Data Analysis: Academic Library Collections, Services, and Expenditures

Susan Van Alstyne

Statistics for Educational Research—EDTC 810

Dr. Glazer

May 3, 2019

Contents

Introduction

The Academic Libraries Survey (ALS) now referred to as the American Libraries (AL) component provides information about library collections and operations for all academic libraries in the United States. The data collection process went through many iterations since 1966 when the data collection was through National Center for Data Statistics (NCES), and now the academic libraries data is a component of the Integrated Postsecondary Education Data System (IPEDS), and this data is now collected annually. There were so many changes in the data collection process when in the past, data was collected either biennially and even triennially; however, since 2014, data continues to be collected annually. With so many changes, a researcher questions this reliability of this data as well the difficulty of comparing year over year especially when this data is collected to plan, evaluate, and create policies. Data about library staffing, operating expenses, the collection such as total volumes, circulation of items, reference transactions, "electronic services" and gate count (number of visitors in the library) is collected. The challenge is that many libraries have different methods of collecting this data such as counting electronic circulation. Congress uses academic library statistics to assess fund allocation and grants and evaluate the condition of academic libraries. There are many uses for IPEDS data, such as State agencies and accrediting bodies use to assess programs as well as professional associations such as the American Library Association (ALA) for statistics about librarianship as a profession as well as librarians to use as benchmark data (IPEDS Q & A, 2016).

Data Collection

The data collected is a segment from the IPEDS dataset to explore the relationship between library expenditures by full-time equivalents (FTE), and the relation to the circulation of the physical and electronic collection (combined) by FTE. The parameters are public 4-year and 2-year institutions and private for-profit two and four-year institutions in New Jersey for the year of 2017, most of the data is from the public institutions as there are inconsistencies or no data reported for the private for-profit institutions. The Academic Libraries Overview survey instrument (see https://surveys.nces.ed.gov/IPEDS) used to collect data is available via the IPEDS website. A researcher can choose a variety of parameters from the IPEDS dataset. Researchers can download files for over 7,000 colleges and 250 variables (IPEDS, 2019). A researcher would need to be well versed in this dataset and focus on the variables of the study as a researcher could spend some time experimenting with many combinations and testing relationships between the variables. Also, a researcher needs to determine whether the accuracy the recency of the data when researchers need to decide between provisional and final release data as was the case with the NCES data for academic libraries (NCES, 2019). As a researcher, the constant changes in the data collection process for Academic Libraries is challenging because of the different data permutations.

The data sets were downloaded via the IPEDS interface by variables and institution characteristics. Institution characteristics include Public and Private-for-profit two- and four-year institutions. Then the Academic Library Data for total library expenditures per FTE Expenditures include salaries, physical library material purchases, subscriptions, additional materials and operations, maintenance, expenditures. An additional report was run to retrieve the FTE for each institution.

Data Analysis

The hypothesis is that library expenditures have a positive correlation on overall library circulation. Circulation is an indicator of activity and is a significant factor in planning library operations. The alternative hypothesis is that there is no significant relationship between library expenditures and circulation. Therefore:

$$H_0 = \mu \text{ expenditures} = \mu \text{ circulation}$$

$$H_a = \mu \text{ expenditures} \neq \mu \text{ circulation}$$

In this relationship, circulation is the dependent variable and expenditure is the independent variable. The following sections will present descriptive statistics about the dataset and inferential statistics to run the correlation coefficient and regression analysis. The correlation coefficient would be best to examine the strength of the relationship between expenditures and circulation. Regression analysis is one inferential statistic to run on this data because this data is used for strategic planning as well as for making decisions about budgets, and regression analysis would provide appropriate forecasting analysis (Kent State University Libraries, 2017).

Descriptive Statistics

The initial dataset was combined with additional data to retrieve a more cohesive set of variables. With the initial analysis of the public and private-for-profit two and four-year institutions, much of the data was not reported, and the data had to be removed from the dataset for calculations. As a researcher, looking at the first run of data – there is a large range of values as well as null values the decision to keep this data set as in reality, if running a survey, there may be many unreported variables, and a researcher may need to adjust the analysis strategy and data collection if there is time to collect additional data. For this report, the null values for circulation and expenditures were deleted from the analysis. The decision was made following

the initial SPSS data output with *N*=55 as illustrated in Table 1. The data in Table 2 is the breakdown by sector. The researcher also decided to calculate circulations per FTE following the initial run of the data.

Table 1

Descriptives

Descriptives - Descriptive Statistics - May 1, 2019

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
LibraryExpendituresperFTE	36	735	0	735	265.17	34.178	205.067	42052.371	.832	.393	-.470	.768
Totallibrarycirculationsphysicalanddigitalelectronic	38	1249625	275	1249900	235634.37	63148.569	389273.924	151534187704.563	1.939	.383	2.247	.750
Valid N (listwise)	32											

The initial look at the data with the standard error SEM = 34.178 and SEM=63,146, indicates a better selection of variables is needed either by computation as to compare Library Expenditures per FTE with Circulations by FTE, that figure needs to be calculated using another data set. The initial data shows a large variance between the selected variables as there are many outliers in this data. Normalizing the data by breaking down the dataset in percentiles may also be another option. For this report, the relationship between collection use and expenditures will be analyzed to determine if expenditures influence library circulation of material, which is a significant indicator of library usage. It is also clear to the researcher about which sector reports more data, as Table 2 shows that public institutions report more data, $n = 30$ than the private sector with only two institutions reporting valid data. Based on the data, an assumption can be made that public institutions need to meet reporting requirements. The researcher then will look at the individual 55 cases to determine what to include in the data analysis as including null values will not be conducive to determine the relationship between the two variables. The descriptive statistics for the modified dataset are shown in Table 3.

Table 2
Breakdown by Sector

Sector

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Private for-profit, 2-year	1	3.1	3.1	3.1
	Private for-profit, 4-year or above	1	3.1	3.1	6.3
	Public, 2-year	18	56.3	56.3	62.5
	Public, 4-year or above	12	37.5	37.5	100.0
	Total	32	100.0	100.0	

Table 3 *N=33*

Statistics

		Circulation per FTE	Total library expenditures per FTE (DRVAL2017)
N	Valid	33	33
	Missing	0	0
Mean		76.2206	307.61
Std. Error of Mean		47.67915	43.292
Median		12.0912	216.00
Mode		.68[a]	179
Std. Deviation		273.89587	248.695
Variance		75018.948	61849.371
Skewness		5.474	1.522
Std. Error of Skewness		.409	.409
Kurtosis		30.762	1.802
Std. Error of Kurtosis		.798	.798
Range		1577.14	943
Minimum		.68	57
Maximum		1577.82	1000

a. Multiple modes exist. The smallest value is shown

Table 3 shows Circulation per FTE mean (*M*= 76.2206), a median (*Mdn*= 12), and mode of 68, as noted below Table 3, multiple modes exist, and the smallest value is 68. The Circulation per FTE is non-normally distributed with skewness of 5.474. The Library Expenditures per FTE

mean (*M*= $307.61), and mode of $179.00 and *Mdn* = $216.00. The Library Expenditures per FTE kurtosis of 1.802, a kurtosis less than 3 (<3) describes data that is not normally distributed with thinner tails around the mean (Kent State University, 2017).

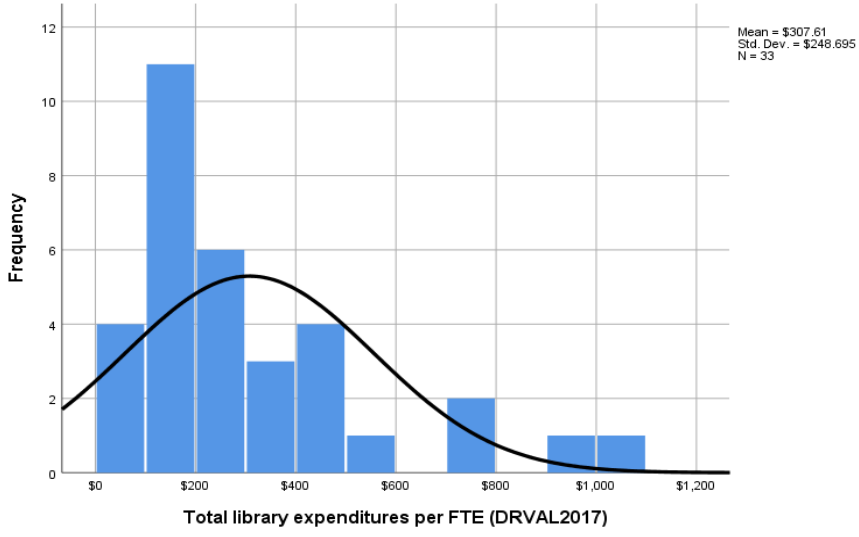
Define the Variables in the Dataset

Table 4 below shows the variable view following the SPSS import of the dataset. Complete definitions are provided in the introduction and data collection sections.

Table 4
Variable View

UnitID	Numeric		Scale
InstitutionName	String	Institution Name	Nominal
Totallibrarycirculationsphysicalanddigitalelectronic AL2017	Numeric	Total library circulations (physical and digital/electronic) (AL2017)	Scale
@12monthfulltimeequivalentenrollment201617DRVEF122017	Numeric	12-month full-time equivalent enrollment: 2016-17 (DRVEF122017)	Scale
Sectorofinstitution	Numeric	Sector of institution	Nominal
TotallibraryexpendituresperFTEDRVAL2017	Dollar	Total library expenditures per FTE (DRVAL2017)	Scale
CirculationperFTE	Numeric	Circulation per FTE	Scale
NumberofDatabasesreported	Numeric	Number of Databases (reported)	Scale

Table 5*: Histogram for Library Expenditure per FTE, N = 33*



The Histogram for Total Library Expenditure per FTE indicates a strong positive skew with the frequency or most of the institutions with a Total Library Expenditure per FTE in alignment with the M=$307.61

The next relationship or set of variables to explore is if there is a significant relationship between the number of databases and the number of circulations per FTE.

Table 6
Descriptive Statistics for the number of databases and circulation per FTE

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
Number of Databases (reported)	33	0	739	103.73	155.978
Circulation per FTE	33	.68	1577.82	76.2206	273.89587
Valid N (listwise)	33				

Correlation

The bivariate Pearson's Correlation Coefficient was run to see if there is a significant relationship within the variables.

Table 7
Pearson's Correlation Coefficient

		Total library expenditures per FTE (DRVAL2017)	Circulation per FTE	Number of Databases (reported)
Total library expenditures per FTE (DRVAL2017)	Pearson Correlation	1	-.125	.252
	Sig. (2-tailed)		.489	.158
	N	33	33	33
Circulation per FTE	Pearson Correlation	-.125	1	-.057
	Sig. (2-tailed)	.489		.751
	N	33	33	33
Number of Databases (reported)	Pearson Correlation	.252	-.057	1
	Sig. (2-tailed)	.158	.751	

N	33	33	33

According to Table 7 there are no significant correlations, the Sig. (2-tailed) alternatively, 2-tailed *p-value* is greater than 0.05 and cannot reject the null hypothesis with this information either.

Regression Analysis

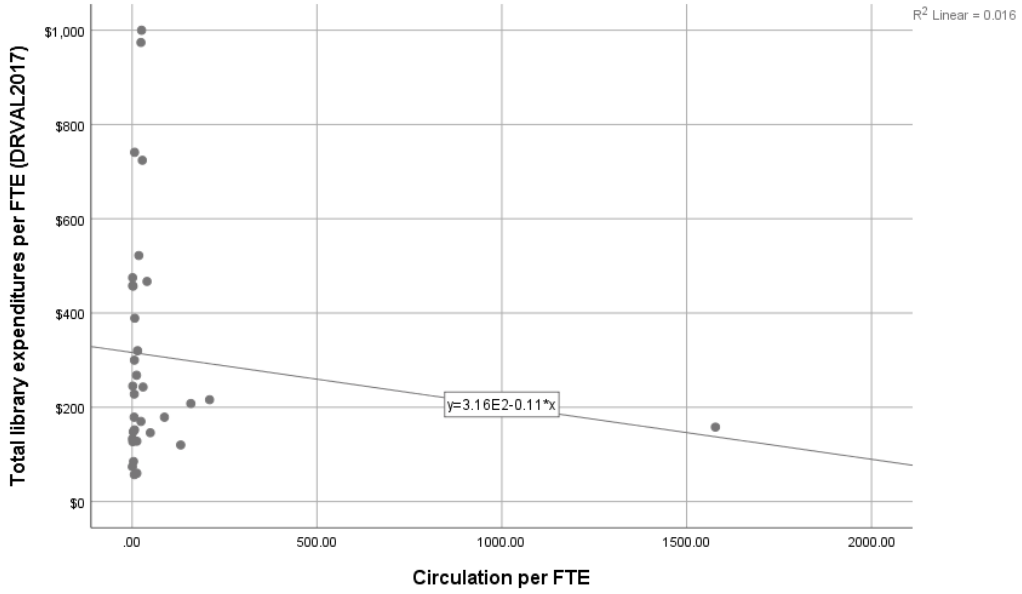


Figure 1. Linear regression analysis with the SPSS output of using the scatter/dot of Circulation and Expenditure with regression as y=3.16E2-0.11*x, which shows a negative relationship. Though most of the data shows data along the same line, both *Figure 1* and *Figure 2* (below) show a significant relationship cannot be established with the current analysis. Though the R^2 =0.016, closer to zero where a value of zero would indicate a strong relationship.

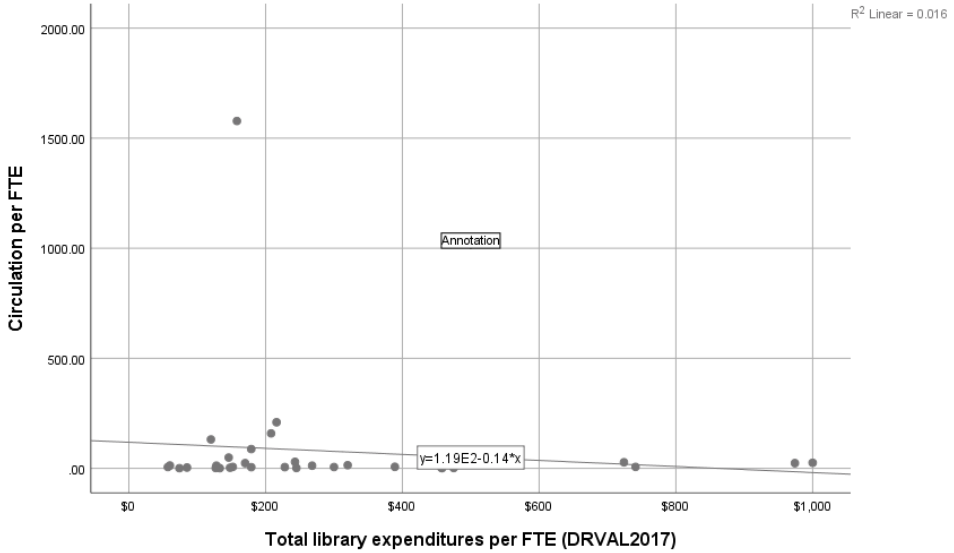


Figure 2. SPSS output

Regression Model

The results of the linear regression analysis using SPSS as shown in Table 9, which is shown below *Figure* 3 All calculations show a significant relationship cannot be defined using this dataset. Though *Figure 3*, the Normal P-P Plot shows a positive cumulative relationship because of standardizing the data and lowering the confidence interval.

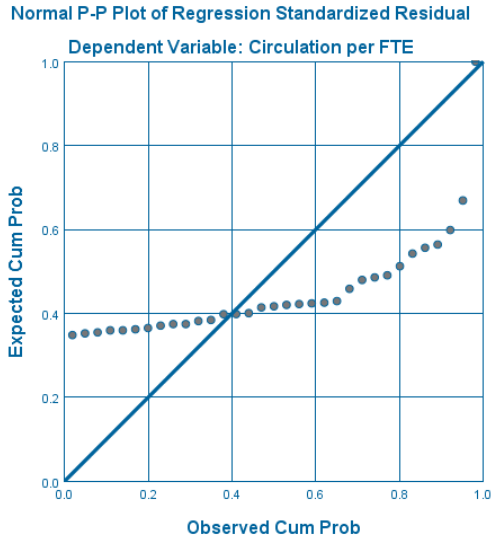


Figure 3

Table 8

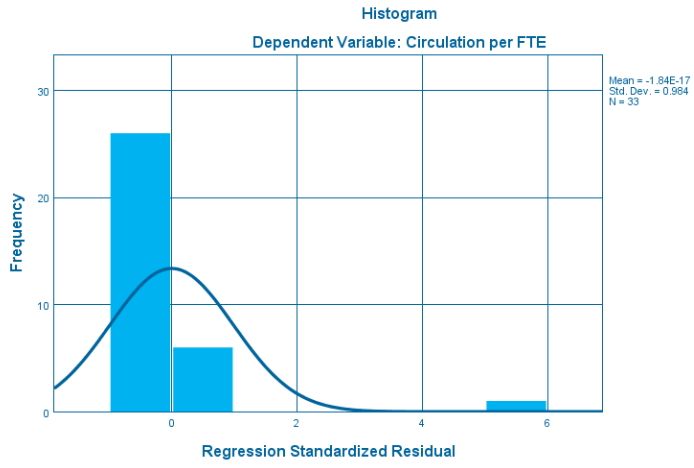


Table 9
SPSS Regression Analysis

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.125[a]	.016	-.016	276.10283

a. Predictors: (Constant), Total library expenditures per FTE (DRVAL2017)

b. Dependent Variable: Circulation per FTE

Regression - ANOVA

ANOVA[a]ANOVA,

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37390.332	1	37390.332	.490	.489[b]
	Residual	2363216.011	31	76232.775		
	Total	2400606.343	32			

a. Dependent Variable: Circulation per FTE

b. Predictors: (Constant), Total library expenditures per FTE (DRVAL2017)

Regression - Coefficients

Coefficients[a]Coefficients,

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	118.500	77.166		1.536	.135
	Total library expenditures per FTE (DRVAL2017)	-.137	.196	-.125	-.700	.489

a. Dependent Variable: Circulation per FTE

Additional Variable Relationships

Another relationship to explore is to see if the number of databases and circulation have a significant relationship. Therefore:

H_0=μ number of databases = μ circulation

Ha= μ number of databases ≠μ circulation

A Chi Test is run to determine the relationship between the number of databases and circulation rates. There are no strong relationships that can be determined by the results in Table 10 and Table 11 of the SPSS Chi-Square Tests.

Table 10

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	957.000[a]	928	.248
Likelihood Ratio	222.452	928	1.000
Linear-by-Linear Association	.105	1	.746
N of Valid Cases	33		

a. 990 cells (100.0%) have expected count less than 5. The minimum expected count is .03.

Table 11

Symmetric Measures

		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Nominal by Nominal	Contingency Coefficient	.983			.248
Interval by Interval	Pearson's R	-.057	.032	-.320	.751[c]
Ordinal by Ordinal	Spearman Correlation	.163	.167	.922	.364[c]
N of Valid Cases		33			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Conclusion

After running several tests, the conclusion is that the variables do not have any significant relationship. There are issues with the dataset, and there are several approaches to correct the data which include but not limited to increasing the sample size, setting parameters such as data for institutions with minimum values, or breaking down the dataset in percentiles, and decreasing the confidence interval. As a researcher, working with this dataset proves the need to ensure reliability in the data to obtain reasonable results to analyze relationships.

Though library use rates differ by size and sector of the institution, the variables of the number of available databases and library expenditure per FTE may be able to determine a relationship with an improved dataset, and that starts with consistent reporting. This analysis

rejects the null hypothesis for the relationship between expenditures and circulation and the null hypothesis of the number of databases and use or circulation rates. The results of the study support further analysis and improved data collection methods to establish the relationship between the variables.

References

IPEDS. (2019). The Integrated Postsecondary Education Data System. Retrieved from https://nces.ed.gov/ipeds/use-the-data

IPEDS Q & A for Library directors and contacts (2015-16). Retrieved from https://nces.ed.gov/ipeds/resource/download/QandAforLibrarians2015-16.docx

Kent State University Libraries. (2017, May 15). SPSS tutorials: Retrieved May 17, 2017, from http://libguides.library.kent.edu/SPSS/

NCES. (2017, May). Academic Libraries Survey (ALS). Retrieved April 30, 2019, from https://nces.ed.gov/surveys/libraries/academic.asp

Appendix A

American Libraries (AL) Survey extracted from the IPEDS (Integrated Postsecondary Education Data System.

UnitID	InstitutionName	Totallibrarycirculations physicalanddigitalelectronicAL2017	@12monthfulltimeequivalentenrollment201617DRVEF122017	Sectorofinstitution	TotallibraryexpendituresperFTEDRVAL2017	CirculationperFTE	NumberofDatabases reported
183655.0	Atlantic Cape Community College	94617	3977	4	$170.00	23.79	41
183743.0	Bergen Community College	301113	10193	4	$243.00	29.54	62
183789.0	Berkeley College-Woodland Park	28286	3995	3	$389.00	7.08	81
183859.0	Brookdale Community College	118425	9414	4	$128.00	12.58	46
183938.0	Camden County College	98487	7673	4	$60.00	12.84	35
184180.0	County College of Morris	1227541	5860	4	$216.00	209.48	78
184205.0	Cumberland County College	6637	2333	4	$148.00	2.84	114
184959.0	Eastwick College-Ramsey	4030	739	3	$179.00	5.45	25
184481.0	Essex County College	12240	7257	4	$127.00	1.69	76
184995.0	Hudson County Community College	74494	6161	4	$268.00	12.09	103
455196.0	Jersey College	10106	2611	6	$85.00	3.87	2
185262.0	Kean University	69504	11985	1	$228.00	5.80	246
185509.0	Mercer County Community College	7990084	5064	4	$158.00	1577.82	60
1855	Middlesex County College	55694	8570	4	$152.00	6.50	97

UnitID	InstitutionName	Totallibrarycirculationsphysicalanddigitalelectronic AL2017	@12monthfulltimeequivalentenrollment201617DRVEF122017	Sectorofinstitution	TotallibraryexpendituresperFTEDRVAL2017	CirculationperFTE	NumberofDatabasesreported
36.0							
185590.0	Montclair State University	115512	18756	1	$300.00	6.16	165
185129.0	New Jersey City University	8976	6768	1	$475.00	1.33	0
185828.0	New Jersey Institute of Technology	24193	9123	1	$457.00	2.65	34
185873.0	Ocean County College	575993	6587	4	$179.00	87.44	44
186034.0	Passaic County Community College	8719	5269	4	$245.00	1.65	76
186201.0	Ramapo College of New Jersey	7951	6043	1	$458.00	1.32	0
186645.0	Raritan Valley Community College	33058	5510	4	$57.00	6.00	71
183877.0	Rowan College at Burlington County	4759	7021	4	$74.00	0.68	52
184791.0	Rowan College at Gloucester County	269962	5489	4	$146.00	49.18	31
184782.0	Rowan University	283531	15749	1	$522.00	18.00	739
186371.0	Rutgers University-Camden	149757	5875	1	$1,000.00	25.49	2
186380.0	Rutgers University-New Brunswick	1345096	48743	1	$724.00	27.60	591
186399.0	Rutgers University-Newark	257023	10824	1	$974.00	23.75	6
186876.0	Stockton University	134401	9220	1	$320.00	14.58	176
247603.0	Sussex County Community College	1548	1833	4	$133.00	0.84	21

UnitID	InstitutionName	Totallibrarycirculationsphysicalanddigitalelectronic AL2017	@12monthfulltimeequivalentenrollment201617DRVEF122017	Sectorofinstitution	TotallibraryexpendituresperFTEDRVAL2017	CirculationperFTE	NumberofDatabasesreported
187134.0	The College of New Jersey	53139	7872	1	$741.00	6.75	122
187198.0	Union County College	1150061	7240	4	$208.00	158.85	102
245625.0	Warren County Community College	166779	1269	4	$120.00	131.43	19
187444.0	William Paterson University of New Jersey	372462	9261	1	$467.00	40.22	106