On April 7, 2016, The New York Times Posted: "**More Than 40% of Student Borrowers Aren't Making Payments**" About 1 in 6 borrowers, or 3.6 million, were in default on $56 billion in student debt, meaning they had gone at least a year without making a payment. Three million more owing roughly $66 billion were at least a month behind.

The College Analysis

Increasing Students Ability to Pay

Meantime, another three million owing almost $110 billion were in "forbearance" or "deferment," meaning they had received permission to temporarily halt payments due to a financial emergency, such as unemployment. The figures exclude borrowers still in school and those with government-guaranteed private loans.

On May 10, 2016, CBS News 8 reported: "**Student loan forgiveness could ease college debt**"

Data Science Certificate Program

Georgetown University

Spring 2016

The Department of Education has launched a new website to help people with student loans find a repayment option that best suits their needs. The move is part of an effort the Obama administration is undertaking to enroll an additional 2 million people into repayment programs such as the Pay As You Earn program, which caps monthly student loan payments at 10 percent of income.

Federal student loan debt exceeds $1.3 trillion, and about one in seven borrowers default on their loans within three years of beginning to repay them.

Team College

Mauricio Botero

Angela Felder

LaRia Rogers

Andrea Wise

On May 11, 2016, Business Insider posted: "**Ivy League borrowers are hitting student-loan lenders where it hurts**" Online lenders initially targeted Ivy League borrowers for their qualities: good credit, relatively high income levels, and high levels of education. Paradoxically, these are the exact traits now hurting the companies. And to entice the students, online lenders have offered lower and lower interest rates. After graduation, students are eager to refinance, further incentivized by online lenders' lack of fees to do so — unlike the federal government or major lenders.

These articles point out the concern for the ongoing increase in student loan debt and students ability to pay
and begin our story…

The Abstract:

The cost of college is steadily increasing. As incoming and prospective students weigh the costs and benefits of higher education, it is important to find schools that can offer the best possible outcomes regarding debt: a burden that impacts everyone. We've utilized key school data provided by the US Department of Education and a selection of distinctive student personal preference criteria to complete an analysis that recommends schools with the best debt-to-earnings ratios based on a time series forecast.

An Introduction:

[The Problem] Student debt has significantly increased, loan borrowers aren't paying off their student loan debt. The cost of a college education is escalating at a rate higher than that of financial aid assistance through grants and scholarships. This is forcing students to borrow more, increasing the student loan debt. More than two-thirds of college graduates graduated with an average debt of just over $35,000.00 in 2014. This is three times more than that in 1993-1994. However, the increase on salaries over the past two decades has remained fairly flat. In 2014, the average salary was $46,481.52, while only increasing by less than two times the 1995 salary of $23,753.53. A graduate's first year out of college is likely to make below the national average. The average starting salary of 2014 college graduates with a bachelor's degree was $45,473.00.

While students are coming out with increasingly more debt and an average salary, the national cohort default rate (CDR) remains high. The most recent cohort default rate published by the US Department of Education for fiscal year 2012 shows a national cohort default rate of 11.8 percent. This means that of the 5.1 million federal student loan borrowers that entered repayment between October 1, 2011, and September 30, 2012, there were about 611,000 that defaulted on their loans before September 30, 2014. This calculation does not account for those borrowers that take advantage of the deferment and forbearance programs during this period reducing chances of default.

There is a great deal of earnings variation within a college, so that students with good outcomes at low-earning schools often perform better than students with poor outcomes at high-earning schools. Earnings variation can be explained at least in part by the program in which students choose to study. However, program-level earnings data is not yet available. This means that a student who graduates from an ivy league school such as Harvard may have a higher debt than someone that goes to an in-state public school. However, a Harvard graduate is more likely to obtain a higher paying job, increasing his/her ability to pay. So…What can we do now to help students make a better decision on selecting a college in order to increase their ability to pay back their student loan debt?

[The Hypothesis] Will providing prospective students with a measurement based on the analysis of the median student loan debt to the median earnings improve the student's ability to pay back their student loans? Once a student is matched to schools based on student

preferences (e.g. public vs. private, major, region, etc) and ranking the matched schools by calculating a debt-to-earnings ratio, we expect an increase to the student's ability to pay. We predict students who receive this analysis, and specifically attend a private four-year university, has a greater chance of making high earnings, therefore a greater chance of being able to pay off their student loans. (The focus of this study is on a standard four-year Bachelor's Degree program of study.)
For a cohort of college graduates the *Debt-to-Earnings Ratio = Total Median Debt / Total Median Earnings*

Debt to income ratio is considered to be manageable if it is less than 10%. If your debt to income ratio exceeds 10% it is considered to be extremely risky. Most research indicates that total student loan debt should not exceed more than 57% of annual earnings.

The Collection and Wrangling of Data:

The data used in this study was collected from the U.S. Department of Education's College Scorecard. The data is a compilation of demographic, program, type, costs, and acceptance information on degree granting institutions. For those schools that participate in Title IV financial aid programs and the students that received Title IV loans, the data include information on student loan debt as well as the median earnings for a given cohort of students.

The sources that make up the College Scorecard data include: National Center of Education Statistics' (NCES), Integrated Postsecondary Education Data System (IPEDS), Federal Student Aid (FSA), and FSA's National Student Loan Data System (NSLDS), and Treasury. The data available through the College Scorecard consisted of 18 fiscal years (1996 - 2013) and 1,720 data elements for each OPEID (Office of Postsecondary Education Identifier) in each fiscal year. There are as many as 7,800 OPEIDs within a given fiscal year. The OPEID is a number assigned by Federal Student Aid to identify a school and a specific location. The first six (6) digits identify a school's primary campus and the last two (2) digits identify a specific location.

The key data elements include school demographic information, school type, average SAT, programs, median loan debt (2003-2011), and earnings(2003-2011) outlined in Appendix B.

The college Scorecard data was downloaded from the site(s) in a .csv file. The .csv file was cleaned and only the values as described in Appendix B were used. Two of the major obstacles was 1) missing data, and 2) duplicate data in the original data sets for earnings and debt. For missing data on Earnings and Debt, we noticed that only the 2003, 2005, 2007, 2009, and 2011 years contained data to be used (the rest of the years were null). After that, duplicate data was found in the schools in which the main campus had several school branches (i.e. University of Texas-Austin had the same data as University of Texas-El Paso). Due to duplication in data, all branches from the main campuses were excluded.

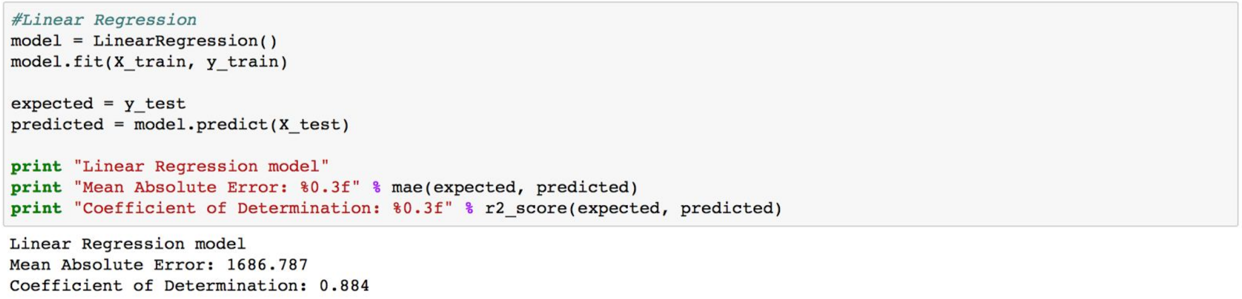
After the data was cleaned, the .csv file was then ingested using Jupyter iPython and loaded into a data-frame using pandas. During the wrangling phase, the general statistics were pulled and scatterplots were created to get a sense of the features and labels we were going to use. Another major problem included all of the earnings (md_earn_wne_p6_*year*), and debt columns (GRAD_DEBT_MDN_*YEAR)* contained several null values (i.e. for the 2003 year md_earn_wne_p6_2003 there were 2030 null values out of 5681 Schools). For this, we calculated the mean value for each of the earnings and debts years (2003, 2005, 2007, 2009, 2011) in order to replace the null values.

The Computation and Analysis:

Once the data was normalized, we decided to do a two-time series forecast: one for earnings and one for debt. For earnings, we took the earnings years (2003-2009) and set them as our features, and 2011 as our label such that →

$\text{Earnings}_{≙2011} = \text{Earnings}_{2009} + \text{Earnings}_{2007} + \text{Earnings}_{2005} + \text{Earnings}_{2003}$

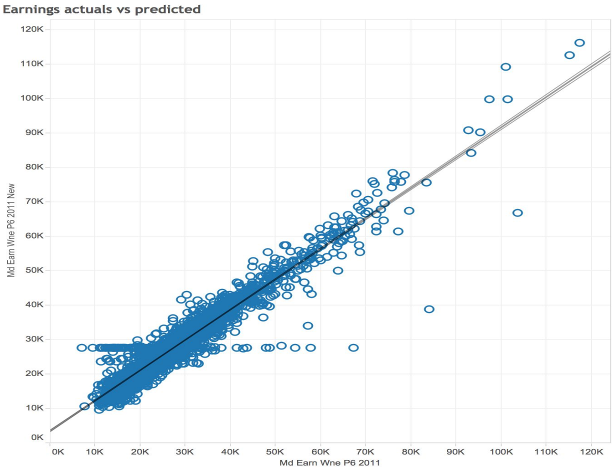
We then moved to set the splits and evaluate different models by calculating its MAE (Mean Absolute Error) and its R^2 (Coefficient of Determination). The model that had the best fit was a linear regression model which yielded the following results (MAE = 1686.7 and R^2 = 88.4%)



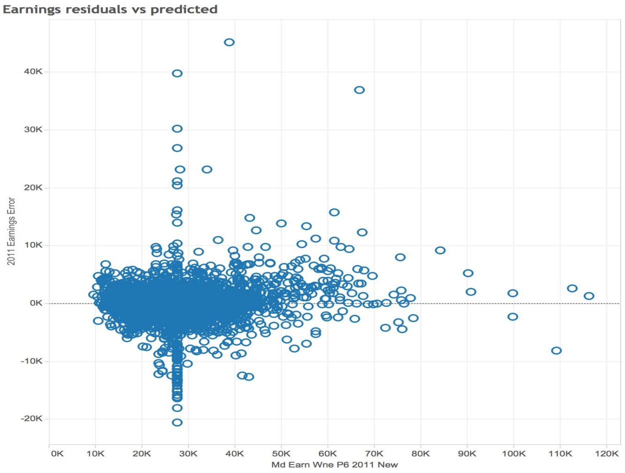
```
#Linear Regression
model = LinearRegression()
model.fit(X_train, y_train)

expected = y_test
predicted = model.predict(X_test)

print "Linear Regression model"
print "Mean Absolute Error: %0.3f" % mae(expected, predicted)
print "Coefficient of Determination: %0.3f" % r2_score(expected, predicted)
```

```
Linear Regression model
Mean Absolute Error: 1686.787
Coefficient of Determination: 0.884
```

Using a linear regression we forecasted $\text{Earnings}_{≙2011}$ and utilized Tableau to plot the following two graphs below:



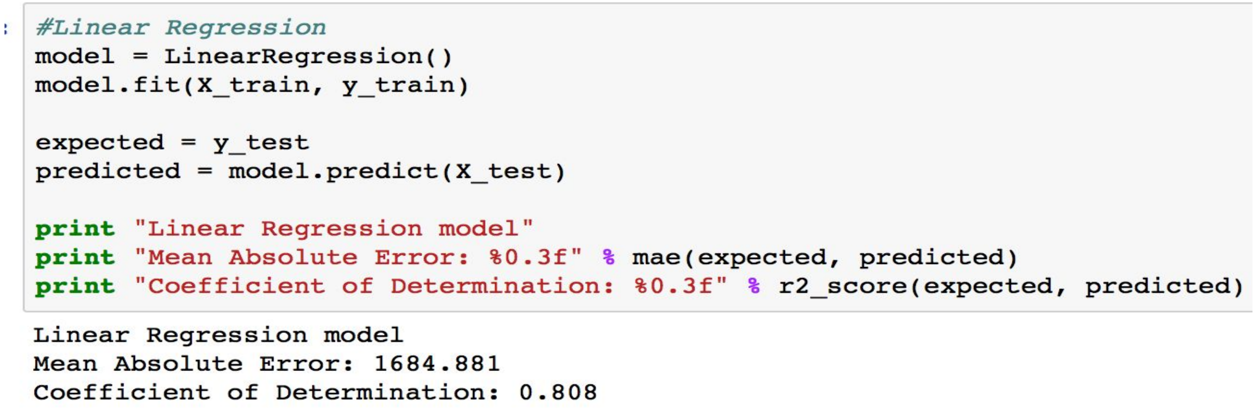
Sum of Md Earn Wne P6 2011 vs. sum of Md Earn Wne P6 2011 New. Details are shown for INSTNM (Earnings Error 2011).



Sum of Md Earn Wne P6 2011 New vs. sum of 2011 Earnings Error. Details are shown for INSTNM (Earnings Error 2011).

The left graph above illustrates the forecasted Earnings$_{\hat{=}2011}$ with the actuals Earnings$_{2011}$. As you can see from the trendline there is a strong correlation between the forecasted vs. the actuals values (ρ = .95). To the right graph, we plotted the Earnings error (Earnings error = Actual earnings - forecasted earnings). The values above 0 indicate that the values were forecasted below the actuals, and the values below 0 indicate that the values were forecasted above the actuals. You can clearly see that the great majority of the forecasted values lie within the 0 horizontal line.

For the debt, we used the same time series as earnings such that → Debt$_{\hat{=}2011}$ = Debt$_{2009}$ + Debt$_{2007}$+ Debt$_{2005}$ + Debt$_{2003}$ setting debt (2003-2009) as our features, and debt 2011 as our label. Then we set the splits and evaluted different models and ended up using a linear regression with the following results (MAE = 1684.8, R^2 = 80.8%) =

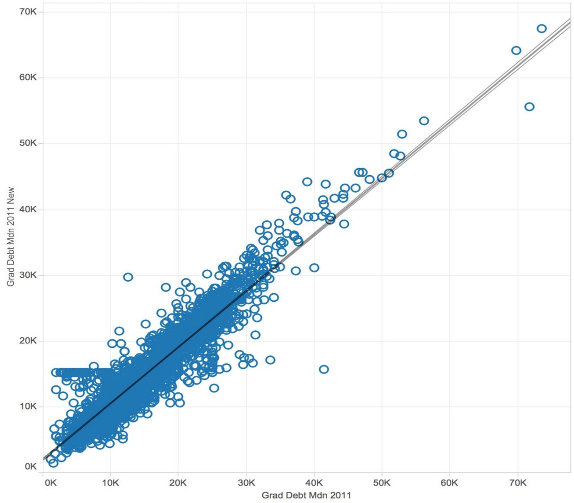


```
#Linear Regression
model = LinearRegression()
model.fit(X_train, y_train)

expected = y_test
predicted = model.predict(X_test)

print "Linear Regression model"
print "Mean Absolute Error: %0.3f" % mae(expected, predicted)
print "Coefficient of Determination: %0.3f" % r2_score(expected, predicted)
```

```
Linear Regression model
Mean Absolute Error: 1684.881
Coefficient of Determination: 0.808
```

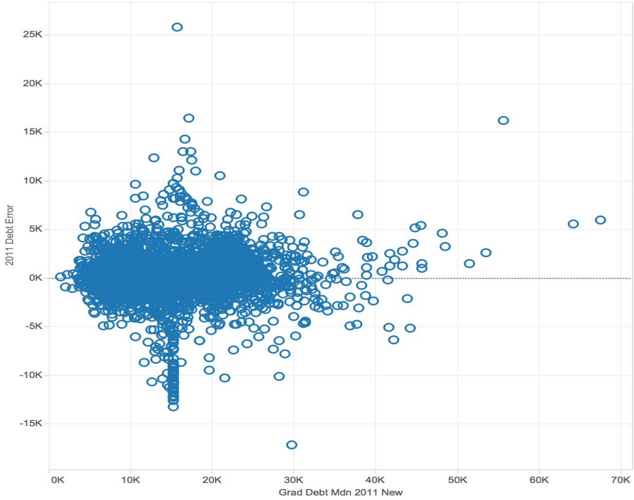
Also using a linear regression we forecasted Debt$_{\hat{=}2011}$ and utilized Tableau to plot the following two graphs below:

Debt actuals vs predicted



Sum of Grad Debt Mdn 2011 vs. sum of Grad Debt Mdn 2011 New. Details are shown for Instnm.

Debt residuals vs predicted

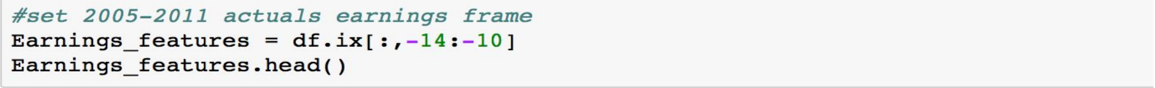


Sum of Grad Debt Mdn 2011 New vs. sum of 2011 Debt Error. Details are shown for Instnm.

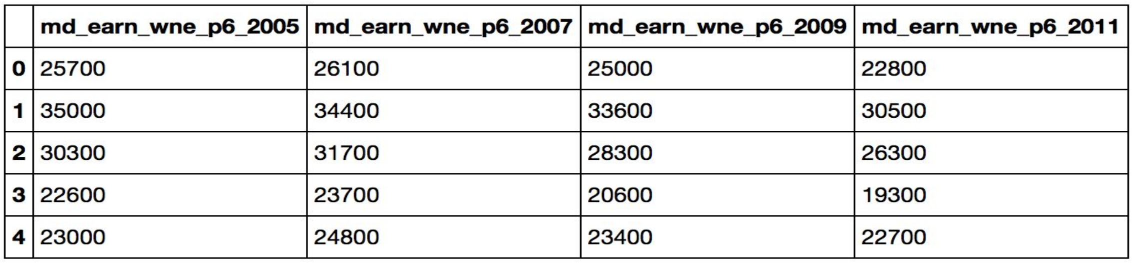
Debt graphs above were the same as earnings , showing a strong *correlation* (ρ = .91) between our forecasted and actual values.

The Modeling and Application:

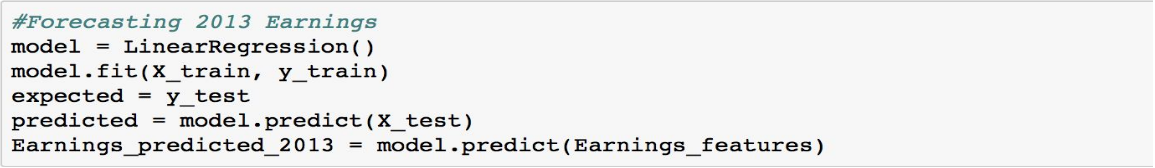
For the modeling, we continued to forecast both earnings and debt for the years 2013, 2015, 2017, 2019, 2021 using a linear regression model. As an example, to forecast 2013 earnings we used 2005, 2007, 2009, and 2011 as our features to then forecast earnings for 2013 as illustrated in the figure below:



```
#set 2005-2011 actuals earnings frame
Earnings_features = df.ix[:,-14:-10]
Earnings_features.head()
```

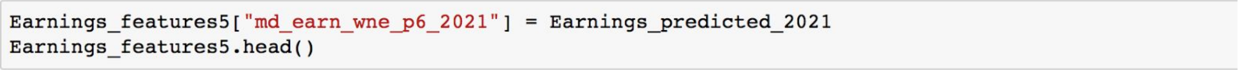


	md_earn_wne_p6_2005	md_earn_wne_p6_2007	md_earn_wne_p6_2009	md_earn_wne_p6_2011
0	25700	26100	25000	22800
1	35000	34400	33600	30500
2	30300	31700	28300	26300
3	22600	23700	20600	19300
4	23000	24800	23400	22700

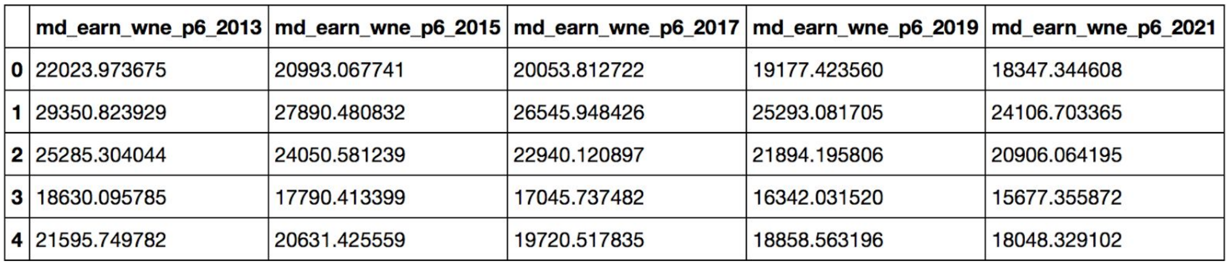


```
#Forecasting 2013 Earnings
model = LinearRegression()
model.fit(X_train, y_train)
expected = y_test
predicted = model.predict(X_test)
Earnings_predicted_2013 = model.predict(Earnings_features)
```

We repeated this same procedure (using the last 4 sets data as our features) to forecast until the year 2021 per figure below:

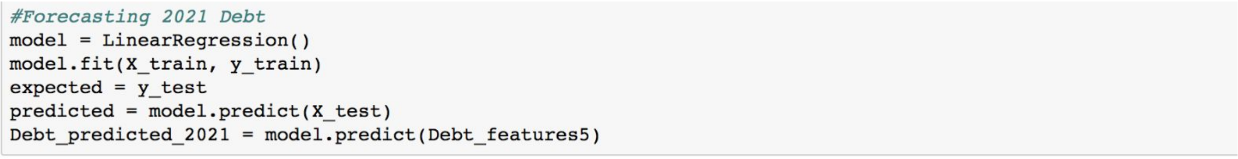


```
Earnings_features5["md_earn_wne_p6_2021"] = Earnings_predicted_2021
Earnings_features5.head()
```

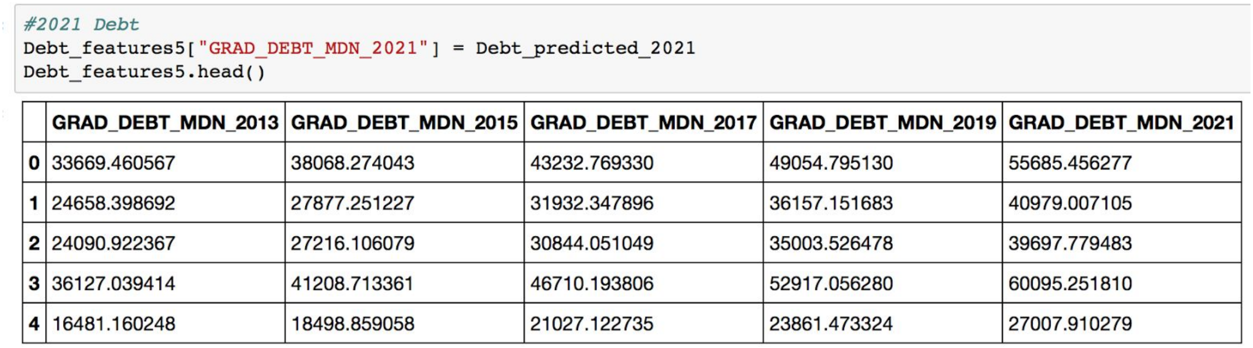


	md_earn_wne_p6_2013	md_earn_wne_p6_2015	md_earn_wne_p6_2017	md_earn_wne_p6_2019	md_earn_wne_p6_2021
0	22023.973675	20993.067741	20053.812722	19177.423560	18347.344608
1	29350.823929	27890.480832	26545.948426	25293.081705	24106.703365
2	25285.304044	24050.581239	22940.120897	21894.195806	20906.064195
3	18630.095785	17790.413399	17045.737482	16342.031520	15677.355872
4	21595.749782	20631.425559	19720.517835	18858.563196	18048.329102

Once earnings was forecasted until 2021, we proceeded with Debt using the same model and methodology as illustrated in the figure below:



```
#Forecasting 2021 Debt
model = LinearRegression()
model.fit(X_train, y_train)
expected = y_test
predicted = model.predict(X_test)
Debt_predicted_2021 = model.predict(Debt_features5)
```

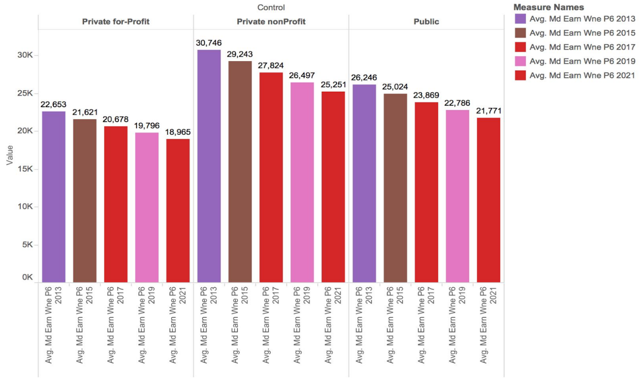


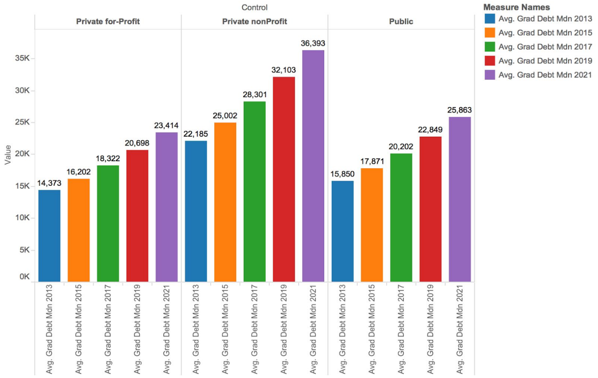
```
#2021 Debt
Debt_features5["GRAD_DEBT_MDN_2021"] = Debt_predicted_2021
Debt_features5.head()
```

	GRAD_DEBT_MDN_2013	GRAD_DEBT_MDN_2015	GRAD_DEBT_MDN_2017	GRAD_DEBT_MDN_2019	GRAD_DEBT_MDN_2021
0	33669.460567	38068.274043	43232.769330	49054.795130	55685.456277
1	24658.398692	27877.251227	31932.347896	36157.151683	40979.007105
2	24090.922367	27216.106079	30844.051049	35003.526478	39697.779483
3	36127.039414	41208.713361	46710.193806	52917.056280	60095.251810
4	16481.160248	18498.859058	21027.122735	23861.473324	27007.910279

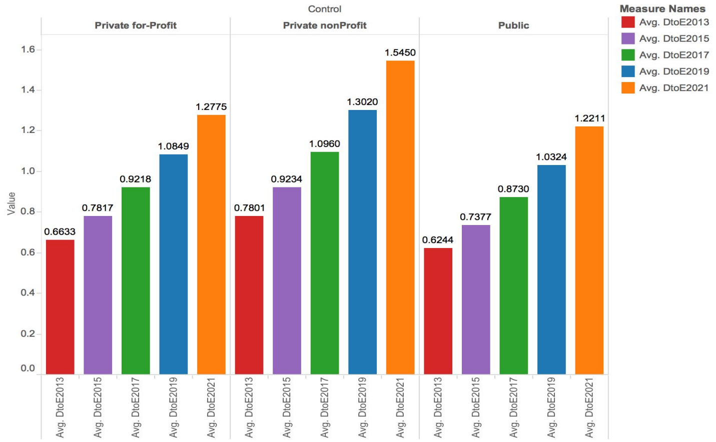
The Reporting and Visualization:

One we got our forecasted values, both for earnings and debt (2013-2021 for even years), we proceeded with the reporting and visualization in Tableau. We loaded the forecasted data and found that earnings are forecasted to drop while debt is forecasted to increase per graphs below (left graph are earnings and right graph are debts):

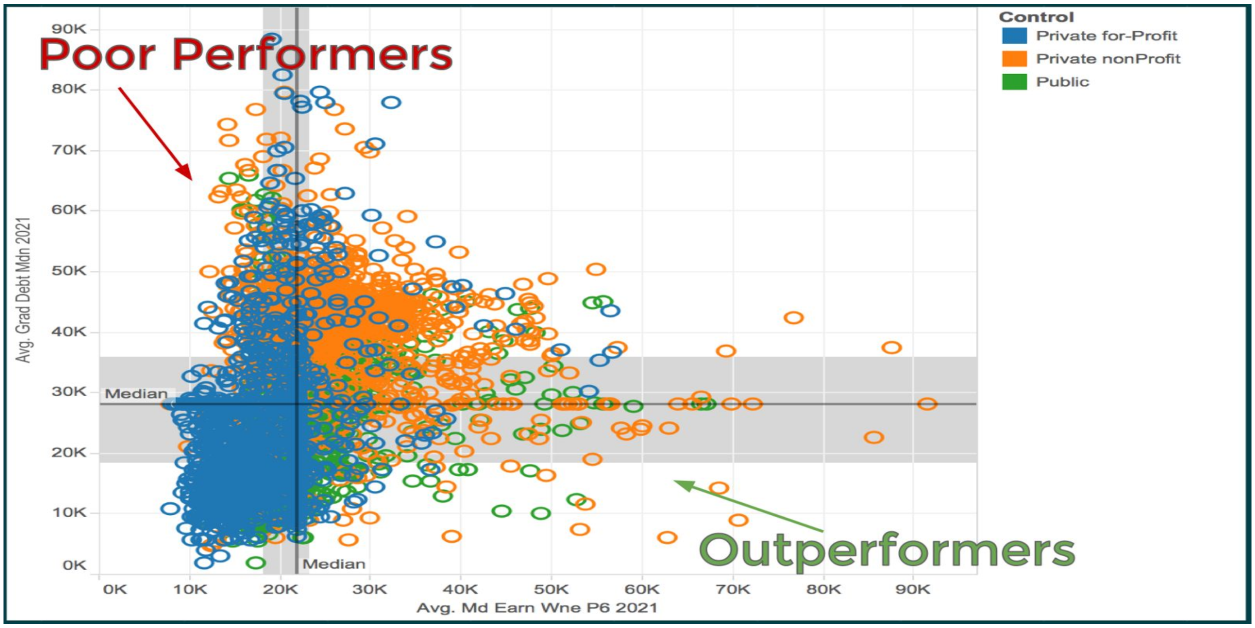




Moreover, you can also see that Private nonProfit Universities have higher earnings but also higher debts than Private for profit, and Public Universities. We then calculated debt to earnings ratios and as suspected they are also set to increase for the three types of Universities with Private nonProfit Universities having the highest ones, and Public Universities the lowest ones per figure below:



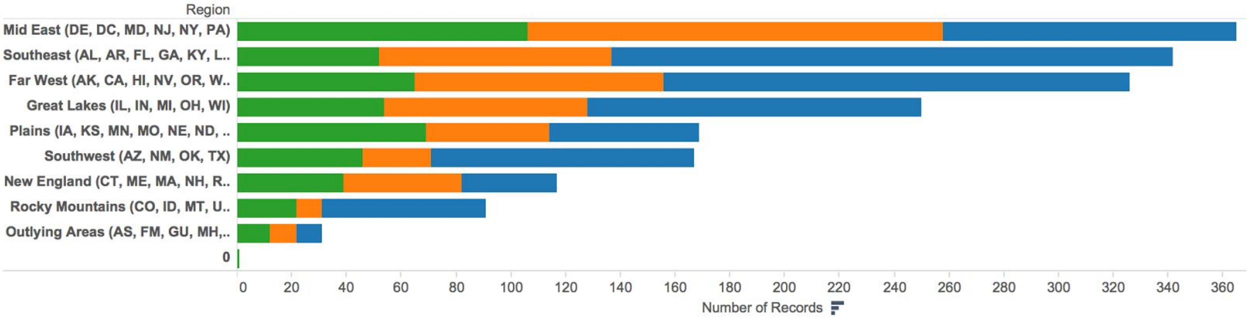
Lastly, we created a scatter plot of 2021 debts (Y-axis) with earnings (X-axis) and set a median quadrant to divide the Universities in 4 and focusing in the ones that were outperformers (above earnings and debts median) and poor performers (below earnings and debt median) per figure below:



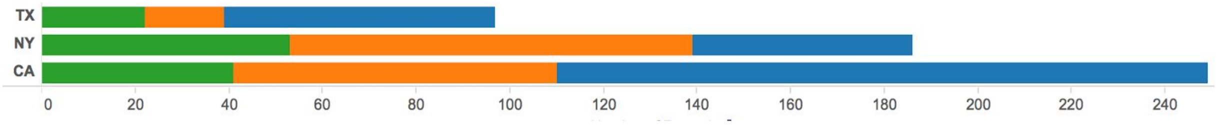
The 3 types of school were also color coded, and at the general level you can see that Private nonProfit universities tend to have higher average median earnings but also higher average median debts. You can also see that Private for-Profit Universities tend to have lower average median debts but also lower average median earnings. We then built two sets one for the outperformers, and the second one for the poor performers to proceed with the final recommendations.

The Making of Decisions:

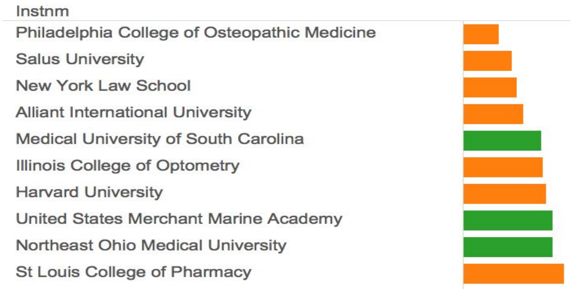
Analyzing our two sets, one for outperformers and the second one for the poor performers, we recommend that students will find more schools that outperform within the Mid East region states per figure below:



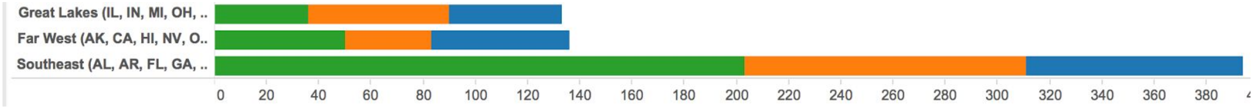
We also recommend for students to look at California, New York, and Texas as the states with the highest number of schools that outperform per figure below:

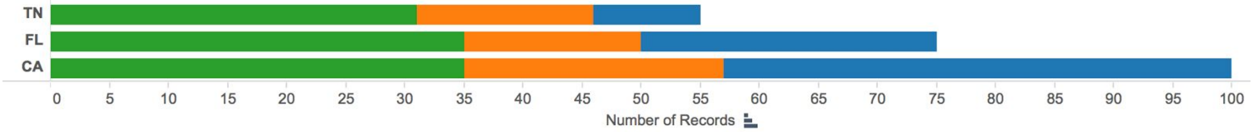


Lastly, below are our top 10 outperform schools:

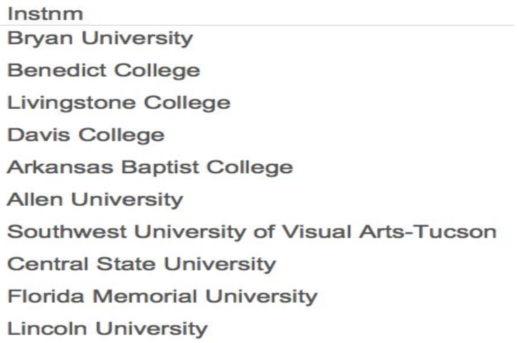


For our poor performers we found that the Southeast states have the largest numbers of poor performers. At the same time, California also has the largest numbers of poor performers followed by Florida and Tennessee per figures below:





Below are our top 10 poor performers schools:



Instnm
Bryan University
Benedict College
Livingstone College
Davis College
Arkansas Baptist College
Allen University
Southwest University of Visual Arts-Tucson
Central State University
Florida Memorial University
Lincoln University

The Conclusion:

In general, we believe we gathered sufficient data and employed a good model to forecast future earnings and debts, which then allowed us to give a fair recommendation to prospective college students. Staying with a time series model, we could have imported the Statsmodel package, and determine a better time series model. We could have also explore other variables to measure successful/non-successful schools such as a MANOVA and get other different results to compare.

The Bibliography:

Why the Student Loan Crisis Is Even Worse Than People Think, Time.com, Mark Kantrowitz, Jan. 11, 2016, http://time.com/money/4168510/why-student-loan-crisis-is-worse-than-people-think/

National Average Wage Index, Social Security Administration, https://www.ssa.gov/oact/cola/AWI.html

Starting Salaries Up 1.2% for Class of 2014 Grads, Society of Human Resource Management, Stephen Miller, April 4, 2014, https://www.shrm.org/hrdisciplines/compensation/articles/pages/salaries-2014-grads.aspx

Cohort Default Rates, US Department of Education, SEPTEMBER 30, 2015, http://www.ed.gov/news/press-releases/cohort-default-rate-continues-drop-across-all-higher-ed-sectors

Using Federal Data to Measure and Improve the Performance of U.S. Institutions of Higher Education, U.S. Department of Education, September 2015, https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf

College Scorecard, Data Insights, U.S. Department of Education, https://collegescorecard.ed.gov/data/

Granite State Management & Resources, Debt-to-Income Calculator, https://www.gsmr.org/calc_debttoincome.asp?submitbutton=submit

How much student loan debt can I afford?, Columbia College, http://web.ccis.edu/offices/financialaid/howmuchdebt.aspx

FinAid, http://www.finaid.org/calculators/scripts/sloanadvisor.cgi

College Tool Kit, http://calculators.collegetoolkit.com/college-calculators/debttoincome.aspx

College Scorecard Data, U.S. Department of Education, https://collegescorecard.ed.gov/data/

Business Insider: http://www.businessinsider.com/sallie-mae-how-america-saves-for-college-2015-2015-6

Appendix A: Project References

GitHub Repository: https://github.com/georgetown-analytics/team-college

Appendix B: Data Dictionary

Attribute Name	Attribute Name Description	Data Type	Fiscal year	Source
OPEID	8-digit OPEID	integer	2011	IPEDS
INSTNM	Institution name	string	2011	IPEDS
GRAD_DEBT_MDN	Median debt for students who have completed	float	2003, 2005, 2007, 2011	NSLDS
STABBR	Institution state	string	2011	IPEDS
CONTROL	School Type: 1 = Public 2 = Private nonprofit 3 = Private for-profit	integer	2011	IPEDS
Region	Institution regional location: 1 = New England (CT, ME, MA, NH, RI, VT) 2 = Mid East (DE, DC, MD, NJ, NY, PA) 3 = Great Lakes (IL, IN, MI, OH, WI) 4 = Plains (IA, KS, MN, MO, NE, ND, SD) 5 = Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV) 6 = Southwest (AZ, NM, OK, TX) 7 = Rocky Mountains (CO, ID, MT, UT, WY) 8 = Far West (AK, CA, HI, NV, OR, WA) 9 = Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI)	integer	2011	IPEDS
INSTURL	URL for the institution's homepage	string	2011	IPEDS
SAT_AVG	Average SAT equivalent score of students admitted	float	2011	IPEDS
Programs	The following values are used for the CIPxxBACHL attributes: 0 = Program not offered 1 = Program offered 2 = Program offered through an exclusive distance-education program		2011	
CIP01BACHL	Agriculture, Agriculture Operations, And Related Sciences.	integer	2011	IPEDS

CIP03BACHL	Natural Resources And Conservation.	integer	2011	IPEDS
CIP04BACHL	Architecture And Related Services.	integer	2011	IPEDS
CIP05BACHL	Area, Ethnic, Cultural, Gender, And Group Studies.	integer	2011	IPEDS
CIP09BACHL	Communication, Journalism, And Related Programs.	integer	2011	IPEDS
CIP10BACHL	Communications Technologies/Technicians And Support Services.	integer	2011	IPEDS
CIP11BACHL	Computer And Information Sciences And Support Services.	integer	2011	IPEDS
CIP12BACHL	Personal And Culinary Services.	integer	2011	IPEDS
CIP13BACHL	Education.	integer	2011	IPEDS
CIP14BACHL	Engineering.	integer	2011	IPEDS
CIP15BACHL	Engineering Technologies And Engineering-Related Fields.	integer	2011	IPEDS
CIP16BACHL	Foreign Languages, Literatures, And Linguistics.	integer	2011	IPEDS
CIP19BACHL	Family And Consumer Sciences/Human Sciences.	integer	2011	IPEDS
CIP22BACHL	Legal Professions And Studies.	integer	2011	IPEDS
CIP23BACHL	English Language And Literature/Letters.	integer	2011	IPEDS
CIP24BACHL	Liberal Arts And Sciences, General Studies And Humanities.	integer	2011	IPEDS
CIP25BACHL	Library Science.	integer	2011	IPEDS
CIP26BACHL	Biological And Biomedical Sciences.	integer	2011	IPEDS
CIP27BACHL	Mathematics And Statistics.	integer	2011	IPEDS
CIP29BACHL	Military Technologies And Applied Sciences.	integer	2011	IPEDS
CIP30BACHL	Multi/Interdisciplinary Studies.	integer	2011	IPEDS
CIP31BACHL	Parks, Recreation, Leisure, And Fitness Studies.	integer	2011	IPEDS
CIP38BACHL	Philosophy And Religious Studies.	integer	2011	IPEDS
CIP39BACHL	Theology And Religious Vocations.	integer	2011	IPEDS
CIP40BACHL	Physical Sciences.	integer	2011	IPEDS
CIP41BACHL	Science Technologies/Technicians.	integer	2011	IPEDS
CIP42BACHL	Psychology.	integer	2011	IPEDS

CIP43BACHL	Homeland Security, Law Enforcement, Firefighting And Related Protective Services.	integer	2011	IPEDS
CIP44BACHL	Public Administration And Social Service Professions.	integer	2011	IPEDS
CIP45BACHL	Social Sciences.	integer	2011	IPEDS
CIP46BACHL	Construction Trades.	integer	2011	IPEDS
CIP47BACHL	Mechanic And Repair Technologies/Technicians.	integer	2011	IPEDS
CIP48BACHL	Precision Production.	integer	2011	IPEDS
CIP49BACHL	Transportation And Materials Moving.	integer	2011	IPEDS
CIP50BACHL	Visual And Performing Arts.	integer	2011	IPEDS
CIP51BACHL	Health Professions And Related Programs.	integer	2011	IPEDS
CIP52BACHL	Business, Management, Marketing, And Related Support Services.	integer	2011	IPEDS
CIP54BACHL	History.	integer	2011	IPEDS
MD_EARN_WNE_p6	Median earnings of students working and not enrolled six (6) years after entry	integer	2003, 2005, 2007, 2011	Treasury