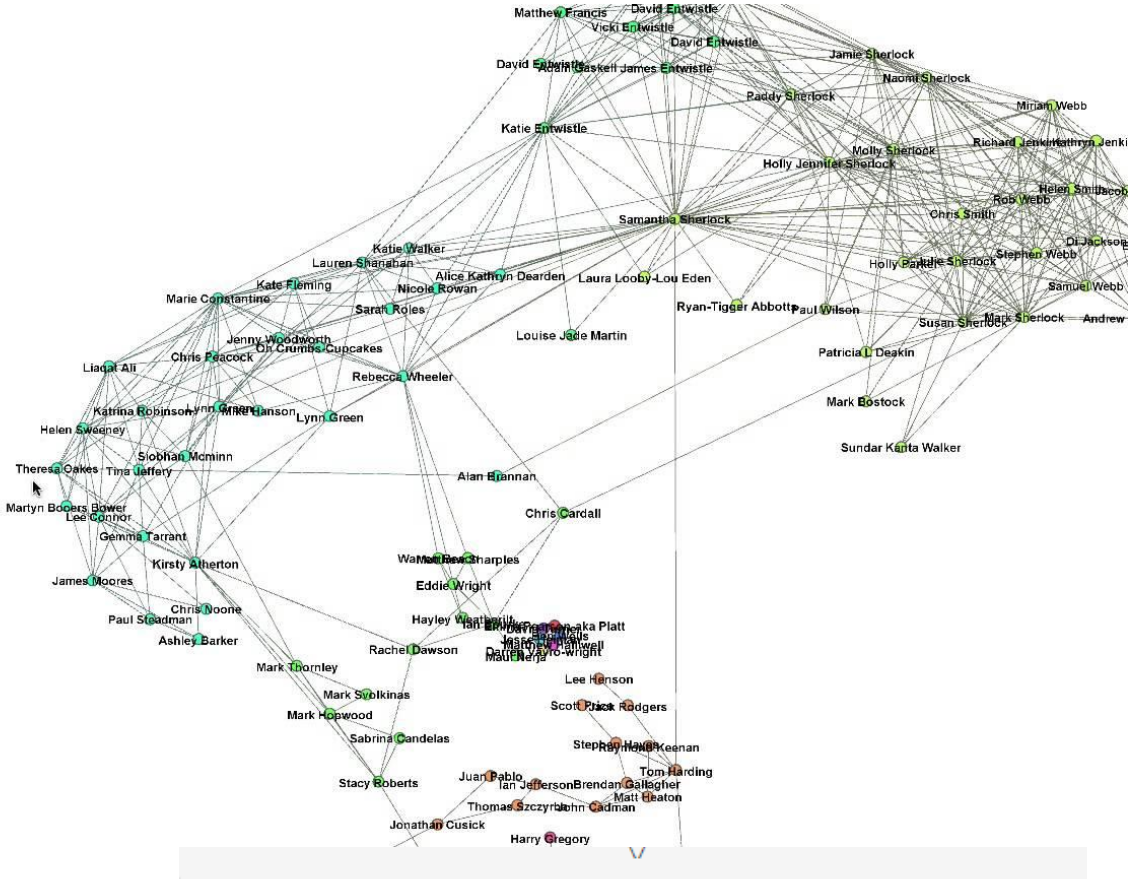
Why should you learn R …



- R is fun!
  - Integrate with Social Media



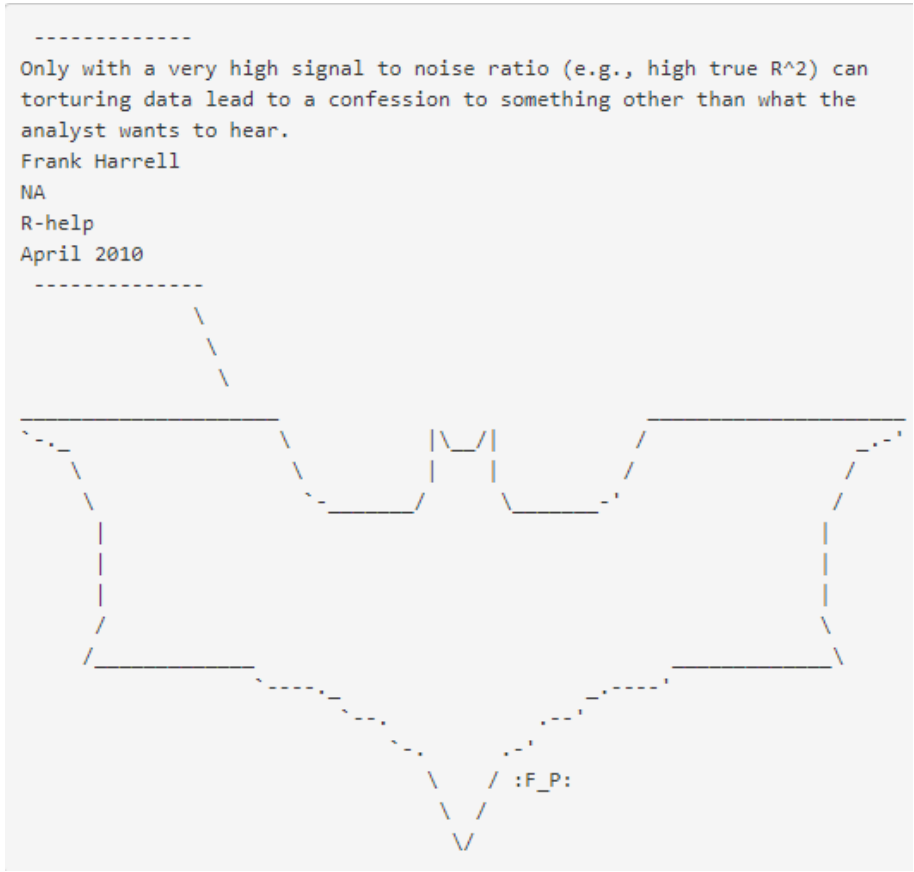
Why should you learn R …



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

Why should you learn R ...

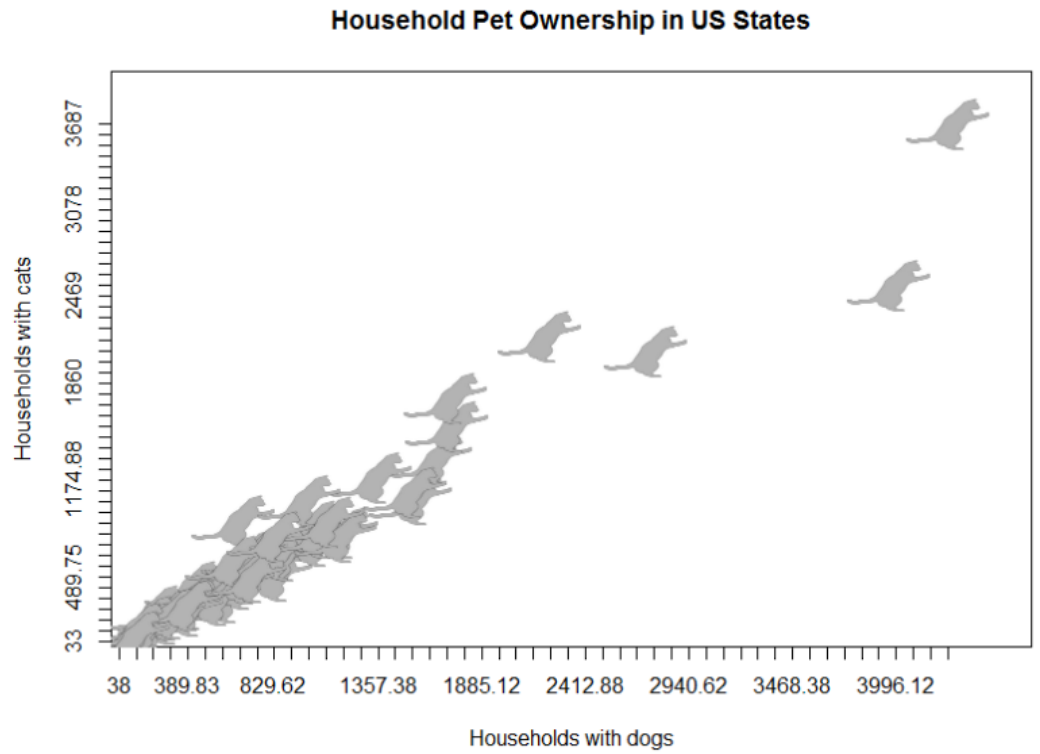
```
 -------------
Only with a very high signal to noise ratio (e.g., high true R^2) can
torturing data lead to a confession to something other than what the
analyst wants to hear.
Frank Harrell
NA
R-help
April 2010
 --------------
```



- 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!

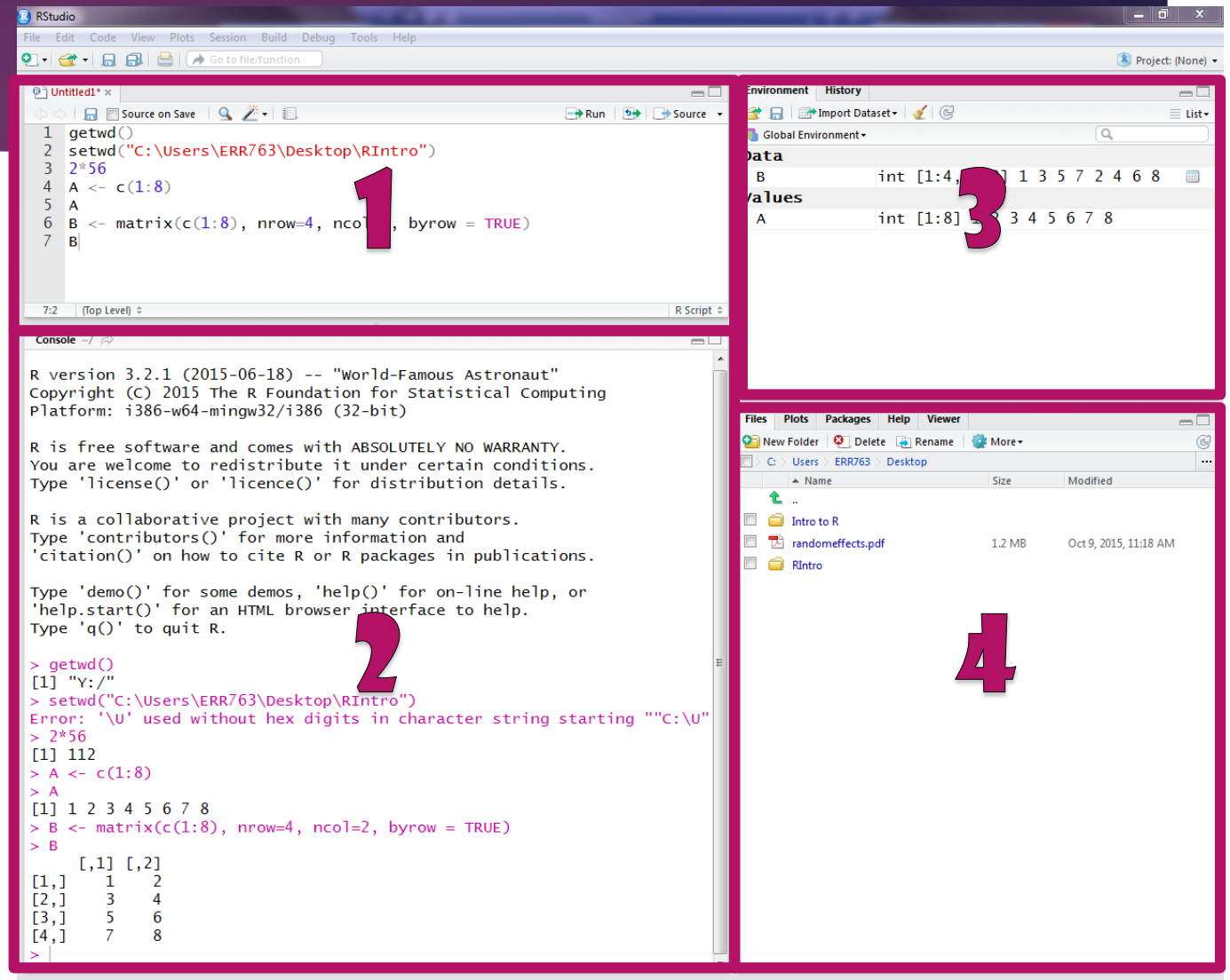


Overview of RStudio

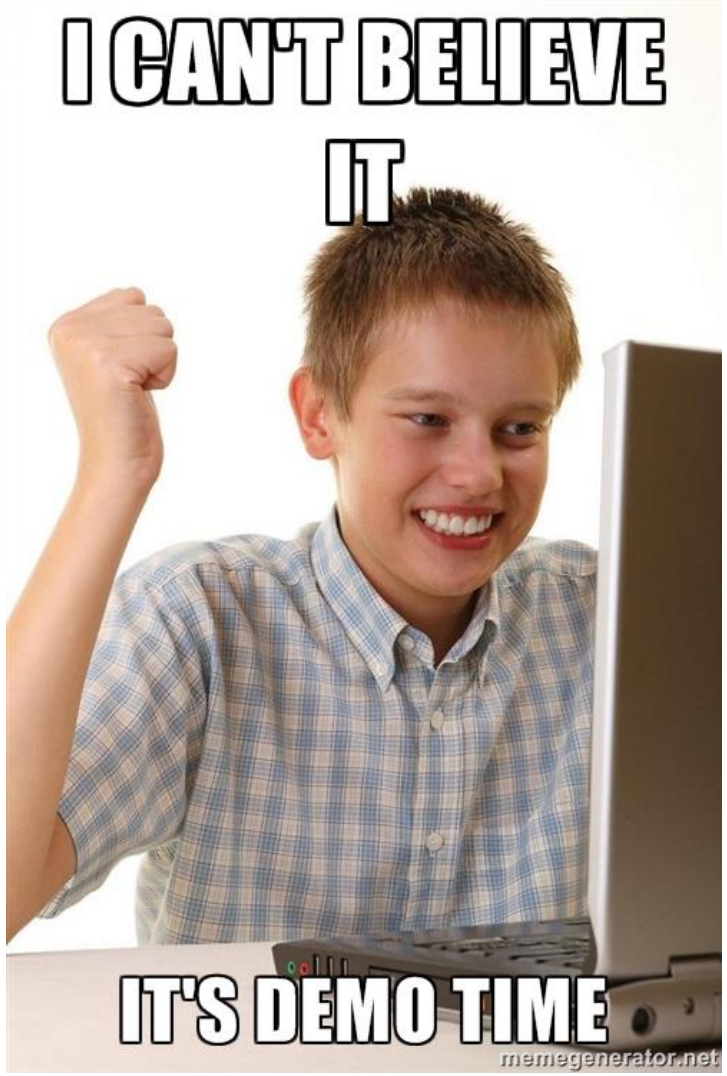
- RStudio is a **user interface**
  - Various tabs to organize graphics, data, file explorer, etc.
  - 4 windows for easy navigation
- RStudio runs R in the **background**
  - Note that you need to install R before installing RStudio
- RStudio works with Shiny and RMarkdown to create interactive web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



Thank You!

EMILY.JONES@SONY.COM

Useful Resources

- WWW.RSTUDIO.COM/RESOURCES/CHEATSHEETS/
- WWW.COURSERA.COM
- WWW.DATACAMP.COM
- R FOR DATA SCIENCE & GGPLOT2 (BY HADLEY WICKHAM)
- AN INTRODUCTION TO STATISTICAL LEARNING (BY GARETH JAMES)
- EMILY.JONES@SONY.COM
- THE INTERNET!