Doing IPEDS Analysis the Easy Way

Michael J. Smith, Portland State University

IPEDS

Data Center Help Desk (866) 558-0658

Start over | Save session | Help | MAIN MENU

Custom Data Files

Final Release Data (Change)

1. Select Institutions | 2. Select Variables | 3. Output

My Comparison Institution - None Selected

Select Variables - Total 0 variables selected

ADD

How would you like to select institutions to include in your data file/report?

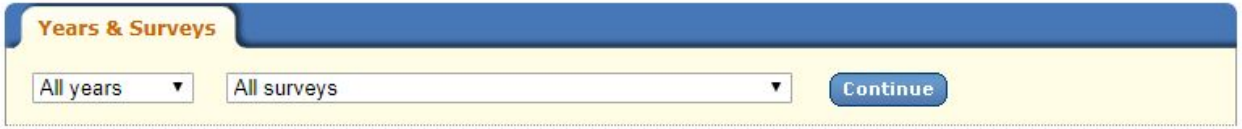
By Names or UnitIDs | By Groups | By Variables | By Uploading a File

Enter either an institution name or UnitID (or a comma separated list of UnitIDs) in the text box below. As you begin typing, a list of matching institutions will appear. You can select a single institution by clicking on it from the list, or, if you want all institutions on the list, click "Select".

Institution Name

Select

The flat files



Data files are available in ZIP format.

Year	Survey	Title	Data File	Stata Data File	Programs	Dictionary
2018	Institutional Characteristics	Directory information	HD2018	HD2018_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Educational offerings, organization, services and athletic associations	IC2018	IC2018_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Student charges for academic year programs	IC2018_AY	IC2018_AY_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Student charges by program (vocational programs)	IC2018_PY	IC2018_PY_STATA	SPSS, SAS, STATA	Dictionary
2018	Institutional Characteristics	Response status for all survey components	FLAGS2018	FLAGS2018_STATA	SPSS, SAS, STATA	Dictionary
2018	12-Month Enrollment	12-month unduplicated headcount: 2017-18	EFFY2018	EFFY2018_STATA	SPSS, SAS, STATA	Dictionary
2018	12-Month Enrollment	12-month instructional activity: 2017-18	EFIA2018	EFIA2018_STATA	SPSS, SAS, STATA	Dictionary

The Access databases

Documentation for IPEDS Access Databases: All IPEDS Access Databases contain metadata tables that describe each data table and provide a list of the variables. The same metadata tables are placed in a WinZip Excel workbook to serve as a standalone reference without having to download an entire database.

Download an IPEDS Access Database:

Database Name	Documentation	Release Type	Release Date
2017-18 Access zip (74.7mb) | decompressed (526mb)	2017-18 Excel (IPEDS201718Tablesdoc.xlsx, 1365kb)	Provisional	January 2019
2016-17 Access zip (64.2mb) | decompressed (553mb)	2016-17 Excel (IPEDS201617Tablesdoc.xlsx, 1384kb)	Final	January 2019
2015-16 Access zip (72.9mb) | decompressed (519mb)	2015-16 Excel (IPEDS201516Tablesdoc.xlsx, 1322kb)	Final	August 2018
2014-15 Access	2014-15 Excel		

However...

FRIENDS DON’T LET FRIENDS USE ACCESS

(for anything)

The *ipeds* package

Credit: Jason Bryer, Ph.D., Excelsior College; https://github.com/jbryer

Available on CRAN, but recommend getting the latest version from Github instead.

How *ipeds* works

1. Downloads the Access database from the NCES web site
2. Extracts all the tables from the Access database and saves them as a compressed R data file
3. Provides tools for extracting collection results from the file

However...

The current version of *ipeds* depends on mdbtools, which is not available on Windows platforms.

Options:

- Install Linux!
- Get hold of the R data files from elsewhere
- Use functions from *ipeds* that don't require mdbtools
- Wait...

Easy way #1: Longitudinal analysis

“Hey, Michael: Can you make me a chart of PSU’s graduation trends, compared to our peer institutions?”

Hard to do the regular way: requires multiple databases or CSV files.

Easy to do the easy way: script the downloads,

Caveat: some variables change name or location over time

Easy way #1: Longitudinal analysis

Foundations: Pull the most recent IC collection to get institution names and IDs

```
directory <- ipeds_survey(table = 'HD', year = 2018, dir=dir)
names(directory) <- tolower(names(directory))

getinstnm <- directory$instnm
names(getinstnm) <- directory$unitid
```

Easy way #1: Longitudinal analysis

Next, grab the 150% graduation rates:

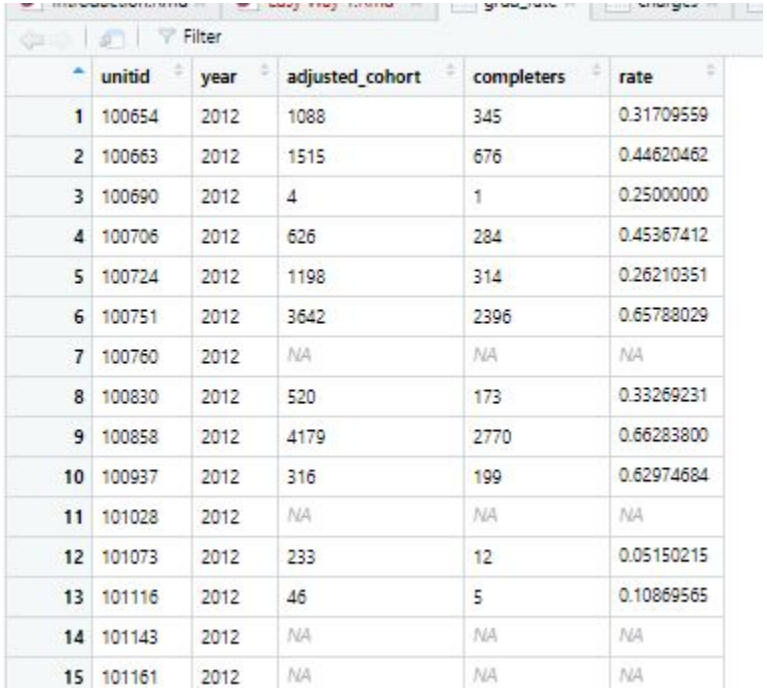
```
grad_rate_get <- function(year) {
  gradrates <- (ipeds_survey(table='GR',year=year, dir=dir))
  names(gradrates) <- tolower(names(gradrates))
  theRates <- dcast(gradrates, unitid ~ grtype, value.var = 'grtotlt')
  names(theRates) <- c('unitid','adjusted_cohort','completers')
  theRates$rate <- theRates$completers/theRates$adjusted_cohort
  theRates$year <- year

  return(theRates[, c('unitid','year','adjusted_cohort','completers','rate')])

}
```

Easy way #1: Longitudinal analysis

Next, grab the 150% graduation rates:

```
years <- (2012:2018)


for (i in seq_along(years)) {
       if (i==1) {grad_rate <- grad_rate_get(years[i])}
      else {grad_rate <- rbind(grad_rate, grad_rate_get(years[i]))}
}
```

Easy way #1: Longitudinal analysis



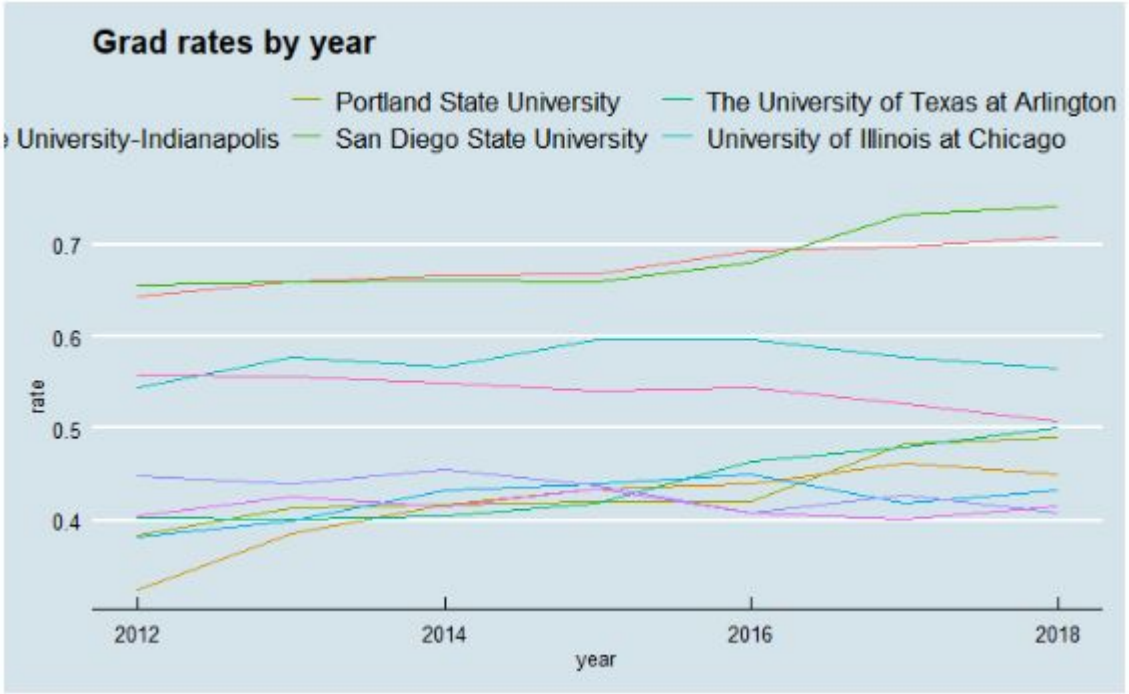
Filter

	unitid	year	adjusted_cohort	completers	rate
1	100654	2012	1088	345	0.31709559
2	100663	2012	1515	676	0.44620462
3	100690	2012	4	1	0.25000000
4	100706	2012	626	284	0.45367412
5	100724	2012	1198	314	0.26210351
6	100751	2012	3642	2396	0.65788029
7	100760	2012	*NA*	*NA*	*NA*
8	100830	2012	520	173	0.33269231
9	100858	2012	4179	2770	0.66283800
10	100937	2012	316	199	0.62974684
11	101028	2012	*NA*	*NA*	*NA*
12	101073	2012	233	12	0.05150215
13	101116	2012	46	5	0.10869565
14	101143	2012	*NA*	*NA*	*NA*
15	101161	2012	*NA*	*NA*	*NA*

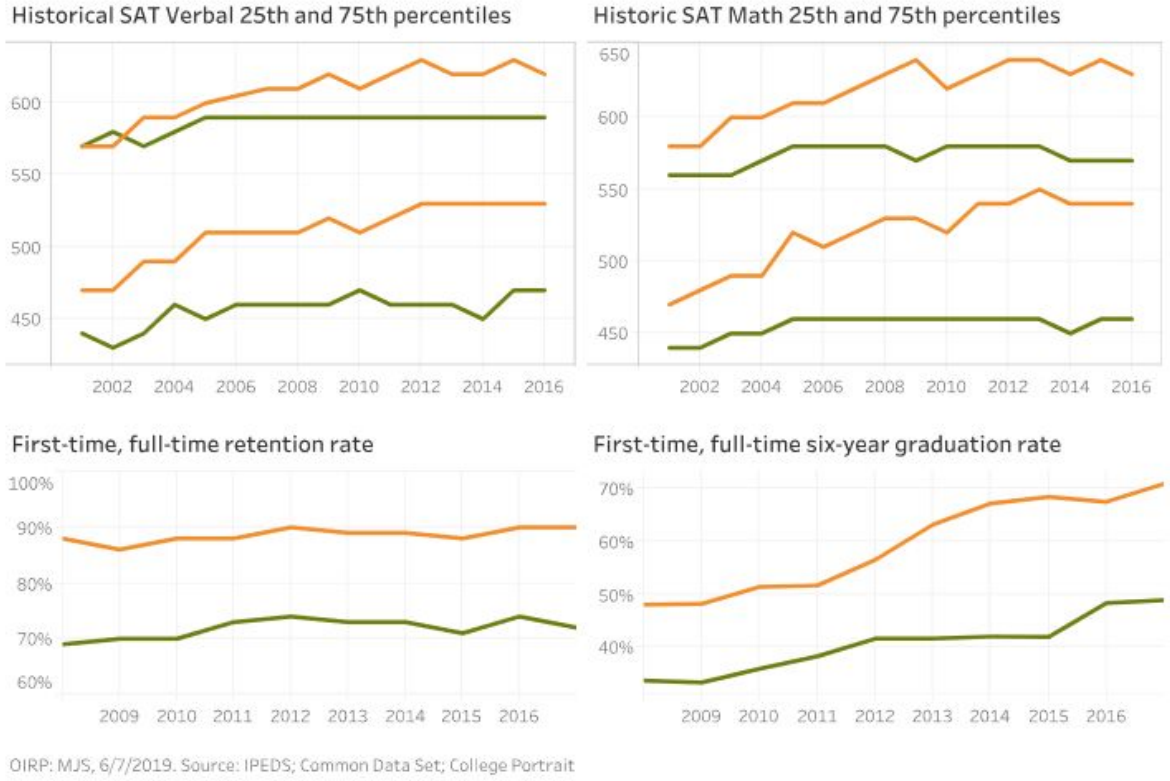
Easy way #1: Longitudinal analysis

```
grad_rate %>%
  filter(unitid %in% the_colleges) %>%
  ggplot(aes(year, rate)) +
  theme_economist() +
  geom_line(aes(colour=factor(getinstnm[as.character(unitid)]) )) +
  ggtitle('Grad rates by year')
```

Easy way #1: Longitudinal analysis



Easy way #1: Longitudinal analysis



Easy way #2: Multivariate regression combining several collections

“Hey, Michael! All else being equal, do quarter schools and semester schools have different graduation rates?”

Again, hard to do the regular way.

- We might want all sorts of different variables (“all else being equal”)
- We probably need to join lots of different collections to get them

We’re also going to have to put it into R (or an equivalent) to do the analysis anyway, so why not start there, too?

Easy way #2: Multivariate regression combining several collections

We'll use these IVs in our analysis:

- Size of graduation cohort
- Selectivity of the institution
- Tuition $
- Control (public/private)
- Locale (city/town/suburb/rural)
- Calendar system (semester/quarter)

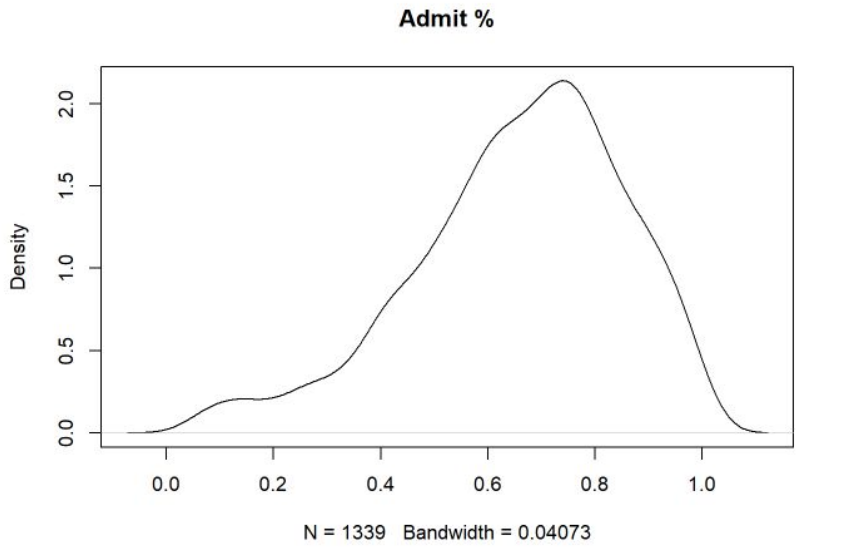
For this, we need (deep breath) the IC header, the IC itself, Admissions, Graduation Rates, and Fall Enrollment.

Easy way #2: Multivariate regression combining several collections

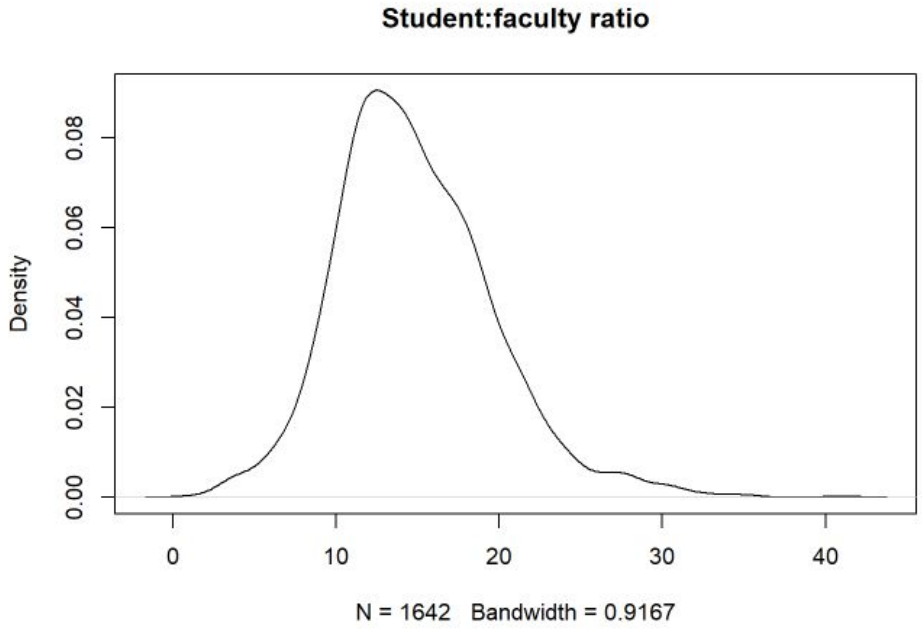
```
directory <- ipeds_survey(table='HD',year=2017, dir=dir)
names(directory) <- tolower(names(directory))
charges <- ipeds_survey('IC_AY', year=2017, dir=dir)
names(charges) <- tolower(names(charges))

charges = charges[,c('unitid',
 'tuition1', 'fee1', 'hrchg1',  #In-district average tuition for full-time undergraduates
 'tuition2', 'fee2', 'hrchg2',  #In-state average tuition for full-time undergraduates
 'tuition3', 'fee3', 'hrchg3',  #Out-of-state average tuition for full-time undergraduates
 'tuition5', 'fee5', 'hrchg5', #In-district average tuition full-time graduates
 'tuition6', 'fee6', 'hrchg6', #In-state average tuition full-time graduates
 'tuition7', 'fee7', 'hrchg7')] #Out-of-state average tuition full-time graduates


dirCharges = merge(charges, directory, by='unitid', all.x=TRUE)

ic <- ipeds_survey(table='IC',year=2017, dir=dir)
names(ic) <- tolower(names(ic))
dirCharges <- merge(dirCharges, ic, by='unitid', all.x=TRUE)
```

Easy way #2: Multivariate regression combining several collections

```
plot(density(d1[which(!is.na(d1$select)),]$select), main="Admit %")
```



Easy way #2: Multivariate regression combining several collections



Easy way #2: Multivariate regression combining several collections

```
table(d1$calsys,d1$control)
```

```
##
##              1   2   3
##   Semester 625 802  52
##   Quarter   50  29  84
```

##recode locale

```
d1$locale2 <- substr(d1$locale,1,1)
d1$locale2 <- as.factor(d1$locale2)
levels(d1$locale2) <- c('City','Town','Suburb','Rural')
table(d1$locale2, d1$locale)
```

```
##
##           11  12  13  21  22  23  31  32  33  41  42  43
##   City   366 209 234   0   0   0   0   0   0   0   0   0
##   Town     0   0   0 315  51  36   0   0   0   0   0   0
##   Suburb   0   0   0   0   0   0  57 163 112   0   0   0
##   Rural    0   0   0   0   0   0   0   0   0  57  25  17
```

```
summary(theLM)
```

```
##
## Call:
## lm(formula = rate ~ calsys + as.integer(tuition1) + control +
##     select + locale2 + adjusted_cohort, data = d1, family = gaussian)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47957 -0.07582  0.00375  0.07103  0.51447
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.820e-01  2.017e-02  23.902  < 2e-16 ***
## calsysQuarter         5.639e-02  1.609e-02   3.506  0.00047 ***
## as.integer(tuition1)  1.035e-05  3.408e-07  30.362  < 2e-16 ***
## control              -9.204e-02  9.529e-03  -9.659  < 2e-16 ***
## select               -1.113e-01  1.779e-02  -6.257 5.27e-10 ***
## locale2Town           3.395e-02  8.293e-03   4.093 4.51e-05 ***
## locale2Suburb         1.171e-02  8.919e-03   1.313  0.18946
## locale2Rural         -2.622e-02  1.541e-02  -1.702  0.08906 .
## adjusted_cohort       5.625e-05  3.346e-06  16.810  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1207 on 1327 degrees of freedom
##   (306 observations deleted due to missingness)
## Multiple R-squared:  0.5617, Adjusted R-squared:  0.5591
## F-statistic: 212.6 on 8 and 1327 DF,  p-value: < 2.2e-16
```

- Email me: mjs26@pdx.edu
- Github: https://github.com/mikesmith2468/pnairp_ipeds_presentation
  - These slides
  - Installation instructions for *ipeds*
  - Code for all the analyses
  - Downloadable R data files for IPEDS back to 2011-12, plus their documentation