Postsecondary Education Data Management and Analysis

By: Itai Gerchikov, Chaewon Lee, Zixin Ouyang, Michael Varga

Introduction

For many high school students today, choosing a postsecondary institution can be a stressful process. In this data report, we will combine several datasets that contain information about postsecondary institutions so that students can have a comprehensive list of options to help them during their college selection process. This data report uses two datasets that were created by the Integrated Postsecondary Education Data System, and were accessed via the data.gov website. The first dataset was found at: https://inventory.data.gov/dataset/032e19b4-5a90-41dc-83ff-6e4cd234f565/resource/38625c3d-5388-4c16-a30f-d105432553a4 and contains 7,770 observations and 66 variables. This dataset contains information about 7,770 postsecondary institutions in the United States and its territories. This information includes location characteristics such as state and city, as well as university name and ID, and the name and title of the school's chief executive officer. The second dataset was found at: https://catalog.data.gov/dataset/college-scorecard/resource/3f8438a9-ba64-4085-95b4-2be89994413e?inner_span=True and contains 7,704 observations and 96 variables. The variables from this dataset that we focused on were university ID, and test scores broken down into categories, such as ACT math, ACT science, or SAT reading, and SAT math. Both datasets are CSV files, and were merged by their university ID.

Methods

For our first dataset, we were able to compile over 7,500 observations from a post-secondary university survey from universities in 2013. From this dataset, we imported raw data variables into SAS; these variables and observations included student IDs, locations by ZIP code, state, and city, institution executives, and institution names. This data set allowed us to create a backbone for a follow up data set, as well as a future merge.

For our second data set, we brought in a data set that allowed us to import raw data containing SAT and ACT scores by university ID. The SAT scores were broken down by reading, math, writing, and an average score. The ACT scores were broken down by science, English, and math; we created an average ACT score variable because there was no ACT average variable in the dataset. This dataset contained test score information, which the first dataset did not have. Both datasets contain university ID, so we will merge the datasets by university ID to create a comprehensive dataset containing both university location characteristics and test score information.

To begin the merge, we started out by sorting the first dataset by the variable UNITID and listing the variables we kept; we did this using the proc sort function. After this, we sorted out the second dataset, again using proc sort, listing test scores by UNITID. Once our two datasets were sorted, we created a merged data set using the merge function. There were institutions on the second data set that weren't on the first data set; so, for the sake of the merge, we deleted these institution observations out of our merged dataset. Following the merge, we used the proc contents procedure to see that there were 7,769 observations, 16 variables, and our variable list in alphabetical order.

To start the clean, we deleted the extra 4 digits in zip codes. We then followed this up by checking for missing values. We used the conditional output, 'where', to check missing values by printing observations that were missing for the variables UNITID, INSTNM, CITY, STABBR, CHFNM, CHFTITLE, and ZIP5. We then used the proc freq function to see if there

were executive title labels or state labels, under variables CHFTITLE and STABBR, that were incorrectly repeated, and in need of formatting. Once we saw which titles needed proper labeling, we went ahead and used ‘if’ statements to adjust the titles in need.

Following the cleanup, we created a do loop to assess test scores by state. Before we could run the loop, we had to delete the observations of the schools that do not release test scores. After this, we created a new dataset, Tests_by_state, and used do loops to find average test scores by state. We followed this by formatting and labeling the variables, so they could be adequately printed out.

We also used proc sql in order to create a table of schools with the highest average ACT score per state. We included variables for state, institution name, and average ACT score, and grouped the observations by state. Then we used conditional statements to output only the school with the highest average ACT score per state,

Results

After the two datasets were merged and zip code was changed to only include the first five digits, the proc print procedure with conditional output was used to find missing values. This procedure was used in order to help us during the cleaning process. The results of this procedure are:

Obs	UNITID	INSTNM	CITY	STABBR	CHFNM	CHFTITLE	ZIP5
341	114372	Fashion Careers College	San Diego	CA			92110
382	116040	LA College International	Los Angeles	CA			90010
621	126410	Boulder College of Massage Therapy	Boulder	CO			80301

858	136011	Everest Institute-Hialeah	Hialeah	FL			33012
941	138983	Augusta State University	Augusta	GA			30904
971	139773	Gainesville State College	Oakwood	GA			30566
989	140322	Macon State College	Macon	GA			31206
992	140401	Georgia Health Sciences University	Augusta	GA			30912
994	140483	Middle Georgia College	Cochran	GA			31014
1009	140997	South Georgia College	Douglas	GA			31533
1020	141307	Waycross College	Waycross	GA			31503
1632	160968	Bangor Theological Seminary	Bangor	ME			44020
1725	164207	Washington Bible College-Capital Bible Seminary	Lanham	MD			20706
2598	191384	Global Business Institute	Far Rockaway	NY			11691
2599	191393	Global Business Institute	New York	NY			10035
2686	194204	Olean Business Institute	Olean	NY			14760
3123	204741	Ohio State School of Cosmetology & Experts Barber School	Columbus	OH			43224
3158	205692	Lincoln College of Technology-Franklin LCT	Franklin	OH			45005
3160	205717	Lincoln College of Technology-Vine Street	Cincinnati	OH			45202
4158	233499	Saint Pauls College	Lawrenceville	VA			23868
4890	379463	Lawton Career Institute-Oak Park Campus	Oak Park	MI			48237
4924	381796	Kaplan College-Denver	Thornton	CO			80229
4928	382054	Texas School of Business-Southwest	Houston	TX			77057
4962	384184	Newbridge College-Santa Ana	Santa Ana	CA			92705

4963	384193	Kaplan College-Stockton	Stockton	CA			95207
5069	406060	Southeastern Beauty School-Columbus Midtown	Columbus	GA			31906
5074	406264	Sanford-Brown College-Hazelwood	Hazelwood	MO			63042
5138	410034	Academy of Somatic Healing Arts	Norcross	GA			30092
5224	417390	Michael's School of Beauty	Augusta	GA			30909
5622	438036	Lincoln Technical Institute-Cromwell	Cromwell	CT			6416
5640	438638	Everest Institute-Decatur	Decatur	GA			30035
5649	438887	Newbridge College-San Diego East	El Cajon	CA			92020
5763	440970	Everest College-Vancouver Massage Therapy	Vancouver	WA			98684
5809	441681	Centro de Capacitacion y Asesoramiento Vetelba	Arecibo	PR			612
5845	442310	Everest College-Arlington	Arlington	VA			22203
5858	442499	Westwood College-Ft Worth	Fort Worth	TX			76137
5859	442505	Westwood College-Dallas	Dallas	TX			75243
5931	443605	Texas School of Business-East	Houston	TX			77029
6056	445355	Carrington College California-Emeryville	Emeryville	CA			94608
6062	445416	Kaplan University-Council Bluffs Campus	Council Bluffs	IA			51503
6133	446428	Performance Training Institute	Toms River	NJ			8755
6162	446783	Lincoln Technical Institute-Suffield	Suffield	CT			6078
6302	448655	Kaplan College-Milwaukee	Milwaukee	WI			53212
6345	449278	Kaplan Career Institute-Detroit	Detroit	MI			48202
6398	450012	ATI College-Santa Ana	Santa Ana	CA			92701

6456	450775	Newbridge College-Long Beach	Long Beach	CA			90815
6546	451893	Lawton Career Institute-Warren Main Campus	Warren	MI			48089
6815	456995	Provo College-American Fork	American Fork	UT			84003
6817	457013	Kaplan Career Institute-Boston	Boston	MA			2215
6922	458238	Kaplan College-Pembroke Pines	Pembroke Pines	FL			33026
7181	461467	CCIC Beauty College	Peoria	AZ			85345
7258	462257	Lincoln College of Technology-Columbus	Columbus	OH			43215
7271	463621	Everest College-Milwaukee	Milwaukee	WI			53212
7391	475246	Kaplan College-Chesapeake	Chesapeake	VA			23320
7637	481216	Mid-America Baptist Theological Seminary	Cordova	TN			38016
7759	482972	Texas Covenant Education	San Antonio	TX			78201

There were 56 observations with missing values for chief name and chief title. These institutions either didn't have a chief executive for their institution, or they didn't provide this information to the IPEDS system. We used if statements with do loops to clean these observations so that all missing values for chief name and chief title were changed to "N/A". Proc freq was used to validate our data and look for missing or incorrectly repeated chief executive title labels or state labels, under variables CHFTITLE and STABBR. The results of this procedure were:

Number of Variable Levels				
Variable	**Label**	**Levels**	**Missing Levels**	**Nonmissing Levels**
CHFTITLE	Chief Title	646	1	645
STABBR	State or Region	59	0	59

State or Region				
STABBR	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
AK	12	0.15	12	0.15
AL	93	1.20	105	1.35
AR	88	1.13	193	2.48
AS	1	0.01	194	2.50
AZ	146	1.88	340	4.38
CA	789	10.16	1129	14.53
CO	141	1.81	1270	16.35
CT	100	1.29	1370	17.63
DC	25	0.32	1395	17.96
DE	21	0.27	1416	18.23
FL	402	5.17	1818	23.40
FM	1	0.01	1819	23.41
GA	199	2.56	2018	25.98
GU	3	0.04	2021	26.01
HI	28	0.36	2049	26.37
IA	96	1.24	2145	27.61
ID	44	0.57	2189	28.18
IL	320	4.12	2509	32.30
IN	139	1.79	2648	34.08

KS	96	1.24	2744	35.32
KY	117	1.51	2861	36.83
LA	128	1.65	2989	38.47
MA	197	2.54	3186	41.01
MD	105	1.35	3291	42.36
ME	43	0.55	3334	42.91
MH	1	0.01	3335	42.93
MI	204	2.63	3539	45.55
MN	149	1.92	3688	47.47
MO	226	2.91	3914	50.38
MP	1	0.01	3915	50.39
MS	64	0.82	3979	51.22
MT	31	0.40	4010	51.62
NC	199	2.56	4209	54.18
ND	30	0.39	4239	54.56
NE	53	0.68	4292	55.25
NH	44	0.57	4336	55.81
NJ	173	2.23	4509	58.04
NM	52	0.67	4561	58.71
NV	53	0.68	4614	59.39
NY	481	6.19	5095	65.58
OH	383	4.93	5478	70.51
OK	149	1.92	5627	72.43

OR	99	1.27	5726	73.70
PA	406	5.23	6132	78.93
PR	158	2.03	6290	80.96
PW	1	0.01	6291	80.98
RI	24	0.31	6315	81.28
SC	113	1.45	6428	82.74
SD	31	0.40	6459	83.14
TN	190	2.45	6649	85.58
TX	479	6.17	7128	91.75
UT	88	1.13	7216	92.88
VA	178	2.29	7394	95.17
VI	1	0.01	7395	95.19
VT	28	0.36	7423	95.55
WA	128	1.65	7551	97.19
WI	127	1.63	7678	98.83
WV	79	1.02	7757	99.85
WY	12	0.15	7769	100.00

The results from the frequency tables indicate that other than the 56 missing values for chief executive title and chief executive name, there are no missing values for any of the non test related variables.

Next, the proc contents procedure was used to summarize our cleaned data set:

Data Set Name	WORK.MERGED2	**Observations**	7769

Member Type	DATA	**Variables**	15
Engine	V9	**Indexes**	0
Created	05/11/2017 23:21:51	**Observation Length**	440
Last Modified	05/11/2017 23:21:51	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	SOLARIS_X86_64, LINUX_X86_64, ALPHA_TRU64, LINUX_IA64		
Encoding	utf-8 Unicode (UTF-8)		

Engine/Host Dependent Information	
Data Set Page Size	131072
Number of Data Set Pages	27
First Data Page	1
Max Obs per Page	297
Obs in First Data Page	288
Number of Data Set Repairs	0
Filename	/saswork/SAS_work3A4800011DA6_odaws04-prod-us/SAS_workBB7C00011DA6_odaws04-prod-us/merged2.sas7bdat
Release Created	9.0401M3
Host Created	Linux
Inode Number	15728660
Access Permission	rw-r--r--
Owner Name	mpvarga20

File Size	4MB
File Size (bytes)	3670016

Alphabetic List of Variables and Attributes					
#	**Variable**	**Type**	**Len**	**Format**	**Label**
10	ACTCMMID	Num	8		ACT Science Median
11	ACTENMID	Num	8		ACT English Median
12	ACTMTMID	Num	8		ACT Math Median
14	ACT_AVG_ALL	Num	8	2.	ACT Average
3	CHFNM	Char	100		Chief Name
4	CHFTITLE	Char	100		Chief Title
2	CITY	Char	30		City
1	INSTNM	Char	100		Institution Name
8	SATMTMID	Num	8		SAT Math Median
7	SATVRMID	Num	8		SAT Reading Median
9	SATWRMID	Num	8		SAT Writing Median
13	SAT_AVG_ALL	Num	8		SAT Average
6	STABBR	Char	8		State or Region
5	UNITID	Num	8		Institute ID
15	ZIP5	Char	30		Zip Code

Our cleaned and validated dataset contains 7,769 observations and 15 variables.

Next, an iterative do loop was used to find the average SAT and ACT score by state in order to conclude which states had the students with the highest and lowest average test scores. In order

to do this, we had to clean the data to include only the schools that release test scores. The results from this iterative do loop are:

Obs	State	Average ACT Per State/Region	Average SAT Per State/Region
1	AK	21	1054
2	AL	22	1033
3	AR	23	1045
4	AZ	24	1100
5	CA	23	1114
6	CO	23	1076
7	CT	14	1087
8	DC	20	1175
9	DE	21	980
10	FL	21	1057
11	GA	21	1018
12	HI	22	1019
13	IA	23	1074
14	ID	22	1027
15	IL	23	1078
16	IN	22	1057
17	KS	22	1021
18	KY	23	1051
19	LA	22	1033
20	MA	18	1131

21	MD	19	1060
22	ME	20	1051
23	MI	23	1067
24	MN	23	1109
25	MO	24	1097
26	MS	21	989
27	MT	23	1068
28	NC	20	1006
29	ND	23	1046
30	NE	23	1065
31	NH	19	1114
32	NJ	16	1061
33	NM	22	1023
34	NV	22	1030
35	NY	17	1113
36	OH	23	1064
37	OK	22	1029
38	OR	20	1083
39	PA	20	1062
40	RI	22	1171
41	SC	16	1007
42	SD	22	1045
43	TN	23	1048
44	TX	21	1038

45	UT	25	1135
46	VA	19	1054
47	VI	18	796
48	VT	20	1058
49	WA	20	1128
50	WI	23	1056
51	WV	21	984
52	WY	24	1118

It is important to note that the dataset produced from the iterative do loop only contains 52 observations because several territories and states didn't publish their average test scores. We can conclude from this dataset that Connecticut has the lowest average ACT score with a score of 14, and Utah had the highest average ACT score with a score of 25. Washington DC had the highest average SAT score with a score of 1175, while the US Virgin Islands had the lowest average SAT score with a score of 796.

We also used the proc sql procedure to create a table of the school in each state that had the highest average ACT score. This table displays the school with the highest ACT score in each state so that high-scoring students in each state know which school is the best school for them.

State or Region	ACT Average	Institution Name
AK	21	Alaska Pacific University
AL	27	Birmingham Southern College
AL	27	University of Alabama in Huntsville
AL	27	Auburn University

AR	27	Hendrix College
AZ	25	Arizona State University-Tempe
CA	35	California Institute of Technology
CO	30	Colorado School of Mines
CT	25	Quinnipiac University
DC	30	George Washington University
DE	27	University of Delaware
FL	30	University of Miami
GA	32	Georgia Institute of Technology-Main Campus
HI	25	Brigham Young University-Hawaii
IA	32	Grinnell College
ID	24	The College of Idaho
IL	34	University of Chicago
IN	30	Rose-Hulman Institute of Technology
KS	25	University of Kansas
KY	29	Centre College
LA	31	Tulane University of Louisiana
MA	34	Massachusetts Institute of Technology
MD	33	Johns Hopkins University
ME	30	Colby College
MI	31	University of Michigan-Ann Arbor
MN	31	Macalester College
MO	33	Washington University in St Louis
MS	26	Millsaps College
MT	25	Montana Tech of the University of Montana

MT	25	Carroll College
NC	33	Duke University
ND	26	University of Jamestown
NE	27	Creighton University
NH	25	University of New Hampshire-Main Campus
NJ	33	Princeton University
NM	26	New Mexico Institute of Mining and Technology
NV	23	University of Nevada-Reno
NY	33	Columbia University in the City of New York
OH	32	Case Western Reserve University
OK	29	University of Tulsa
OR	31	Reed College
PA	33	University of Pennsylvania
RI	32	Brown University
SC	26	Wofford College
SD	26	Augustana College
TN	33	Vanderbilt University
TX	33	Rice University
UT	29	Brigham Young University-Provo
VA	32	Washington and Lee University
VI	18	University of the Virgin Islands
VT	27	University of Vermont
WA	28	University of Washington-Seattle Campus

WI	29	University of Wisconsin-Madison
WV	23	West Virginia University
WY	24	University of Wyoming

The school with the highest available average ACT score is MIT in Massachusetts with an average score of 34. The school with the lowest maximum available average ACT score is the University of the Virgin Islands in the US Virgin Islands with an average score of 18.