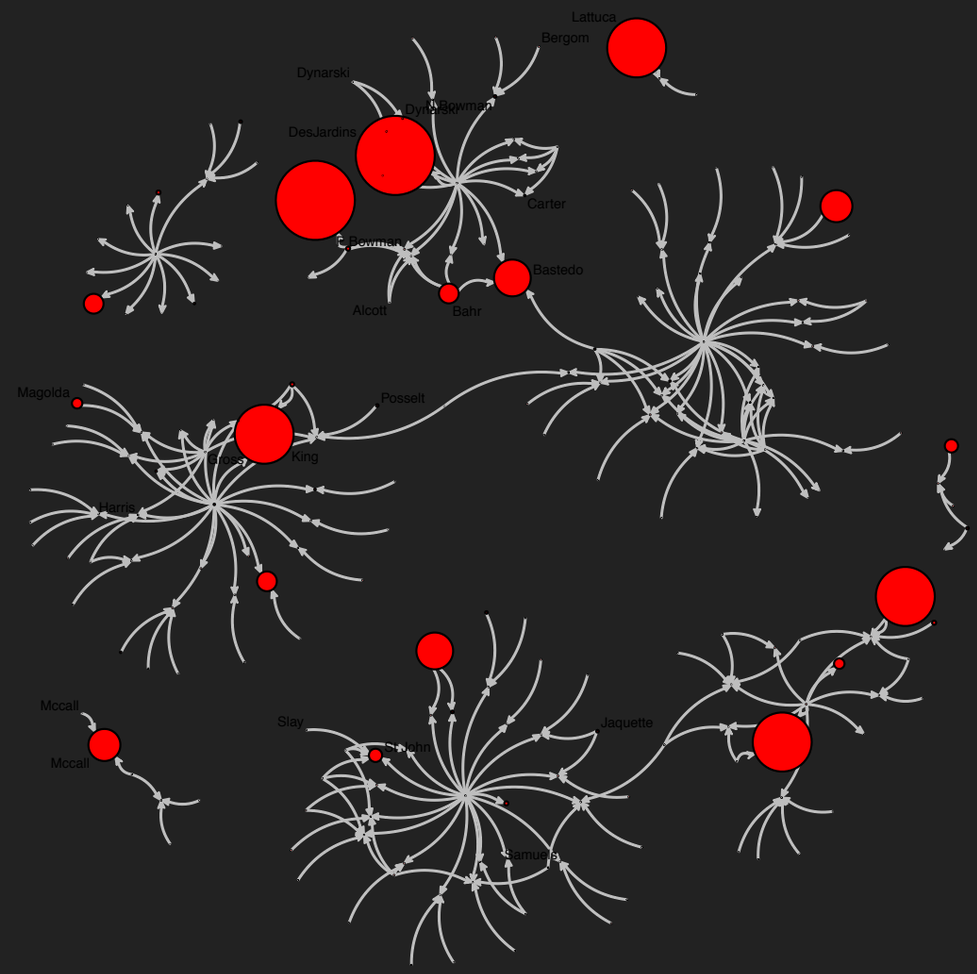


A QUICK GUIDE TO

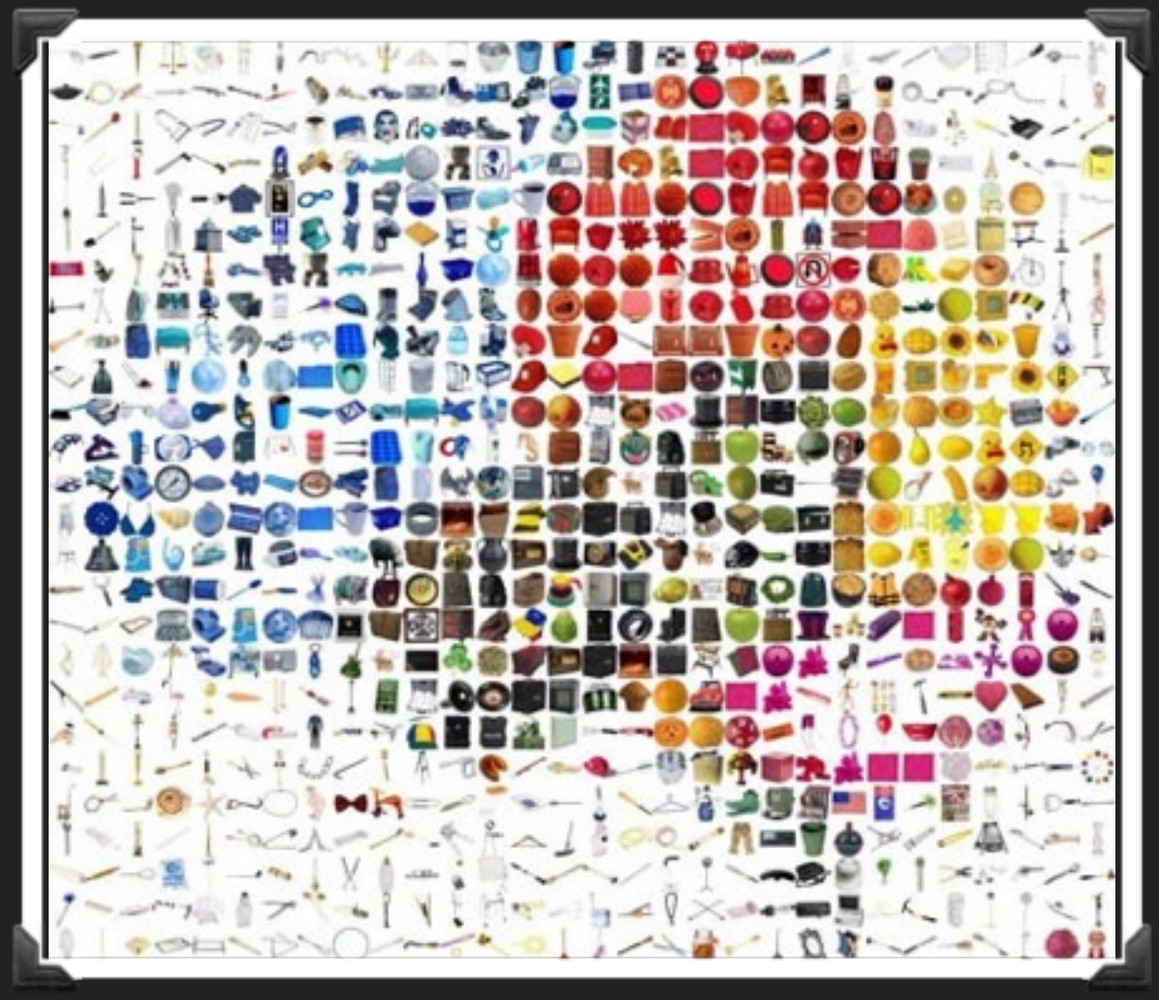
R & SNA

AGENDA

- Quick Guide to R & R Studio
  - Object Oriented Programming
  - Collegescorecard Example
  - Beyond Regression
- Quick Guide to Social Network Analysis
  - Data collection, organization, and analysis
  - CSHPE Faculty example

WHAT ARE THE BENEFITS OF R?

- Object oriented programming
  - Can store multiple data frames simultaneously and make individual calls to the data
- Store data of lots of different types
- Can store graphs and visualizations as you create them



Stata/MP 13.0 - http://stata-press.com/data/r13/nhanes2d.dta - [Results]

File Edit Data Graphics Statistics User Window Help

Review

Command
webuse nhanes2d
svy: nbreg highbp weight a

Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Generalized linear models
Tree
Ensembles
Sample-selection models
Exact statistics
Nonparametric analysis
Time series
Multivariate time series
Longitudinal/panel data
Multilevel mixed-effects models
Survival analysis
Epidemiology and related
SEM (structural equation modeling)
Survey data analysis
Multiple imputation
Multivariate analysis
Power and sample size
Resampling
Postestimation
Other

Stata can generally handle one dataset

at a time

```
t age age2 female black
ion sample)

                                =      10351
                                =  117157513
                                =         31
                    F(   5,    27)  =     246.13
                    Prob > F     =     0.0000

Linearized
 Std. Err.      t    P>|t|     [95% Conf. Interval]

 .0007974    25.91   0.000     .0190356    .0222881
 .0072088     7.92   0.000     .0423697    .0717745
 .0000739    -3.98   0.000    -.0004454   -.0001438
 .0285662    -3.22   0.003    -.1501412   -.0336188
 .0482098     3.27   0.003     .0593438    .2559928
 .1948737   -22.82   0.000    -4.844691   -4.049796
```

Setup and utilities
Tables
Means, proportions, ratios, totals
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Survival models
Endogenous covariates
Sample-selection models
Generalized linear models
DEFF, MEFF, and other statistics

Poisson regression
Negative binomial regression
Generalized negative binomial regression
Zero-inflated Poisson regression
Zero-inflated negative binomial regression
Truncated Poisson regression

Command
svy: nbreg highbp weight age a

Variables

Variable	Label
sampl	unique case identi...
strata	stratum identifier, ...
psu	primary sampling ...
region	1=NE, 2=MW, 3=S...
smsa	1=SMSAcity,2=S...
location	stand number, 1-64
houssiz	# persons in hous...
sex	1=male, 2=female
race	1=white, 2=black, ...

Properties

Variables

Name	sampl
Label	unique case ident
Type	long
Format	%9.0g
Value Label	
Notes	

Data

Filename	nhanes2d.dta
Label	
Notes	
Variables	58
Observations	10,351
Size	1.07M
Memory	64M

C:\Users\stata

CAP NUM OVR

File Edit Code View Project Workspace Plots Tools Help

Go to file/function

Project: (None)

diamondPricing.R* | formatPlot.R | diamonds

Source on Save | Source

```
1  library(ggplot2)
2  source("plots/formatPlot.R")
3
9  clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12             data=diamonds, color=clarity,
13             xlab="Carat", ylab="Price",
14             main="Diamond Pricing")
15
```

R allows you to store multiple data sets as well as

as host of other objects simultaneously

15:1 (Top Level) R Script

Workspace | History

Load | Save | Import Dataset | Clear All

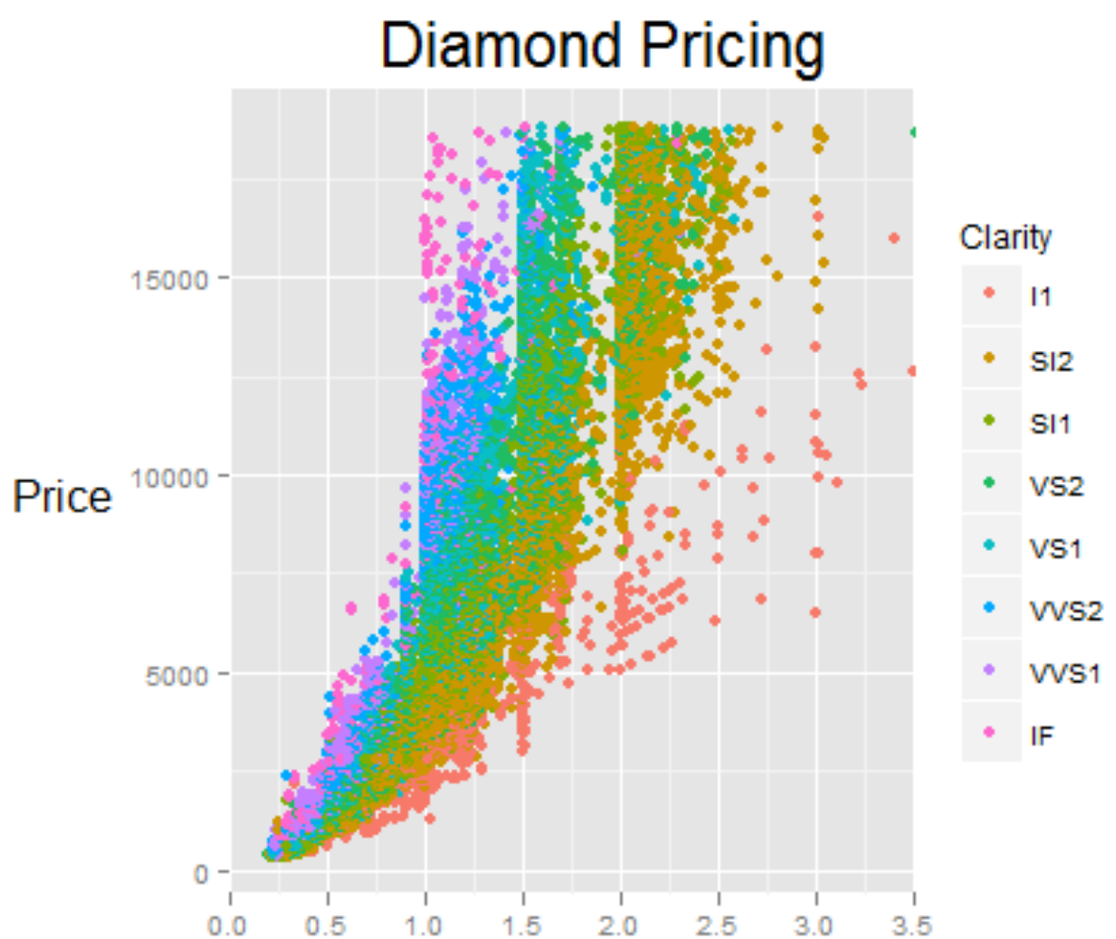
Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]
p	ggplot[8]
Functions	
format.plot(plot, size)	

Files | Plots | Packages | Help

Zoom | Export | Clear All

Console ~/

```
       x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
```



File Edit Code View Project Workspace Plots Tools Help

Go to file/function

Project: (None)

diamondPricing.R* | formatPlot.R | diamonds

Source on Save | Source

```
1  library(ggplot2)
2  source("plots/formatPlot.R")
3
9  clarity <- levels(diamonds$clarity)
10
11 p <- qplot(carat, price,
12             data=diamonds, color=clarity,
13             xlab="Carat", ylab="Price",
14             main="Diamond Pricing")
15
```

R was built with visualizations in mind so your

output can have more variety and flair.

15:1 (Top Level) R Script

Console ~/

```
      x                y                z
 Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
 1st Qu.: 4.710   1st Qu.: 4.720   1st Qu.: 2.910
 Median : 5.700   Median : 5.710   Median : 3.530
 Mean   : 5.731   Mean   : 5.735   Mean   : 3.539
 3rd Qu.: 6.540   3rd Qu.: 6.540   3rd Qu.: 4.040
 Max.   :10.740   Max.   :58.900   Max.   :31.800
> summary(diamonds$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    326     950    2401    3933    5324   18820
> aveSize <- round(mean(diamonds$carat), 4)
> clarity <- levels(diamonds$clarity)
> p <- qplot(carat, price,
+            data=diamonds, color=clarity,
+            xlab="Carat", ylab="Price",
+            main="Diamond Pricing")
>
> format.plot(p, size=24)
```

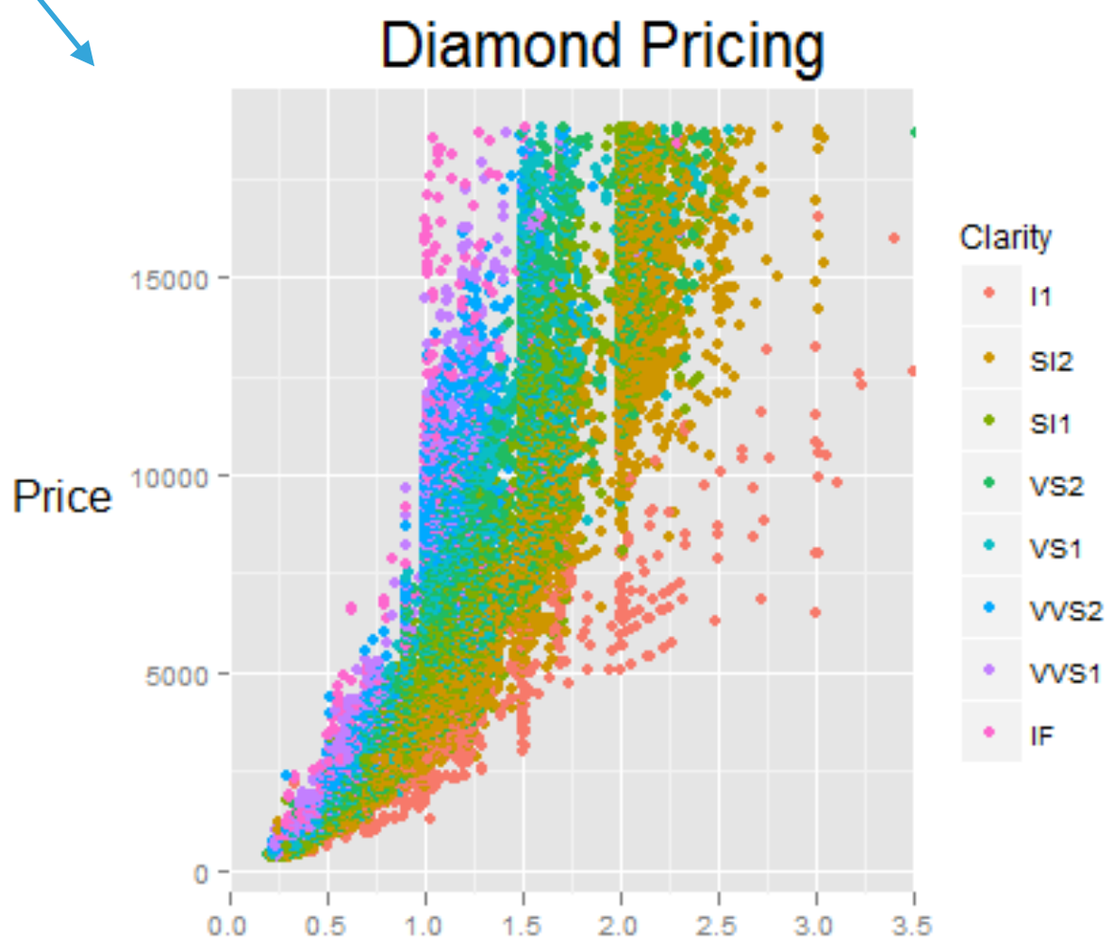
Workspace | History

Load | Save | Import Dataset | Clear All

Data	
diamonds	53940 obs. of 10 variables
Values	
aveSize	0.7979
clarity	character[8]
p	ggplot[8]
Functions	
format.plot(plot, size)	

Files | Plots | Packages | Help

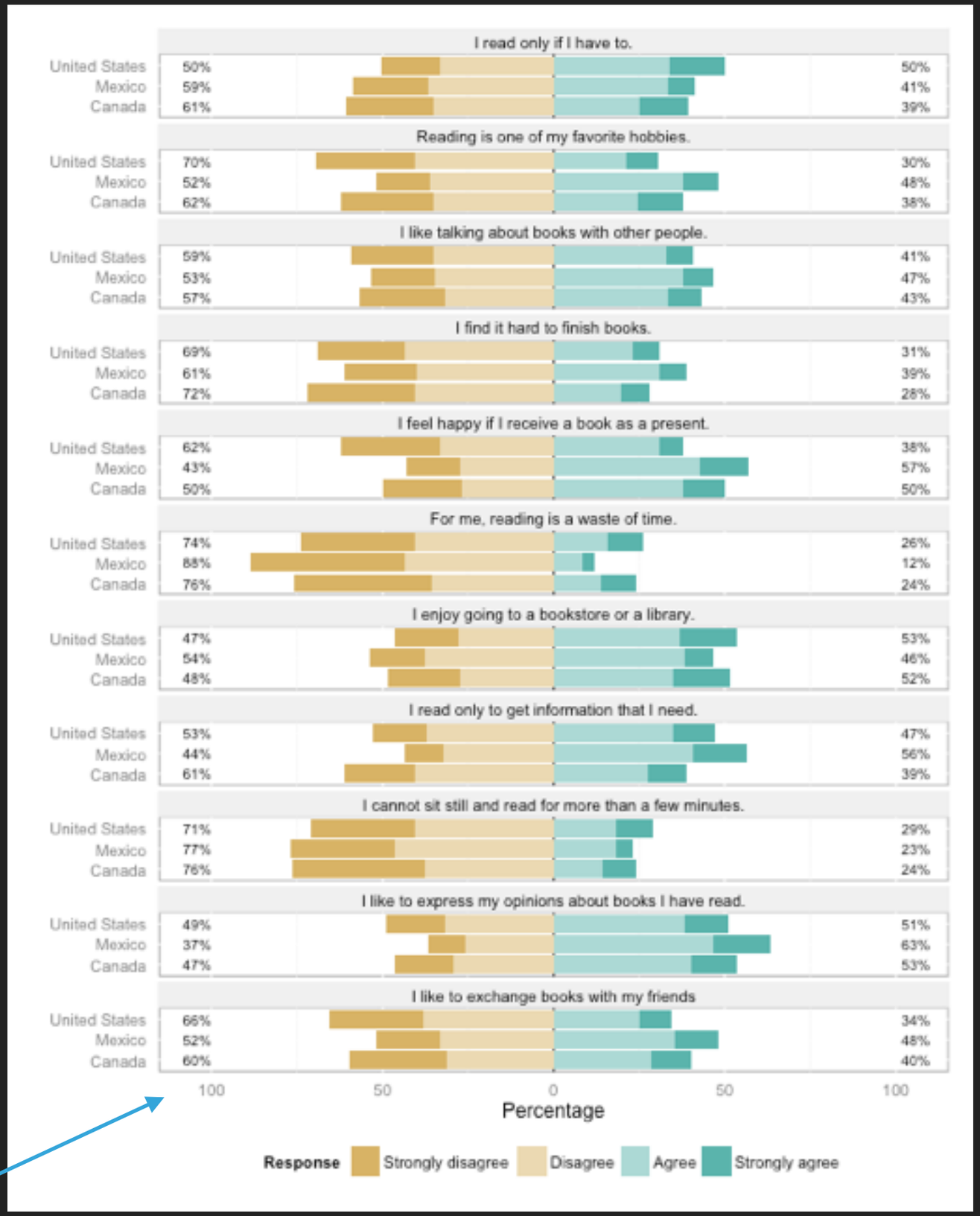
Zoom | Export | Clear All



R IS OPEN SOURCE

- Which means that researchers and programmers write their own packages
  - Collegescorecard
  - IPEDS
  - Likert
  - Stanford Natural Language Processing
  - Statnet

Likert package by Jason Bryer



BUT THE BEST OF ALL

IS EXPLAINR CREATED BY HILLARY PARKER AT ETSY

According to Google Image Search this is Hillary Parker

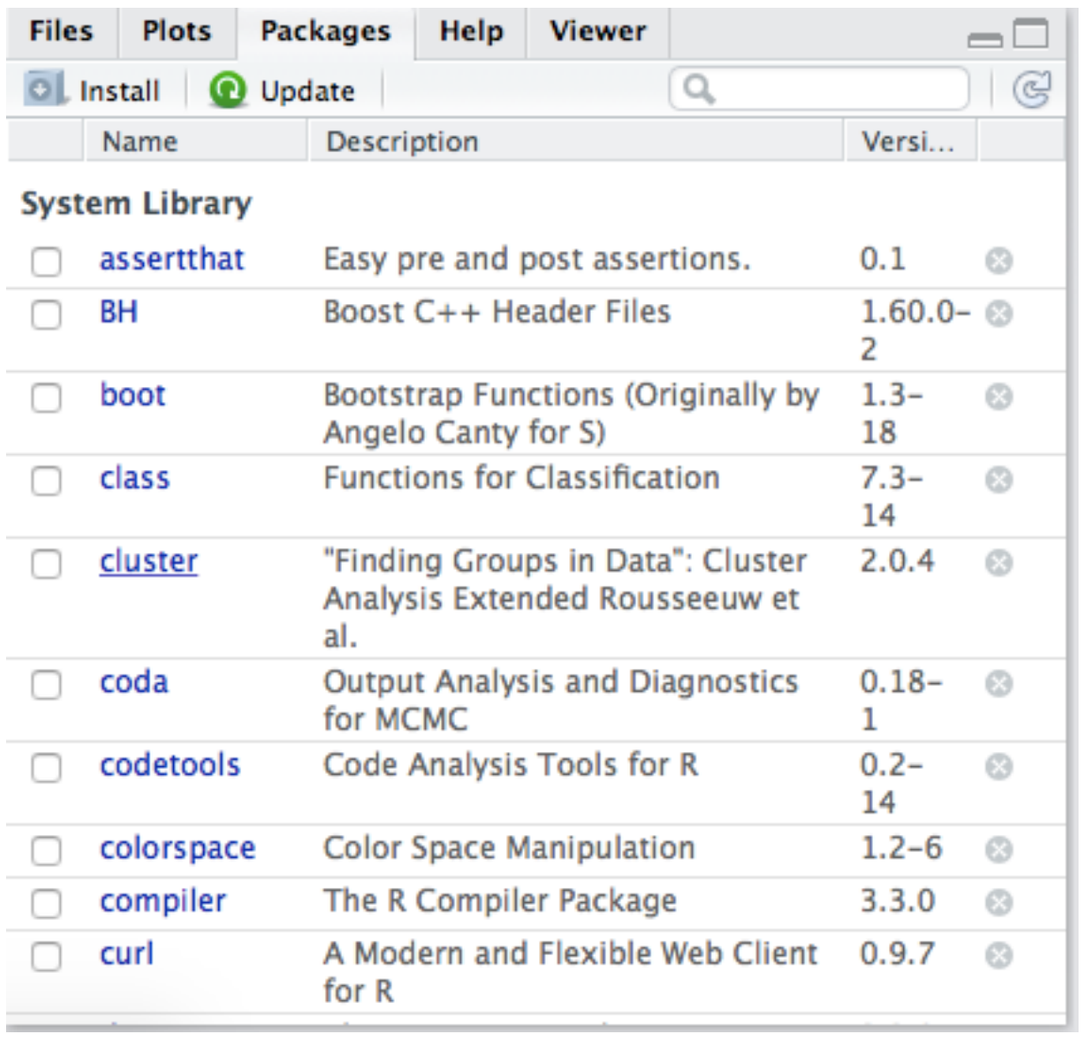
LET'S PLAY IN R

GO TO:

MICHAELBROWN.WORK/RWORKSHOP

STARTING WITH PACKAGES

```
install.packages(devtools)
library(devtools)
devtools::install_github("hilaryparker/explainr")
library(explainr)
```



#RStudio comes with lots of pre-loaded packages. Click on Packages in the bottom right hand pane for some examples.

#You could just click on the package you want to attach, if you're not a fan of writing scripts

#Try clicking the script for the cluster package and see what Studio does.

LET'S DO A PROPORTIONS TEST

R Code

```
#Let's try and run a proportions test
prop.test(x = 500, n = 1008) # Cool, but what if we want to keep those results for later?
ptest <- prop.test(x = 500, n = 1008) #We have stored the results of our 1-sample proprotions test in the R Environment
ptest #So when we call the results of that test, they're stored in memory
explain(ptest) # And now we get an explanation of what we just did
#YOU MAY NEVER NEED A GSI AGAIN!
```

What the R Code is doing

PROP.TEST(X=500, N=1008) #RUN A PROPORTION TEST WITH 1008 OBSERVATIONS
PTEST<-PROP.TEST(. . . . . #STORE RESULTS OF THE P-TEST FOR LATER USE
PTEST #WILL RECALL RESULTS AND DISPLAY THEM IN THE OUTPUT WINDOW

EXPLAIN(PTEST) #TELLS R TO USE THE EXPLAINR TEMPLATE AND PROVIDE US WITH AN INTERPRETATION OF THE RESULTS WE STORED IN THE 'PTEST' OBJECT

EXPLAINR(PTEST)

This was a one-sample proportion test of the null hypothesis that the true population proportion is equal to 0.5. Using a significance level of 0.05, we do not reject the null hypothesis, and cannot conclude that true population proportion is different than 0.5. The observed sample proportion is 0.496031746031746 (500 events out of a total sample size of 1,008).

The confidence interval for the true population proportion is (0.464746, 0.5273481). This interval will contain the true population proportion 95 times out of 100.

The p-value for this test is 0.8254979. This, formally, is defined as the probability -- if the null hypothesis is true -- of observing a sample proportion that is as or more extreme than the sample proportion from this data set. In this case, this is the probability -- if the true population proportion is 0.5 -- of observing a sample proportion that is greater than 0.503968253968254 or less than 0.496031746031746.

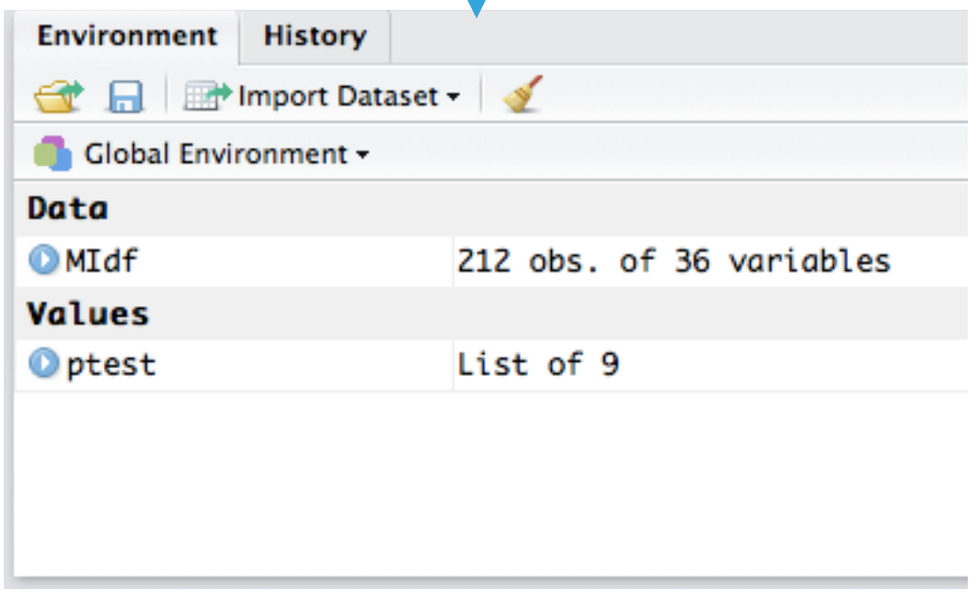
LOADING DATA

EXAMPLE OF HOW TO LOAD FROM THE WEB:

```
#Let's actually load some data and play around in R
library(foreign)
MIdf<-read.csv("http://michaelbrown.work/wp-content/uploads/2016/05/MIdf.csv", header=T, sep=",") # This downloads the resources from my website.
#In the Environments Panel you should now have a file named MIdf with 212 observations of 36 variables.
summary(MIdf) #This provides quick descriptives for each variable in the dataset.
```

EXAMPLE OF HOW TO LOAD FROM YOUR COMPUTER:

```
library(xlsx)
library(foreign)
setwd("~/Box Sync/Mike_Research_Project/STATS250")
IDKEY<-read.xlsx("stats250IDKEY.xlsx",1)
setwd("~/Downloads")
Tailor<-read.csv("ECoachWN2016Stats250TailoringData.csv")
IDKEY$username<-tolower(IDKEY$UM_UNQNM)
IDKEY$EZ_id <-match(IDKEY$username, Tailor$username)
setwd("~/Box Sync/Mike_Research_Project/STATS250")
write.csv(IDKEY, file="stats250IDKEY52016.csv")
rm()
```



College Affordability and Transparency Center

College Scorecard

Austin Community College District

Austin, TX
Primarily associate's degree granting
Undergraduate enrollment: 45,100

Back to Search | More Information | Print Profile

Costs

$6,884 / yr

What does it typically cost to attend Austin Community College District?

The average net price for undergraduate in-state students is $6,884 per year. Net price is what undergraduate students pay after grants and scholarships (financial aid you don't have to pay back) are subtracted from the institution's cost of attendance.

The average net price has decreased 3.2% from 2007 to 2009.

Click here to see listings of changes in college costs.

Click here to go to the Net Price Calculator for a better estimate of what your costs would be.

Graduation Rate

What percentage of students graduate?

4.2% of full-time students graduated within 150% of the expected time for completion and 38.5% transferred to another institution. Graduation rate data are based on undergraduate students who enrolled full-time and have never enrolled in college before. This may not represent all undergraduates that attend this institution.

Loan Default Rate

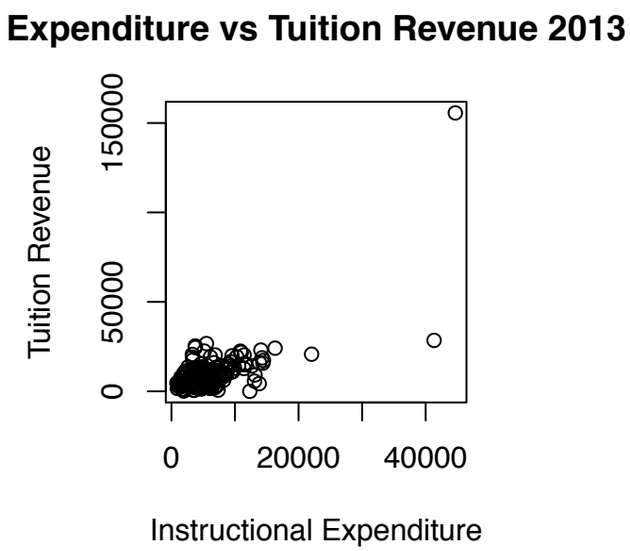
Are students able to repay their loans after they graduate?

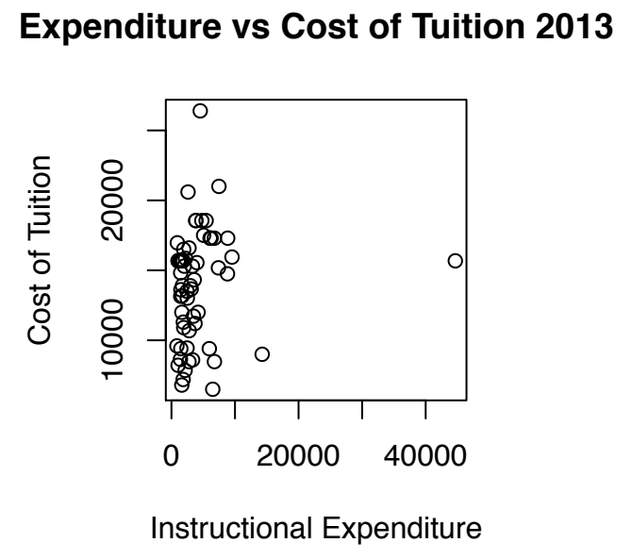
17.1% of borrowers defaulted on their Federal student loans within three years of entering repayment.

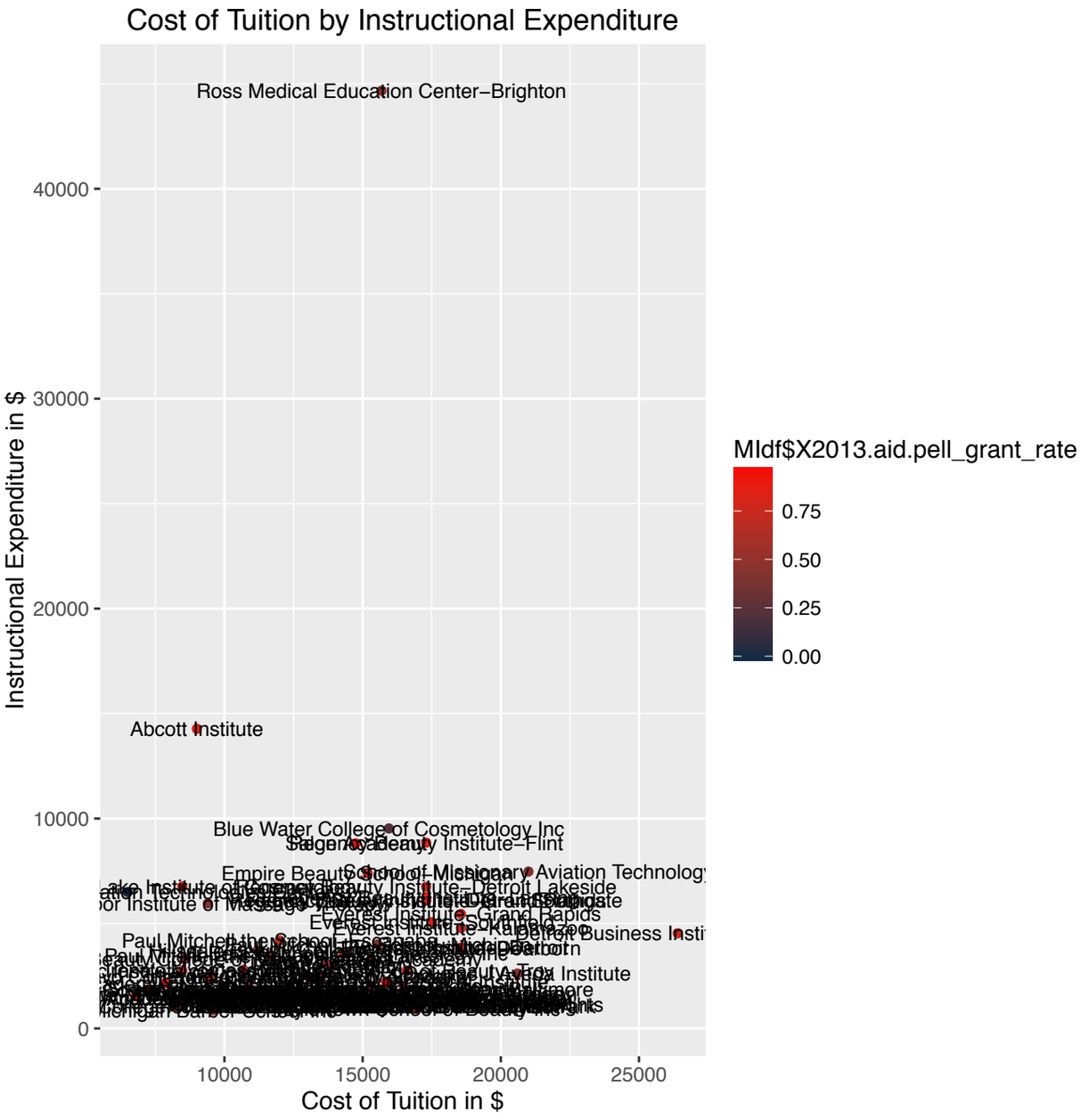
COLLEGE SCORE CARD

TUITION AND EXPENDITURES

EXPENDITURES AND TUITION





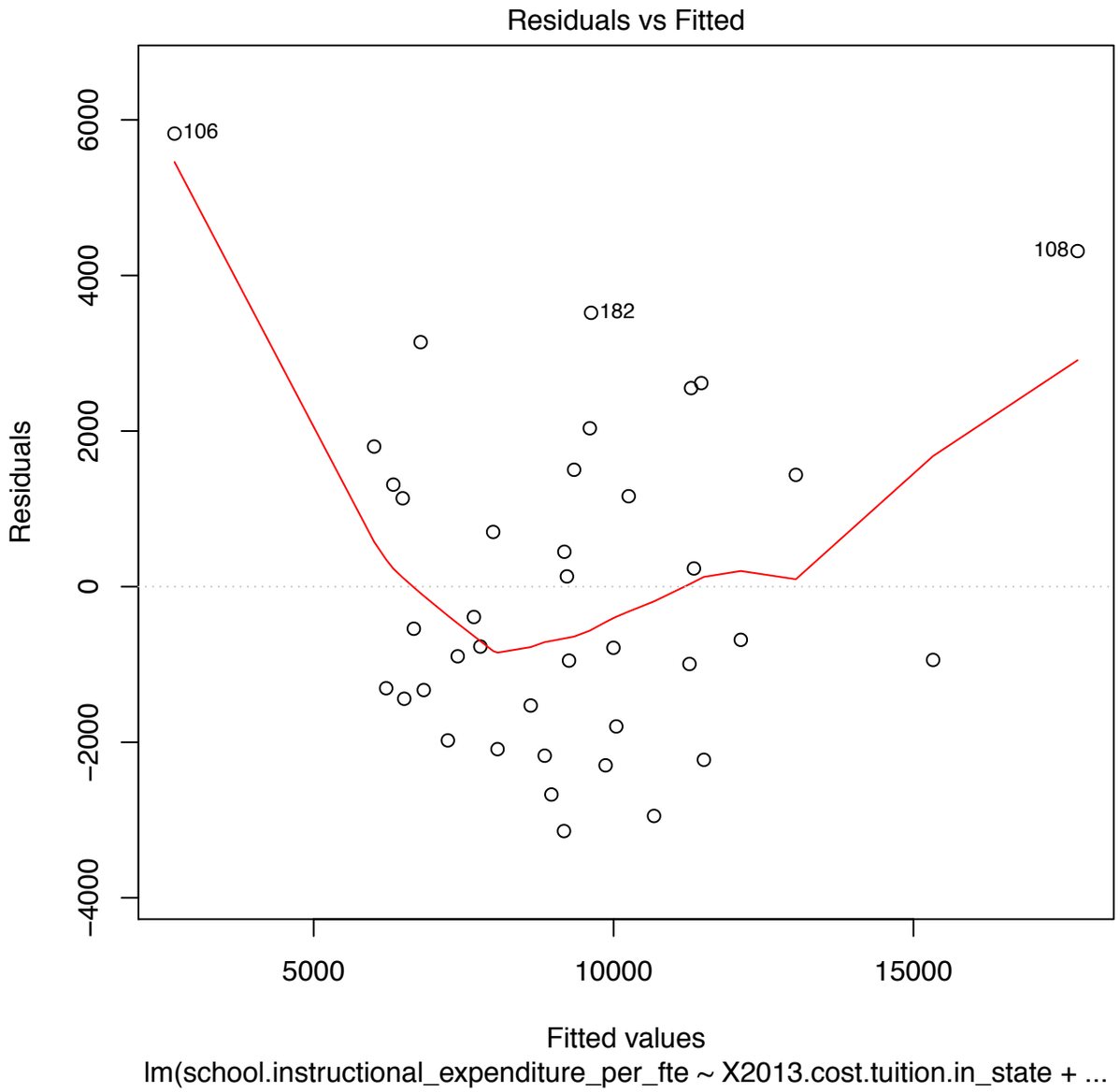


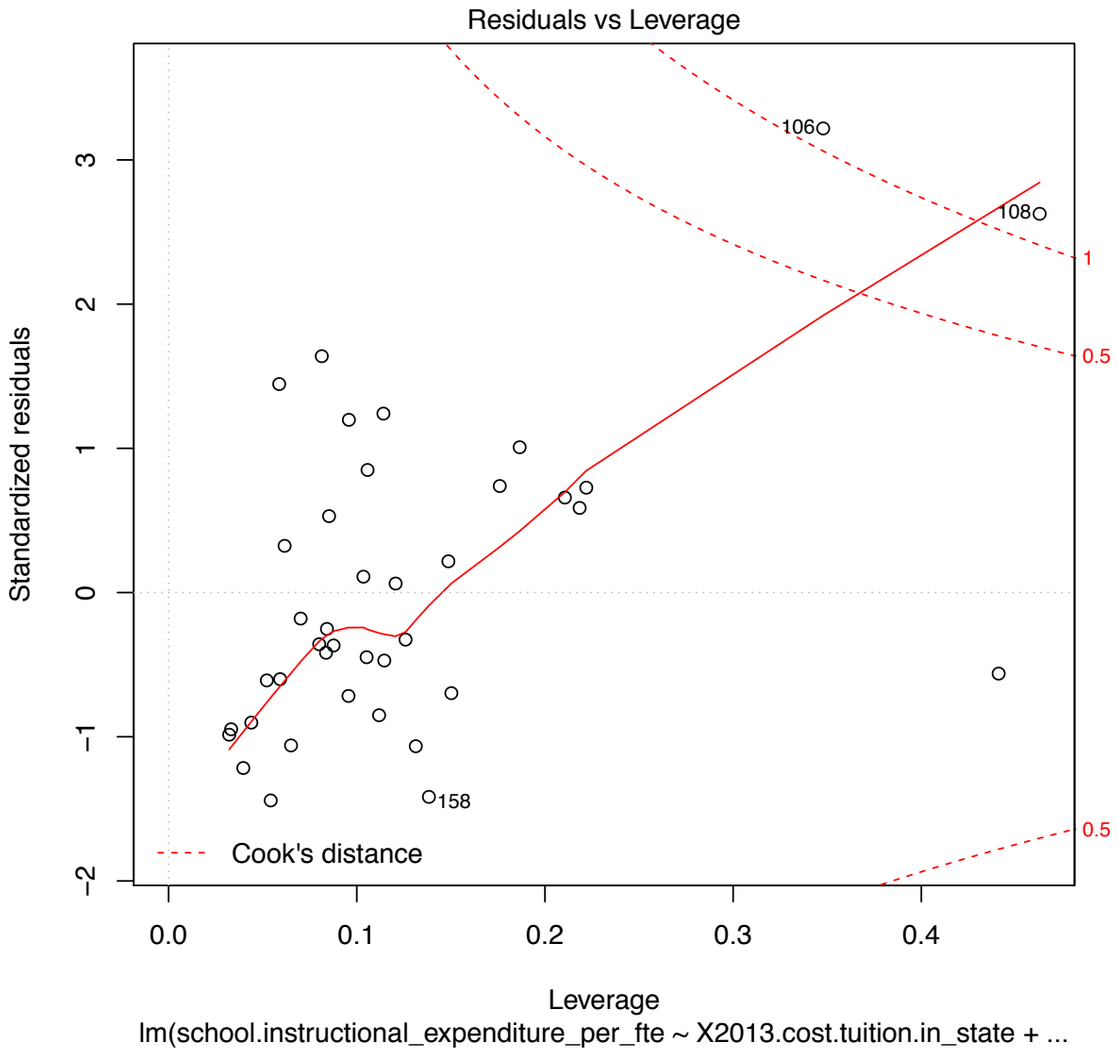
RUNNING A REGRESSION MODEL

```
#DV: school.instructional_expenditure_per_fte (Instructional Expenditure per student)
#IV: X2013.cost.tuition.in_state, X2013.admissions.admission_rate.overall, X2013.admissions.sat_scores.average.overall, X2013.student.siz
CostModel <- lm(school.instructional_expenditure_per_fte  ~ X2013.cost.tuition.in_state + X2013.admissions.admission_rate.overall +
                  X2013.admissions.sat_scores.average.overall + X2013.student.size, data=MIdf, x=T)
summary(CostModel)
```

```
Call:
lm(formula = school.instructional_expenditure_per_fte ~ X2013.cost.tuition.in_state +
    X2013.admissions.admission_rate.overall + X2013.admissions.sat_scores.average.overall +
    X2013.student.size, data = MIdf, x = T)

Residuals:
    Min      1Q  Median      3Q     Max
-3141.4 -1484.5  -683.6  1374.4  5824.2

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                -9.935e+03  5.253e+03  -1.891 0.067151 .
X2013.cost.tuition.in_state                 1.056e-01  6.013e-02   1.756 0.088092 .
X2013.admissions.admission_rate.overall    -2.589e+03  3.254e+03  -0.796 0.431786
X2013.admissions.sat_scores.average.overall 1.636e+01  4.456e+00   3.672 0.000821 ***
X2013.student.size                          1.791e-01  6.093e-02   2.939 0.005875 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2241 on 34 degrees of freedom
  (173 observations deleted due to missingness)
Multiple R-squared:  0.6211,    Adjusted R-squared:  0.5765
F-statistic: 13.93 on 4 and 34 DF,  p-value: 7.899e-07
```

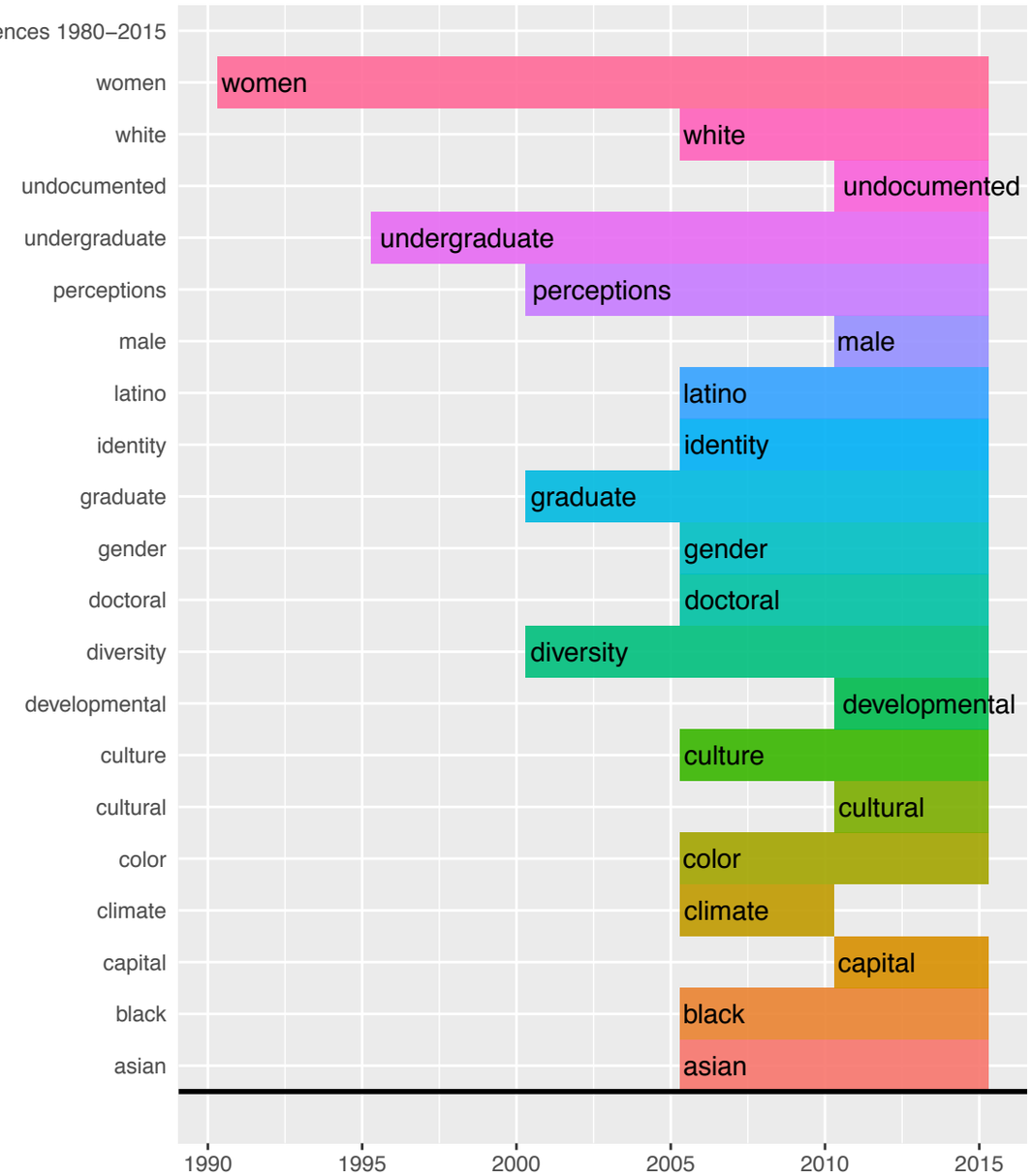
REGRESSION DIAGNOSTICS

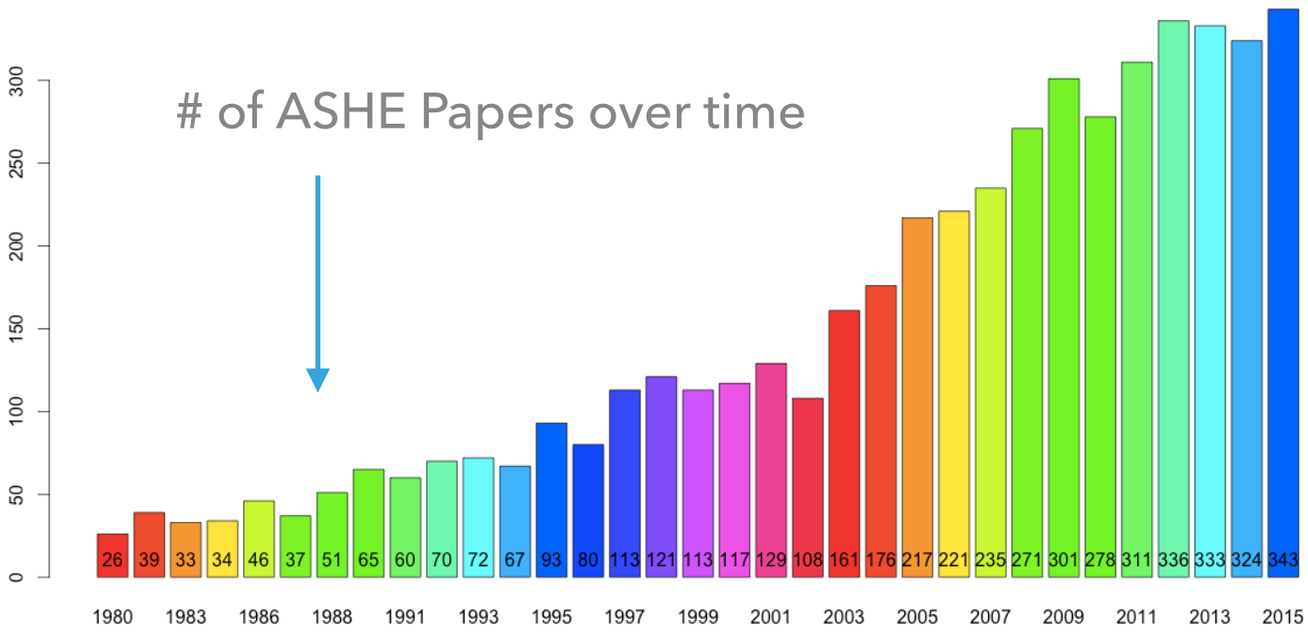




REAL BASIC TOPIC MODELING (HERNANDEZ-HAMED & BROWN, IN PREPARATION*)

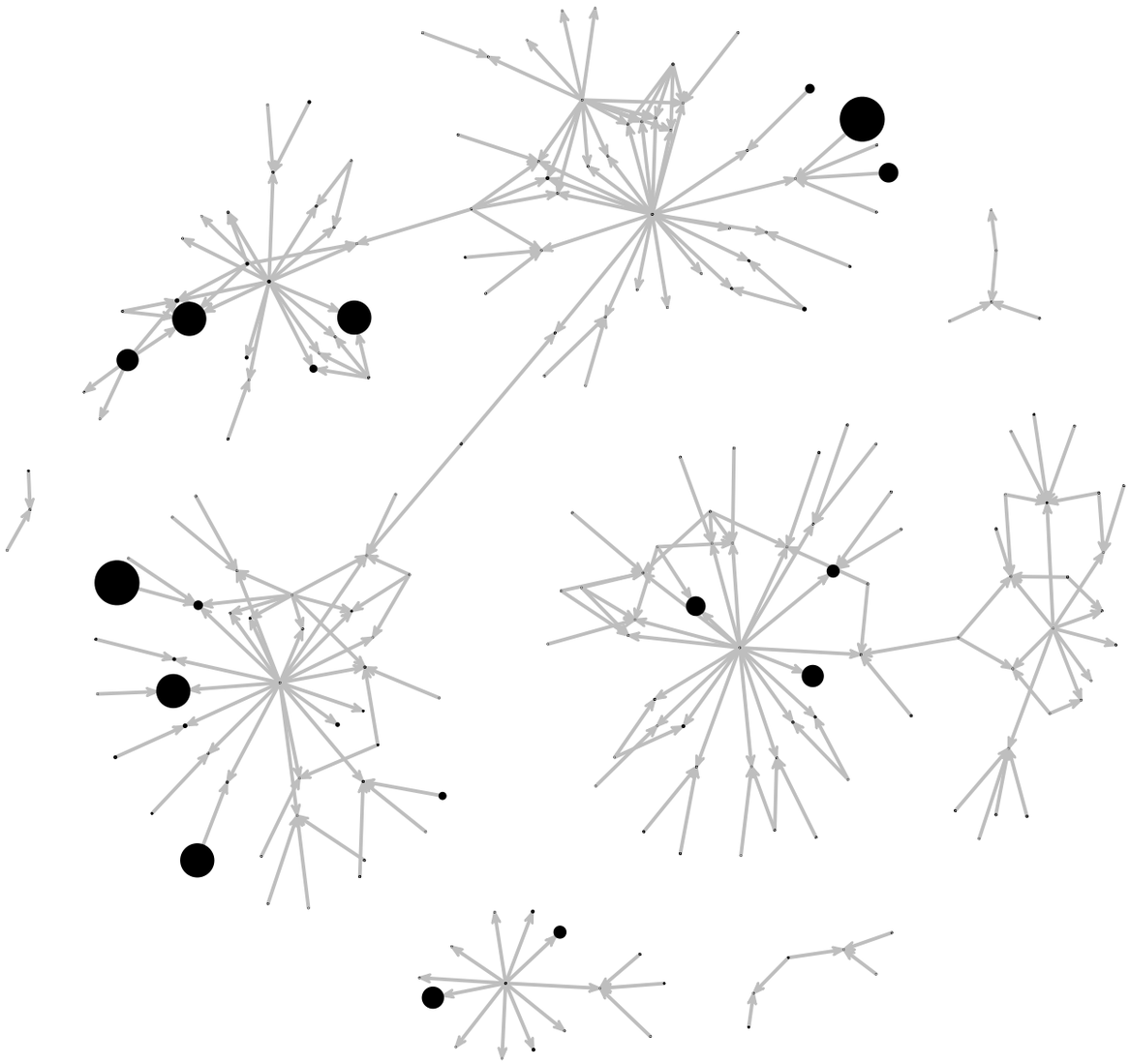
Frequent Title Keywords from papers about identity presented at ASHE between 1990-2015





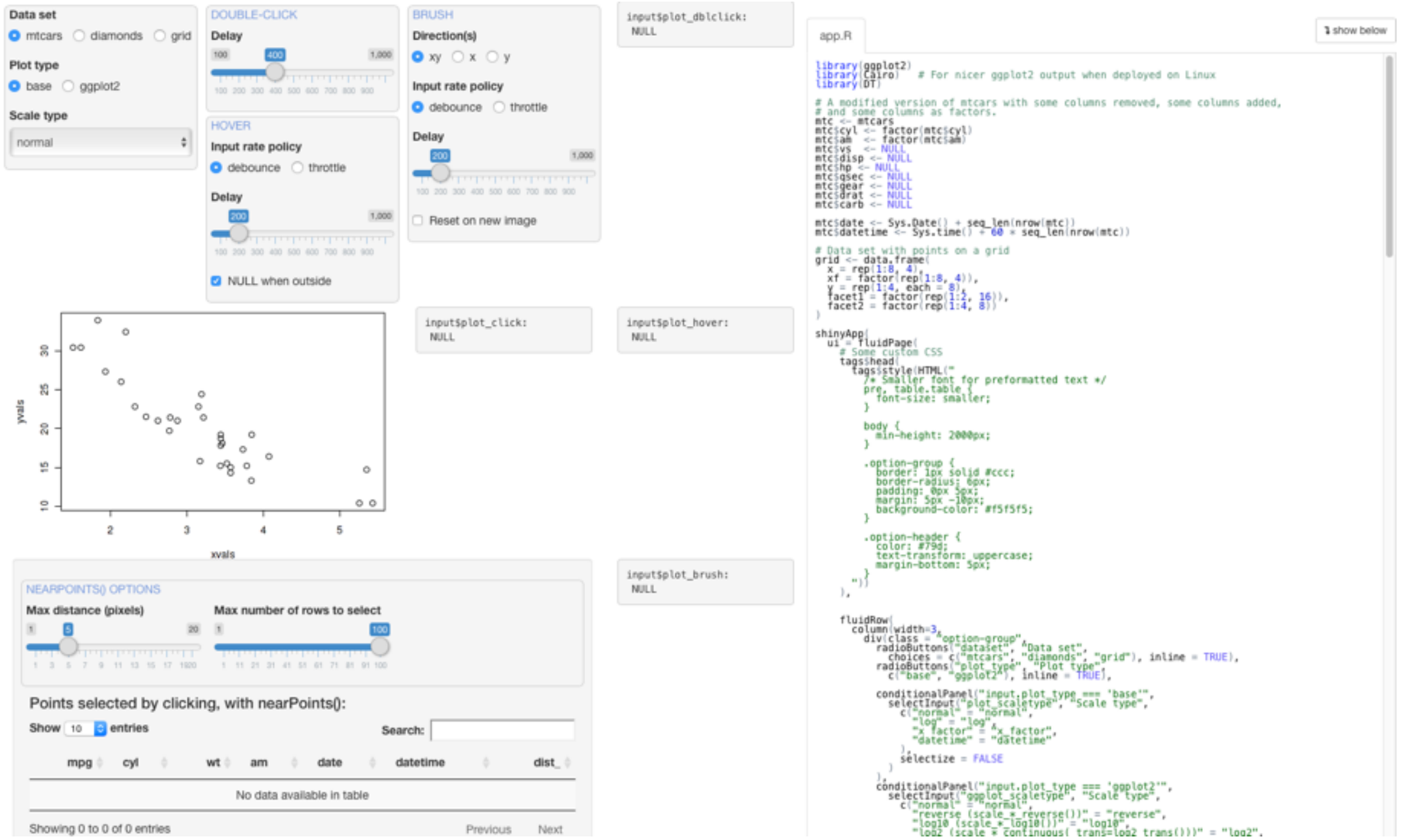
*G-d willing

CITATION ANALYSIS OF SCHOLARLY PRODUCTIVITY (CSHPE)



Faculty Collaboration Network

GOOGLE VIZ AND RSHINY APPS



SOME USEFUL RESOURCES FOR LEARNING R

- TryR.codeschool.com
- Fox's *R Companion to Applied Regression*
- *Discovering Statistics Using R*
- *Learning Analytics with SNA and MPIA using R*
- QuickR

BREAK