Policy Relevant Visualization and Analysis of LDS Data With Open Source Tools

Jared Knowles

Policy Research Advisor
Wisconsin Department of Public Instruction

January 20, 2012 / MIS Conference

Outline

Outline

Outline

Defining Terms

Policy relevant analysis is answering questions that inform policy in a timely fashion and presenting results in an accessible and engaging fashion.

The Big Questions

- States and LEAs have an abundance of data, but how do we extract meaning from it?
- Can we do data analysis fast enough to inform decisions and improve outcomes?
- Can we produce analyses that are approachable to policy makers and the public so that they galvanize change?
- Can we do these things in a time of reduced staffing, decreased budgets, and a shortage of time?

Example

The state chief school officer asks:
"Do our state bilingual-bicultural programs provide any benefit to our students? Should we focus on ESL more or keep our BLBC programs?"

Options?

Option

- Contract with university faculty
- Consult existing literature
- Call the REL

Caveats

- This will take months. Budget proposal is due in three weeks. Costly.
- No studies in our state. State legislators not impressed.
- Study will take months.

Or Ask Jen

Not Enough Jen's

Outline

How do we do more?

What tools exist to help us turn data into usable information that informs decisions?

Examples from One State

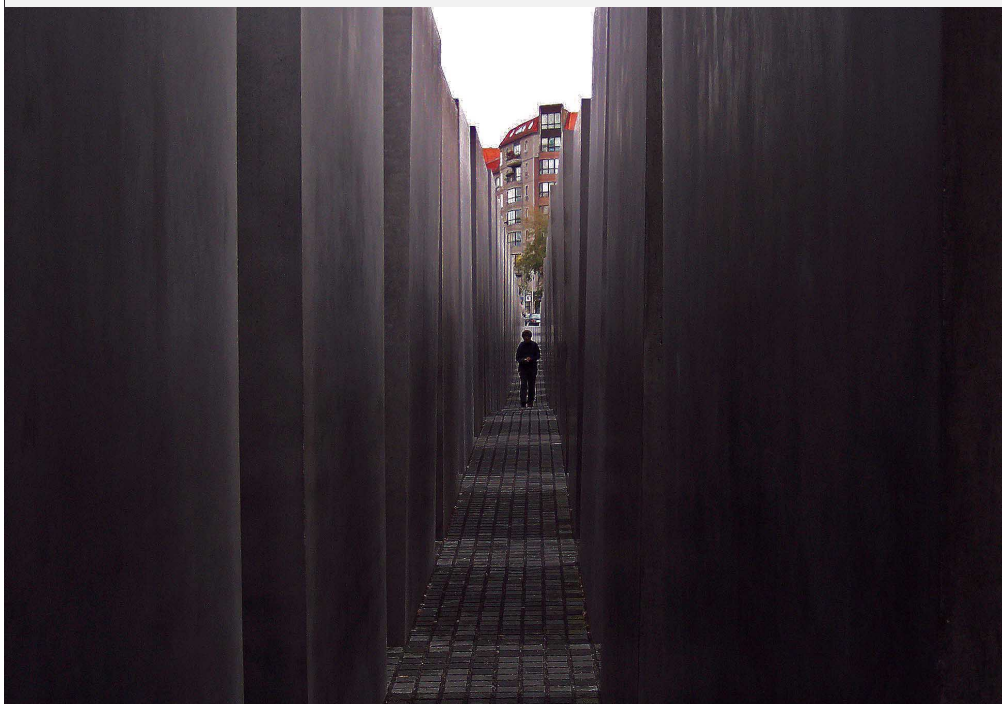
In Wisconsin we have done a few analyses that have helped us make decisions.

- An analysis of the effectiveness of bilingual-bicultural programs
- An analysis of reading performance and markers of struggling literacy
- An analysis of concentration in dropouts
- Data visualization to demonstrate problem scope for grants and press materials

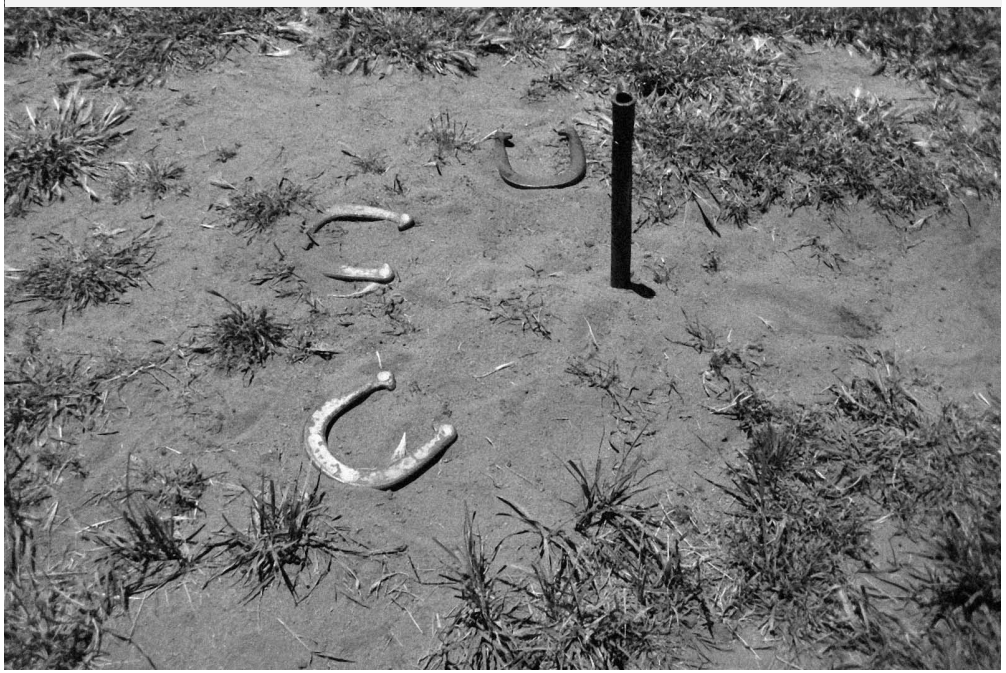
How?

Policy Relevant Analysis is FAST

Policy Relevant Analysis is FOCUSED



Policy Relevant Analysis is APPROXIMATE



Outline

Outline

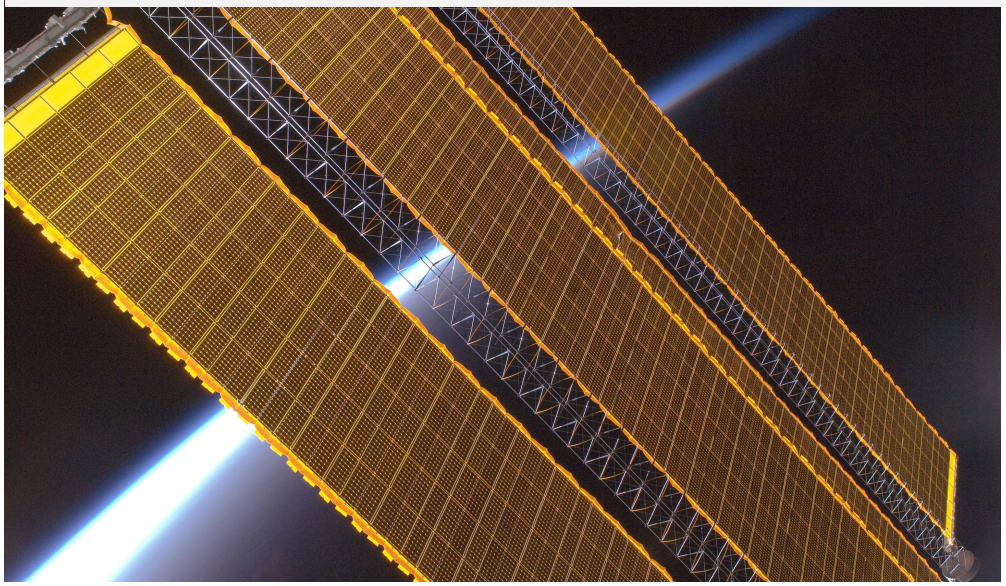
Gathering More Data

- States and districts collect hundreds of attributes about millions of students
- Data is collected before children reach school age and after they have moved to a college or a career
- Patterns can tell us how choices in policy will actually affect the population

Data is like ore

Analysis concentrates its value

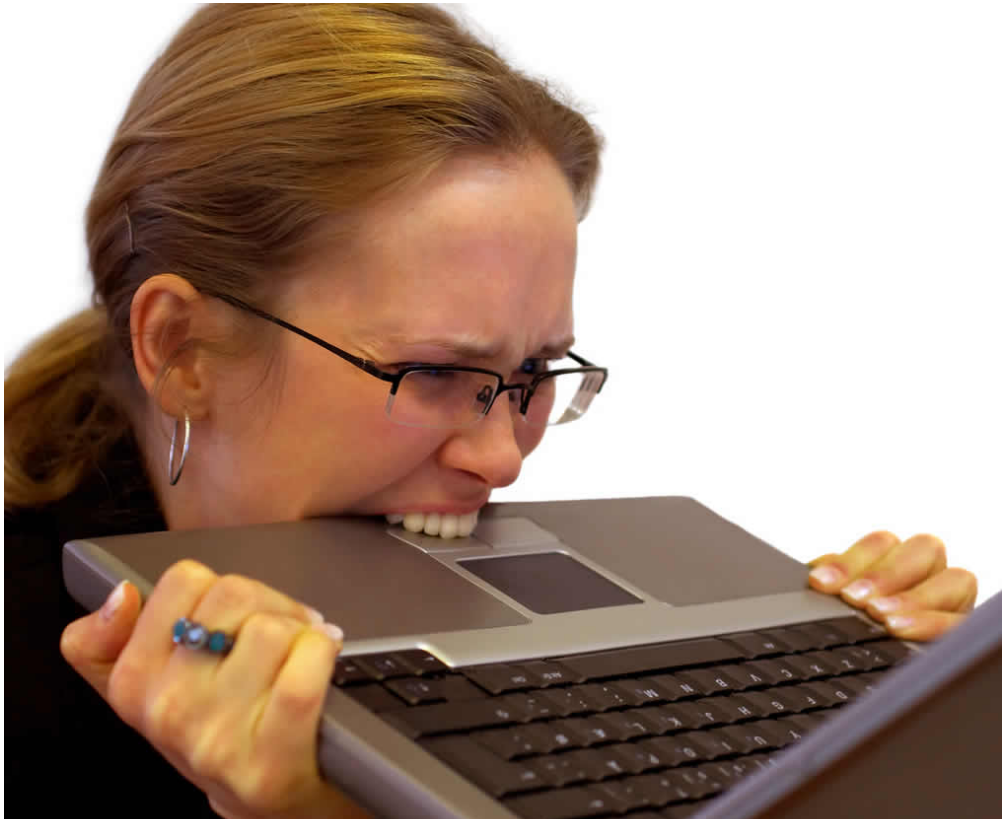
And it can be used to produce something



Outline

Institutional Frustrations

We just need to get our jobs done. We need to do them efficiently, but also transparently and in a reproducible manner. This is currently costly in time, money, and management resources.



Institutional Frustrations

We just need to get our jobs done. We need to do them efficiently, but also transparently and in a reproducible manner. This is currently costly in time, money, and management resources.

- Acquiring proprietary tools from vendors takes agreements, legal documents, and lag time
- Sharing data with external researchers requires legal agreements, levels of management approval, and planning time to specify narrow scope

Institutional Frustrations

- Analyses are often done in proprietary tool sets, poorly documented, and unable to be reproduced with updated data later

Analyses Don't Get Used

Often when we do an in house analysis it does not get used or only gets used once.

- In house analysis often relies on the expertise of one or two staff who are obligated elsewhere.
- Analysis are done using ad-hoc tools distributed among expertise of individual staff with no comprehensive standard.
- The information we have is dependent on individual staff and the analysis projects they undertake and their tenure supporting these efforts.
- Staff turnover threatens to change the information available to make decisions as knowledge leaves the agency, breaking continuity with previous reports

Incoherence

Irrelevant

Outline

AYP as an Example

- Think about how AYP is calculated within your agency for SEAs, or how school performance reports are distributed to schools for LEAs
- Who does this task? Could someone else step in and replace them and produce the exact same results if necessary?
- How many other employees could use the same tools and recreate this work?
- Is this risk acceptable for this pivotal function?

R As the Solution

Objections to Data Analysis

- Costly
- Slow and Time Consuming
- Technical and complex
- Opaque and not actionable

The R Solution

- R is free and open source
- R allows reproducible and sharable analysis across researchers
- R can be scripted to do common tasks
- R is a lingua franca that standardizes common tasks

More Support for R

- R is a common tool among data experts at major universities
- No need to go through procurement, R can be installed in any environment on any machine and used with no licensing or agreements needed
- R source code is very readable to increase transparency of processes
- R code is easily borrowed from and shared with others
- R is incredibly flexible and can be adapted to specific local needs
- R is under incredibly active development, improving greatly, and supported wildly by both professional and academic developers

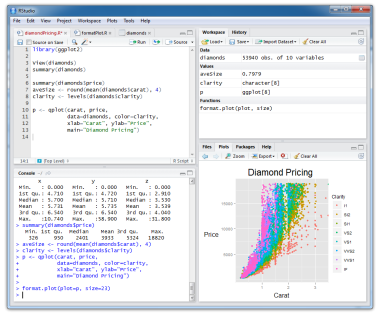
Outline

Outline

- R is an Open Source (and freely available) environment for statistical computinga nd graphics
- Available for Windows, Mac OS X, and Linux
- R is being actively developed with two major releases per year and dozens of releases of add on packages
- R can be extended with 'packages' that contain data, code, and documentation to add new functionality

Using R

- R can be used with an excellent Integrated Development Environment
- RStudio makes many of the basic tasks in R much easier like
  - Importing data
  - Previewing plots
  - Version control
  - Collaboration
- Greatly increases ease of use



Pros and Cons of R

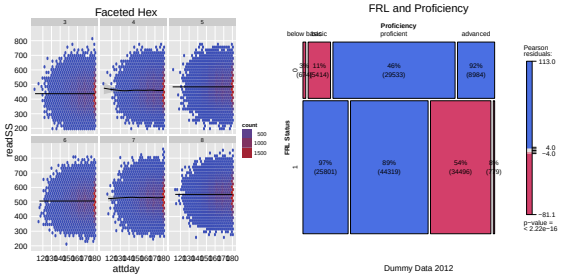
Pros of **R**

- Open source and freely available on all platforms
- Scripting for reproducible and transparent analyses
- Extensible to fit skills, needs, and cutting edge techniques
- Excellent graphical and output capabilities

Cons

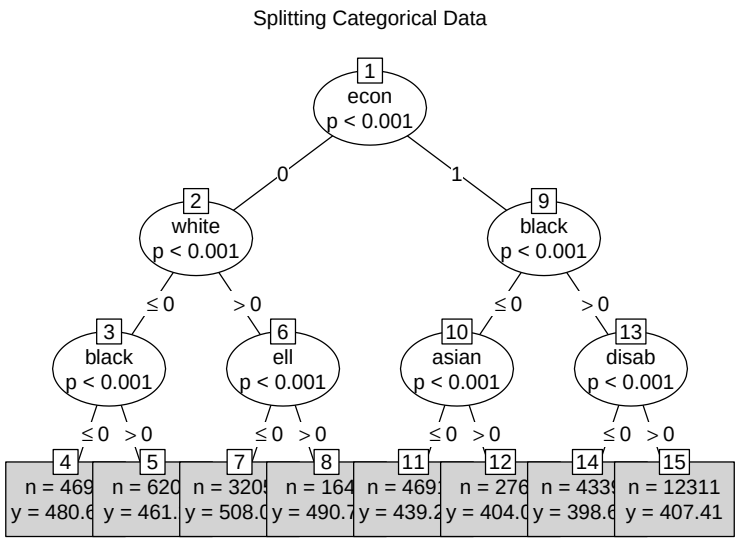
- Steep learning curve and command line interface
- Requires specific inputs to get desired results
- Unforgiving of misspecification of inputs
- Data input can be tricky at first

Examples of R Figures

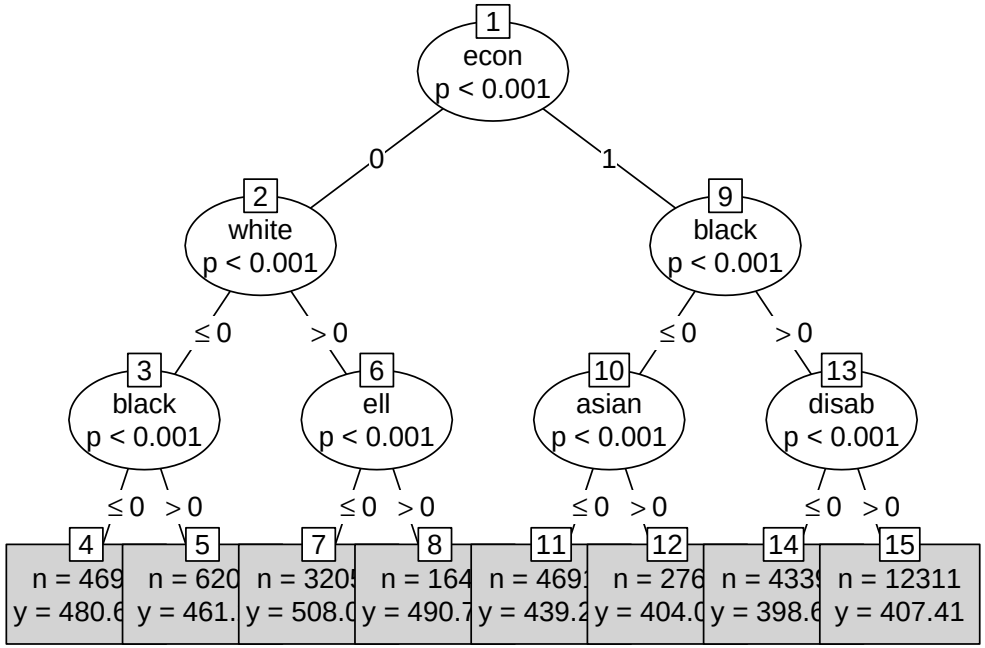


Outline

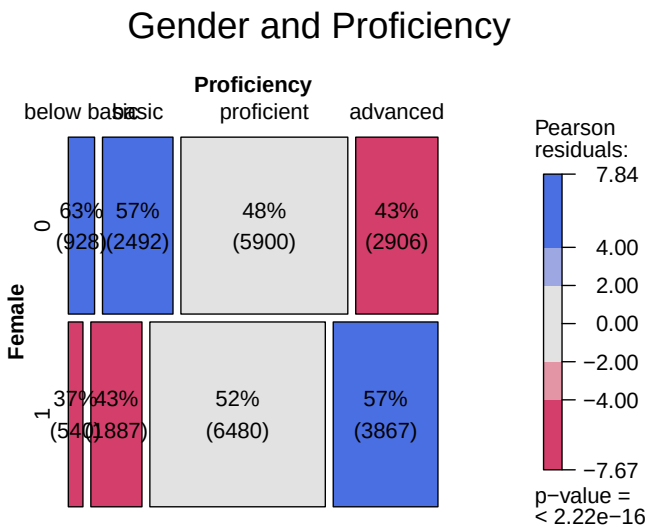
Inference Trees



Inference Trees

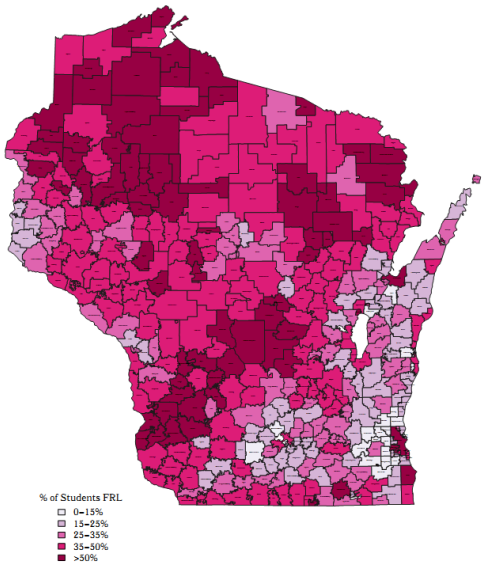


Visual Crosstabs

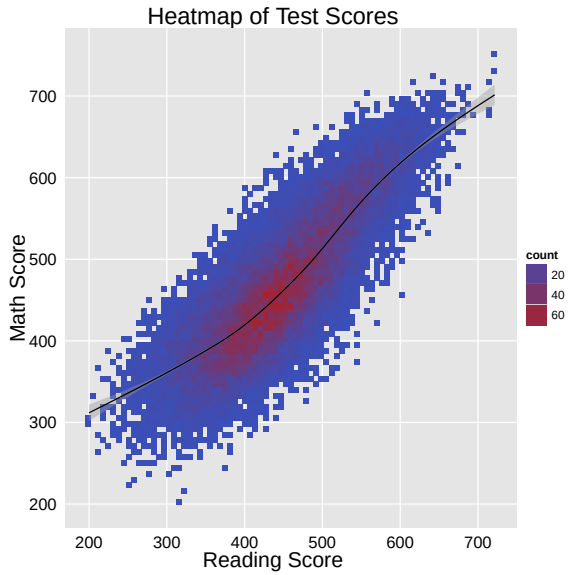


Dummy Data 2012

Maps

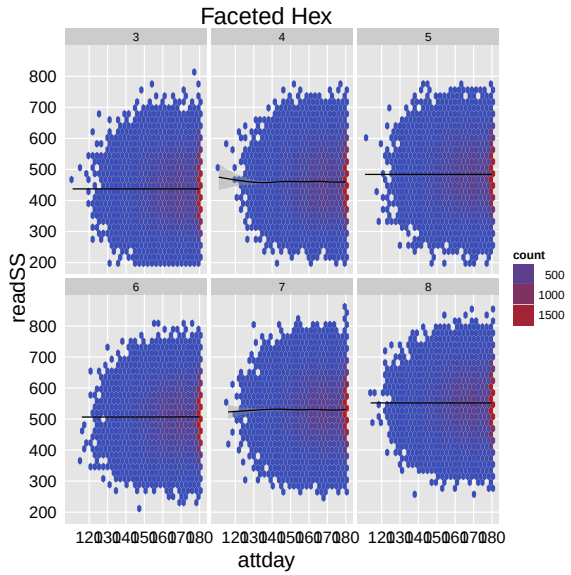


Heatmap

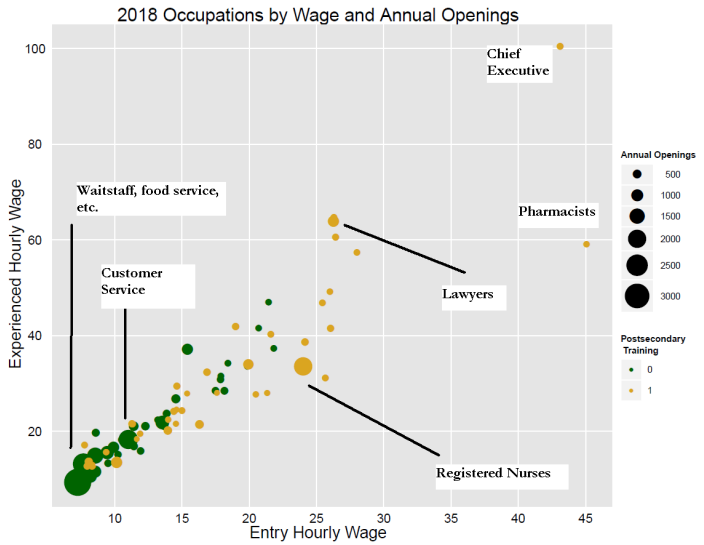


Conditional Density

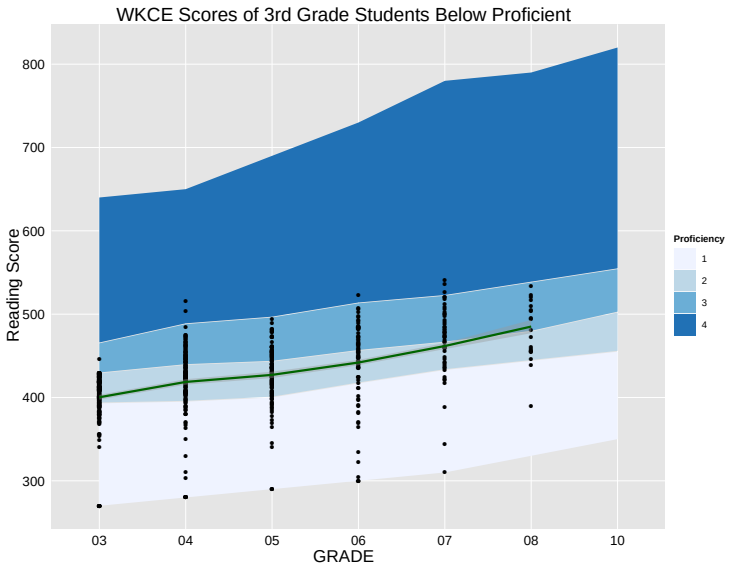
Faceted Hex



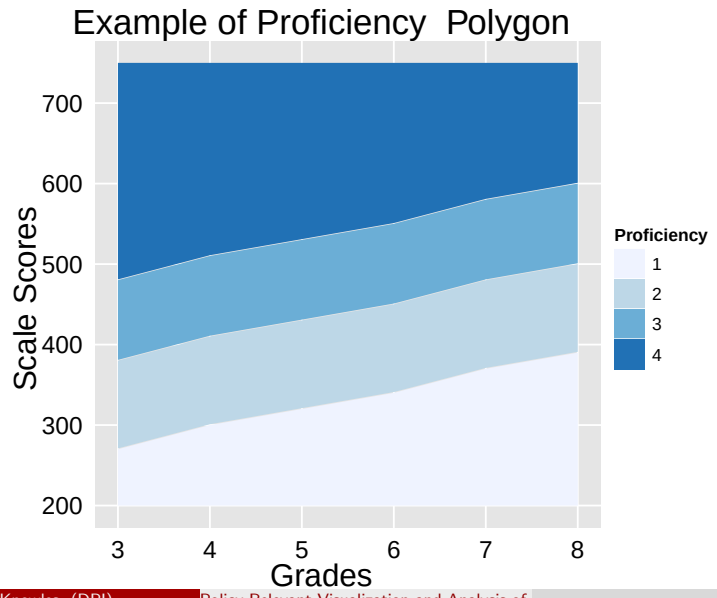
Polished Scatter



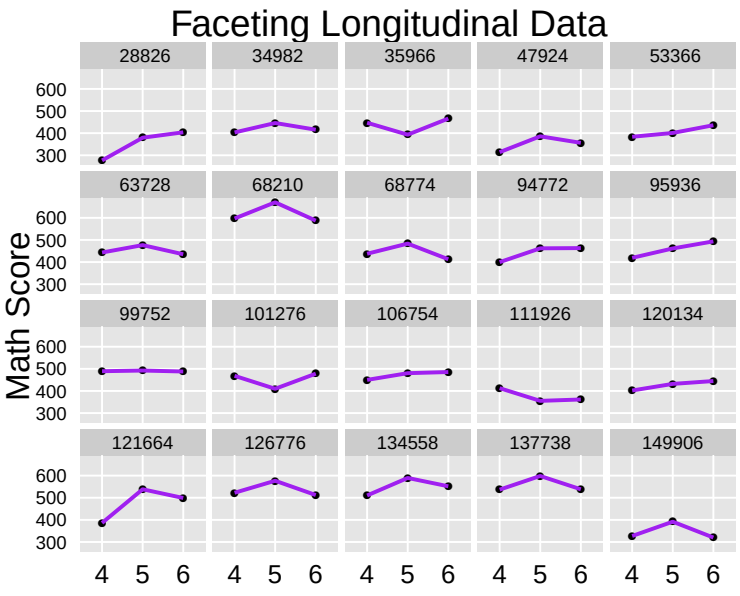
Longitudinal Data



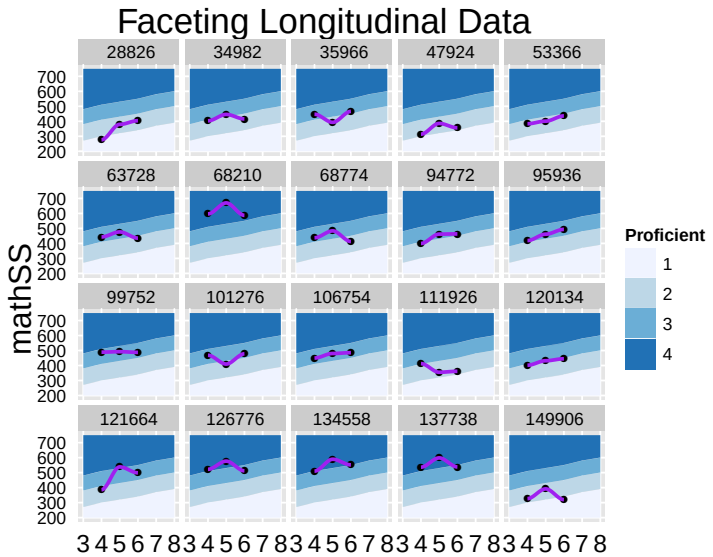
Proficiency Polygon



Individual Growth Trajectories



Individual Growth Trajectories



Outline

The Command Line

- R can be tricky because it uses command lines.
- This is powerful, but requires a learning curve.
- Some simple calculations can give a feel for how R works

```
> 2+2
[1] 4
> 7*4
[1] 28
> exp(3)
[1] 20.08554
> pi
[1] 3.141593
```


Deconstruct R Commands

```
> summary(student_long[,28:31])
     readSS          mathSS             proflvl
 Min.   :200.0   Min.   :200.0   below basic: 37618
 1st Qu.:430.3   1st Qu.:419.5   basic      : 85322
 Median :495.6   Median :480.9   proficient :195231
 Mean   :495.1   Mean   :483.7   advanced   :131829
 3rd Qu.:559.5   3rd Qu.:545.4
 Max.   :866.9   Max.   :839.9
 race
 A:  8802
 B:185748
 H: 50025
 I:  3732
 W:201693
```

- **summary** is the function
- **student_ long** is the data object

Simple R Operations

```
> with(student_long,mean(readSS[year=='2001' & grade==4]))
[1] 445.0533
> with(student_long,median(readSS[year=='2001' & grade==4]))
[1] 442.5596
> with(student_long,max(readSS[year=='2001' & grade==4]))
[1] 721.4892
> with(student_long,min(readSS[year=='2001' & grade==4]))
[1] 200
> with(student_long,summary(readSS[year=='2001' & grade==4]))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  200.0   393.3   442.6   445.1   494.7   721.5
```

Crosstabs

Let's test for balance among some categories of students

```
> with(subset(student_long,year=='2001'
+             & grade==3),table(female,race))

      race
female    A    B    H    I    W
     0  234 5087 1414   99 5504
     1  255 5209 1381  113 5704

> #As proportions
> with(subset(student_long,year=='2001'
+             & grade==3),round(prop.table
+             (table(female,race))*100),4)

      race
female  A  B  H  I  W
     0  1 20  6  0 22
     1  1 21  6  0 23
```


Crosstabs

We can even output the results of R commands into a print-ready format. As we have below.

	A	B	H	I	W
0	1.00	20.00	6.00	0.00	22.00
1	1.00	21.00	6.00	0.00	23.00

Outline

Outline

Doing More than the Basics

- R can routinize basic functions like tables, crosstabs, and visualization of data
- R can also be extended to do more advanced analyses like multilevel modeling, spatial error modeling, Bayesian data analysis, forecasting, and simulation
- R can do advanced graphical functions as well
- R can even be expanded to incorporate additional programming languages like Python, C++, and Java

The downside of this is that these functions can have a steep learning curve.

ANOVA

We can also do statistical tests using both Bayesian and Frequentist methods.

```
> nova1<-aov(readSS~female*race*econ,data=novaset)
> summary(nova1)
                    Df   Sum Sq  Mean Sq  F value   Pr(>F)
female               1  2097543  2097543  728.507  < 2e-16
race                 4 40225804 10056451 3492.748  < 2e-16
econ                 1 22330214 22330214 7755.600  < 2e-16
female:race          4    81042    20261    7.037 1.17e-05
female:econ          1     6797     6797    2.361   0.1244
race:econ            4   189430    47358   16.448 1.83e-13
female:race:econ     4    23672     5918    2.055   0.0838
Residuals        24980 71923350     2879

female              ***
race                ***
econ                ***
female:race         ***
female:econ
race:econ           ***
female:race:econ
```


Pretty Output

We can also do print-ready model outputs with R's extensible formatting

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
female	1	2097543.47	2097543.47	728.51	0.0000
race	4	40225803.97	10056450.99	3492.75	0.0000
econ	1	22330213.97	22330213.97	7755.60	0.0000
female:race	4	81042.06	20260.51	7.04	0.0000
female:econ	1	6797.30	6797.30	2.36	0.1244
race:econ	4	189430.29	47357.57	16.45	0.0000
female:race:econ	4	23671.91	5917.98	2.06	0.0838
Residuals	24980	71923350.40	2879.24		

Outline

A simple OLS Model I

```
> mod1<-lm(readSS~female*race*econ+grade*year,data=student_long)
> summary(mod1)

Call:
lm(formula = readSS ~ female * race * econ + grade * year, data = s

Residuals:
     Min       1Q   Median       3Q      Max
-232.761  -34.179    0.123   34.270  264.081

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       373.8444     1.4086 265.397  < 2e-16 ***
female             16.3185     1.8846   8.659  < 2e-16 ***
raceB             -39.3492     1.4405 -27.315  < 2e-16 ***
raceH             -35.1787     1.5318 -22.966  < 2e-16 ***
raceI               3.6151     2.4163   1.496   0.1346
```


A simple OLS Model II

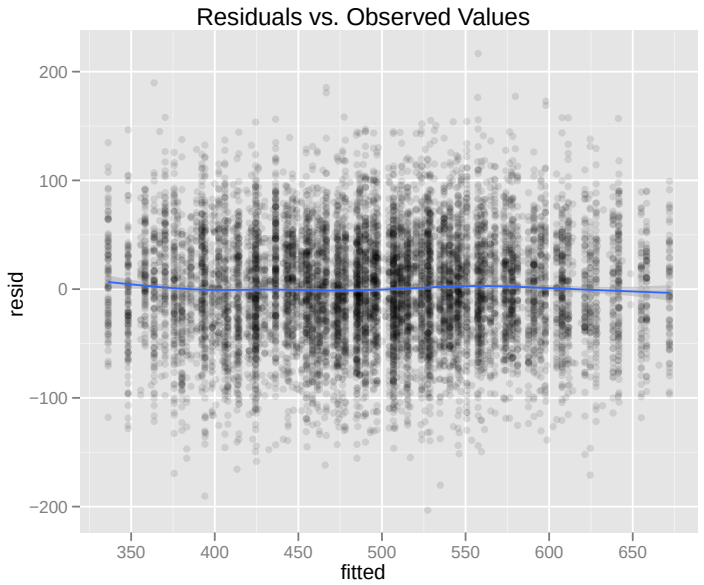
```
raceW                8.9380     1.3582   6.581 4.68e-11 ***
econ1              -77.9826     1.6455 -47.391  < 2e-16 ***
grade               22.0576     0.0776 284.258  < 2e-16 ***
year2001             2.7031     0.6320   4.277 1.89e-05 ***
year2002           140.6231     0.6320 222.506  < 2e-16 ***
female:raceB         1.9210     2.0292   0.947   0.3438
female:raceH        -1.5938     2.1502  -0.741   0.4585
female:raceI         4.6054     3.4470   1.336   0.1815
female:raceW        -2.6560     1.9121  -1.389   0.1648
female:econ1        -1.2084     2.3147  -0.522   0.6016
raceB:econ1         13.5581     1.7386   7.798 6.28e-15 ***
raceH:econ1         28.8879     1.8427  15.677  < 2e-16 ***
raceI:econ1          2.5158     3.0038   0.838   0.4023
raceW:econ1         14.8205     1.6773   8.836  < 2e-16 ***
grade:year2001       8.2691     0.1097  75.353  < 2e-16 ***
grade:year2002      -5.1974     0.1097 -47.361  < 2e-16 ***
female:raceB:econ1  -5.0485     2.4468  -2.063   0.0391 *
```


A simple OLS Model III

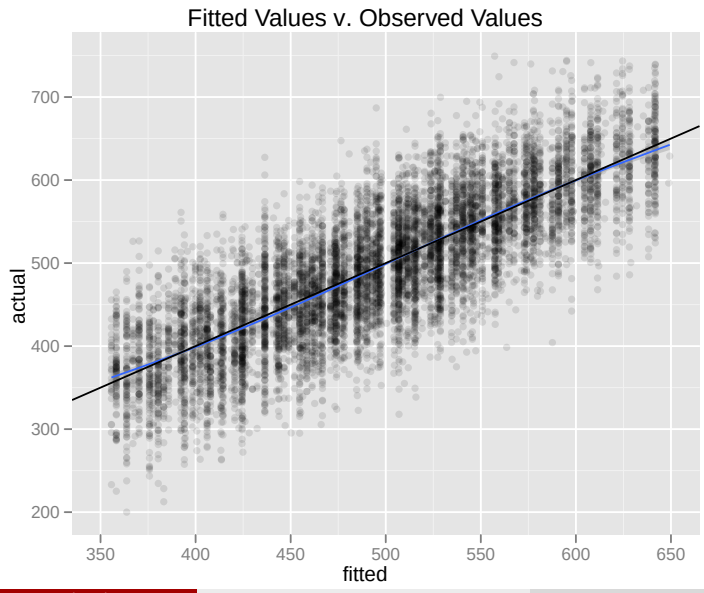
```
female:raceH:econ1  -2.2068     2.5869  -0.853   0.3936
female:raceI:econ1  -4.4296     4.2397  -1.045   0.2961
female:raceW:econ1   0.2727     2.3594   0.116   0.9080
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂ

Residual standard error: 51.33 on 449975 degrees of freedom
Multiple R-squared: 0.6852,  Adjusted R-squared: 0.6852
F-statistic: 4.081e+04 on 24 and 449975 DF,  p-value: < 2.2e-16
```


Model Evaluation

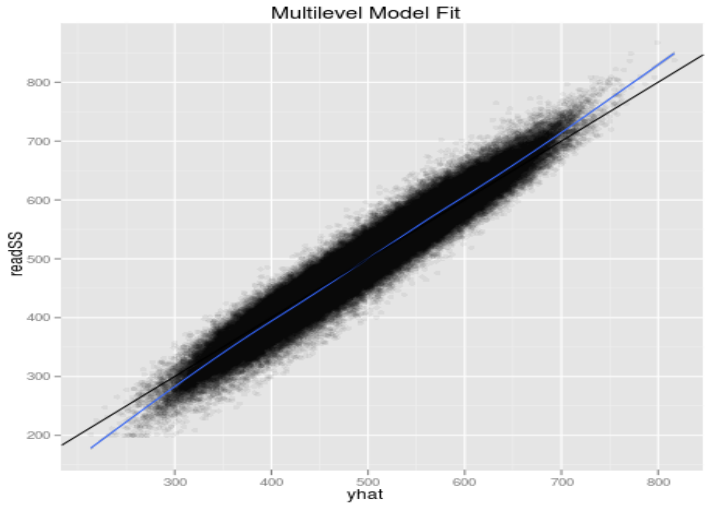


Model Evaluation Part 2



Better Fitting

Using advanced techniques we can greatly improve our model fit over the OLS model



Compare OLS and Mixed Effects

A simple mixed-effects model estimated with an R package can outperform the simple OLS without much additional effort