

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



Review

Command

webuse nhanes2d

svy: nbreg highbp weight a

- Summaries, tables, and tests
- Linear models and related
- Binary outcomes
- Ordinal outcomes
- Categorical outcomes
- Count outcomes
- Generalized linear models
- Time series
- Trivariate
- En
- 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

at age age2 female black  
(on sample)

= 10351  
= 117157513  
= 31  
= 246.13  
= 0.0000

F( 5, 27)  
Prob > F

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

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

Command

svy: nbreg highbp weight age a

```
diamondPricing.R* x formatPlot.R x diamonds x  
1 library(ggplot2)  
2 source("plots/formatPlot.R")  
3  
4 view(diamonds)  
5  
6 R allows you to store multiple data sets as well as  
7 as host of other objects simultaneously  
8  
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
```

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]

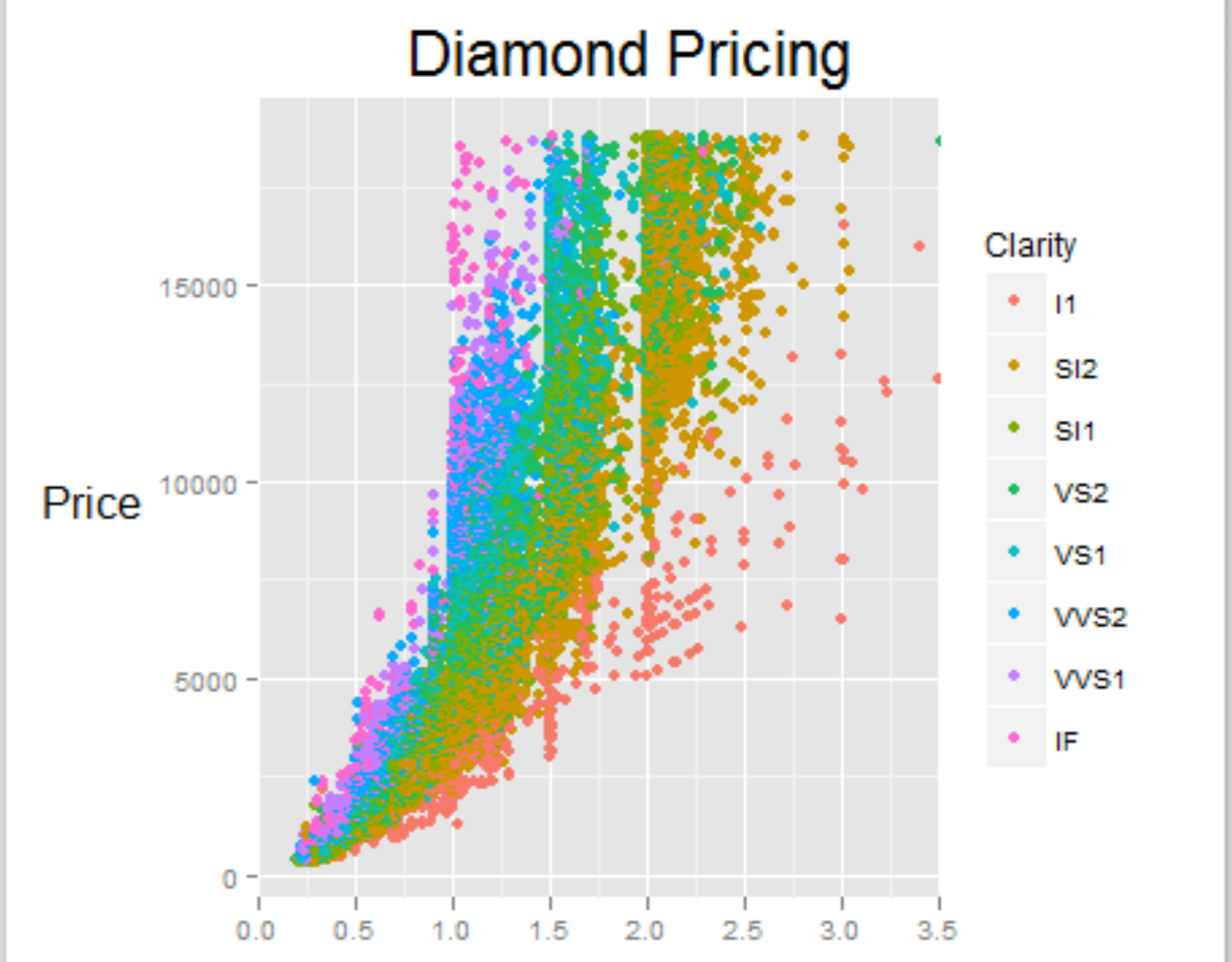
Functions  
format.plot(plot, size)

Files Plots Packages Help

Zoom Export Clear All

Console

```
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)
```



```
diamondPricing.R* x formatPlot.R x diamonds x  
1 library(ggplot2)  
2 source("plots/formatPlot.R")  
3  
4 view(diamonds)  
5  
6 R was built with visualizations in mind so your  
7 output can have more variety and flair.  
8  
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
```

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]

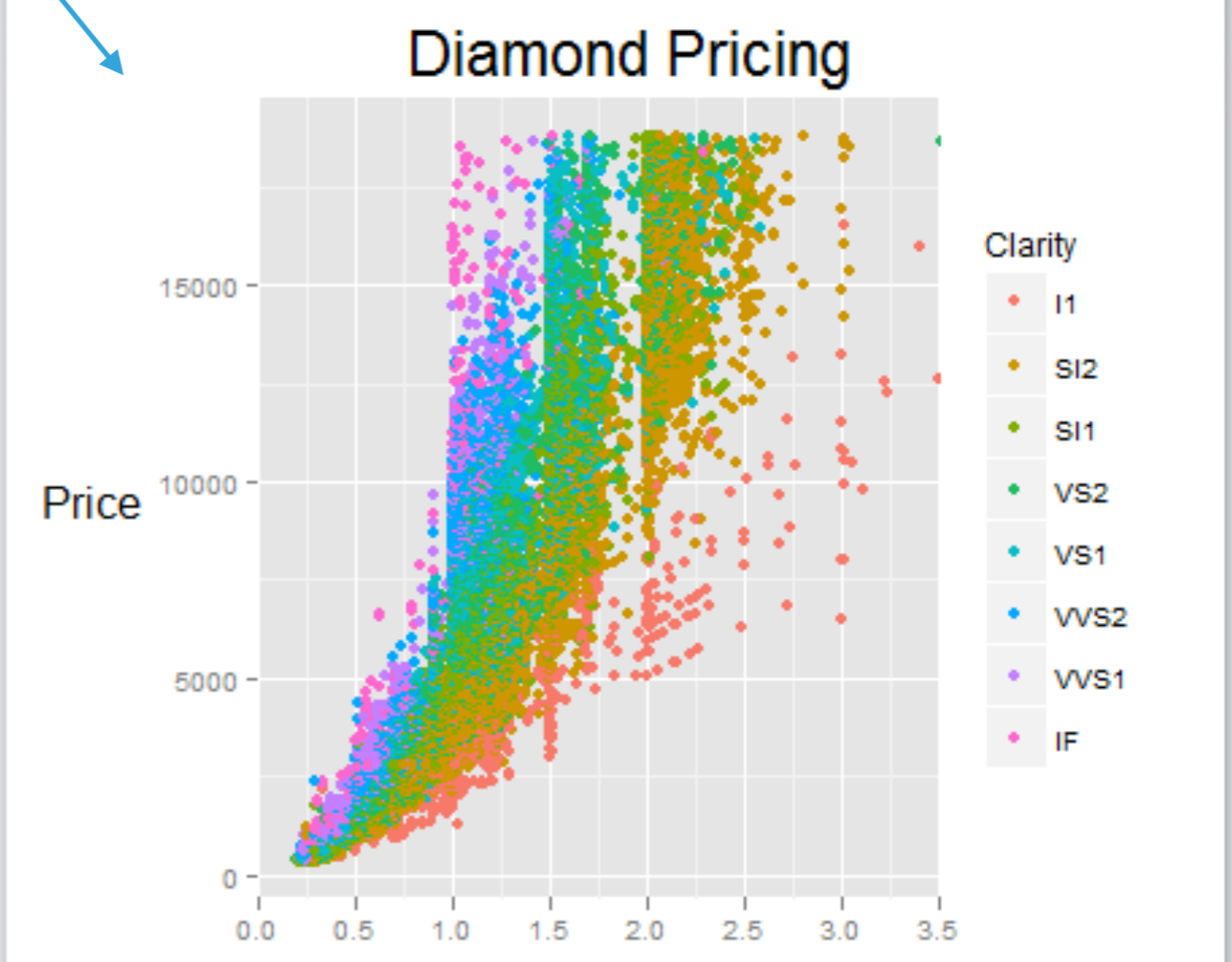
Functions  
format.plot(plot, size)

Files Plots Packages Help

Zoom Export Clear All

Console

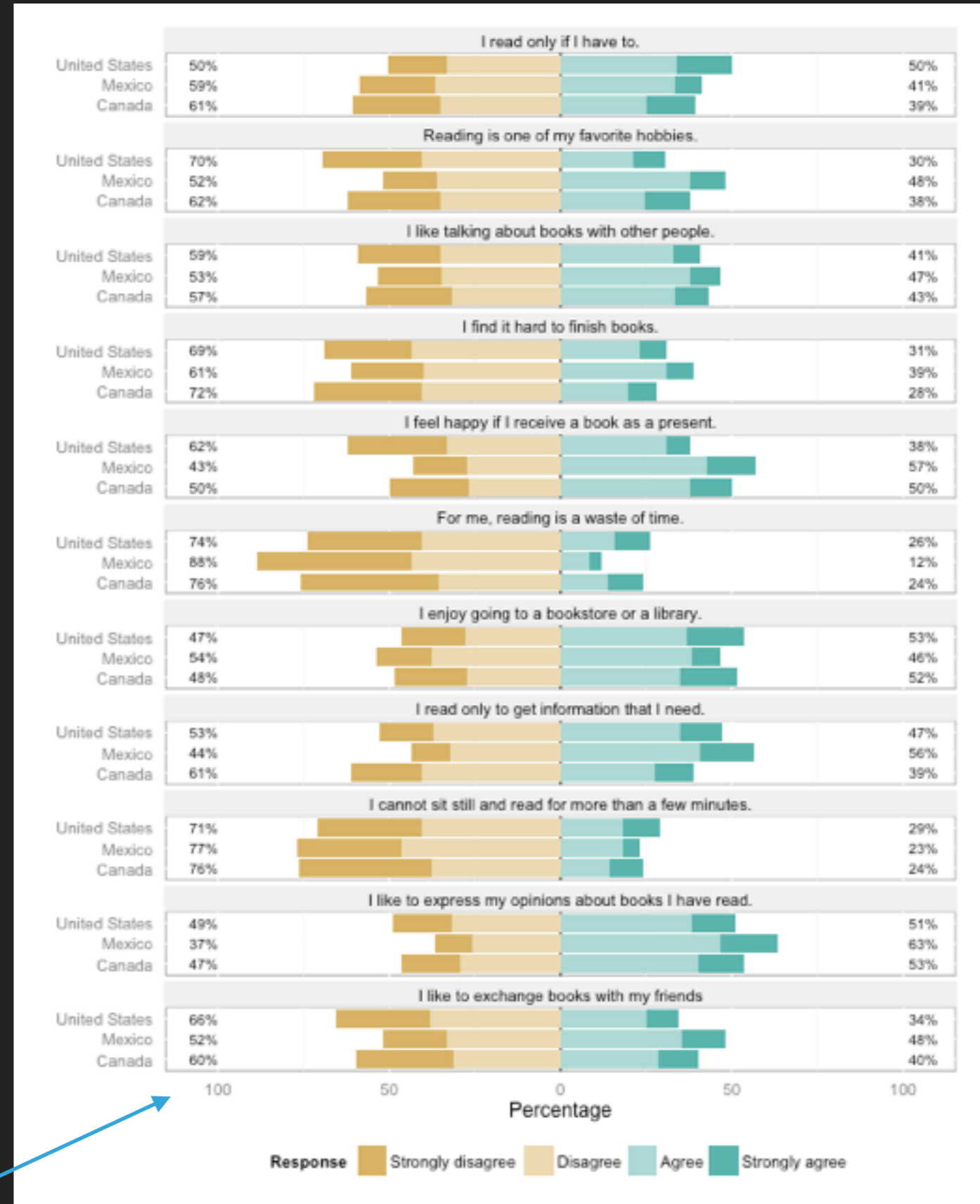
```
Min.      : 0.000   Min.      : 0.000   Min.      : 0.000  
1st Qu.  : 4.710   1st Qu.  : 4.720   1st Qu.  : 2.910  
Median   : 5.700   Median   : 5.710   Median   : 3.530  
Mean     : 5.731   Mean     : 5.735   Mean     : 3.539  
3rd Qu.  : 6.540   3rd Qu.  : 6.540   3rd Qu.  : 4.040  
Max.     :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)
```



# 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**

Michael Geoffrey Brown

Some stuff about Teaching, Learning, and Technology in American Higher Education

MY RESEARCH

MY PUBLICATIONS

MY TEACHING

BIO



## Bio



I study social learning and instructional technology in undergraduate STEM courses using learning analytics and network analysis. I am interested in how social learning experiences shape student outcomes, how students become engaged in their coursework, and how instructors organize their courses using digital social technologies. [\[CV\]](#)

GO TO:

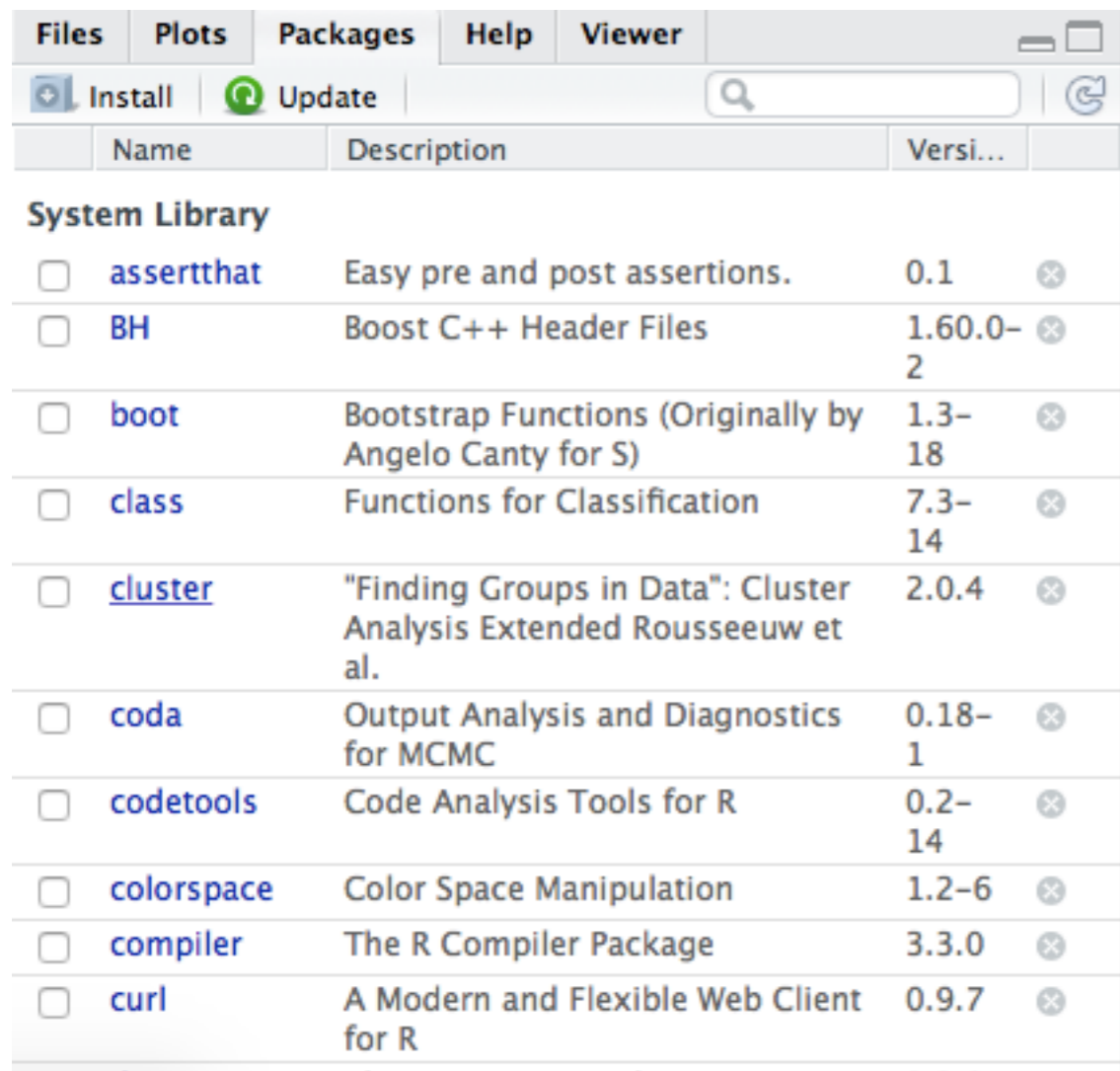
---

MICHAELBROWN.WORK/  
RWORKSHOP

## STARTING WITH PACKAGES

---

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



The screenshot shows the RStudio interface with the 'Packages' pane open. The pane has tabs for 'Install' and 'Update'. Below the tabs is a search bar and a refresh icon. The main area displays a table of packages. The table has columns for 'Name', 'Description', and 'Versi...'. The packages listed are:

Name	Description	Versi...
<b>System Library</b>		
<input type="checkbox"/> <a href="#">assertthat</a>	Easy pre and post assertions.	0.1
<input type="checkbox"/> <a href="#">BH</a>	Boost C++ Header Files	1.60.0-2
<input type="checkbox"/> <a href="#">boot</a>	Bootstrap Functions (Originally by Angelo Canty for S)	1.3-18
<input type="checkbox"/> <a href="#">class</a>	Functions for Classification	7.3-14
<input type="checkbox"/> <a href="#">cluster</a>	"Finding Groups in Data": Cluster Analysis Extended Rousseeuw et al.	2.0.4
<input type="checkbox"/> <a href="#">coda</a>	Output Analysis and Diagnostics for MCMC	0.18-1
<input type="checkbox"/> <a href="#">codetools</a>	Code Analysis Tools for R	0.2-14
<input type="checkbox"/> <a href="#">colorspace</a>	Color Space Manipulation	1.2-6
<input type="checkbox"/> <a href="#">compiler</a>	The R Compiler Package	3.3.0
<input type="checkbox"/> <a href="#">curl</a>	A Modern and Flexible Web Client for R	0.9.7

#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 proportions 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.

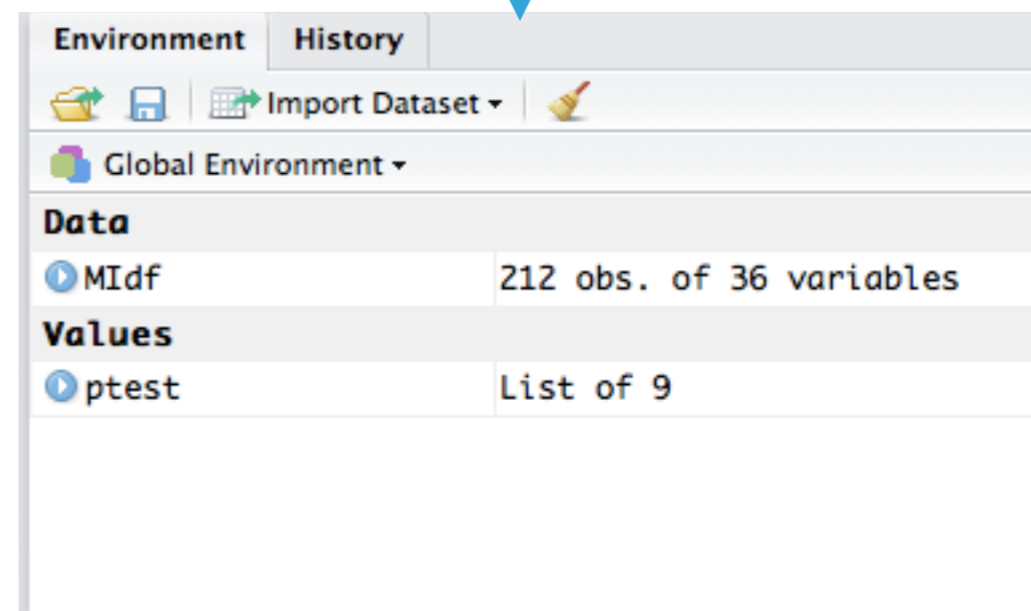
# LOADING DATA

## 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$E2_id <-match(IDKEY$username, Tailor$username)
setwd("~/Box Sync/Mike_Research_Project/STATS250")
write.csv(IDKEY, file="stats250IDKEY52016.csv")
rm()
```



The screenshot shows the RStudio Environment panel with the following content:

Environment	
Global Environment ▾	
<b>Data</b>	
▶ MIdf	212 obs. of 36 variables
<b>Values</b>	
▶ ptest	List of 9

A blue arrow points from the text above to the 'MIdf' entry in the Environment panel.



**Austin Community College District**

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

[Back to Search](#) [More Information](#) [Print Profile](#)

**Costs**



**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.](#)

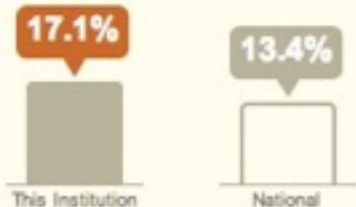
**Graduation Rate**



**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.

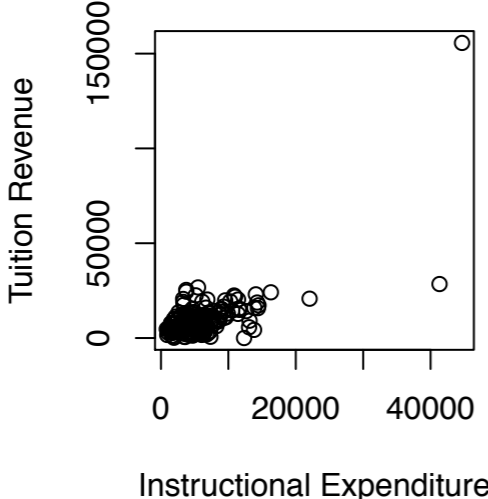
COLLEGE SCORE CARD

TUITION AND EXPENDITURES

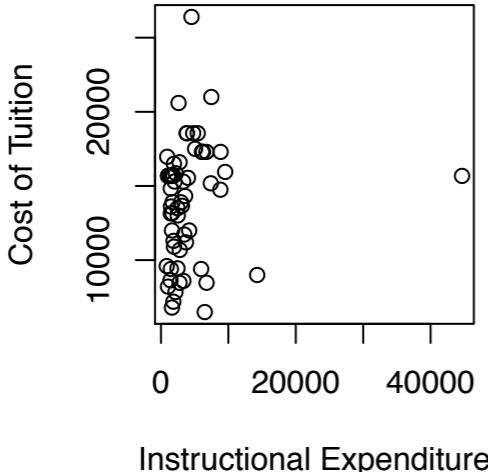


# EXPENDITURES AND TUITION

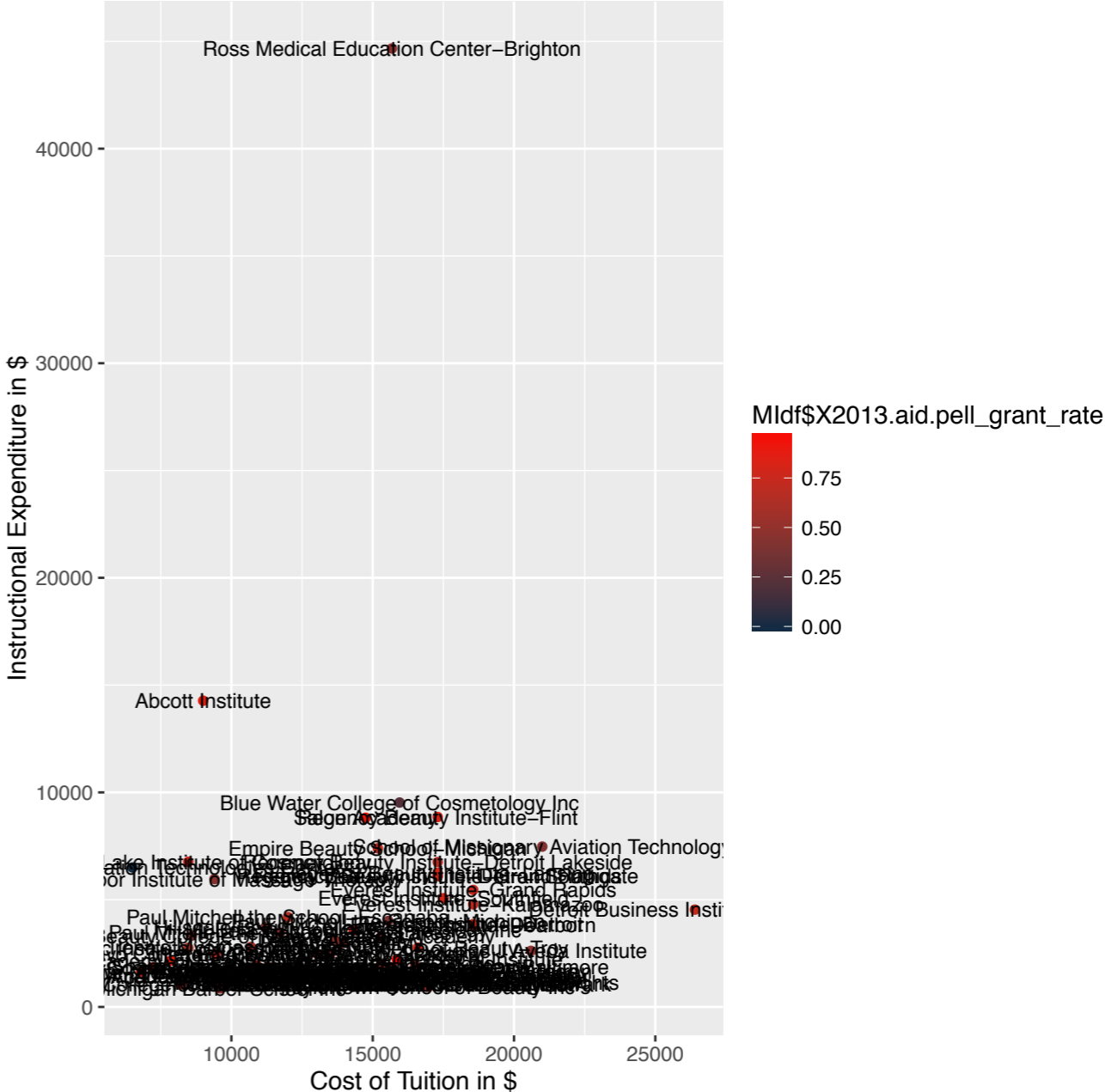
Expenditure vs Tuition Revenue 2013



Expenditure vs Cost of Tuition 2013



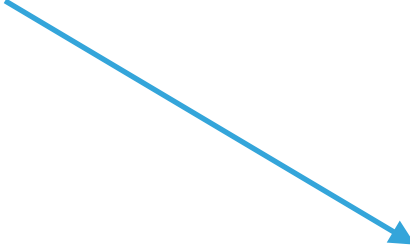
Cost of Tuition by Instructional Expenditure



---

# 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.size
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
```

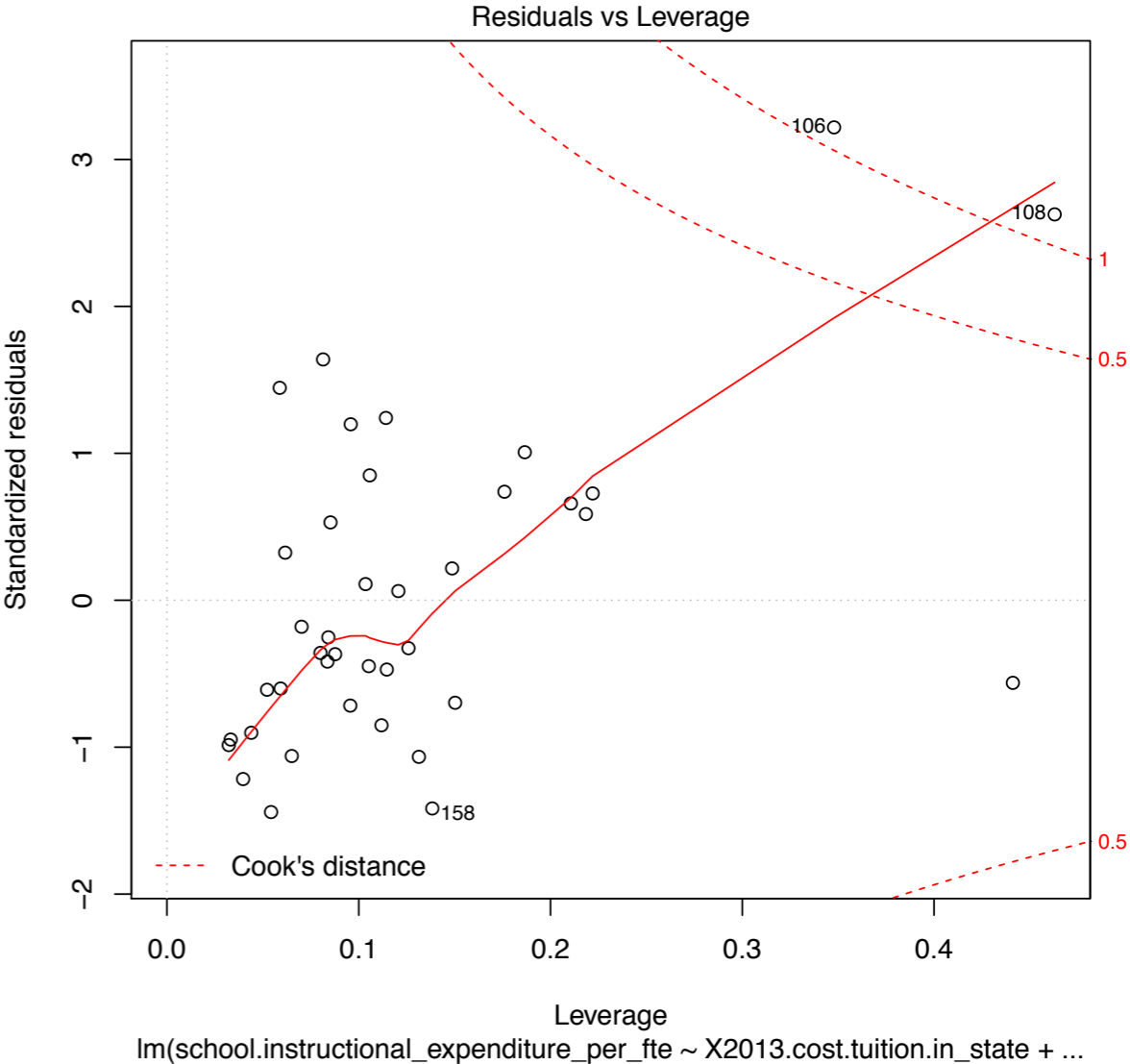
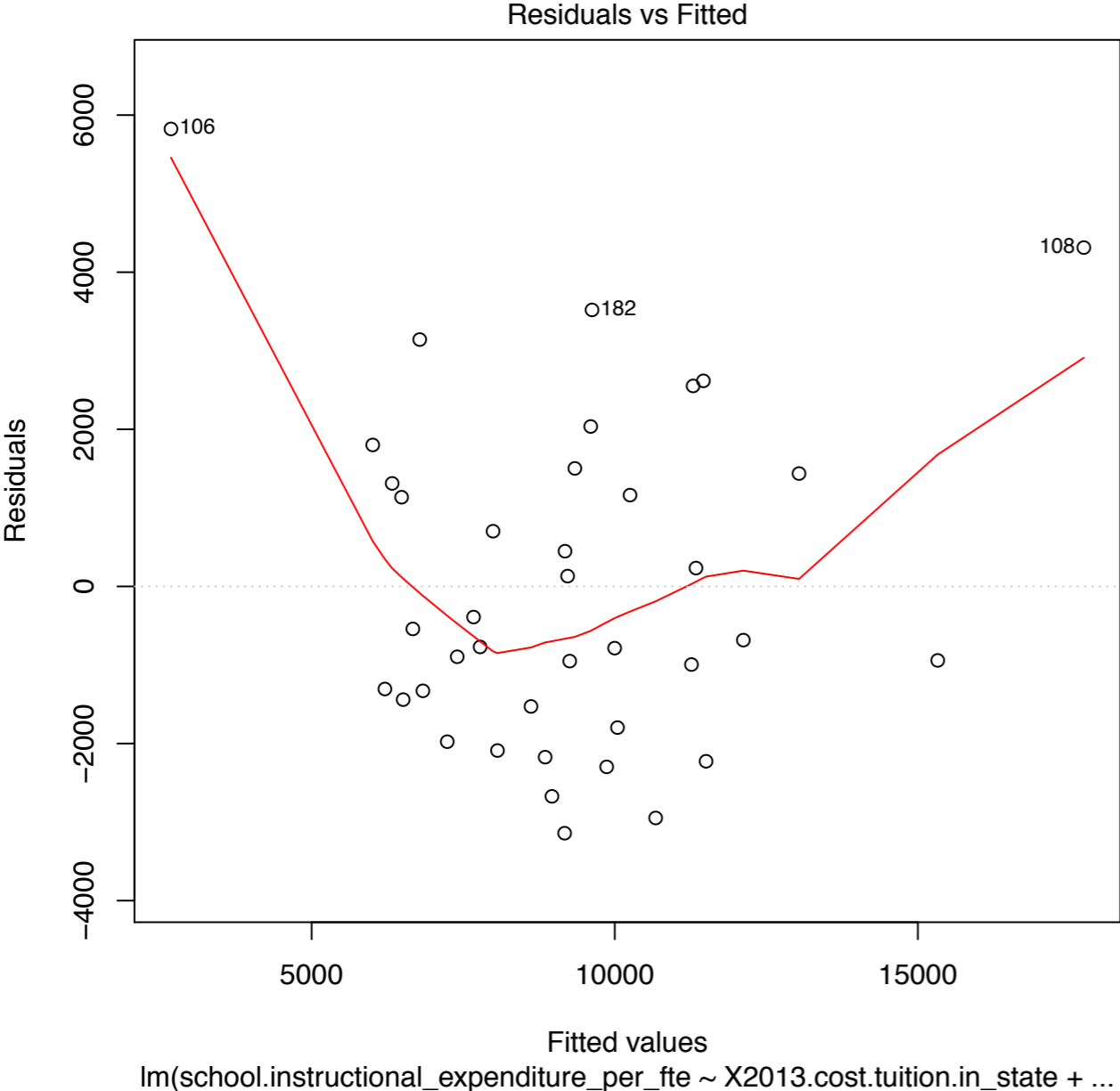
```
Coefficients:
```

	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
```

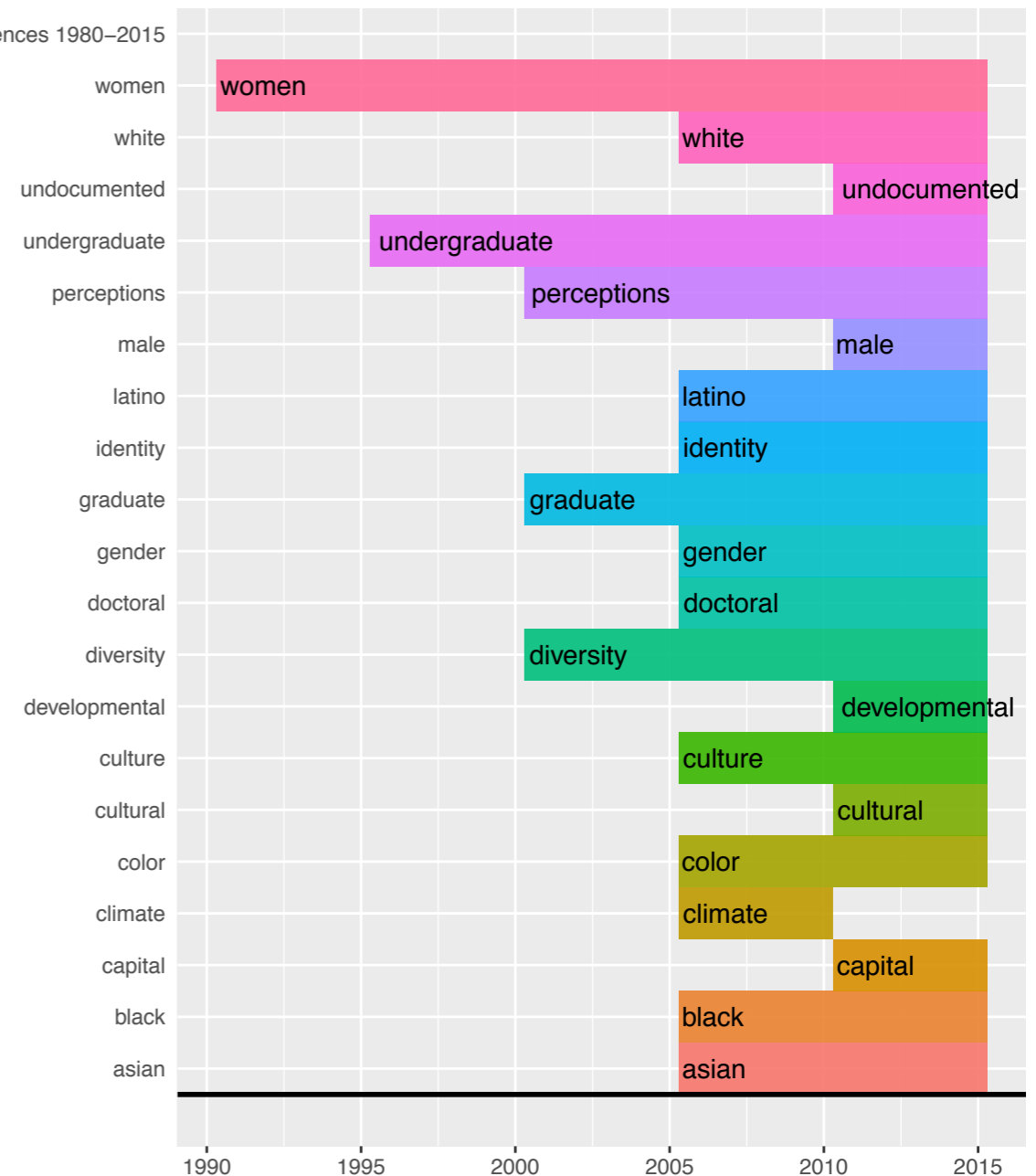
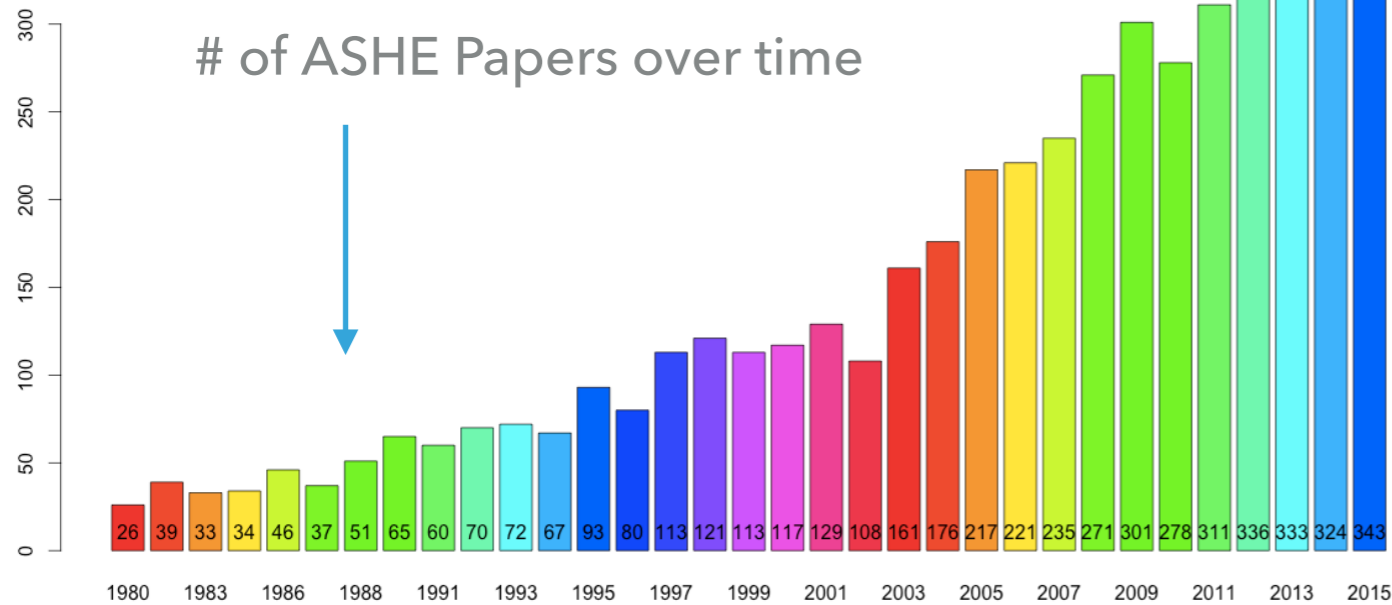
# REGRESSION DIAGNOSTICS



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

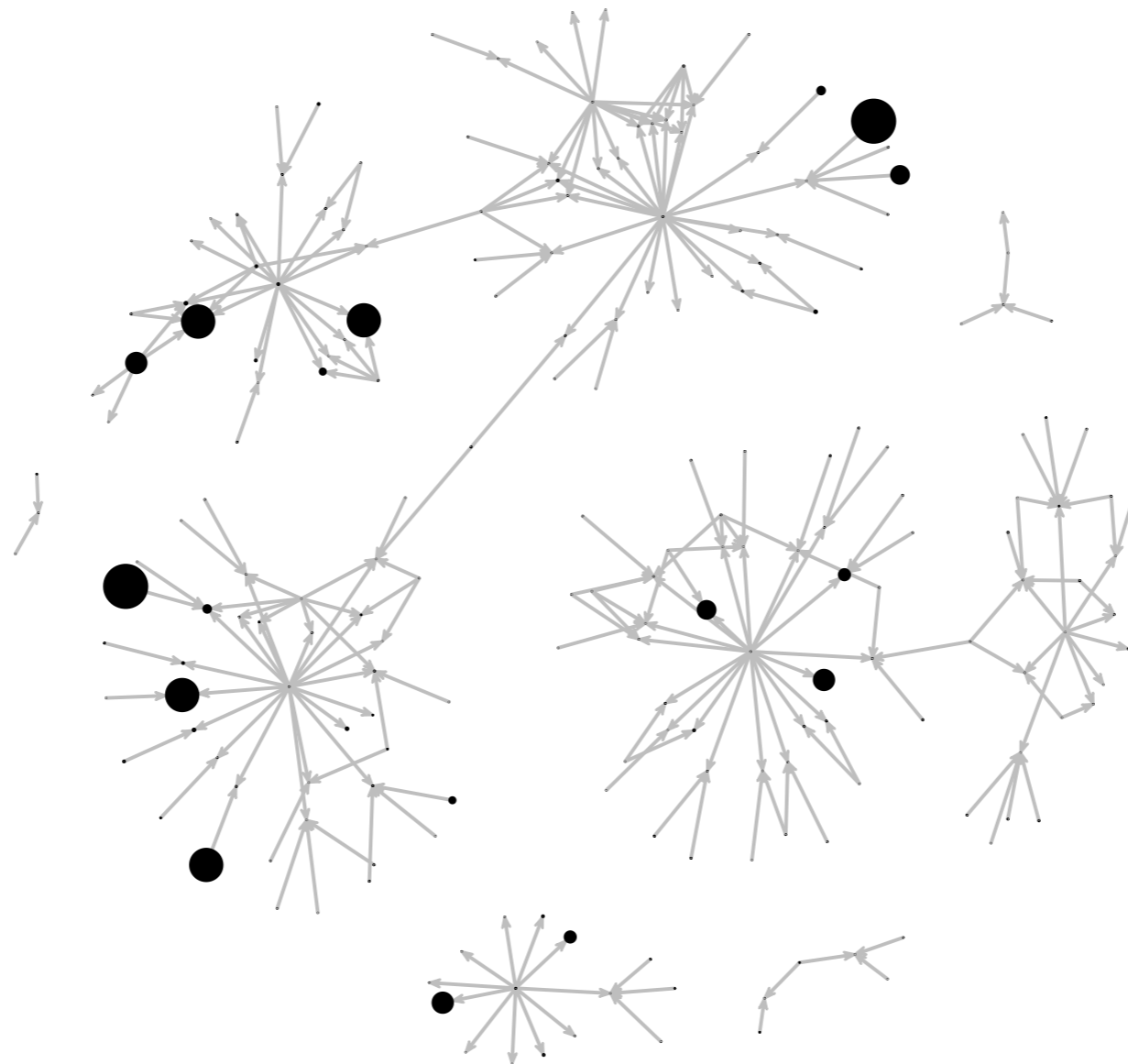
Frequent Title Keywords from papers  
about identity presented  
at ASHE between 1990-2015

ASHE Conferences 1980-2015



\*G-d willing

# CITATION ANALYSIS OF SCHOLARLY PRODUCTIVITY (CSHPE)



Faculty Collaboration Network

## GOOGLE VIZ AND RSHINY APPS

**Data set**

mtcars  diamonds  grid

**Plot type**

base  ggplot2

**Scale type**

normal

**DOUBLE-CLICK**

Delay: 400

**HOVER**

Input rate policy:  debounce  throttle

Delay: 200

NULL when outside

**BRUSH**

Direction(s):  xy  x  y

Input rate policy:  debounce  throttle

Delay: 200

Reset on new image

input\$plot\_dbclick: NULL

input\$plot\_click: NULL

input\$plot\_hover: NULL

input\$plot\_brush: NULL

```

app.R
library(ggplot2)
library(cairo) # For nicer ggplot2 output when deployed on Linux
library(DT)

# A modified version of mtcars with some columns removed, some columns added,
# and some columns as factors.
mtc <- mtcars
mtc$cyl <- factor(mtc$cyl)
mtc$am <- factor(mtc$am)
mtc$vs <- NULL
mtc$disp <- NULL
mtc$hp <- NULL
mtc$sec <- NULL
mtc$gear <- NULL
mtc$drat <- NULL
mtc$carb <- NULL

mtc$date <- Sys.Date() + seq_len(nrow(mtc))
mtc$datetime <- Sys.time() + 60 * seq_len(nrow(mtc))

# Data set with points on a grid
grid <- data.frame(
  x = rep(1:8, 4),
  xf = factor(rep(1:8, 4)),
  y = rep(1:4, each = 8),
  facet1 = factor(rep(1:2, 16)),
  facet2 = factor(rep(1:4, 8))
)

shinyApp(
  ui = fluidPage(
    # Some custom CSS
    tags$head(
      tags$style(HTML("
        /* Smaller font for preformatted text */
        pre, table.table {
          font-size: smaller;
        }
      "))
    ),
    body(
      min-height: 2000px;
    ),
    .option-group {
      border: 1px solid #ccc;
      border-radius: 6px;
      padding: 0px 5px;
      margin: 5px -10px;
      background-color: #f5f5f5;
    },
    .option-header {
      color: #79d;
      text-transform: uppercase;
      margin-bottom: 5px;
    }
  )
),
  server = function(input, output, session) {
    fluidRow(
      columnWidth=3,
      div(class = "option-group",
        radioButtons("dataset", "Data set",
          choices = c("mtcars", "diamonds", "grid"), inline = TRUE),
        radioButtons("plot_type", "Plot type",
          c("base", "ggplot2"), inline = TRUE),
        conditionalPanel("input.plot_type == 'base'",
          selectInput("plot_scaletype", "Scale type",
            c("normal" = "normal",
              "log" = "log",
              "x factor" = "x factor",
              "datetime" = "datetime")
          ), selectize = FALSE
        ),
        conditionalPanel("input.plot_type == 'ggplot2'",
          selectInput("ggplot_scaletype", "Scale type",
            c("normal" = "normal",
              "reverse (scale * reverse())" = "reverse",
              "log10 (scale * log10())" = "log10",
              "log2 (scale * continuous( trans=log2 trans()))" = "log2".
            )
          )
        )
      ),
      columnWidth=3,
      div(class = "option-group",
        sliders("double_click_delay", "DOUBLE-CLICK", "Delay", 100, 1000, 400),
        sliders("hover_delay", "HOVER", "Delay", 100, 1000, 200),
        sliders("brush_delay", "BRUSH", "Delay", 100, 1000, 200),
        sliders("nearpoints_max_dist", "NEARPOINTS() OPTIONS", "Max distance (pixels)", 1, 20, 5),
        sliders("nearpoints_max_rows", "NEARPOINTS() OPTIONS", "Max number of rows to select", 1, 100, 10)
      ),
      div(
        p("Points selected by clicking, with nearPoints():")
        p("Show 10 entries Search: ")
        p("mpg cyl wt am date datetime dist_")
        p("No data available in table")
        p("Showing 0 to 0 of 0 entries Previous Next")
      )
    )
  }

```

## SOME USEFUL RESOURCES FOR LEARNING R

- ▶ [TryR.codeschool.com](https://tryr.codeschool.com)
- ▶ *Fox's R Companion to Applied Regression*
- ▶ *Discovering Statistics Using R*
- ▶ *Learning Analytics with SNA and MPIA using R*
- ▶ QuickR

**BREAK**