Using R to Automate IPEDS Reporting

Jason M. Bryer

Excelsior College
jason@bryer.org
www.bryer.org

May 23, 2011
AIR Annual Forum

Agenda

1. Getting Started with the IPEDS Package
2. Using the IPEDS Package for Data Reporting
3. Using the IPEDS Package for Data Analysis
4. Conclusions

Installing the ipeds Package

The ipeds package is hosted by R-Forge. It can be downloaded from http://ipeds.r-forge.r-project.org or directly within R with the following command:

```
> install.packages("ipeds", repos = c("http://R-Forge.R-project.org"
       "http://lib.stat.cmu.edu/R/CRAN"), dep = TRUE)
```

Installing the ipeds Package

The ipeds package is hosted by R-Forge. It can be downloaded from http://ipeds.r-forge.r-project.org or directly within R with the following command:

```
> install.packages("ipeds", repos = c("http://R-Forge.R-project.org"
       "http://lib.stat.cmu.edu/R/CRAN"), dep = TRUE)
```

And to load the package:

```
> library(ipeds)
```

Functions

```
> ls("package:ipeds")
 [1] "completions"                  "downloadAllSurveys"
 [3] "downloadHelp"                 "downloadIPEDSSurvey"
 [5] "formatYear"                   "getIPEDSSurvey"
 [7] "ipedsHelp"                    "mapIPEDSFields"
 [9] "recodeAwardLevel"             "recodeControl"
[11] "recodeDirectory"              "recodeGeographicRegion"
[13] "recodeGraduateOffering"       "recodeHighestDegreeOffered"
[15] "recodeHighestLevelOfOffering" "recodeImputation"
[17] "recodeInstitutionSize"        "recodeLevelOfInstitution"
[19] "recodeOpenPublic"             "recodeSector"
[21] "recodeTitleIVEligibility"     "recodeTitleIVIndicator"
[23] "recodeUndergraduateOffering"  "twelveMonthEnrollment"
```

1. Getting Started with the IPEDS Package

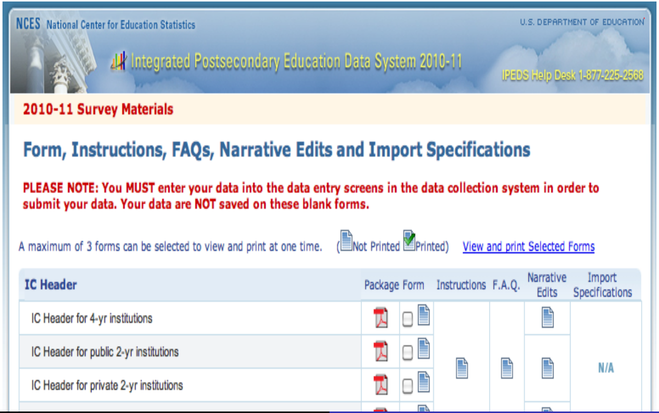
2. Using the IPEDS Package for Data Reporting

3. Using the IPEDS Package for Data Analysis

4. Conclusions

Generating Survey Files for IPEDS Reporting

- IPEDS provides the ability to upload data files for many of the surveys. Details of the file formats are available here: `https://surveys.nces.ed.gov/ipeds/VisResults.aspx`



NCES National Center for Education Statistics

U.S. DEPARTMENT OF EDUCATION

Integrated Postsecondary Education Data System 2010-11

IPEDS Help Desk 1-877-225-2568

2010-11 Survey Materials

Form, Instructions, FAQs, Narrative Edits and Import Specifications

PLEASE NOTE: You MUST enter your data into the data entry screens in the data collection system in order to submit your data. Your data are NOT saved on these blank forms.

A maximum of 3 forms can be selected to view and print at one time. (🗎Not Printed ✓Printed) View and print Selected Forms

IC Header	Package	Form	Instructions	F.A.Q.	Narrative Edits	Import Specifications
IC Header for 4-yr institutions	🗎	□ 🗎	🗎	🗎	🗎	N/A
IC Header for public 2-yr institutions	🗎	□ 🗎			🗎	
IC Header for private 2-yr institutions	🗎	□ 🗎				

Example: Completions

```
> grads = read.csv("grads.fy2010.csv")
> names(grads)

 [1] "DEGREE_CODE"          "CONCENTRATION_CODE"
 [3] "SPECIALIZATION_CODE" "AGE_AT_GRADUATION"
 [5] "ETHNICITY_SUMMARY"   "ENROLL_DATE"
 [7] "COUNTRY_CODE"        "STATE_CODE"
 [9] "ZIP_CODE"            "GENDER"
[11] "DIVISION_CODE"       "SVC_PROGRAM_CODE"
[13] "DEGREE_TYPE"         "X2010.CIP.code"
[15] "EC.Program.Name"     "AwardLevel"
[17] "residence"           "ProgramType"

> nrow(grads)

[1] 6170
```

Example: Completions

```
> grads = read.csv("grads.fy2010.csv")
> names(grads)

 [1] "DEGREE_CODE"         "CONCENTRATION_CODE"
 [3] "SPECIALIZATION_CODE" "AGE_AT_GRADUATION"
 [5] "ETHNICITY_SUMMARY"   "ENROLL_DATE"
 [7] "COUNTRY_CODE"        "STATE_CODE"
 [9] "ZIP_CODE"            "GENDER"
[11] "DIVISION_CODE"       "SVC_PROGRAM_CODE"
[13] "DEGREE_TYPE"         "X2010.CIP.code"
[15] "EC.Program.Name"     "AwardLevel"
[17] "residence"           "ProgramType"

> nrow(grads)

[1] 6170

> mapping = mapIPEDSFields(programCIPCode = "X2010.CIP.code",
       educationalProgramType = "ProgramType",
       educationalProgramName = "EC.Program.Name",
       educationalAwardLevel = "AwardLevel",
       ethnicityColumn = "ETHNICITY_SUMMARY",
       genderColumn = "GENDER")
```

Example: Completions

Once the data is prepared, call the `completions` function to generate the XML file.

```
> xml = completions(grads, mapping, institutionId = 196680,
      institutionName = "Excelsior College")
```

Example: Completions

Once the data is prepared, call the completions function to generate the XML file.

```
> xml = completions(grads, mapping, institutionId = 196680,
      institutionName = "Excelsior College")
```

The XML object can be further manipulated if necessary, or simply save to a file.

```
> cat(saveXML(xml), file = "completions.xml")
```

That's it! Just upload the file to IPEDS and verify the numbers through the web interface.

CIP Codes

Classification of Instructional Programs (CIP) codes are necessary for many of the surveys. There are two additional data frames available for your convenience, namely cipcodes and crosswalk.

```
> data(cipcodes)
> names(cipcodes)
```

```
[1] "CIPFamily"        "CIPCode"
[3] "Action"           "TextChange"
[5] "CIPTitle"         "CIPDefinition"
[7] "CrossReferences" "Examples"
```

```
> data(crosswalk)
> names(crosswalk)
```

```
[1] "Original.code"  "Original.title"
[3] "Action"         "Text.change"
[5] "Current.code"   "Current.title"
```

1 Getting Started with the IPEDS Package

2 Using the IPEDS Package for Data Reporting

3 Using the IPEDS Package for Data Analysis

- Getting Started
- Comparison of Peer Institutions
- Completions
- Relationship Between SAT Scores and Retention
- Download Surveys not in `surveys`

4 Conclusions

Getting Data

- The IPEDS package provides an interface to the IPEDS Data Center http://nces.ed.gov/ipeds/datacenter/DataFiles.aspx
- The list of surveys currently available is in the surveys data frame.

```
> data(surveys)
```

- There are currently 36 surveys available.
- Two pieces of information are required to download a survey: the SurveyID from the surveys data frame and the year.

Available Surveys

```
   SurveyID                                              Title
1        HD                              Directory information
2        IC Educational offerings, organization, admissions, s
3     IC_AY         Student charges for academic year programs
4     IC_PY   Student charges by program (vocational programs)
5     FLAGS          Response status for all survey components
6     EFEST                               Estimated enrollment
7       EFA Race/ethnicity, gender, attendance status, and lev
8     EFANR Race/ethnicity, gender, attendance status, and lev
9       EFB Age category, gender, attendance status, and level
10      EFC Residence and migration of first-time freshman: Fa
11      EFD     Total entering class and retention rates: Fall
12     EFFY                    12-month unduplicated headcount
```

Available Surveys (cont.)

```
      SurveyID                                              Title
13        EFD1     12-month unduplicated enrollment prior to 2002
14        EFIA                    12-month instructional activity
15        EFD2      12-month instructional activity prior to 2002
16       FLAGS          Response status for all survey components
17         C_A Awards/degrees conferred by program (6-digit CIP c
18       FLAGS          Response status for all survey components
19       SAL_A Salaries of full-time instructional faculty, by co
20       SAL_B Fringe benefits of full-time instructional faculty
21 SAL_FACULTY Tenure status of full-time instructional faculty i
22   SAL_A_LT9 Number of full-time instructional faculty with les
23       FLAGS          Response status for all survey components
24       S_ABD Employees by primary occupation, salary categories
```

Available Surveys (cont.)

```
   SurveyID                                              Title
25      S_F Full-time instruction/research/public service staf
26      S_G New hires by primary occupation, race/ethnicity, a
27     S_CN Employees by primary occupation, race/ethnicity, a
28    FLAGS          Response status for all survey components
29      EAP Employees by faculty status, primary function/occu
30    FLAGS          Response status for all survey components
31    F_F1A       Public institutions - GASB 34/35 Fiscal year
32     F_F2 Private not-for-profit institutions or Public inst
33     F_F3        Private for-profit institutions Fiscal year
34       GR Graduation rate data, 150% of normal time to compl
35    GR_L2 Graduation rate data, 150% of normal time to compl
36    GR200 Graduation rate data, 200% of normal time to compl
```

Getting Help

The ipedsHelp will download the data dictionary for the given survey and year.

```
> ipedsHelp("HD", 2009)
```

Getting Help

The ipedsHelp will download the data dictionary for the given survey and year.

```
> ipedsHelp("HD", 2009)
```

Note that beginning in 2010 IPEDS changed the format of the data dictionaries from HTML to Excel. As such, you will need to have a program that can read Excel files (e.g. MS Excel, Open Office) to view the data dictionary.

Peer Comparisons

The IPEDS institution IDs were predetermined but could also be retrieved by name if necessary. The result of bellow is a new data frame called peers that contains a row for each peer institution.

```
> directory = getIPEDSSurvey("HD", 2009)
> names(directory) = tolower(names(directory))
> directory = recodeDirectory(directory)
> unitid = directory[which(directory$instnm ==
      "Excelsior College"), "unitid"]
> peerIds = c(105668, 127918, 128780, 144777,
      163204, 183257, 187046, 202806, 206279,
      223816, 260901, 413413, 433387, 444158,
      445027, 449339)
> peers = directory[which(directory$unitid %in%
      peerIds), ]
```

Peer Comparisons

We'll use the xtable package to create a LaTeXtable with basic information about the peer institutions.

```
> library(xtable)
> p = peers[, c("instnm", "webaddr", "stabbr",
      "control")]
> names(p) = c("Institution", "Web Address",
      "State", "Sector")
> x = xtable(p, caption = NULL)
```

Peer Comparisons

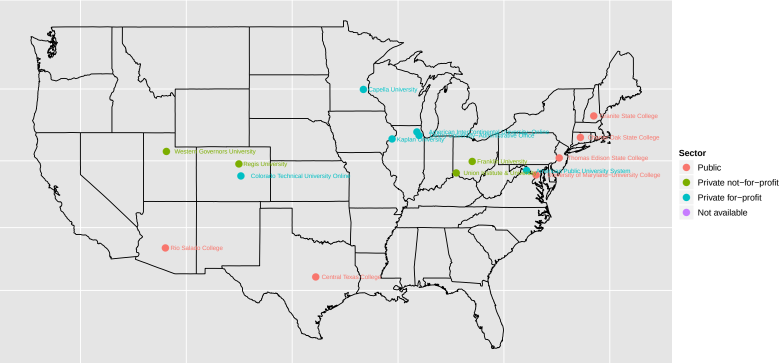
```
> print(x, include.rownames = FALSE, size = "small")
```

Institution	Web Address	State	Sector
Rio Salado College	www.rio.maricopa.edu	AZ	Public
Regis University	www.regis.edu/	CO	Private not-for-profit
Charter Oak State College	www.cosc.edu	CT	Public
DeVry University-Administrative Office	www.devry.edu	IL	Private for-profit
University of Maryland-University College	www.umuc.edu	MD	Public
Granite State College	www.granite.edu	NH	Public
Thomas Edison State College	www.tesc.edu	NJ	Public
Franklin University	www.franklin.edu	OH	Private not-for-profit
Union Institute & University	www.myunion.edu	OH	Private not-for-profit
Central Texas College	www.ctcd.edu	TX	Public
Kaplan University	www.kucampus.edu	IA	Private for-profit
Capella University	www.capella.edu	MN	Private for-profit
Western Governors University	www.wgu.edu	UT	Private not-for-profit
Colorado Technical University Online	www.ctuonline.edu	CO	Private for-profit
American InterContinental University-Online	www.aiuonline.edu	IL	Private for-profit
American Public University System	www.apus.edu	WV	Private for-profit

Mapping Peer Institutions

```
> library(maps)
> usa = data.frame(map("state", plot = FALSE)[c("x",
      "y")])
> p = ggplot(usa, aes(x = x, y = y)) + geom_path() +
      coord_map() + geom_point(data = peers,
      aes(x = longitud, y = latitude, colour = control),
      size = 4) + geom_text(data = peers, aes(x = longitud,
      y = latitude, label = instnm, colour = control),
      size = 2.5, hjust = -0.1) + labs(colour = "Sector") +
      xlab(NULL) + ylab(NULL) + opts(axis.text.x = theme_blank(),
      axis.text.y = theme_blank(), axis.ticks = theme_blank())
```

Mapping Peer Institutions


Capella University
American InterContinental University−Online
DeVry University−Administrative Office
Kaplan University
Western Governors University
Regis University
Colorado Technical University Online
Franklin University
Union Institute & University
American Public University System
University of Maryland−University College
Granite State College
Charter Oak State College
Thomas Edison State College
Rio Salado College
Central Texas College
Sector
● Public
● Private not−for−profit
● Private for−profit
● Not available

Comparison of Enrollments

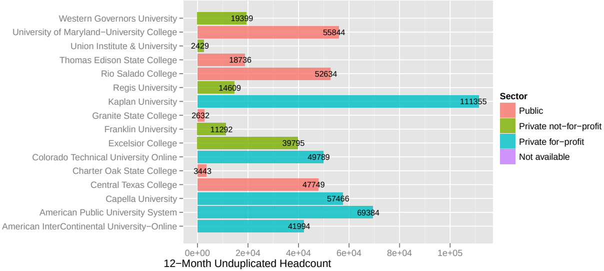
```
> enrollment = getIPEDSSurvey("EFFY", 2010)
> names(enrollment) = tolower(names(enrollment))
> enrollment = enrollment[, c("unitid", "effylev",
      "efytotlt", "efytotlm", "efytotlw", "efynralm",
      "efynralw", "efynralt", "efyunknm", "efyunknw",
      "efyunknt", "dveyhsm", "dveyhsw", "dveyhst",
      "dveyaim", "dveyaiw", "dveyait", "dveybkm",
      "dveybkw", "dveybkt", "dveyapm", "dveyapw",
      "dveyapt", "dveywhm", "dveywhw", "dveywht",
      "efy2morm", "efy2morw", "efy2mort")]
> names(enrollment) = c("unitid", "Level", "Total",
      "Total men", "Total women", "Nonresident Alien Men",
      "Nonresident Alien Women", "Nonresident Alien",
      "Race/ethnicity unknown men", "Race/ethnicity unknown women",
      "Race/ethnicity unknown", "Hispanic men",
      "Hispanic women", "Hispanic", "American Indian or\nAlaska Native men",
      "American Indian or\nAlaska Native women",
      "American Indian or\nAlaska Native", "Black or African American men",
      "Black or African American women", "Black or\nAfrican American",
      "Asian/Native Hawaiian/\nPacific Islander men",
      "Asian/Native Hawaiian/\nPacific Islander women",
      "Asian/Native Hawaiian/\nPacific Islander",
      "White men", "White women", "White", "Two or more races men",
      "Two or more races women", "Two or more races")
```

Comparison of Enrollments

```
> enrollment.peers = enrollment[which(enrollment$unitid %in%
      peerIds), ]
> enrollment.peers = rbind(enrollment[which(enrollment$unitid ==
      unitid), ], enrollment.peers)
> enrollment.peers = merge(enrollment.peers,
      directory[, c("unitid", "instnm", "control")],
      by = "unitid", all.x = TRUE, sort = FALSE)
> enrollment.peers = enrollment.peers[which(enrollment.peers$Level ==
      1), ]
> enrollment.peers = enrollment.peers[, c(ncol(enrollment.peers),
      1:(ncol(enrollment.peers) - 1))]
```

Comparison of Enrollments

```
> print(ggplot(enrollment.peers, aes(y = Total,
        x = factor(instnm), label = Total, fill = control)) +
        geom_bar(stat = "identity", alpha = 0.8) +
        coord_flip() + xlab("") + ylab("12-Month Unduplicated Headcount")
        geom_text(hjust = 0.7, vjust = 0.5, angle = 0,
            size = 3) + labs(fill = "Sector"))
```



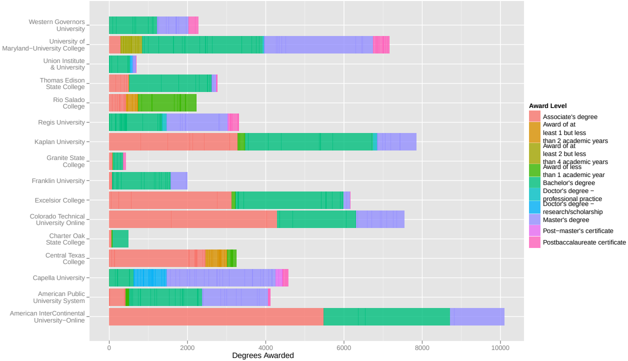
Comparison of Completions

```
> graduates = getIPEDSSurvey("C_A", 2010)
> graduates = graduates[, c("UNITID", "CIPCODE",
       "AWLEVEL", "CTOTALM", "CTOTALW", "CTOTALT",
       "CNRALM", "CNRALW", "CNRALT", "CUNKNM",
       "CUNKNW", "CUNKNT", "DVCAIM", "DVCAIW",
       "DVCAIT", "DVCAPM", "DVCAPW", "DVCAPT",
       "DVCBKM", "DVCBKW", "DVCBKT", "DVCHSM",
       "DVCHSW", "DVCHST", "DVCWHM", "DVCWHW",
       "DVCWHT")]
> graduates$AWLEVEL = recodeAwardLevel(graduates$AWLEVEL)
> names(graduates) = c("unitid", "cipcode",
       "awardlevel", "Men", "Women", "Total",
       "Nonresident alien men", "Nonresident alien women",
       "Nonresident alien", "Race/ethnicity unknown men",
       "Race/ethnicity unknown women", "Race/ethnicity unknown",
       "American Indian or Alaska Native men",
       "American Indian or Alaska Native women",
       "American Indian or Alaska Native", "Asian/Native Hawaiian/Other Pacific Islander men",
       "Asian/Native Hawaiian/Other Pacific Islander women",
       "Asian/Native Hawaiian/Other Pacific Islander",
       "Black or African American men", "Black or African American women",
       "Black or African American", "Hispanic or Latino men",
       "Hispanic or Latino women", "Hispanic or Latino",
       "White men", "White women", "White")
```

Comparison of Completions

```
> graduates.peers = graduates[which(graduates$unitid %in%
      peerIds), ]
> graduates.peers = rbind(graduates[which(graduates$unitid ==
      unitid), ], graduates.peers)
> graduates.peers = merge(graduates.peers, directory[,
      c("unitid", "instnm")], by = "unitid",
      all.x = TRUE, sort = FALSE)
> graduates.peers$cipcode = format(graduates.peers$cipcode,
      width = 7, nsmall = 4)
> graduates.peers$cip2 = unlist(strsplit(as.character(graduates.peers[,
      "cipcode"]), "\\."))[seq(1, 2 * nrow(graduates.peers),
      by = 2)]
> graduates.peers$instnm = breakColumn(graduates.peers$instnm)
> graduates.peers$awardlevel = breakColumn(graduates.peers$awardlevel,
      mod = 4)
> aggr = aggregate(graduates.peers[, c("Total")],
      by = list(graduates.peers$instnm, graduates.peers$cip2),
      FUN = "sum")
> graduates.peers2 = graduates.peers[-which(graduates.peers$cip2 ==
      "99"), ]
```

Comparison of Completions



Relationship Between SAT Scores and Retention

The first step is to download the necessary survey data.

```
> directory = getIPEDSSurvey("HD", 2009)
> admissions = getIPEDSSurvey("IC", 2009)
> retention = getIPEDSSurvey("EFD", 2009)
```

Relationship Between SAT Scores and Retention

The first step is to download the necessary survey data.

```
> directory = getIPEDSSurvey("HD", 2009)
> admissions = getIPEDSSurvey("IC", 2009)
> retention = getIPEDSSurvey("EFD", 2009)
```

The following commands will subset the data so that we only have the variables we are interested in.

```
> directory = directory[, c("unitid", "instnm",
       "sector", "control")]
> retention = retention[, c("unitid", "ret_pcf",
       "ret_pcp")]
> names(retention) = c("unitid", "FullTimeRetentionRate",
       "PartTimeRetentionRate")
```

Relationship Between SAT Scores and Retention

The first step is to download the necessary survey data.

```
> directory = getIPEDSSurvey("HD", 2009)
> admissions = getIPEDSSurvey("IC", 2009)
> retention = getIPEDSSurvey("EFD", 2009)
```

The following commands will subset the data so that we only have the variables we are interested in.

```
> directory = directory[, c("unitid", "instnm",
      "sector", "control")]
> retention = retention[, c("unitid", "ret_pcf",
      "ret_pcp")]
> names(retention) = c("unitid", "FullTimeRetentionRate",
      "PartTimeRetentionRate")
```

Now merge the separate data frames.

```
> ret = merge(directory, admissions, by = "unitid")
> ret = merge(ret, retention, by = "unitid")
```

Relationship Between SAT Scores and Retention

```
> ret2 = ret
> ret2 = ret2[-which(ret2$FullTimeRetentionRate <
      20), ]
> ret2$SATMath = (ret2$SATMath75 + ret2$SATMath25)/2
> ret2$SATWriting = (ret2$SATWriting75 + ret2$SATWriting25)/2
> ret2$SATTotal = ret2$SATMath + ret2$SATWriting
> ret2$AcceptanceTotal = ret2$AdmissionsTotal/ret2$ApplicantsTotal
> ret2$UseAdmissionTestScores = as.factor(as.character(ret2$UseAdmissio
```

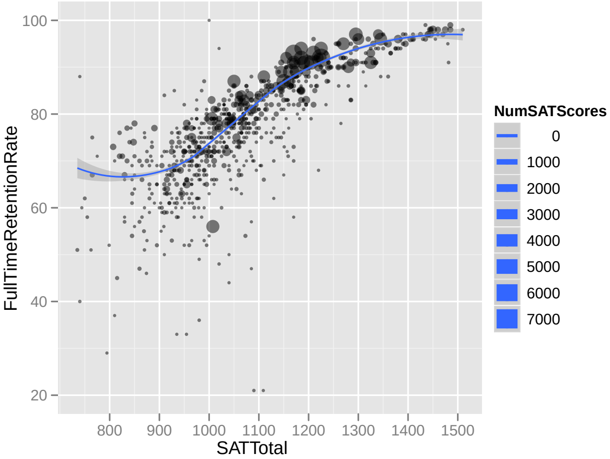
Relationship Between SAT Scores and Retention

```
> ret2 = ret
> ret2 = ret2[-which(ret2$FullTimeRetentionRate <
      20), ]
> ret2$SATMath = (ret2$SATMath75 + ret2$SATMath25)/2
> ret2$SATWriting = (ret2$SATWriting75 + ret2$SATWriting25)/2
> ret2$SATTotal = ret2$SATMath + ret2$SATWriting
> ret2$AcceptanceTotal = ret2$AdmissionsTotal/ret2$ApplicantsTotal
> ret2$UseAdmissionTestScores = as.factor(as.character(ret2$UseAdmissio
```

Finally, we can plot the relationship with ggplot2.

```
> p = ggplot(ret2, aes(x = SATTotal, y = FullTimeRetentionRate,
      size = NumSATScores)) + geom_point(alpha = 0.5) +
      geom_smooth(aes(weight = NumSATScores))
```

Relationship Between SAT Scores and Retention



Download Data not in surveys

IPEDS occasionally changes the file naming convention on the Data Center (`http://nces.ed.gov/ipeds/datacenter/DataFiles.aspx`). Though all of the current years are available, there may (*likely*) be surveys from older years that are not represented in the `surveys` data frame. However, it is easy to add them! Here are the columns in `surveys`:

Download Data not in surveys

IPEDS occasionally changes the file naming convention on the Data Center (http://nces.ed.gov/ipeds/datacenter/DataFiles.aspx). Though all of the current years are available, there may (*likely*) be surveys from older years that are not represented in the surveys data frame. However, it is easy to add them! Here are the columns in surveys:

```
> names(surveys)
```

```
[1] "SurveyID"    "Survey"      "Title"
[4] "DataFilePre" "DataFilePost" "YearFormat"
```

Download Data not in surveys

IPEDS occasionally changes the file naming convention on the Data Center (http://nces.ed.gov/ipeds/datacenter/DataFiles.aspx). Though all of the current years are available, there may (*likely*) be surveys from older years that are not represented in the surveys data frame. However, it is easy to add them! Here are the columns in surveys:

```
> names(surveys)

[1] "SurveyID"      "Survey"        "Title"
[4] "DataFilePre"   "DataFilePost" "YearFormat"
```

- SurveyID - Unique id within this table. This is what is passed to the various functions (e.g. getIPEDSSurvey, ipedsHelp).
- Survey - The survey category.
- Title - The survey description.
- DataFilePre - The portion of the filename before the year.
- DataFilePost - The portion of the filename after the year.
- YearFormat - Either 2 or 4 for whether the year is in 2 or 4 digits.

1 Getting Started with the IPEDS Package

2 Using the IPEDS Package for Data Reporting

3 Using the IPEDS Package for Data Analysis

4 Conclusions

Conclusions

My goals were to:

- Streamline analysis that utilizes IPEDS data.
- Provide a framework to automate such analysis.
- Facilitate the process of reporting data to IPEDS.

Conclusions

My goals were to:

- Streamline analysis that utilizes IPEDS data.
- Provide a framework to automate such analysis.
- Facilitate the process of reporting data to IPEDS.

There is still much work to be done. I have focussed on data extraction and data reporting for 4-year private institutions. However, the framework to expand the utility of this package for all institutions is there.

If you are interested in contributing to this project, please contact me at `jason@bryer.org`.

Thank You

Jason Bryer (jbryer@bryer.org)

`http://ipeds.r-forge.r-project.org`

`http://bryer.org`