

# Instant Comparison: IPEDS Survey Files Working for You

---

Emma Morgan

Tufts University

Office of Institutional Research & Evaluation

June 3, 2016

# Glancing ahead (and in the review)



Special thanks to  
**Kathleen Foley** and **Kate  
Aloisio** in the Office of  
Institutional Research at  
Smith College for sharing  
original code!


# Motivating Question

---

How can we use publicly available data from IPEDS to create longitudinal peer comparison data sets?

# Challenge: Downloading data

Publications & Products | Surveys & Programs | Data & Tools | Fast Facts | School Search | News & Events | About Us



Data Center Help Desk (866) 558-0658

[Start over](#) [Save session](#) [Help](#) [MAIN MENU](#)

## Look up an institution

Provisional Release Data ([Change](#))

### 1. Select Institutions

My Comparison Institution - None 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**

168148: Tufts University, Medford, MA

[Select](#)

[S](#) | [Privacy Policy](#)  
[IStats](#) | [ChildStats](#)



worldofstatistics.org

# Custom data files

Step 1:  
Select your  
institutions

**ies** INSTITUTE OF EDUCATION SCIENCES

**NATIONAL CENTER FOR EDUCATION STATISTICS**

Enter search terms here

Publications & Products | Surveys & Programs | Data & Tools | Fast Facts | School Search | News & Events | About Us

## IPEDS

Data Center Help Desk (866) 558-0658

Start over | Save session | Help | MAIN MENU

### Custom Data Files

Provisional Release Data (Change)

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

My Comparison Institution - Tufts University **i** CHANGE REMOVE

Select Institutions - You have selected 11 institution(s) VIEW/MODIFY

Select Variables - Total 0 variables selected ADD NEW VARIABLES VIEW/MODIFY

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

When you have finished selecting institutions, CONTINUE to Step 2 - Select Variables.

#### My Institutions

MODIFY

ID	Institution Name	City	State
164924	Boston College <b>i</b>	Chestnut Hill	MA
215062	University of Pennsylvania	Philadelphia	PA
217156	Brown University	Providence	RI
182670	Dartmouth College	Hanover	NH
147767	Northwestern University	Evanston	IL
131496	Georgetown University	Washington	DC
179867	Washington University in St Louis	Saint Louis	MO
162928	Johns Hopkins University	Baltimore	MD
190415	Cornell University	Ithaca	NY
198419	Duke University	Durham	NC
190150	Columbia University in the City of New York	New York	NY

# Custom data files

Step 2:  
Select (a) year and  
(b) variables of interest

1. Select Institutions 2. Select Variables 3. Output

**My Comparison Institution** - Tufts University ⓘ CHANGE REMOVE

**Select Institutions** - You have selected 11 institution(s) VIEW/MODIFY

**Select Variables** - Total 0 variables selected ADD NEW VARIABLES VIEW/MODIFY

In order to get a custom data set, select data to include in your data set by first selecting a year, then browsing that year's tree for variables. You can select data from multiple years before clicking continue.

Continuous variable  Alpha/String variable  Discrete variable

Search for variable(s)  Search When you have finished selecting variables from the tree, click Continue Continue

Available Year(s)

2014	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001
2000	1999	1998	1997	1996	1995	1994	1993	1992	1991	1990	1989	1988	1987
1986	1985	1984	1980										

- Frequently used/Derived variables
- Institutional Characteristics
- Admissions and Test Scores
- Fall Enrollment**
  - Frequently used enrollment variables: Fall 2014
  - Gender, attendance status, and level of student: Fall 2014
    - Gender, attendance status, and level of student
      - Select Qualifying Variable(s) (Recommend) ⓘ
      - Level of student

Select from the List of Variables

Select All | Unselect All

- Grand total ⓘ
- Total men ⓘ
- Total women ⓘ
- Full time total ⓘ
- Full time men ⓘ
- Full time women ⓘ
- Part time total ⓘ
- Part time men ⓘ
- Part time women ⓘ

# Custom data files

## Step 3: Download the data file

A	B	C	D	E	F	G	H	I	J
unitid	institution name	year	EF2014.Level of student	EF2014.Grand total	EF2014.Total men	EF2014.Total women	EF2014.Full time total	EF2014.Full time men	EF2014.Full time women
164924	Boston College	2014	All students total	14317	6509	7808	13024	5905	7119
164924	Boston College	2014	All students, Undergraduate total	9856	4576	5280	9526	4406	5120
164924	Boston College	2014	All students, Undergraduate, Degree/certificate-seeking total	9483	4393	5090	9345	4313	5032
164924	Boston College	2014	All students, Undergraduate, Degree/certificate-seeking, First-time	2295	1016	1279	2284	1013	1271
164924	Boston College	2014	All students, Undergraduate, Other degree/certificate-seeking	7188	3377	3811	7061	3300	3761
164924	Boston College	2014	All students, Undergraduate, Other degree/certificate-seeking, Transfer-ins	142	69	73	142	69	73
164924	Boston College	2014	All students, Undergraduate, Other degree/certificate-seeking, Continuing	7046	3308	3738	6919	3231	3688
164924	Boston College	2014	All students, Undergraduate, Non-degree/certificate-seeking	373	183	190	181	93	88
164924	Boston College	2014	All students, Graduate	4461	1933	2528	3498	1499	1999
217156	Brown University	2014	All students total	9181	4499	4682	8756	4357	4399
217156	Brown University	2014	All students, Undergraduate total	6548	3150	3398	6255	3049	3206
217156	Brown University	2014	All students, Undergraduate, Degree/certificate-seeking total	6264	3057	3207	6241	3041	3200
217156	Brown University	2014	All students, Undergraduate, Degree/certificate-seeking, First-time	1559	764	795	1559	764	795
217156	Brown University	2014	All students, Undergraduate, Other degree/certificate-seeking	4705	2293	2412	4682	2277	2405
217156	Brown University	2014	All students, Undergraduate, Other degree/certificate-seeking, Transfer-ins	55	33	22	55	33	22
217156	Brown University	2014	All students, Undergraduate, Other degree/certificate-seeking, Continuing	4650	2260	2390	4627	2244	2383
217156	Brown University	2014	All students, Undergraduate, Non-degree/certificate-seeking	284	93	191	14	8	6
217156	Brown University	2014	All students, Graduate	2633	1349	1284	2501	1308	1193
190150	Columbia University in the City of New York	2014	All students total	27589	13596	13993	23268	11409	11859
190150	Columbia University in the City of New York	2014	All students, Undergraduate total	8100	4247	3853	7496	3926	3570
190150	Columbia University in the City of New York	2014	All students, Undergraduate, Degree/certificate-seeking total	8100	4247	3853	7496	3926	3570
190150	Columbia University in the City of New York	2014	All students, Undergraduate, Degree/certificate-seeking, First-time	1533	785	748	1490	759	731
190150	Columbia University in the City of New York	2014	All students, Undergraduate, Other degree/certificate-seeking	6567	3462	3105	6006	3167	2839
190150	Columbia University in the City of New York	2014	All students, Undergraduate, Other degree/certificate-seeking, Transfer-ins	585	361	224	512	323	189
190150	Columbia University in the City of New York	2014	All students, Undergraduate, Other degree/certificate-seeking, Continuing	5982	3101	2881	5494	2844	2650
190150	Columbia University in the City of New York	2014	All students, Graduate	19489	9349	10140	15772	7483	8289
190415	Cornell University	2014	All students total	21679	11172	10507	21602	11144	10458
190415	Cornell University	2014	All students, Undergraduate total	14282	7024	7258	14269	7021	7248
190415	Cornell University	2014	All students, Undergraduate, Degree/certificate-seeking total	14195	6979	7216	14182	6976	7206
190415	Cornell University	2014	All students, Undergraduate, Degree/certificate-seeking, First-time	3225	1577	1648	3225	1577	1648
190415	Cornell University	2014	All students, Undergraduate, Other degree/certificate-seeking	10970	5402	5568	10957	5399	5558
190415	Cornell University	2014	All students, Undergraduate, Other degree/certificate-seeking, Transfer-ins	554	275	279	553	275	278
190415	Cornell University	2014	All students, Undergraduate, Other degree/certificate-seeking, Continuing	10416	5127	5289	10404	5124	5280
190415	Cornell University	2014	All students, Undergraduate, Non-degree/certificate-seeking	87	45	42	87	45	42
190415	Cornell University	2014	All students, Graduate	7397	4148	3249	7333	4123	3210

# Custom data files

Step 4:

Repeat for every year of interest and  
set of variables



# Custom data files

## Challenges & frustrations

- What if we use different comparison institutions?
  - Example: Engineering has a different set of comparison institutions than is traditionally used; grad/professional schools might be interested in different comparison institutions; we might want the flexibility of looking at different institutions?
- What if we are interested in different variables?
  - Example: I downloaded total enrollment and full-time enrollment by gender, but now I also am interested in enrollment by full-time/part-time status *and* race/ethnicity
- What if we need additional years of data?

# Alternative: Complete data files

## Step 1: Choose year and survey

**ies** INSTITUTE OF EDUCATION SCIENCES

NATIONAL CENTER FOR EDUCATION STATISTICS

Enter search terms here

Publications & Products | Surveys & Programs | Data & Tools | Fast Facts | School Search | News & Events | About Us

**IPEDS**  
Data Center Help Desk (866) 558-0658

Start over | Save session | Help | **MAIN MENU**

Complete Data Files Provisional Release Data ([Change](#))

**Years & Surveys**

2014 | Fall Enrollment | **Continue**

Data files are available in ZIP format.

U.S. Department of Education  
Institute of Education Sciences  
National Center for Education Statistics

[NewsFlash](#) | [Staff](#) | [Contact](#) | [Help](#) | [Blog](#) | [RSS](#) | [Privacy Policy](#)  
[Statistical Standards](#) | [ED Data Inventory](#) | [FedStats](#) | [ChildStats](#)

**Tufts** Office of Institutional Research and Evaluation

# Complete data files

## Step 2: Download **data file** and **dictionary**



Data Center Help Desk (866) 558-0658



[Start over](#)



[Save session](#)

[Help](#)



[MAIN MENU](#)

### Complete Data Files

Provisional Release Data ([Change](#))

#### Years & Surveys

2014



Fall Enrollment



[Continue](#)

Data files are available in ZIP format.

Year	Survey	Title	Data File	Stata Data File	Programs	Dictionary
2014	Fall Enrollment	Race/ethnicity, gender, attendance status, and level of student: Fall 2014	<a href="#">EF2014A</a>	<a href="#">EF2014A STATA</a>	<a href="#">SPSS</a> , <a href="#">SAS</a> , <a href="#">STATA</a>	<a href="#">Dictionary</a>

# Complete data files

Step 4:

Repeat for every year of interest

# Complete data files

## Challenges & frustrations

- One file per year
  - Longitudinal comparisons necessitate compiling data files
- Size of complete survey files
  - Complete survey files are huge and can be unwieldy to work with
  - For any institution, much of this data will not provide useful peer comparisons
    - Example: Enrollment at Tufts
- Column headers
  - Options: constantly reference the dictionary, or take time to rename
- Certain data columns have additional reference tables
  - Example: Student Level (EFALEVEL)

# Complete data files: Example

Enrollment by race/ethnicity, gender,  
attendance status, and level of student:  
Fall 2014

# Compiling Data

---

R to the rescue

Function: `IPEDS_clean_merge()`

# Goals of the program

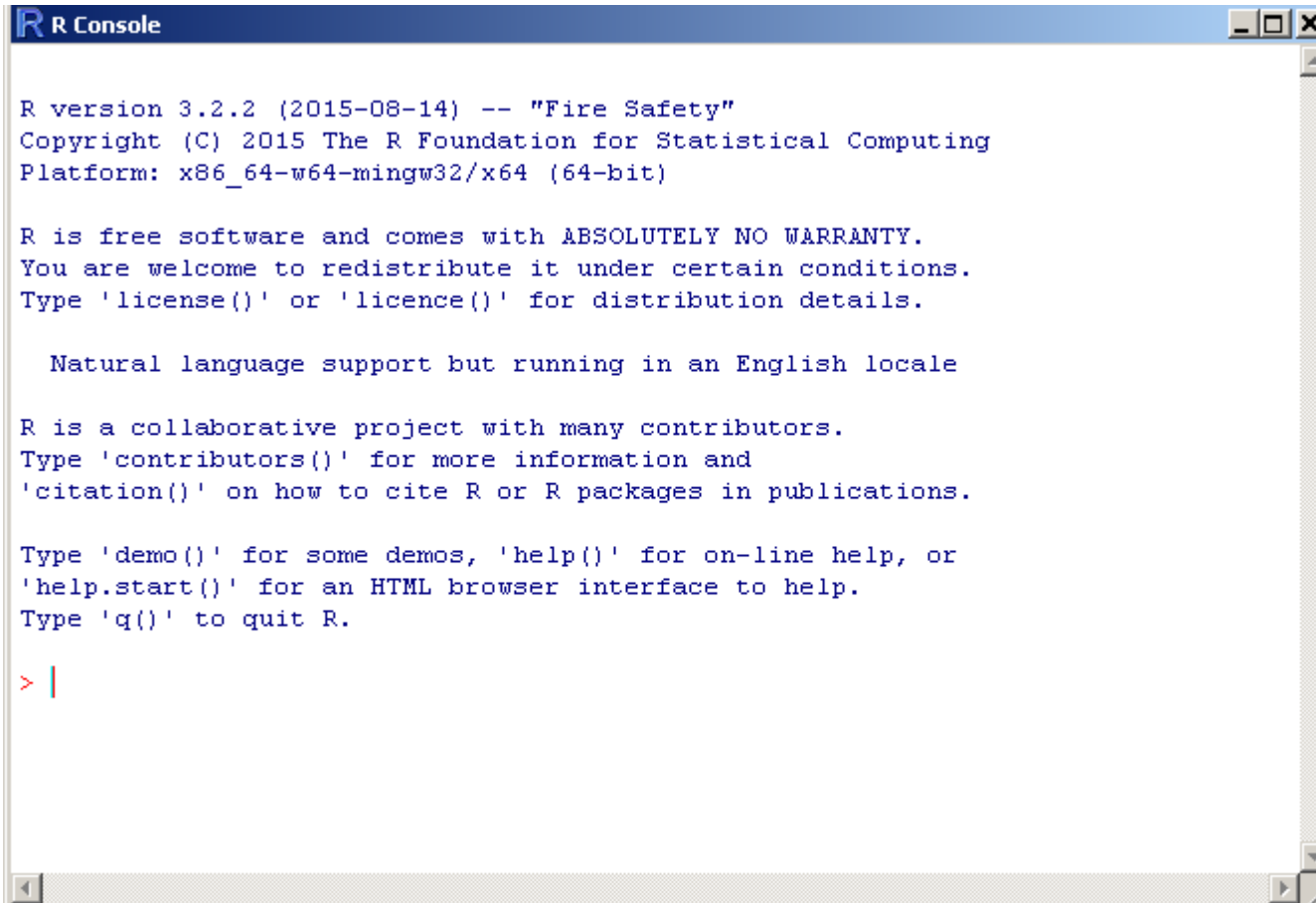
- Create a more manageable data file by keeping only institutions and metrics that are of interest
- Add institution *names* in addition to UNITID for easier reference
- Replace column headers with text description
- Replace any reference columns (i.e. student level for fall enrollment) with reference *description*
  - Fall Enrollment: EFALEVEL 1 → “All Students, Total”
- Merge across years to create longitudinal data set
- Automate so that files can be updated when new data are released



Don't fear the code!



# Getting started with R and R Studio



```
R Console

R version 3.2.2 (2015-08-14) -- "Fire Safety"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Command line only – not very conducive to learning or exploring

# Getting started with R and R Studio

## View and edit code

The screenshot displays the RStudio interface. The main editor window shows R code for loading and processing data. The Environment pane on the right lists loaded objects: 'peerFile' (12 obs. of 2 variables) and 'varDictionary' (65 obs. of 8 variables). The console at the bottom shows the execution of the code, including the output of the `?which` command.

```
110 # EF_A - Fall Enrollment
111 #*****
112
113 inputDirectory <- "Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer
114 outputDirectory <- "Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Pe
115 datedDirectory <- "Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer
116
117
118 setwd(file.path("Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer C
119 varDictionary <- read.csv("EF_A Compiled Dictionary.csv", check.names=FALSE)
120
121
122 setwd(file.path("Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer C
123 peerFile <- read.csv("UndergradPeers_IDandNames.csv", check.names=FALSE)
124
125 freq_table = TRUE
126 lookup_dict <- "efaLevel_DICT_revised.csv"
127
128 IPEDS_clean_merge(inputDirectory, varDictionary, peerFile, outputDirectory,
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

```
> ?read.csv
> ?which
> inputDirectory <- "Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer Com
parative Data/IPEDS Data Center/IPEDS Surveys/EF_A - Fall Enrollment"
> outputDirectory <- "Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer Co
mparative Data/IPEDS Data Center/Compiled Survey Files from R"
> datedDirectory <- "Q:/Staff/President, Provost, Trustees/TAAC Dashboard/Peer Com
parative Data/IPEDS Data Center/R output for Tableau/Dated Files"
>
> ?which
> |
```

See what is loaded and what functions and/or data are ready for use

Load packages, get help, see graphs, etc.

## Command Line

# Necessary input

```
IPEDS_clean_merge <- function (inputDirectory,  
                                varDictionary,  
                                peerFile,  
                                outputDirectory,  
                                outputName,  
                                datedDirectory,  
                                freq_table,  
                                lookup_dict) {
```

# Necessary input

- **inputDirectory**
  - Filepath to folder containing complete survey files (.csv)
- **varDictionary**
  - “varlist” page from data dictionary, saved as .csv and read into R
- **peerFile**
  - List of comparison institution names and UNITID (IPEDS identifier)
- **outputDirectory**
  - Filepath to destination folder of compiled dataset
- **outputName**
  - Name of the compiled file
- **datedDirectory**
  - For record keeping only; location of files from every day the code is run
- **freq\_table (True/False)**
  - Does the dictionary have a “Frequencies” tab with additional lookup values?
- **lookup\_dict (optional)**
  - If an additional lookup table exists, data dictionary “Frequencies” tab saved as .csv and read into R

# Input: varDictionary

## Sample Format

varnumber	varname	DataType	Fieldwidth	format	imputationvar	varTitle
1	UNITID	N	6	Cont		Unique identification number of the institution
20166	EFALEVEL	N	2	Disc		Level of student
21985	LINE	N	2	Disc		Level of student (original line number on survey form)
21990	SECTION	N	1	Disc		Attendance status of student
21991	LSTUDY	N	1	Disc		Level of student
20286	EFTOTLT	N	6	Cont	XEFTOTLT	Grand total
20241	EFTOTLM	N	6	Cont	XEFTOTLM	Grand total men
20246	EFTOTLW	N	6	Cont	XEFTOTLW	Grand total women
24432	EFAIANT	N	6	Cont	XEFAIANT	American Indian or Alaska Native total
24434	EFAIANM	N	6	Cont	XEFAIANM	American Indian or Alaska Native men
24436	EFAIANW	N	6	Cont	XEFAIANW	American Indian or Alaska Native women

Source: File documentation for enrollment by race/ethnicity, gender, attendance status, and level of student: Fall 2014, "varlist"

# Necessary input

- **inputDirectory**
  - Filepath to folder containing complete survey files (.csv)
- **varDictionary**
  - “varlist” page from data dictionary, saved as .csv and read into R
- **peerFile**
  - List of comparison institution names and UNITID (IPEDS identifier)
- **outputDirectory**
  - Filepath to destination folder of compiled dataset
- **outputName**
  - Name of the compiled file
- **datedDirectory**
  - For record keeping only; location of files from every day the code is run
- **freq\_table (True/False)**
  - Does the dictionary have a “Frequencies” tab with additional lookup values?
- **lookup\_dict (optional)**
  - If an additional lookup table exists, data dictionary “Frequencies” tab saved as .csv and read into R

# Input: freq\_table

## Sample Format

varnumber	varname	codevalue	Valuelabel
20166	EFALEVEL	1	All students total
20166	EFALEVEL	2	All students, Undergraduate total
20166	EFALEVEL	3	All students, Undergraduate, Degree/certificate-seeking total
20166	EFALEVEL	4	All students, Undergraduate, Degree/certificate-seeking, First-time
20166	EFALEVEL	5	All students, Undergraduate, Other degree/certificate-seeking
20166	EFALEVEL	19	All students, Undergraduate, Other degree/certificate-seeking, Transfer-ins
20166	EFALEVEL	20	All students, Undergraduate, Other degree/certificate-seeking, Continuing
20166	EFALEVEL	11	All students, Undergraduate, Non-degree/certificate-seeking
20166	EFALEVEL	12	All students, Graduate
20166	EFALEVEL	21	Full-time students total
20166	EFALEVEL	22	Full-time students, Undergraduate total

...continuing for all values of EFALEVEL, LINE, SECTION, LSTUDY

Source: File documentation for enrollment by race/ethnicity, gender, attendance status, and level of student: Fall 2014, "frequencies"



# Function: IPEDS\_clean\_merge

1. Specify where to find data files (inputDirectory)
2. Set up a dummy that will hold the “cleaned” data files from each year until it’s ready to compile

```
13 IPEDS_clean_merge <- function (inputDirectory, varDictionary, peerFile,  
14     outputDirectory, outputName, datedDirectory, freq_table, lookup_dict) {  
15  
16     # Specify where R will find your file set  
17     setwd(inputDirectory)  
18  
19     # Set up necessary blank lists  
20     # List to hold survey data from various years  
21     ds_list <- list()  
22  
23     # For each file in directory:  
24     # (1) read in the file  
25     # Create list of rows to keep based on peers/comparison schools  
26     # Create list of columns to drop (imputation columns or blank fields for Tufts)  
27     # store "good" data in file_list
```

# Function: IPEDS\_clean\_merge

## 3. For each data file...

- a) Read in the data file
- b) Identify which columns to keep
  - Depending on your purposes, you might change this
  - At Tufts, we delete
    - (a) columns for imputation variables
    - (b) columns that Tufts has left blank
- c) Identify which rows to keep

Keep rows for Tufts and comparison institutions

```
# Read in all years of survey files; capitalize to standardize column titles
for (i in 1:length(list.files())) {
  fileName <- list.files()[i]
  ds <- read.csv(fileName, check.names=FALSE)

  #Some years of data have lowercase headers, so change all to uppercase for consistency
  names(ds) <- toupper(names(ds))
  dropCol <- c()

  # below, we will select rows for comparison institutions
  # this will use UNITID %in% list from peerFile$unitid

  # identify columns to delete (imputation & blank for Tufts)
  tuftsRow <- which(ds$UNITID == "168148")

  #we will only want to delete based on "blanks" if we have one row for Tufts;
  #if there are multiple Tufts rows, then only remove imputation variables

  # when there is only one row per institution, remove imputation rows
  # and any column for which Tufts (or your main institution) has a null value

  if (length(tuftsRow) == 1) {
    myRow <- tuftsRow
    for (k in 1:ncol(ds)) {
      #drop imputation columns
      if (substring(names(ds)[k],1,1) == "X") {
        dropCol <- c(dropCol,k)
      }

      #drop columns for which primary institution is null/blank
      else if (is.na(ds[myRow,k])) {
        dropCol <- c(dropCol,k)
      }
      else if (ds[myRow,k] == ".") {
        dropCol <- c(dropCol,k)
      }
    }
  }

  #If we have multiple rows per file, only remove imputation columns and leave all others

  else if (length(tuftsRow > 1)) {
    for (k in 1:ncol(ds)) {
      if (substring(names(ds)[k],1,1) == "X") {
        dropCol <- c(dropCol,k)
      }
    }
  }

  # create a new ds with only the "good" rows and columns
  ds_good <- ds[ds$UNITID %in% peerFile$unitid, -dropCol]
```

# Function: IPEDS\_clean\_merge

## 3. For each data file (cont)...

### d) Add survey year

Columns for fall and fiscal year, generated using the function IPEDS\_FY that generates year from the survey file name

### e) Add institution name

Use peerFile to lookup institution names based on UNITID

### f) Store the cleaned data as an element of the list

```
# We now want to add columns for Fiscal Year, Fall, and Institution Name
# Create placeholders first
ds_good$FY <- ""
ds_good$Fall <- ""
ds_good$Institution_Name <- ""

# Add FY and "Fall" for joining purposes in Tableau
# FY is based on the type of survey file and when data is collected
# IPEDS_FY function is defined below

survey_details <- IPEDS_FY(fileName)

ds_good$FY <- survey_details[[1]]
ds_good$Fall <- ds_good$FY-1

ds_good$`Survey Code` <- survey_details[[2]]
ds_good$`Survey Title` <- survey_details[[3]]
survey <- survey_details[[2]]

# Add Institution Name
for (l in 1:nrow(ds_good)) {
  p <- which(ds_good$UNITID[l] == peerFile$unitid)
  ds_good$Institution_Name[l] <- as.matrix(peerFile$`institution name`)[p]
}

# Store the good info in our new list
ds_list[[i]] <- assign(paste("ds",i,sep=""), ds_good)
}
```

# Function: IPEDS\_clean\_merge

4. Compile all years of data into a single dataset

Each variable column now includes entries for each institution in each year

```
full_ds <- as.data.frame(rbind.fill.matrix(ds_list))
```

# Function: IPEDS\_clean\_merge

4. Compile all years of data into a single dataset

Each variable column now includes entries for each institution in each year

```
full_ds <- as.data.frame(rbind.fill.matrix(ds_list))
```

# Function: IPEDS\_clean\_merge

5. If the survey dictionary included a “frequency” table with additional lookup values...
  - a. Generate reference tables for these variables using the frequency table
  - b. Create additional columns as necessary with descriptions of coded values

```
if (freq_table==TRUE) {  
  full_ds <- lookup(full_ds, lookup_dict)  
}
```

# Function: IPEDS\_clean\_merge

6. Replace column name variable codes with descriptive titles  
Use the variable dictionary to substitute full text descriptions for each variable
7. Set the output directory
8. Export the compiled data set from R as a .csv

```
# Replace variable codes with full header titles

for (m in 1:ncol(full_ds)) {
  if (names(full_ds)[m] %in% varDictionary$varname) {
    a <- which(names(full_ds)[m] == as.matrix(varDictionary$varname))
    names(full_ds)[m] <- as.matrix(varDictionary$varTitle)[a]
  }
}

# Set directory to where you want the file to output
setwd(outputDirectory)

# Write full .csv to output directory
write.csv(full_ds, paste(outputName, ".csv", sep=""), row.names = FALSE, na="")

# Set dated directory to save copies of files
setwd(datedDirectory)
write.csv(full_ds, file = paste(outputName, sys.Date(), ".csv", sep=""), row.names = FALSE, na="")
}
```

Your longitudinal peer comparison data file is ready to use!





# Continuing Challenges

- Format of IPEDS data files may change
  - Example: Undergraduate admissions
    - Pre-2014, undergraduate admissions statistics included in Institutional Characteristics data file
    - In 2014, undergraduate admissions released as independent data file
    - The variable names are consistent, so we can still merge on these fields
- Automation requires checking for consistency
- Variables may change
  - Example: Race/ethnicity categorization before and after 2008
- Older documentation is available as a web link, but is not downloadable as an Excel file
  - For historical data, it may be necessary to create a compiled dictionary if older variable names differ from what appears in the current documentation file

# Possible Next Steps

- Use IPEDS directory file to generate peer lists
  - Carnegie Classification, size, region, etc.
- Extend to other data sources
  - College Scorecard
- Use the Shiny package in R to make the program more interactive and user-friendly
  - User can select data file directory, choose the correct dictionary and specify where to export files from a dialogue box rather than specifying all file paths in text

# Questions?

Emma Morgan

Research Analyst, Tufts University

[emma.morgan@tufts.edu](mailto:emma.morgan@tufts.edu)

Special thanks to **Kathleen Foley** (Associate Director for Analytics) and **Kate Aloisio** (Research Analyst) from the Office of Institutional Research at Smith College for their ideas and code that jumpstarted this project.