Higher Education Institutions and Student Enrollment

An Introductory Analysis on IPDES 2012-2020 Data

By Carlotta Amaduzzi
Data Science Certificate Candidate

University of Maryland Baltimore County
DATA 606 – Capstone Course
Dr. Chaojie Wang

Abstract

Drawing from openly available data published and maintained by IPDES (The Integrated Postsecondary Education Data System), I set forth to analyze whether it would be possible to determine if publicly available information disclosed patterns in students' enrollment decisions in Higher Education Institutions (HEIs).

Given the nature of the data selected, primarily categorical in nature, and the time frame of reference (years 2012-2020), some interesting trends emerged: a) while it is in fact true that there has been a contraction in aggregate students' enrollment over this time, the HEIs affected have been mostly those at the extremes of the market –the ones that are exceptionally large and those that are exceptionally small, with those in between experiencing greater changes in the composition of their student bodies; b) there have been shifts within the population of students enrolling in public institutions that reflect structural changes occurring within American society; c) publicly disclosed information that traditionally has been considered important in students' admission into HEIs has slowly been changing; HEIs are changing the importance they place on traditional academic characteristics, however in order to determine if these changes are reactionary or intentional would require additional research. HEIs have the option of using some of these features to differentiate themselves from their competitors and use them more intentionally to attract students and promote their educational services.

From a machine learning perspective, given the categorical nature of the data, the presence of numerous outliers, and the inhomogeneous nature of the data, traditional clustering and regression algorithms were non-performing. Only thanks to the adoption of XGBoost (a gradient descent algorithm used for predictive purposes) was I able to achieve good levels of student enrollment prediction. Furthermore, using Principal Component Analysis in connection with Density Based Spatial Clustering of Applications with Noise (DBSCAN) I was able to identify the primary features characterizing HEIs by dimensional group.

It should be noted that the selection of non-traditional features for the analysis was intentional and based on the premise that this was an academic exercise.

Summary of Findings

Institutions of Higher Education (HEIs) are a numerous and heterogenous group of organizations that have a significant impact on society as a whole. While I did not focus on quantifying this impact, we can all agree that they play a significant role on the long term growth of our society, any society for that matter.

In the USA there traditionally is a wide variety of institutions that address a variety of educational needs, from traditionally academic needs to more job-related ones. In the past these different segments seemed to maintain greater distances between each other. Recently, due to changes in enrollment patterns and a reduction in enrollment overall, HEIs seem to have been competing more aggressively not only within their clusters but also intra-clusters to attract student enrollments.

The information collected by IPDES due to reporting standards and requirements placed on HEIs by the Federal Government allows us to see a shift in characteristics HEIs used to consider essential for them to both open their doors to students and to distinguish themselves from the competition. These characteristics are evolving over time and disclosing some of the challenges HEIs are facing to continue to grow (in some case) and to continue to thrive (in most case).

With this research work I set out to see if the information publicly shared by HEIs allowed us to confirm the existence of differences between HEIs that ultimately have an affect of students' enrollment decisions. I chose to use this enrollment metric as a proxy of students' interest, thinking it might be more accurate, draw greater attention from HEIs as it directly impacts their bottom line, and because it more accurately distinguishes HEIs size.

Our work confirms that that reported features' importance is evolving over time and highlighting a shift in strategies by HEIs. It is unclear if these changes are more reactive in nature or not but our results indicate that changes are occurring and that they are occurring in part in parallel to societal changes as well. For example, while there is great debate around the reduction of enrollment numbers aggregately in terms of students who identify as white, there is little mention of the increases in numbers of students enrolling who are Latino or of more than one race or even African American, to a smaller extent. While these increases are not enough to offset the aggregate reduction in enrollment across the USA, they do reflect societal changes that are taking place at the same time as this debate is, and they reflect policy changes that were first enacted by the Obama administration aimed at greater equity and access in education.

HEIs have a wide array of features that can use to differentiate themselves and compete with other HEIs to encourage student enrollment. I only analyzed a small portion of publicly available information and have definitely found differences among HEIs and especially how these differences translate into a different student body, at least so far. Further research would be necessary to explore more deeply these characteristics and to explore them in connection to the extent of their impact on student enrollment decisions. However, our initial finding support our research questions, although there certainly are many other elements at play.

The choices I made in selecting our machine learning algorithms proved correct in connection to both the type of data I had to work with and the objectives I had set out for this work, since ultimately, I was able to support my initial assumptions. Further research and an extension to analyze more in depth the effect of time on these outcomes would be warranted.

Introduction

Access to education has always been considered one of the most important factors that can lead to more equitable and just social structures within societies. Higher education in particular is considered to be an essential mean through which to open greater opportunities for professional, and consequently financial, success, ultimately leading to greater social mobility and thus greater fairness for all members of the population. In other words, admissions into Higher Education institutions, and ultimately enrollment, can open the door to productive and meaningful futures, even when taking into consideration reduced social mobility.

Higher Education Institutions (HEI) are, for all practical purposes, rather large organizations that have significant impact on the economies of the States they are located in (and beyond) – both directly (via the jobs and contracts they create) and indirectly (thanks to the professional work-force they contribute to). They are organizations that need to remain healthy and productive for the long term growth and health of our society.

In recent years, HEI have been reporting a decline in enrollment. Publicly available information seems to confirm this trend, in particular for undergraduate programs. (NSCRC, 2021)

Given the wide impact HEI have on the long term growth potential of a Community, a State, and more broadly of the Nation, it is important to understand the factors that influence students towards enrollment in HEI. Using only publicly available information reported by HEI themselves, and as an exclusively academic exercise, I intend to take a closer look at HEIs reported features to try to identify those having the greatest impact on students' undergraduate enrollment and possibly open avenues for further research and/or interventions for HEI to test.

Research Questions

The questions I looked into with this project are the following:

1) Based on publicly reported information regarding HEIs, do any of the reported features seem to affect student enrollment choice?
2) With an eye to I.D.E.A. (Inclusion, Diversity, Equity, and Access), do HEIs with different structural characteristics, fare differently across the US? Has there been a change over time in the features affecting enrollment decisions?
3) Are new policies adopted by HEIs, such as standardized-tests-blind admission policies, having an effect on students' enrollment?

Data

One of the primary difficulties in working with education data is that most of the data that is openly accessible is observational in nature. In addition, when, and if, any randomized control trials are conducted, sample sizes are relatively small and usually not publicly available. Furthermore, education data acquires greater meaning when analyzed over a certain period of time, a certain number of years. In fact, this type of analysis allows to uncover trends or changes over time that would otherwise not be immediate.

Recent research studies have suggested that using auxiliary observational data (sometimes referred to as remnant data, ie data that did not get used in a randomized control trial) in conjunction with machine learning techniques, can successfully enhance/improve the conclusions reached through a randomized trial, without negatively affecting bias, precision, or introducing additional assumptions. (Gagnon-Bartsch et al. 2021) This result, as pointed out by the researchers themselves, has tremendous potential for research in education where the characteristic of having access to large observational datasets is the norm, and not the exception.

As part of a purely academic exercise then, I have decided to use publicly available observational data to try to identify the elements that influence undergraduate student enrollment the most and to see if there have been any changes in these elements over time, in parallel to policy changes.

The data used for this project is publicly available structured data from IPDES (The Integrated Postsecondary Education Data System) which is a centralized data repository made available through the National Center for Education Statistics (NCES) (https://nces.ed.gov/ipeds/ and https://nces.ed.gov/ipeds/use-the-data), which is collected via direct reporting by the HEIs themselves, sometimes begrudgingly.

HEIs report to IPDES by completing mandatory surveys. These reporting requirements are the direct result of the HEIs participation in Federally funded programs authorized via the Higher Education Act of 1965 (Title IV), the Civil Rights' Act of 1964 (Title VI), and/or the Education Amendments of 1972 (Title IX), and later amendments. When not compliant, HEIs can be fined and suffer further consequences.

The existence of this data is particularly interesting, given the fact that it has been collected over time and thus can given us an instrument to highlight the changes occurring in HEIs enrollments over time as well as the changes, if present, on the elements that seem to influence student enrollment decisions.

It is important to highlight that while it is compulsory for HEIs to report on a number of features back to IPDES, the data reported is often fraught with errors, omissions, or inaccuracies that hinder an accurate analysis. Researchers part of IPDES are unable to follow through to verify the accuracy of the data reported in many cases, given the sheer number of institutions that exist throughout the US.

This being the case, a great deal of effort has been put into preparing the data and getting it ready for its descriptive analysis as well as its use in targeted machine learning algorithms in an attempt to respond to the research questions.

Project Structure & Approach

The approach embraced in analyzing the data was to first focus on a single year of data (2020) and then use the strategies applied to the 2020 data to extend the analysis across the nine years of data collected.

The primary reason for this approach was to test, as is customary with Machine Learning Projects, the reasoning adopted on a smaller data set and only once determined that the approach offers interpretable conclusions, apply the analysis to a wider data set taking into consideration the necessary characteristics which usually require adaptations. In this specific case, given the fact that the extended dataset is a time series, the extension of the analysis was adjusted accordingly.

The phases the project can be divided into are the following:

1) Data collection, cleaning and merging

2) Exploratory Analysis on both the 2020 data and then data from 2012-2020

3) Application of Machine Learning algorithms to the 2020 data and then the 2012-2020 data and drawing of relevant conclusions

Phase I - Data collection, cleaning, and merging

The IPDES data sets available are relatively broad and provide a lot of different information regarding HEIs and their student admission and enrollment characteristics.

The HEIs data is reported on a yearly basis and maintained in publicly available files that are subdivided into different groups based on the data they contain. Throughout the years, the subdivision of the information is occasionally modified, which poses some challenges.

Every file is accompanied by a relatively up to date dictionary file that explains the data collected and provides supporting information which is essential for further analysis.

IPDES's mandatory survey is a yearly exercise that is completed in the fall of each academic year and collects/refers to data from the previous academic year. As mentioned, the files collected and analyzed referred to the years from 2012 to 2020 included.

It is important to note that, given the fact that the data collected refers to the previous academic year, data for 2020 was not affected by the pandemic which hit the country in early 2020, since it presented HEIs' situation as of the year 2019.

The dictionary and data files that were collected and a brief description of their information follows:

- *Institutional Characteristics* – including directory information for each HEI subject to the reporting requirements subdivided into about 123 different variables (files starting with 'ic')
- *Institutional Characteristics* – including information pertaining to the educational offerings, organization, admissions, services and athletic associations subdivided into about 73 different variables ('hd' files)
- *12-Month- Enrollment* – including data pertaining to the 12-month unduplicated headcount by race/ethnicity, gender and level of student subdivided into about 74 different variables ('effy' files)
- *Admissions and test Scores* – including data pertaining to the admission considerations, applications, enrollees, and test scores reported subdivided into about 68 different variables ('adm' files)

Please see the Appendix for an indicative list of all variables included in each file based on the data dictionary for the year 2020.

To make the analysis as reliable as possible, only data reported directly by the HEIs to IPDES was used. In other words, any values that were recorded by IPDES researchers were not included in the analysis. Furthermore, missing values and not reported information was also eliminated from the data set.

The preparation of the data process started off by selecting only data for HEIs that were open and fully functioning in 2019 (based on the institutions' filings in the Fall of 2020). All other data files were adjusted based on the initial short list of HEIs still active in 2020.

After several rounds of data cleaning, merging, and reconciliation, the final variables-list was composed of a total of 54 variables ranging from Institution ID (unique identified for each institution) to the total student enrollment per year (see appendix for an exhaustive list of the features selected).

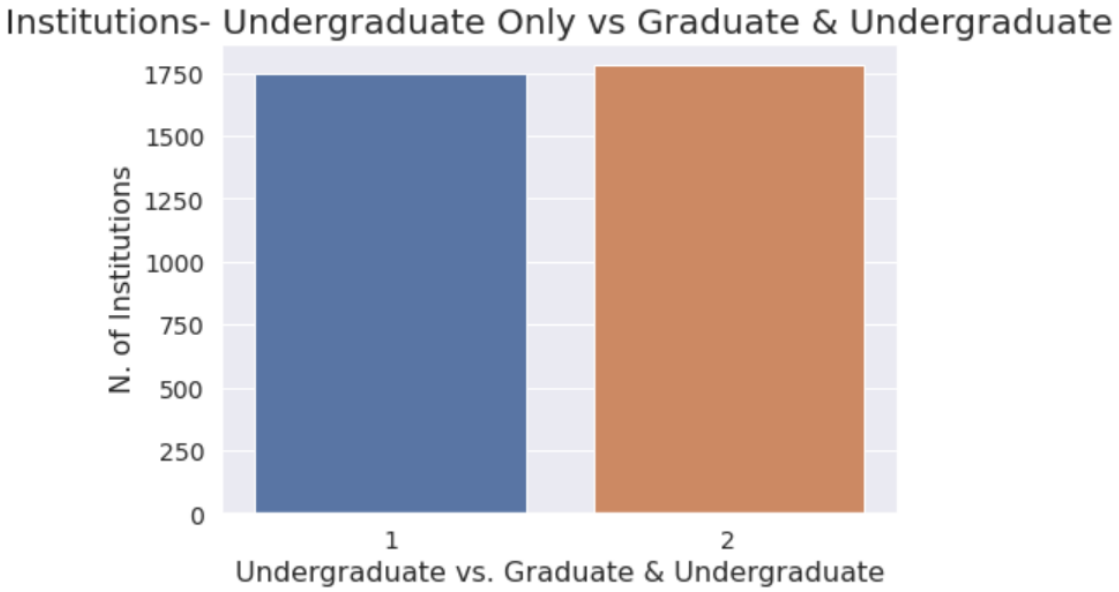
It is important to note that the analysis is focused on only undergraduate student enrollment in institutions of higher education that were still operational in 2020 and that offered undergraduate programs. While the data spans from 2012 to 2020, any institutions that were closed or otherwise suspended operations during this time and were not operational in 2019 were excluded from our analysis. However, institutions that started operations after 2012 and were still operational in 2020 were still part of our analysis. Furthermore, institutions offering exclusively graduate level programs were excluded from our work, however, institutions offering both undergraduate and graduate programs at the same time were not excluded.

This preliminary work led to a “short-list” of a total of about 3500 HEIs distributed over 59 United States and Territories, and an initial short list of features that added up to a total of 65 variables (when including the data relative to student enrollment disaggregated by race) and only 54 variables without taking into account this disaggregation.

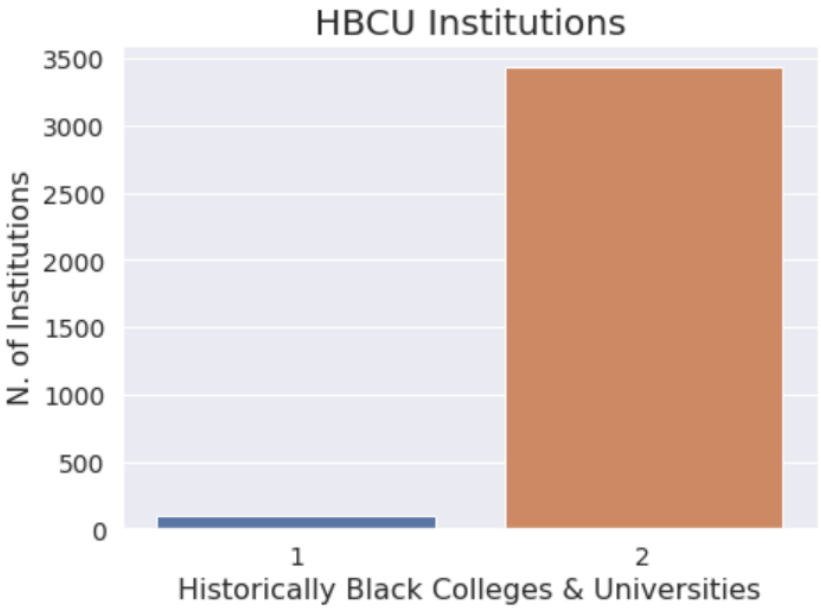
Phase II - Exploratory Analysis on both the 2020 data and then data from 2012-2020

Preliminary data analysis included both a focus on 2020 data and the data pertaining to the whole period from 2012 to 2020.

One of the primary characteristics of this data that must be taken into consideration is that it is highly unbalanced. Practically all features, except for the grouping of the short-listed HEIs by program offering, are highly unbalanced. Only the number of institutions offering both undergraduate and graduate degrees are roughly equal to those offering exclusively undergraduate degrees, at least for 2020 data.

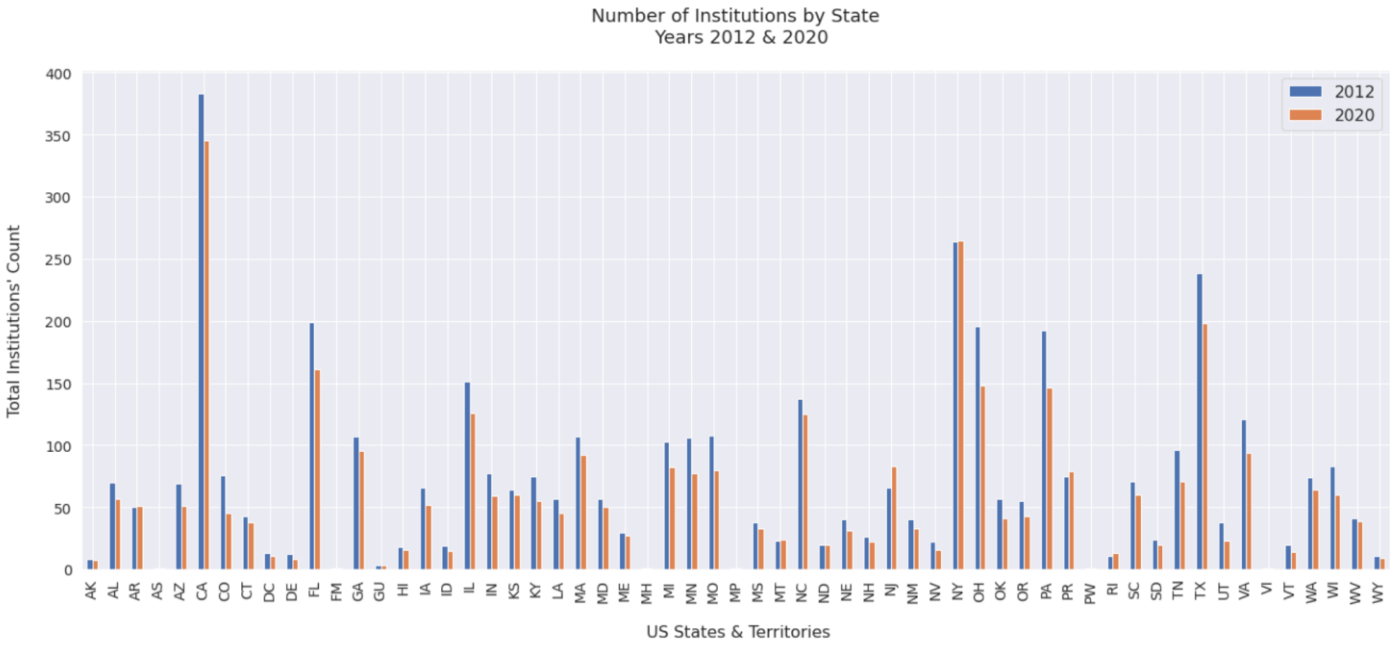


Taking a look at the data from the perspective of Higher Education Institutions first, we can elicit some interesting trends from the data overall. First of all, and possibly the most unbalanced of all of the features drawn from the data is the number of Undergraduate level HEIs classified as HBCU (Historically Black Colleges and Universities) that are present across the USA. These institutions are a significant minority of the overall HEIs in our set – not so surprising. (2020 Data)

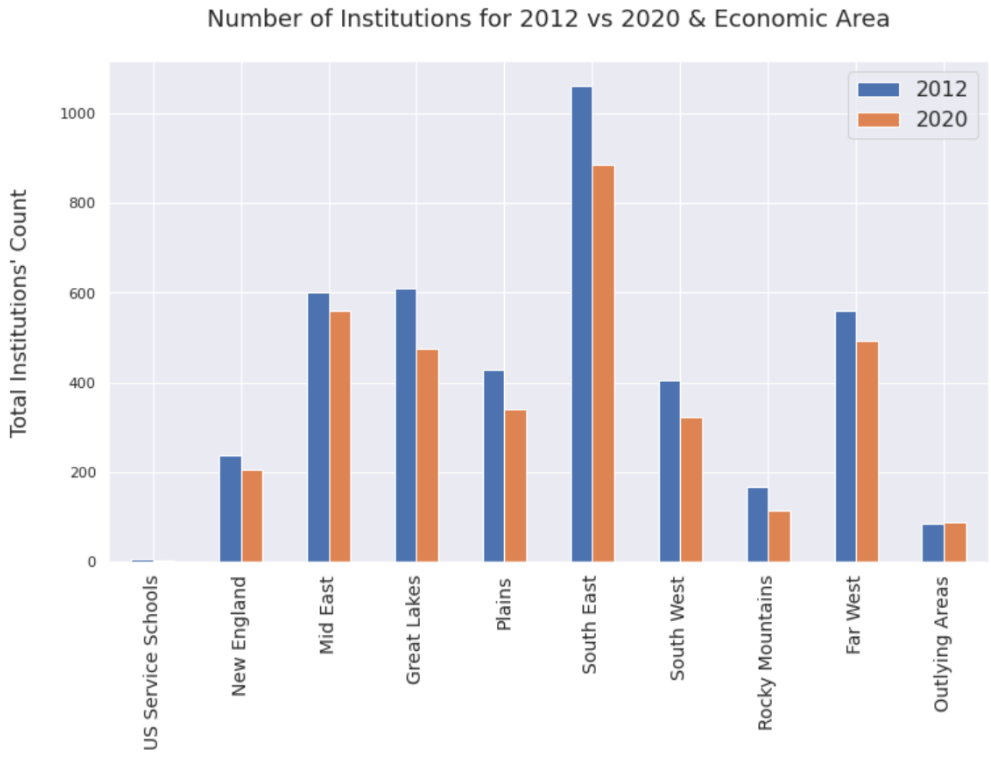


The presence of a reduced number of HBCU institutions across the US makes recent initiatives aimed at directly supporting funding for these organizations the more interesting, especially when looking at the general state of federal funding towards HEIs across the US, which many sources denounce as being in dire condition and experiencing contracting levels in recent years. (Mitchell et al, 2019, Marcus 2019, Oliff 2019)

Taking a look at the distribution of the institutions by State, it is immediately apparent that there is quite a significant imbalance in the distribution of HEIs across the US and its territories. This is in part due to the difference in size and populations of each State, with the notoriously larger States, in fact, confirming a wider count of HEIs (California, Texas, Florida)



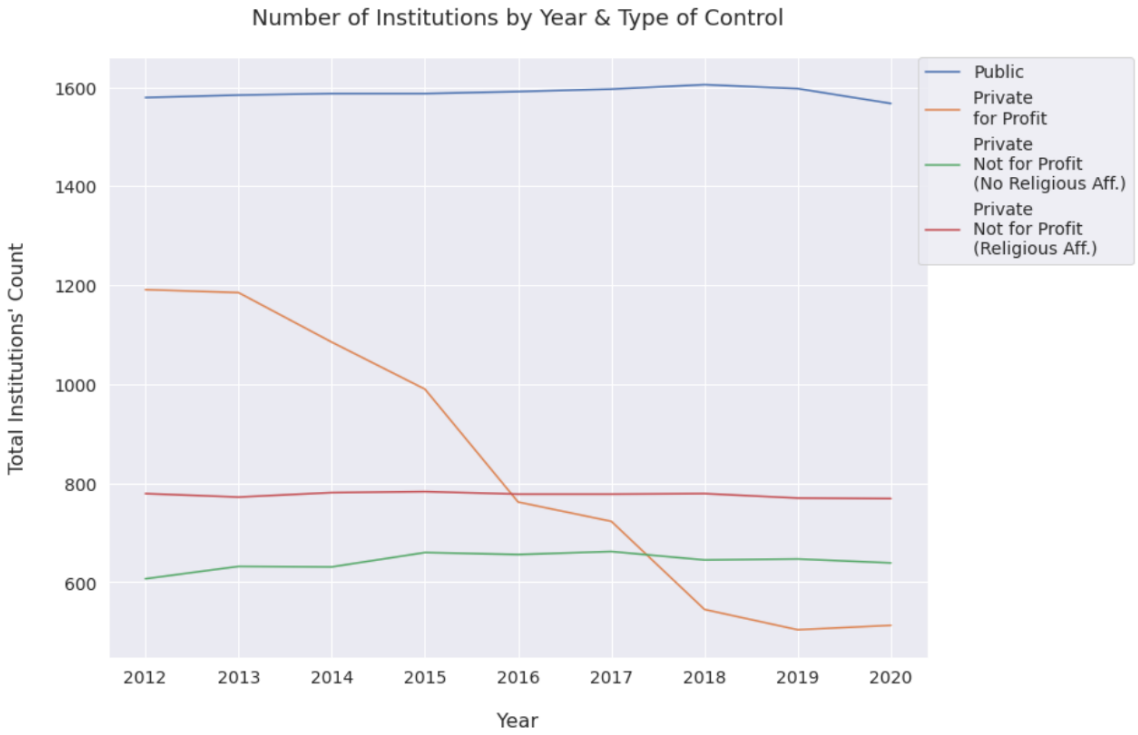
Interestingly, however, over the period of interest, in most States there has been a contraction in the total number of institutions operating between 2012 and 2020, at least in terms of those required to report back to the Federal Government due to financial aid they receive or their direct participation in federal level programs.



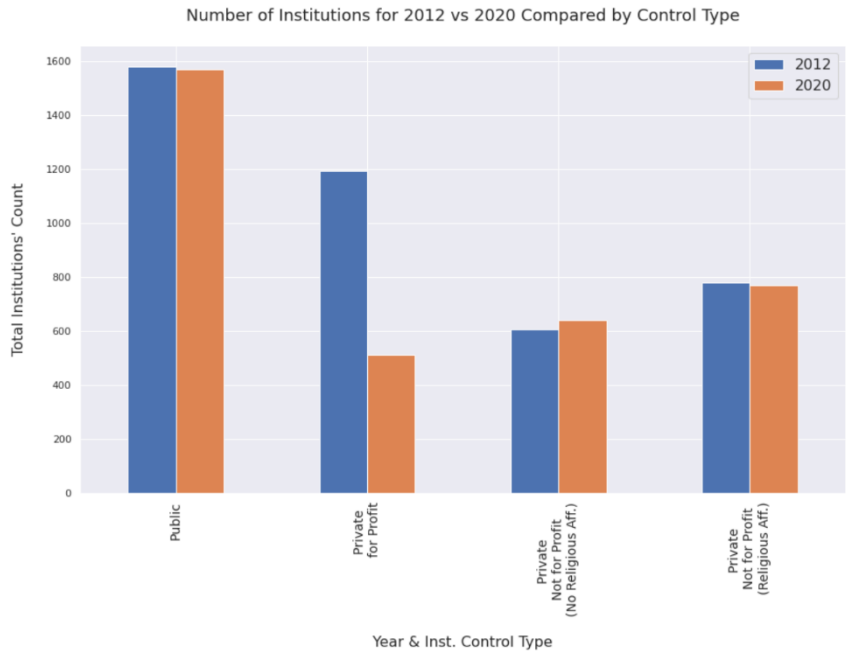
This contraction is consistent across all major economic areas, as defined and identified by the Bureau of Economic Analysis, including the middle eastern portion of the US, where Maryland is located. The most significant contraction occurred in the South East Region. Interestingly, only non-continental US Territories seem to have maintained constant in terms of count over the period.

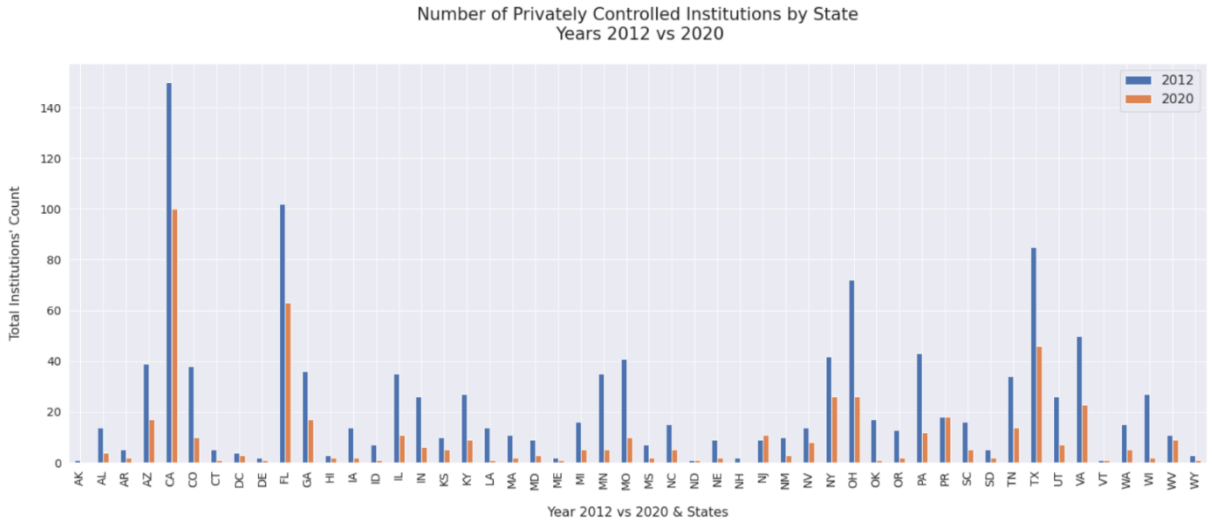
When we analyze what type of control the HEIs that have been experiencing a reduction over the period have had we can see that the greatest contraction has been experienced in the segment of HEIs that are categorized as private for-profit institutions. This may be an indication that the "business of education" is becoming a less profitable one and thus attracting reduced investments. When reading this information in parallel to the stated reduction in financial support in general towards HEIs, then maybe the situation may not be as dire as suggested by some.

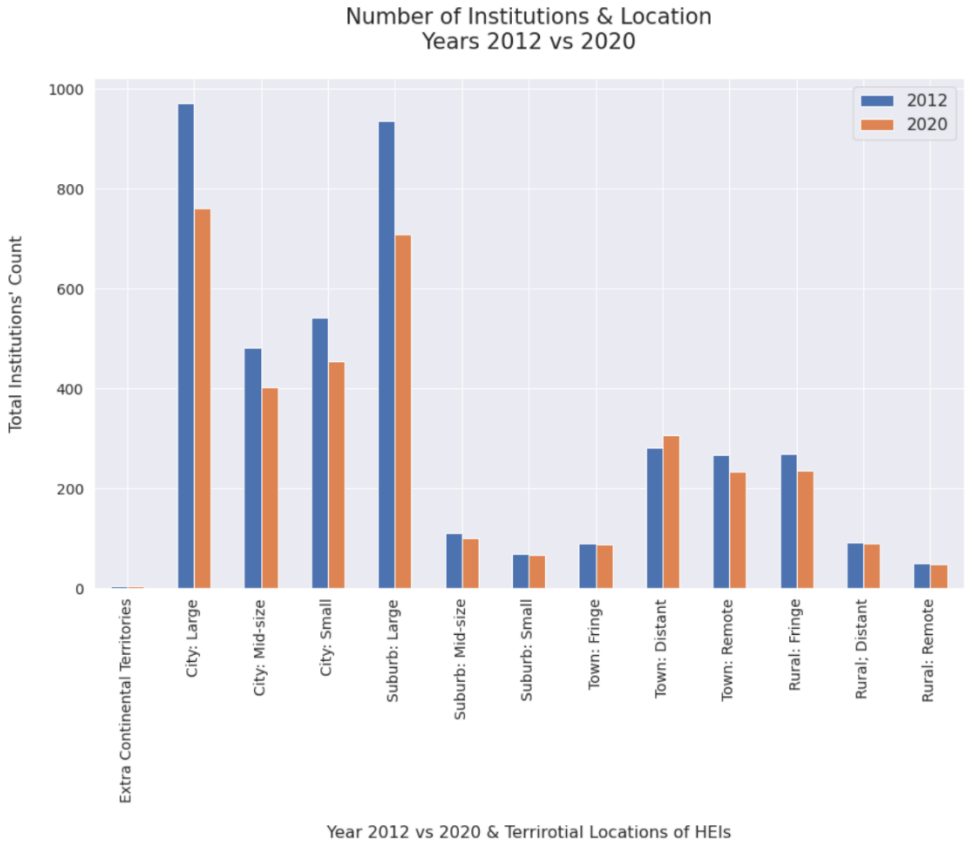
This contraction in total number of privately controlled HEIs is not a relatively new occurrence, on the contrary. When taking a look at the trends over the whole period in fact it becomes apparent that this contraction has exacerbated recently but has started in 2013.



When taking a look at the distribution of Privately Controlled HEIs by State and how these reduced in number since 2012. it is interesting to note how the contractions have occurred across practically all States, without exceptions other than New Jersey and Puerto Rico.



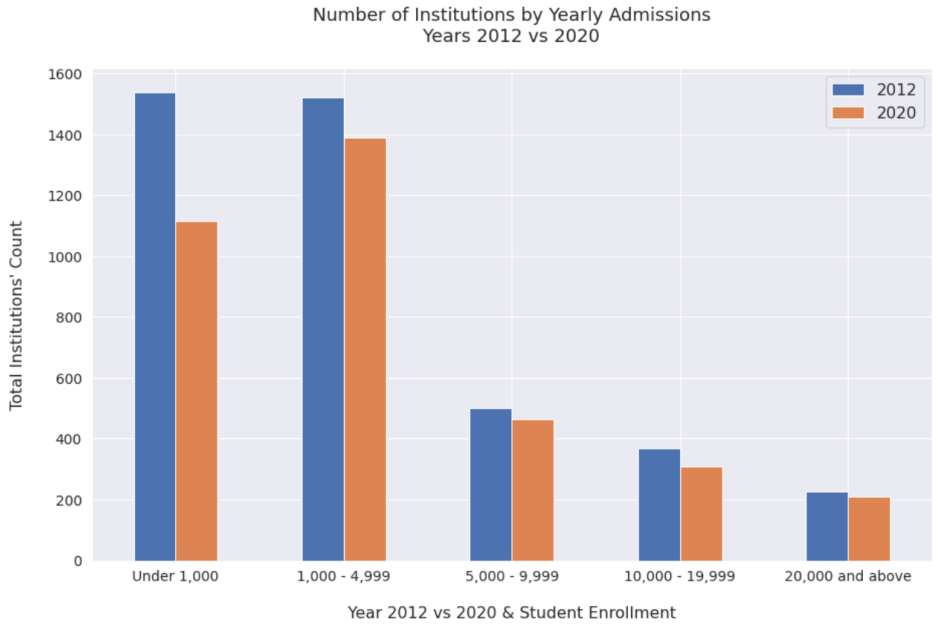




Taking a look at the distribution of institutions according to the type of area they are located it, generally referred to as “Locale’, it is not surprising to see that most of them are located in Urban area settings and that this characteristic has not really changed when we compare 2012 data with 2020 data, as we can see below.

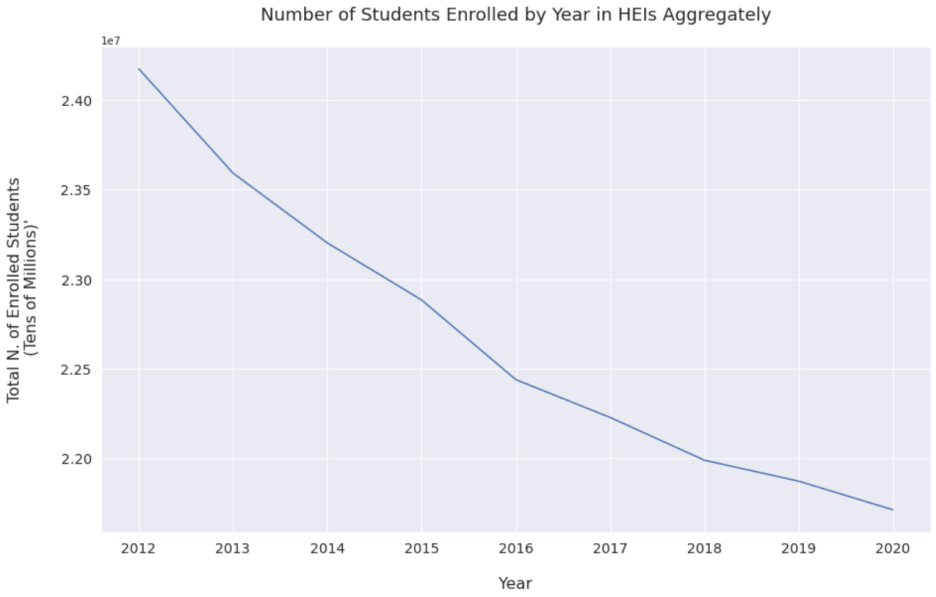
Furthermore, the graph seems to imply that the greatest number of institutions lost over the period were located in urban settings. If we read this information in parallel to the fact that most of these HEIs that have closed over the period were also for profit and smaller in size (as we can see below in our next graph), then it is really surprising to note that they were also located in urban areas, since they were most likely organizations set up with a broader objective than education alone.

When taking a look at the size of the institutions that have experienced the highest contraction in numbers, it is not very surprising to see that mostly smaller institutions enrolling a maximum of 1,000 students per year and those enrolling between 1,000 and 5,000 are those that have suffered the most between 2012 and 2020.

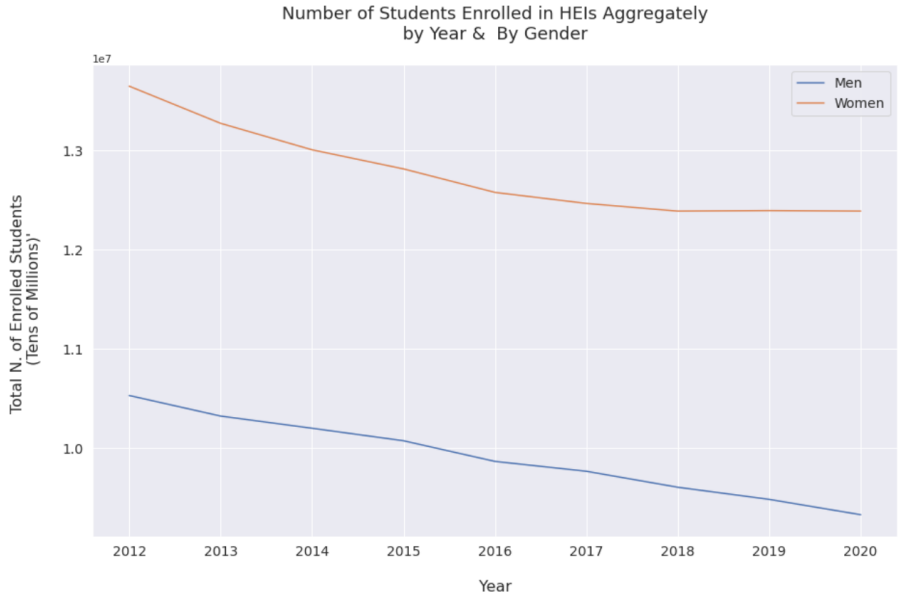


This graph also gives an idea of how much more numerous are the HEIs that are smaller in size (admitting up to 5000 Students per year) but there are a number of institutions roughly 200 per year, that are clearly exceptional in terms of size and accommodate more than 20,000 students per year.

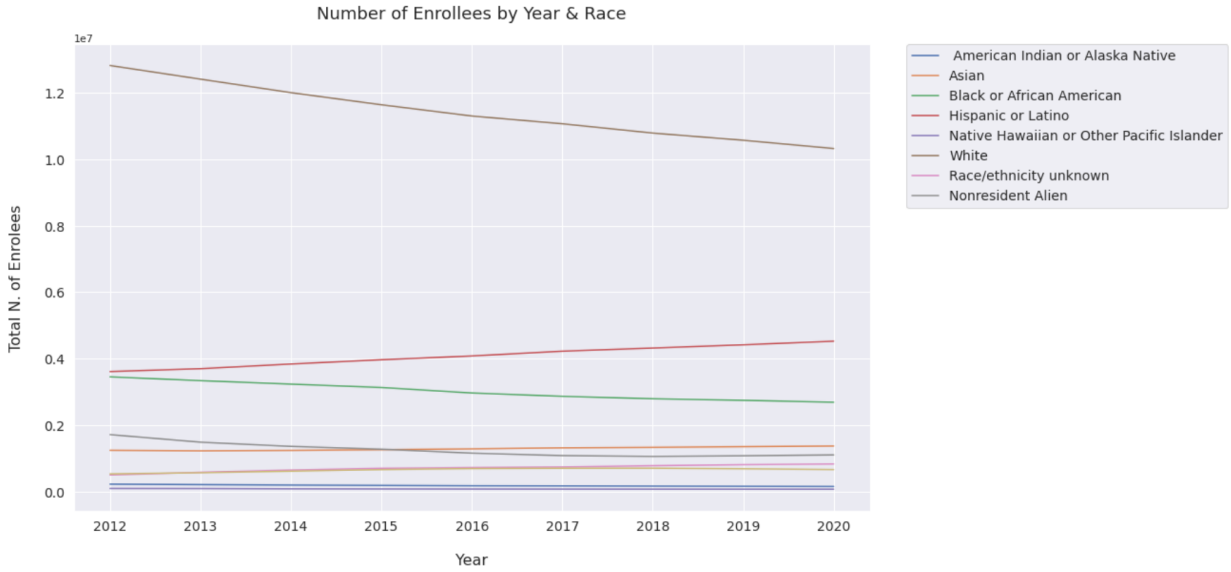
Taking a look at the institutions from the perspective of student enrollment levels, there definitely seems to be a consistent contraction in total number of enrollments as often decried by HEIs and public officials. As the graph below shows, the contraction has been of about 20% overall over the period between 2012-2020



Furthermore, when taking a look at the changes in composition of student enrollments by gender, it is interesting to note that the most significant contraction has affected Male students – which can in part be fueling the general sense of unease that sociologists have been denouncing for men across our society over recent years.



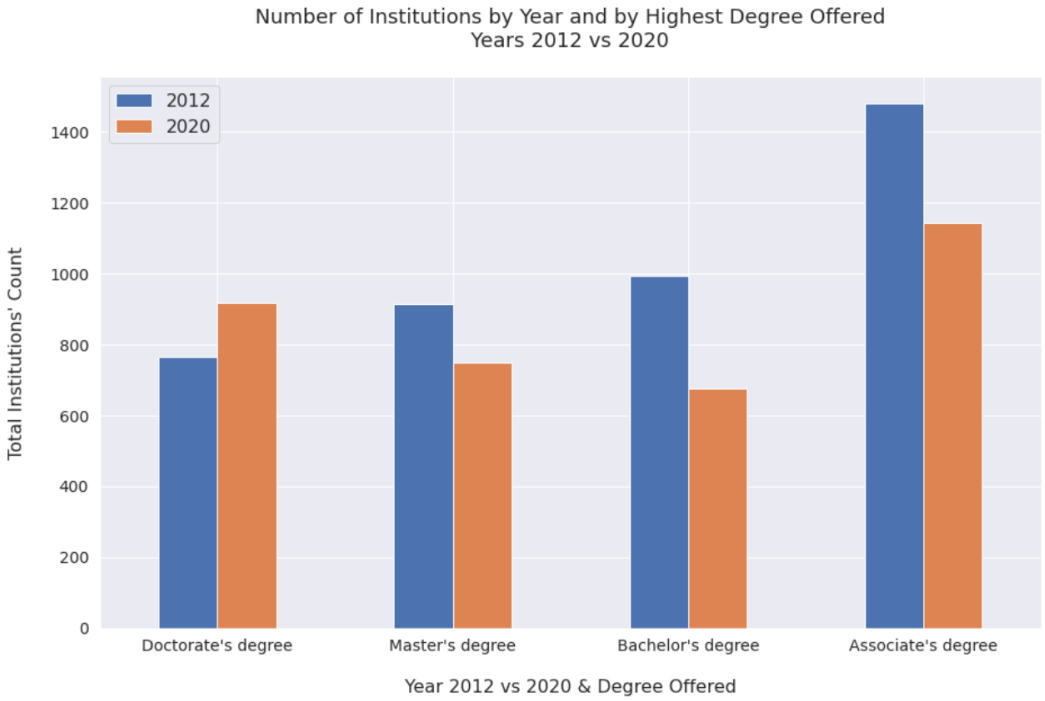
Looking at the same information by race, again, the data confirms the greatest reduction in number of students enrolling are those identifying as White. It is worth noting that this contraction is not counterbalanced by increases in enrollments by students identifying by other races (as confirmed by the graphs above). At the same time, from the perspective of attempting to reduce the gap in access to HEIs by the Latino and Hispanic population, the numbers seem to indicate that this ethnic group did in fact make improvements over these past years in parallel to the changes in policy aimed at increasing the educational opportunities available to minority students put in motion by the Obama administration through their Department of Education budget priorities since 2011.



Enrollees who identify as African American or Black seem to have also contracted to a certain degree as have the Native Hawaiian or Pacific Islanders, but to a much smaller extent than White students.

Worthy of note is the fact that Race/Ethnic group categorized as Unknown and students who identify as Non-Resident Aliens have had a fairly flat enrollment throughout the period. This seems a little peculiar and while investigating these numbers is not part of my research purpose, I think there may be some accounting error or under-counting error taking place here. Finally looking at the numbers of enrollees that identify as Asian we can see that their ethnic group has slightly increased in number overall.

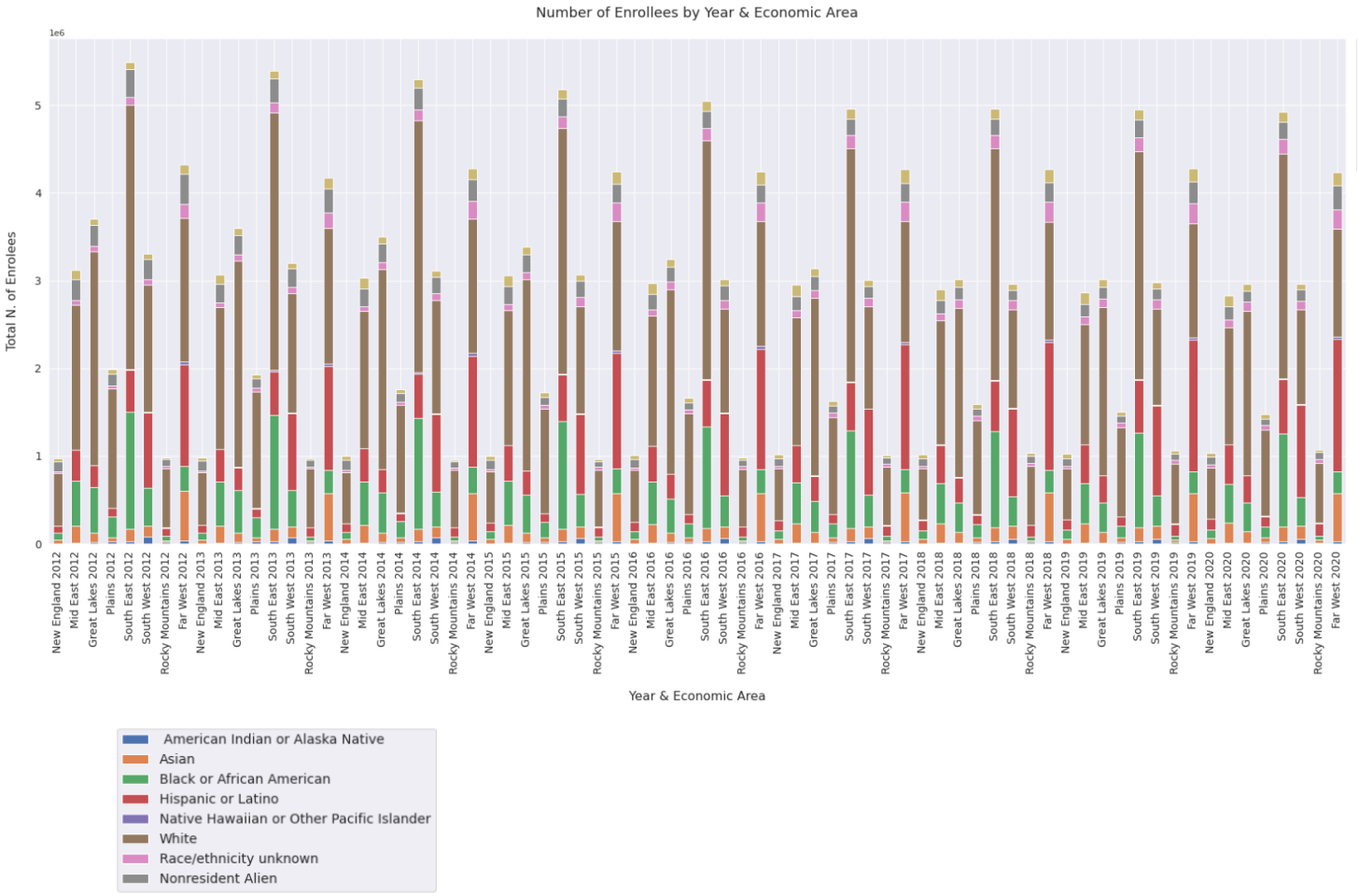
Another point that is interesting to underline is the fact that over the period of reference, the number of institutions offering Doctorate-level degrees actually increased, in spite of decreasing enrollment of undergraduate students overall. This may pose additional challenges for HEIs in the near future.



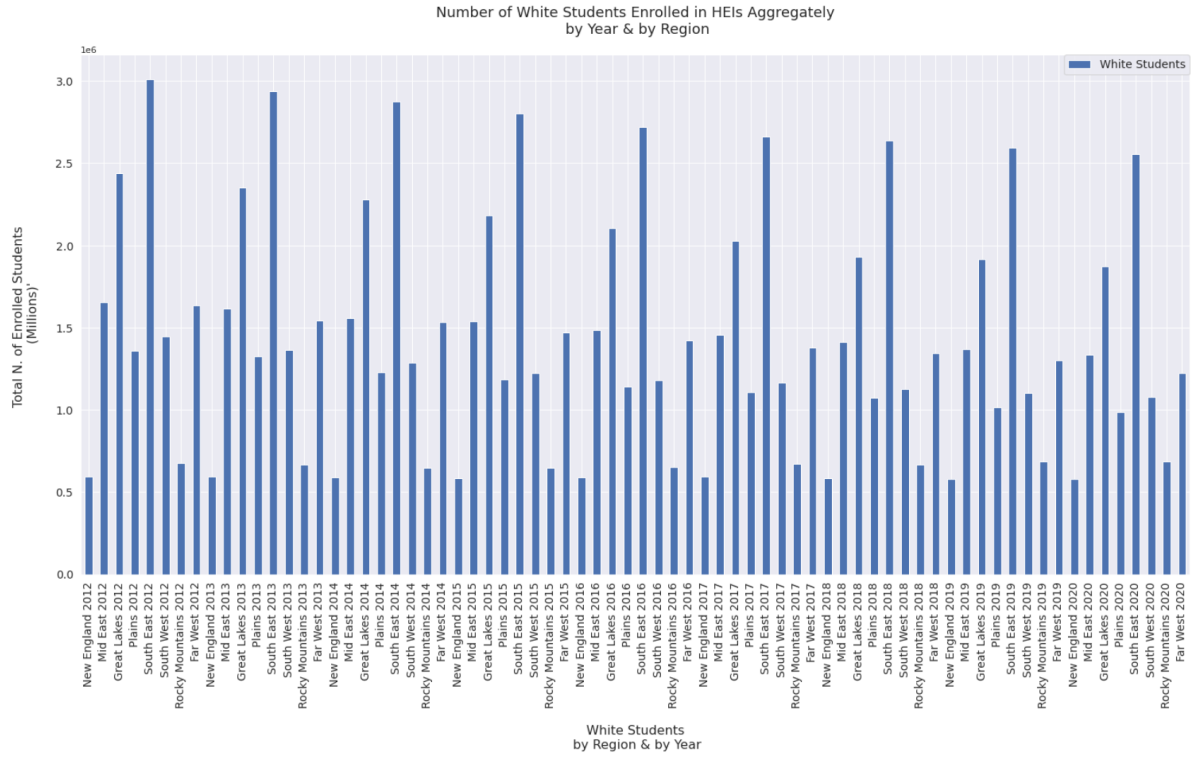
Total number of admissions and total enrollment numbers in HEIs are obviously two highly correlated variables, as are total number of applications. All three features are actually highly correlated among each other, or more accurately one depends on the other. (see appendix for relative graphs)

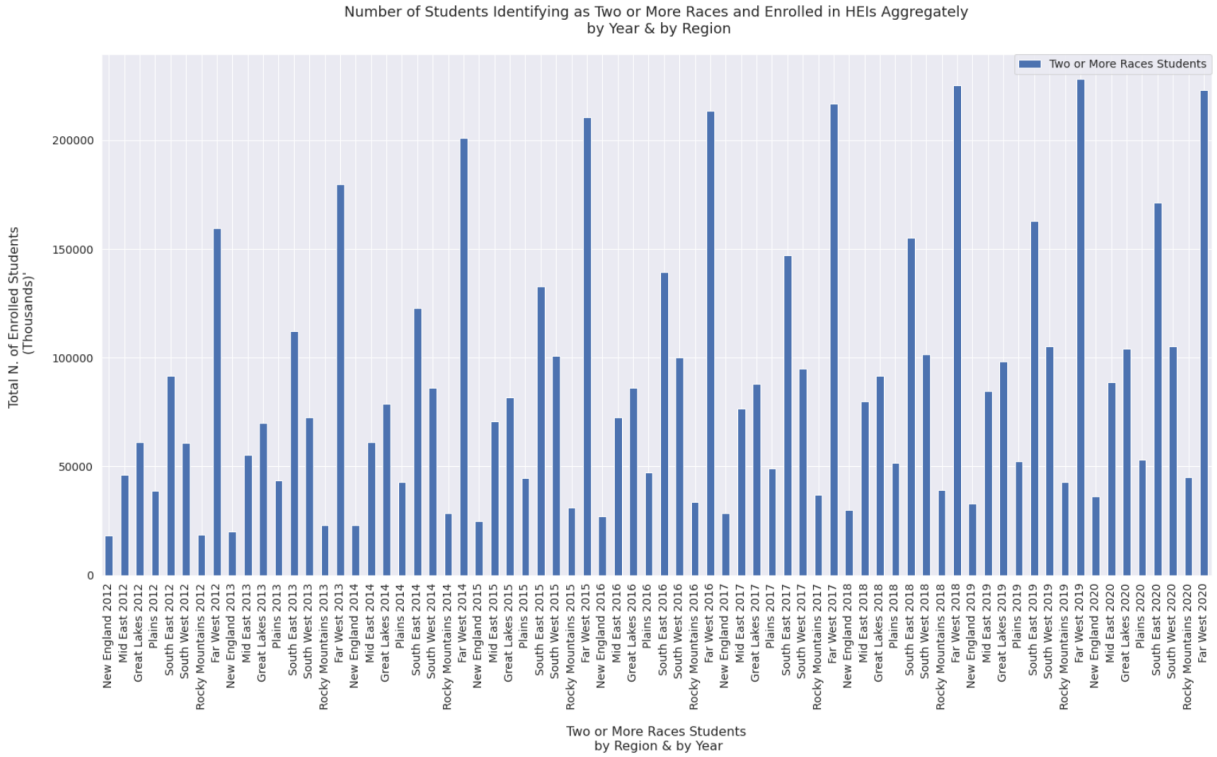
Among these three feature, I chose to continue my research focusing on the feature total number of Enrolled students. This choice is based on the fact that ultimately, from the perspective of HEIs, total enrollment is the primary feature of concern, given that it is this number that affects HEIs bottom line and is the number that HEIs are likely interested in maximizing while maintaining their prestige intact. Therefore, I chose to move forward with my work, attempting to answer whether any of the other publicly disclosed features affected total enrollment numbers. Technically we may argue that total number of applications could be a better variable to indicate the effects of certain policy choices such as asking students to submit SAT scores, but, as explained, given the high dependency between the three variables and the direct link of total enrollment to HEIs bottom line, I selected the latter.

Before proceeding further, let us take a look at the Enrollment levels across Economic Area and years. Aggregately the South Eastern region seems to have experienced a contraction in aggregate levels of enrollment, as has the Great Lakes region,



As we can see below, while White Students enrollment has decreased almost consistently across the regions across the years, students identifying as belonging to Two or More Races have actually increased, reflecting a change in the demographics of our society





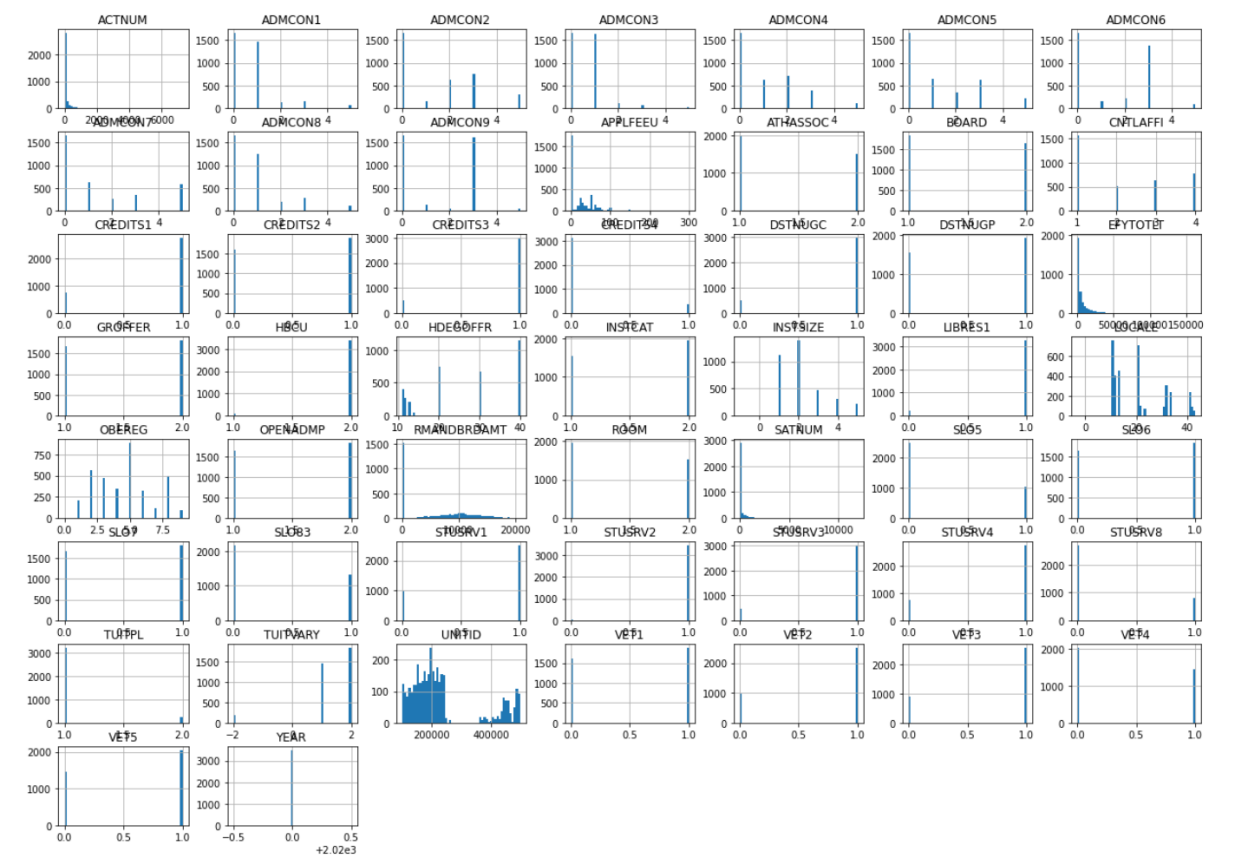
These characteristics pose a challenge to our analysis.

Phase III - Application of Machine Learning algorithms to the 2020 data and then the 2012-2020 data and drawing of relevant conclusions

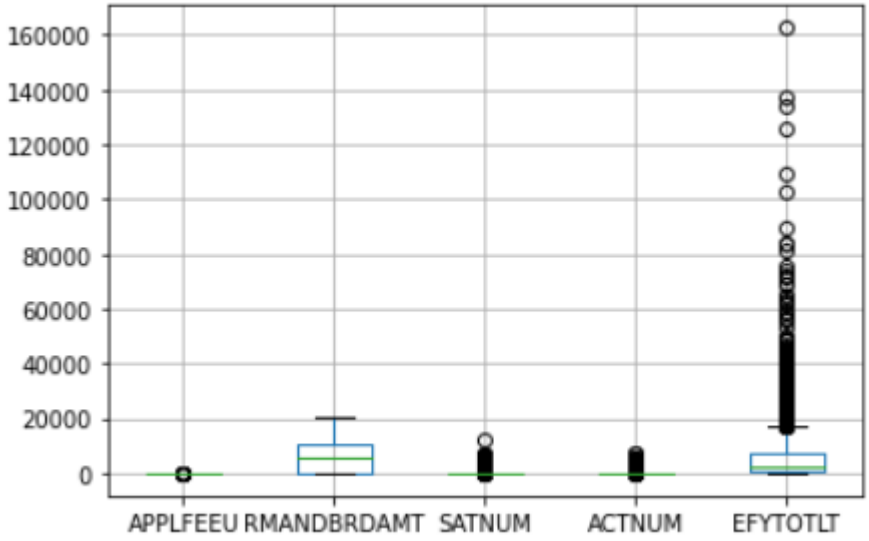
Following this initial descriptive analysis of the data, I decided to proceed by separating enrollment data by race outside of our overall data and to proceed without making significant additional changes to my data set to limit adding any bias to my analysis.

I thus proceeded by selecting exclusively data pertaining to the year 2020 to use as a benchmark for my machine learning (ML) approach.

First and foremost, I took a look at the data's distribution – from which I was able to identify some additional issues that needed to be addressed and I confirmed the lack of balance that data presented overall across all features.



Pulling aside quantitative variables, it soon became apparent that the data needed to be re-scaled



Furthermore, all of the non-quantitative variables were discreet and categorical in nature, which meant that they required to be handled so that our algorithms would interpret them correctly. All of the features, in other words, required preprocessing. However, before preprocessing the data, I chose to split the data in training and testing datasets, which complicated the handling of the algorithms, in order not to influence the Training dataset with information from our Testing dataset, therefore allowing us to build a more valuable model for prediction purposes.

There is debate among experts as to whether this is the proper approach to hold – meaning whether it is necessary to pre-process the data after splitting it, especially considering the fact that the data was categorical in nature. However, this is the decision I made.

The numerical data's mean and standard deviations were very spread apart

APPLFEEU mean: 22.94 standard dev: 28.87

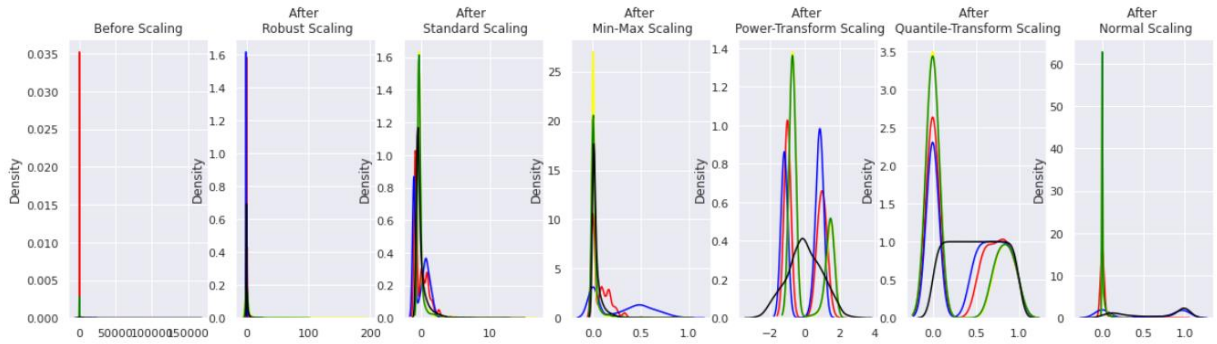
RMANDBRDAMT mean: 5770.26 standard dev: 5779.17

SATNUM mean: 213.82 standard dev: 714.67

ACTNUM mean: 171.44 standard dev: 563.23

EFYTOTLT mean: 6225.02 standard dev: 10659.58

I decided to check the results of different preprocessing procedures to handle the quantitative data and show the various results from the various methods. The scaling methods tested were: RobustScaler, StandardScaler, MinMaxScaler, PowerTransformer, QuantileTransformer. As visible in the image below, the best scaling results for our quantitative data were achieved with Robust Scaling which normalized the data, therefore this method was chosen.



it seems clear that using Robust scaling for the Numerical columns in our df is the best approach. This approach to pre-processing the data takes into consideration the presence of outliers, which our data has many of.

Taking into consideration that the data includes several categorical variables coded with numerical values, the next step was to choose the appropriate preprocessing tool for this type of variable in order to make sure that the algorithms did not assign different quality values to the data, further affecting their conclusions. Thus, One-Hot-Encoding was applied to the categorical values.

Once the data was split and preprocessed, a straightforward linear regression was applied to the 2020 data as a test case set. The performance of the straightforward regression on our test data is not very good. This is not surprising as I have not performed any data manipulation or feature selection to better calibrate our model

Note that the performance is not excellent on neither Testing nor Training data. Our model in fact reaches only about 67% accuracy on Training data and only about a 60% accuracy on Testing data, which, on the data taken as is, is still not bad at all.

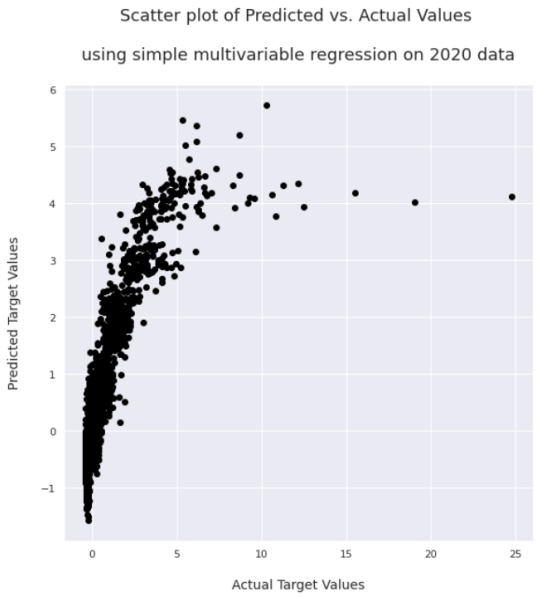
Although the accuracy is relatively low, the measure of the Mean Squared Error is not very high all things considered. This would seem to indicate that the model performs fairly and that the error (the residuals) between actual enrollment and predicted enrollment for our HEIs for the year 2020 are relatively small. However, there are contradicting indicators emerging which raise a red flag. In fact, looking at the explained variance from the model I note that this is actually not too high which also confirms that the model is not performing so well, since higher explained variance would imply a stronger association between the variables and the target variable thus ultimately better predictions.

The Explained Variance score (also known as the Coefficient of Determination or R Squadre) is about 60% (perfect predictions would imply an R Squared of 1 or 100%). Since the R-Squared represents the fraction of response variance captured by the model (a 'standardized' Mean Squared Error), the higher it is the better the model.

Taking a look at the relationship between predicted and actual target values for our training dataset, we can readily see that the model is not performing very well and that there is room for improvement. This was confirmed by our Accuracy scores, and by our Coefficient of

Determination (R Squared) but we can take a look at the residuals' Scatter plot and reach the same conclusion.

The scatter plot of actual target values vs predicted shows a clear trend, instead of being randomly distributed, confirming that our model is working, although it is not excellent in its predictions. Looking at the scatter plot comparing Actual Target values with the Predicted ones, we can see that Predicted values are predicted to be lower than they should be (they mostly lie below the 45 degree line – which would indicate perfect fit of the model).



There are a number of reasons affecting our predictions. First of all, the presence of outliers is affecting our model in spite of the pre-processing I applied on the data. Furthermore, I simply took the data-set and in order not to influence results based on additional assumptions, I simply used the data I had as it was, without eliminating any variables a priori.

Given these results, the next step I took in my analysis was to see if adopting a penalized Linear Regression model such as Ridge would actually lead to better predictions. The selection of the Ridge or L2 model was justified by the fact that the data has outliers, and this model tends to perform better under these circumstances, since it tends to avoid overfitting of the data

After applying the L2 regression on the data I was surprised by the fact that there was no real improvement on the model.

A first approach in trying to improve performance was to adopt the SVR Support Vector Regression algorithm to identify the most significant features affecting the data and thus find the best hyperplane explaining. However, this did not lead to great results. Thus I decided to use the LazyPredict Library to take a look at the performance of various Linear Regression Model options that are open to us in general to see which would perform the best with our 2020 standardized data.

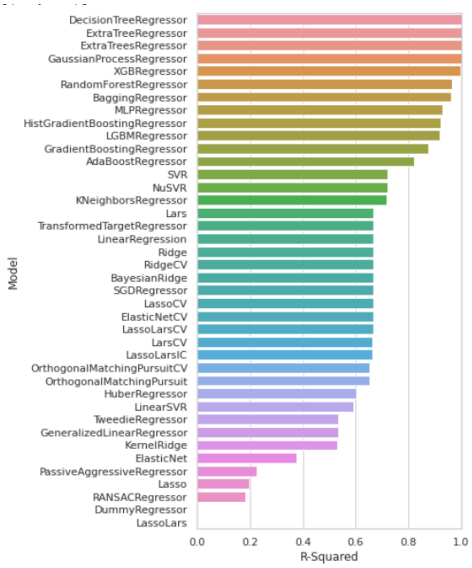
Applying Support Vector Regression (SVR) as a simple feature selection process to the data, the first ten features that emerged as being the most significant on 2020 data in predicting enrollment are listed below. However, given some apparent inconsistencies, I chose to abandon this approach and opted instead towards running LazyPredict on the data to identify the best model.

First ten SVR selected features:

1) 'HDEGOFFR', The level of degree offered by the Academic-Oriented HEI
2) 'INSTSIZE', The HEIs' size
3) 'CNTLAFFI', The type of control the HEIs is under
4) 'CREDITS2', The presence of a policy giving students credit for life experiences
5) 'SLO83', The presence of a teacher certification program approved by the State for initial certification or licensing for teachers
6) 'STUSRV4', the presence of Placement services for completers
7) 'TUITVARY', the adoption of a tuition plan that is diversified based on whether students are in-State or Out of State
8) 'BOARD', the presence of a meal plan
9) 'VET3', credit for military training
10) 'VET4' the presence of a recognized student veteran organization

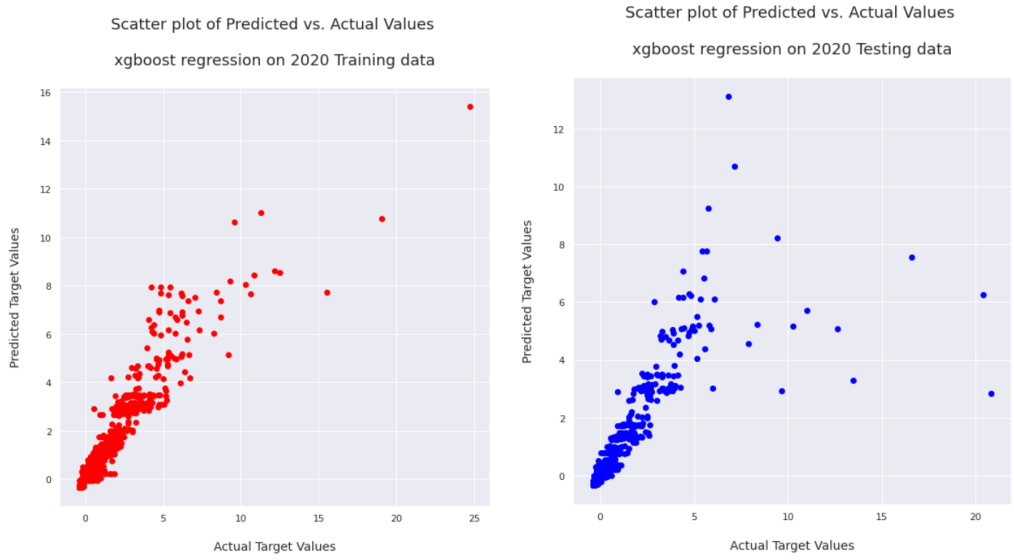
Note: Some of these features are not surprising and likely interrelated. For example, institutional size and the level of degree offered may be overlapping variables and their individual contribution to predictions may not be very clean.

LazyPredict is a library of algorithms that tests various models against our data and returns various indicators on the models' performance with our data. It is a short cut that can be used to try to get a sense of the models results before any additional data processing or fine tuning or decisions.



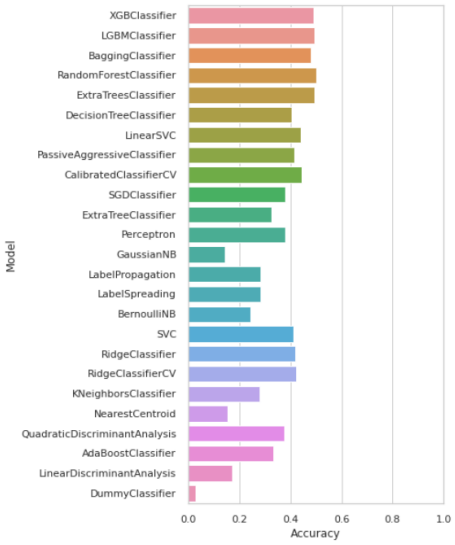
Based on the LazyPredict Library results, I should see significant improvements by utilizing a Random Tree selection model with my data. In particular, one of the best performances seems to be possible based on Xgboost which implements the gradient boosting trees concept, in other words, uses random tree selection and the gradient descent, to handle large datasets and improve predictions. I thus choose to apply this algorithm to the 2020 data . However, the improvement

occurs primarily on our Training data which achieves an accuracy level of approximately 90% - Our testing data set also improves in performance reaching an accuracy of about 68% with an overall improvement of about 13%.



Before proceeding with testing our data on classifier algorithms I choose the target variable to be the State of origin of the HEIs. – As mentioned before, this is an academic exercise thus choosing this variable was primarily done to ground our efforts and evaluate the models' performance.

Running LazyPredict on the 2020 Data to view various algorithms performance in terms of classifiers, yielded the following results:



As we can see there really wasn't a great improvement in the classificiation results even after I made sure that the features would be correctly interpreted as categorical in nature. In addition, I

want to note that the ROC AUC values are NONE in the table summarizing performance below because this is a multi-classification problem rather than a two-class classification problem and the LazyPredict library does not support it.

Unfortunately, the summary table does not include DBSCAN for us - in order to check on the performance of this classifier I will have to implement this algorithm on our 2020 data separately by hand.

Model	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
LGBMClassifier	0.51	0.40	None	0.49	10.29
XGBClassifier	0.49	0.38	None	0.48	41.11
ExtraTreesClassifier	0.50	0.38	None	0.48	1.20
RandomForestClassifier	0.51	0.37	None	0.48	0.87
LinearSVC	0.45	0.35	None	0.44	1.80
BaggingClassifier	0.48	0.35	None	0.46	1.18
DecisionTreeClassifier	0.41	0.33	None	0.42	0.36
CalibratedClassifierCV	0.47	0.31	None	0.42	13.16
Perceptron	0.37	0.31	None	0.36	0.62
PassiveAggressiveClassifier	0.42	0.31	None	0.37	0.91
SGDClassifier	0.38	0.30	None	0.35	1.16
GaussianNB	0.18	0.24	None	0.16	0.14
LabelPropagation	0.27	0.22	None	0.27	0.51
LabelSpreading	0.27	0.22	None	0.27	0.90
BernoulliNB	0.23	0.21	None	0.23	0.30
ExtraTreeClassifier	0.32	0.21	None	0.32	0.23
RidgeClassifierCV	0.43	0.20	None	0.32	0.21
RidgeClassifier	0.43	0.20	None	0.32	0.13
SVC	0.42	0.19	None	0.33	2.97
KNeighborsClassifier	0.28	0.19	None	0.27	0.54
LinearDiscriminantAnalysis	0.19	0.15	None	0.18	0.23
NearestCentroid	0.10	0.12	None	0.11	0.15
QuadraticDiscriminantAnalysis	0.34	0.12	None	0.25	0.28
AdaBoostClassifier	0.27	0.08	None	0.17	1.56
DummyClassifier	0.03	0.02	None	0.03	0.28

Running LazyClassify on the relabeled data does not improve our outcomes particularly. This may be due to the fact that our models are being applied in a multi-class classification problem, which they are not all set up to do.

Traditional classification problems that are set up to handle multi-class classification are (among others): naive_bayes.BernoulliNB; tree.DecisionTreeClassifier; tree.ExtraTreeClassifier; ensemble.ExtraTreesClassifier; naive_bayes.GaussianNB; neighbors.KNeighborsClassifier, and others) However, we can use certain techniques to handle multiclass classifiers in such a way that our models view them as a binary class problem. These strategies can help us use the evaluation functions already available in SkLearn to assess how well our model is performing. Furthermore, and especially, they allow us to use the algorithms correctly so that the outputs can be considered reliable. But first, let us keep in mind that multi-class means that the target variables can be predicted as belonging to one of many different output targets (for example one of the 50 States) while a multi-label classification problem would have the target variable pertaining to one class that could acquire different values (such as low, medium, high level of risk)

Our classification problem is Multi-class since we are trying to see if the data collected across all of our Institutions can in any way help us classify the State from which the Institution is from. Since our data is labelled, technically this is a classification problem, although a multiclass one. I will use both clustering (K-means, DBSCAN) and classification (Nearest Neighbors) techniques to see what the data tells us.

It must be noted that while K-means is considered a clustering algorithm it is a bit of a hybrid as it can work with both unlabeled and labelled data, ultimately creating clusters of uniform groups.

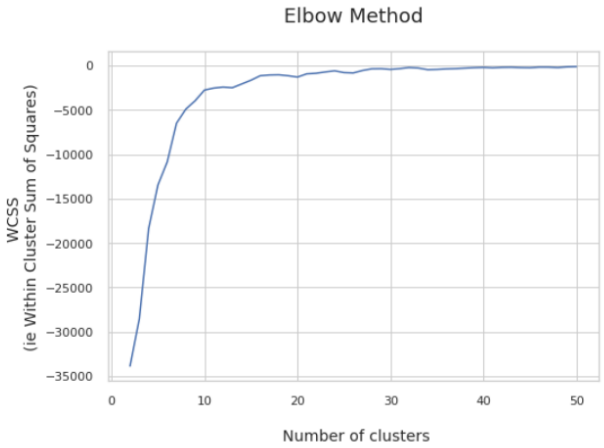
Let us start by running k-means to see if our 2020 data clusters by State, and if the clustering reflects the Bureau of Economic Analysis groupings. Using a k value of 50 leads to a pretty low accuracy level (a little over 1%). This seems like a reasonable result as the institutions' characteristics are not specific to the State the institution is located in.

Changing the number of clusters to 9 (based on the economic areas the US is normally subdivided into), improves the level of overall accuracy practically four-fold (accuracy 3.95%), yet the levels are still too low to be significant. While this may be disappointing it still is not surprising since the features we are looking at are not characteristically tied to the particular location where an institution resides. Many of the features chosen in fact reflect Federal level programs that are adopted across the board by many institutions across the US.

Repeating the clustering exercise for a number of clusters between 2 and 50 the total number of States we see the range of possible results per State (see below)

The graph shows the plot of the gradient of the inertia as a function of the number of clusters. The inertia is the intra-cluster average distance between the data and the centroid of each cluster. The objective when creating clusters is to minimize such distance (because this leads to clusters with high density well separated from each other).

The values of the gradient are negative, and this indicates that the function is decreasing. As the number of clusters increases the gradient values tend to zero, which means that the slope tends to zero, which means that we reached a minimum for the original function.



If we look closely at the graph we see a first change in slope of the graph around n =3 and then a more significant change in slope at n =9

The Homogeneity score confirms that the clusters are NOT very homogenous (the score ranges from 0 to 1) with low values indicating low homogeneity.

Homogeneity score for 3 number of clusters is: 0.0142
Completeness score for 3 number of clusters is: 0.1428

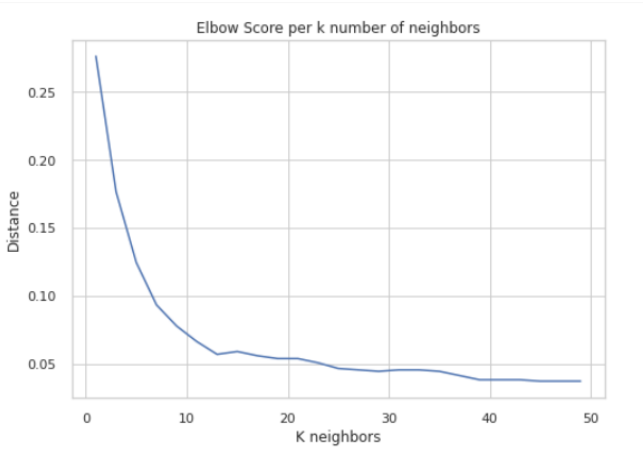
Homogeneity score for 9 number of clusters is: 0.0610
Completeness score for 9 number of clusters is: 0.1645

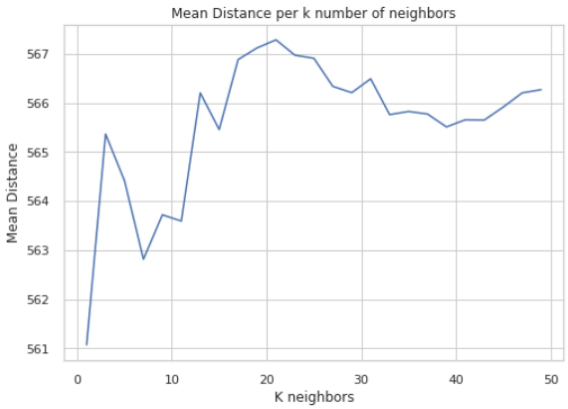
The result does not seem all that surprising since I used States as our targets and as mentioned the institutions tend to adopt the same federal programs across States.

The completeness score which is complementary to the homogeneity score, and also ranges from 0 to 1, indicates how well the algorithm is assigning samples with the same true labels to the same cluster. - Again our algorithm is not performing very well.

Checking these same values for a number of clusters equal to 3 (instead of 9) for comparative purposes confirms that 9 clusters are better with both Completeness Score and Homogeneity Scores improving. However, in order to select the k value that is best, I will use a mathematical method: I will look at the Within Clusters Sum of Squares (which is a measure of how dense the clusters are) and try to minimize this function using gradient descent to identify the best number of clusters the data aggregates into.

Using Nearest Neighbors Classifier we see that for this classifier the aggregation of the data in classes (groups) of about 9 neighbors ~ or about 5-6 clusters of States





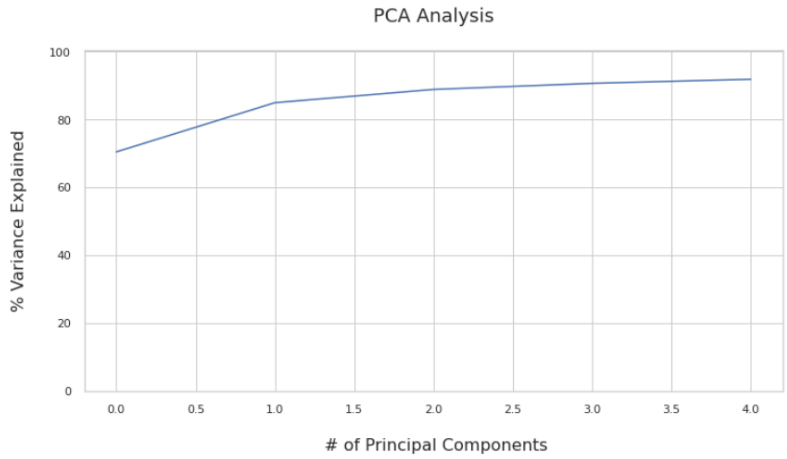
The results seem somewhat incoherent and thus I choose to move on to apply DBSCAN to see if we get a better results on the test data.

Clustering by State

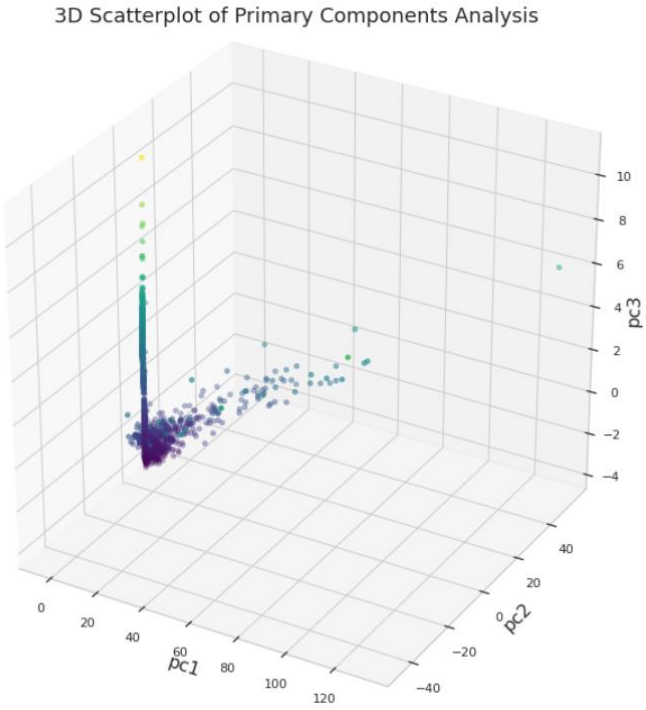
In order to apply DBSCAN or in other words, Density-Based Spatial Clustering of Applications with Noise, I need to select two parameters:
(1) epsilon (based on which's value the total number of clusters will depend); and
(2) the number of samples (this value should be at least equal to the number of features and when the data is characterized by the presence of outliers, it may be useful to select an even higher value for this parameter - as we know, increasing the number of samples increases the sampling distributions' chances to reflect the true population of data)

Trying to implement DBSCAN, let us first try to reduce the dimensions of our data. I will in other words use PCA Principal Component Analysis, to reduce the data dimensionality while not loosing the information from all of our data features. (We are not eliminating features but rather extracting the principal components from all features that affect our output data).

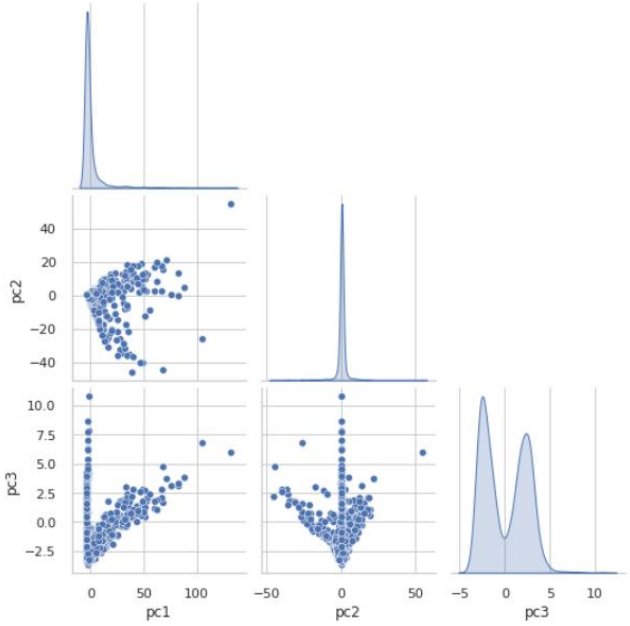
The primary reason for us to use this algorithm is to be then able to visualize the data with DBSCAN and verify if there are meaningful clusters of data formed.



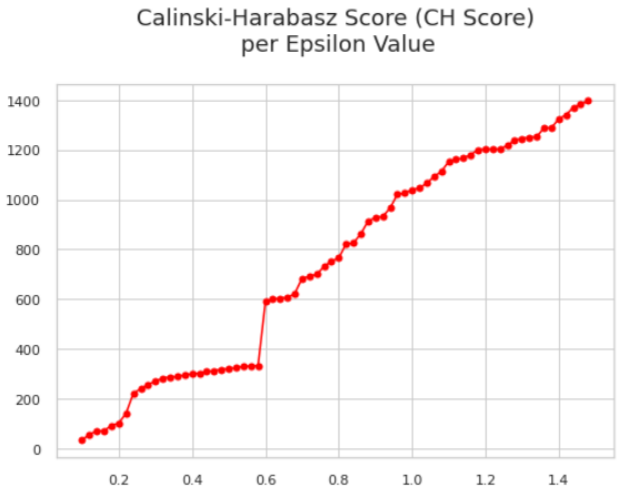
The Principal Component Analysis (PCA) allows us to see that practically 90% of the total variance of our model is explained by the principal three features. As the graph shows, additional features add limited information to our model.



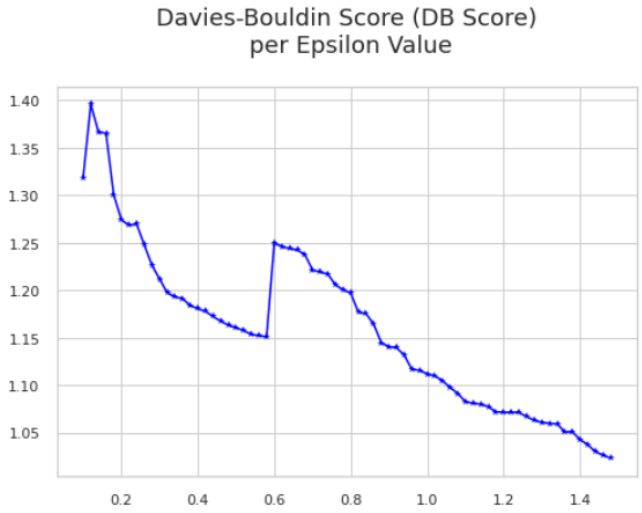
Taking a look at the correlation existing between the first three components of the dataset we definitely see that there is a strong correlation between these features and that there do seem to three clusters the data can be subdivided into as shown from the 3D graph above as well.



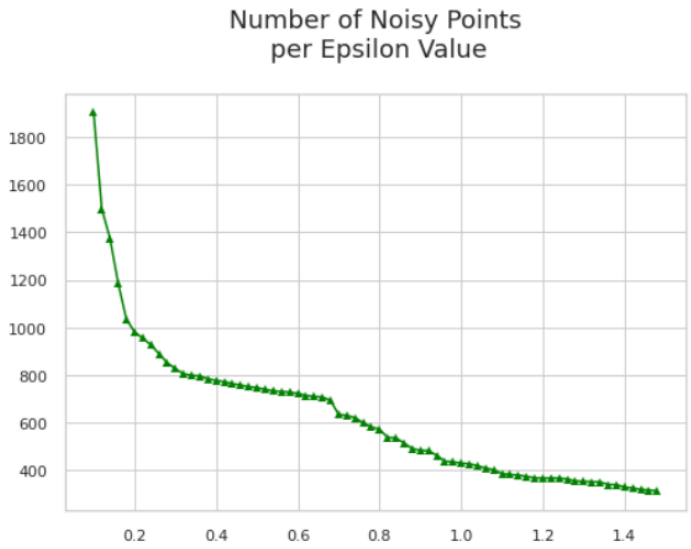
Taking a look at how we can evaluate these results we will use three measures that are typically used in cluster analysis.



The first graph in red represents the Calinski-Harabasz Score (CH Score) which is a measure of cohesion and separation (at the same time) of the clusters. This score is not bounded (other than reaching a limit value based on the data structure), thus generally speaking, higher values of this score indicate a better clustering result.



The second graph in blue represents the Davies-Bouldin Score (DB Score) which is a measure of the amount of separation existing between clusters. Lower values of the DB Score indicate better separation between clusters.



The third graph, in green represents the number of data points classified as noise.

As we can see from all three graphs, there is a change in all three measures when epsilon is around 0.625.

If we look at the number of noisy points when epsilon is 0.625, we see that there is a sharp change in slope of the decreasing function. This indicates that there is a sharp reduction in the number of points classified as noise at this epsilon value.

At the same epsilon value there is a sharp increase in the DB Score and a sharp increase in the CH Score. All three values then indicate that the optimal value of epsilon is in fact 0.625

Checking the algorithm's results with an epsilon value of 0.625 and first using Euclidean distances, and then Minkowski, we see that the performance improves (somewhat) using a Minkowski distance, however, the number of clusters identified is still equal to two, and given the small data set, the overall number of data points classified as noise is still relatively high.

Euclidean Distances
No. of clusters: (2,)
No. of noisy points: 712
CH Score =600.825
DB Score =1.245

Minkowski Distances with p value of 4
No. of clusters: (2,)
No. of noisy points: 633
CH Score =685.389
DB Score =1.221

Clustering by Institutional Size

All of the work we have done has been confirmed by our data. However, as we initially said, we chose a target feature which may have not been the best. That being the case, I decided to re-run this analysis choosing as target variable Institution Size to see if clearer patterns emerge then, from the data.

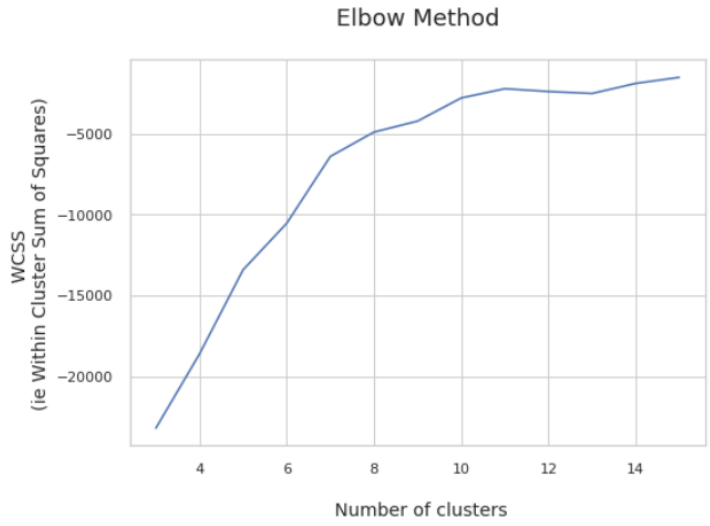
The classification report offers a better result in this case, in fact our accuracy reaches 35% overall. Our classification appears to be mostly suited in identifying HEIs in our extreme classes, which seems to suggest an influence by outliers on one end of the spectrum and influence by the overall lack of balance that the data is characterized by.

```
Classification Report:
                    precision    recall  f1-score   support

1       << 1,000                0.37      1.00      0.54       292
2       1,000 - 4,999           0.00      0.00      0.00       377
3       5,000 - 9,999           0.00      0.00      0.00       134
4       10,000 - 19,999         0.28      0.34      0.31        98
5       20,000 <<               0.64      0.15      0.24        62

          accuracy                           0.35       963
         macro avg       0.26      0.30      0.22       963
      weighted avg       0.18      0.35      0.21       963

Accuracy:  0.34683281412253375
[-23211.20706272 -18626.45440892 -13418.61458282 -10543.8090176
  -6408.39391476  -4902.07411114  -4222.99300745  -2792.44541542
  -2219.945871    -2391.54757732  -2515.55385751  -1892.9195458
  -1516.2906226 ]
```



Checking Homogeneity and Completeness scores against a varying number of clusters, there is no definitive answer – both ratios range between zero and one and are complementary to each other, yet their values do not paint a clear picture as you can see below.

One can note that the Completeness score seems to be stabilized at its lowest value beyond a number of clusters equal to 5, which we would expect using the categories of institutional size as

targets. Yet, the Homogeneity score does not hold a similar behavior. This is probably due to the high variety of characteristics existing between Institutions of the same size, making it highly difficult for our algorithm to identify homogenous clusters.

Homogeneity score for 2 number of clusters is: 0.08
Completeness score for 2 number of clusters is: 0.62

Homogeneity score for 3 number of clusters is: 0.13
Completeness score for 3 number of clusters is: 0.51

Homogeneity score for 4 number of clusters is: 0.14
Completeness score for 4 number of clusters is: 0.42

Homogeneity score for **5** number of clusters is: 0.15
Completeness score for **5** number of clusters is: 0.34

Homogeneity score for 6 number of clusters is: 0.16
Completeness score for 6 number of clusters is: 0.21

Homogeneity score for 7 number of clusters is: 0.18
Completeness score for 7 number of clusters is: 0.22

Homogeneity score for 8 number of clusters is: 0.17
Completeness score for 8 number of clusters is: 0.2

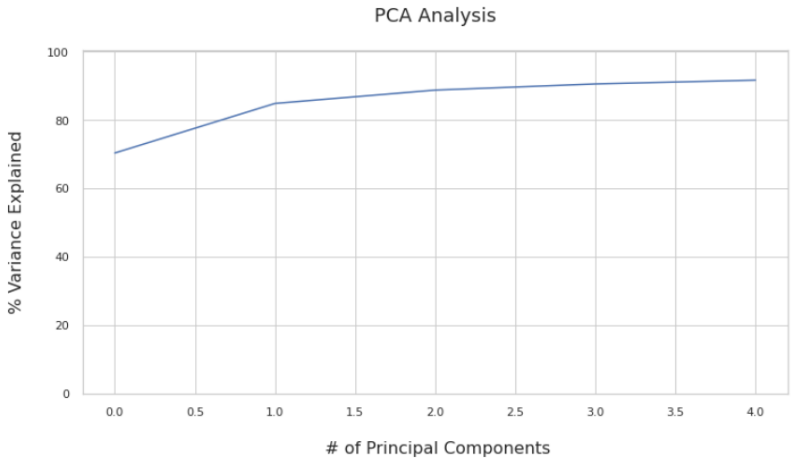
Homogeneity score for 9 number of clusters is: 0.2
Completeness score for 9 number of clusters is: 0.21

Homogeneity score for 10 number of clusters is: 0.2
Completeness score for 10 number of clusters is: 0.21

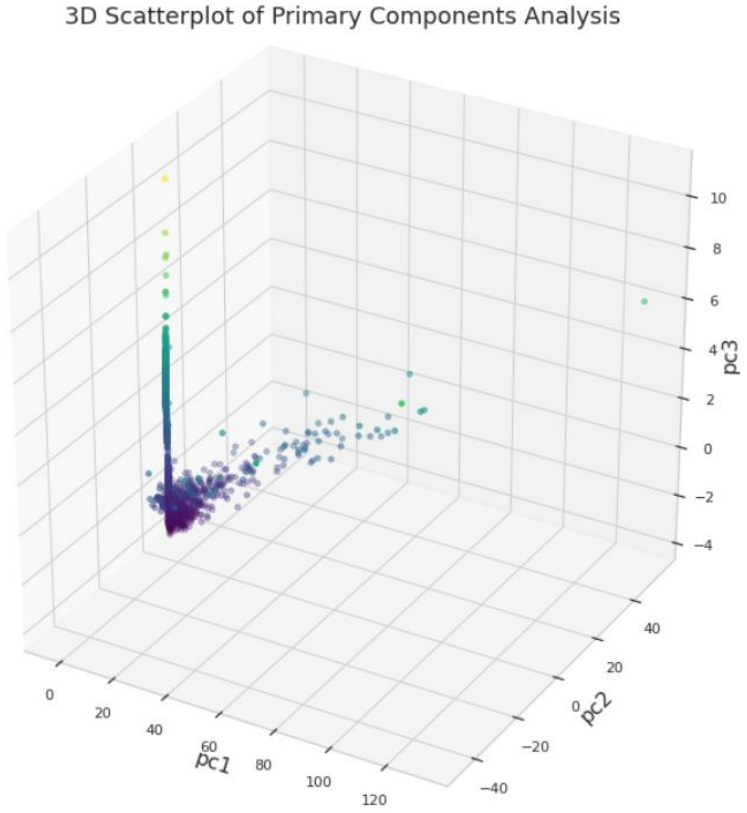
Given the results, we can confirm similar results as those we have reached by clustering our data based on economic area composition.

Principal Component Analysis by Institutional Size

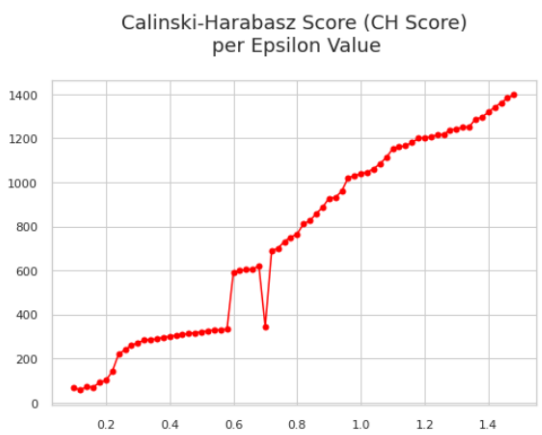
Running our PCA Analysis we get similar results, with 90% of our variance explained by three principal components, even though we may argue that in this case, two may be sufficient:

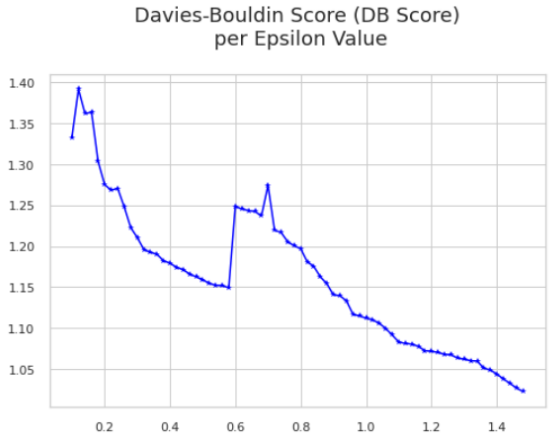


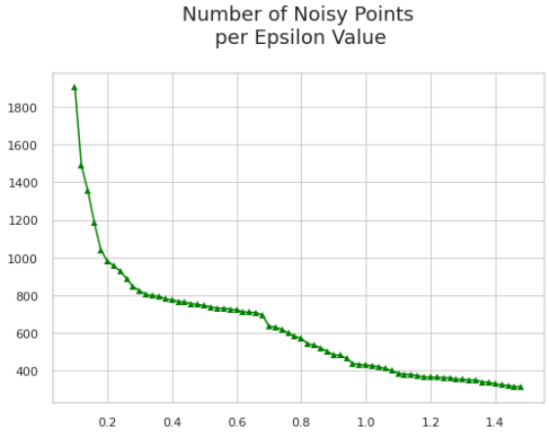
Our 3D clusters look very similar, not surprisingly. Since the data has not changed, we would expect to get similar results if the algorithm is robust.



Interestingly, checking our evaluation measures that help us identify the best epsilon value, we see that an epsilon value of 0.625 is again the best overall, which is a nice confirmation for our work:







Epsilon value of 0.625 and Euclidean distance
No. of clusters: (2,)
No. of noisy points: 712
CH Score =601.044
DB Score =1.245
Counter({0: 1533, -1: 712})

Epsilon value of 0.625 and Minkowski distance w p=4
No. of clusters: (2,)
No. of noisy points: 722
CH Score =590.739
DB Score =1.248
Counter({0: 1523, -1: 722})

Epsilon value of 0.6 and Euclidean distance
No. of clusters: (2,)
No. of noisy points: 722
CH Score =590.739
DB Score =1.248
Counter({0: 1523, -1: 722})

Epsilon value of 0.6 and Minkowski distance w p=4
No. of clusters: (2,)
No. of noisy points: 722
CH Score =590.739
DB Score =1.248
Counter({0: 1523, -1: 722})

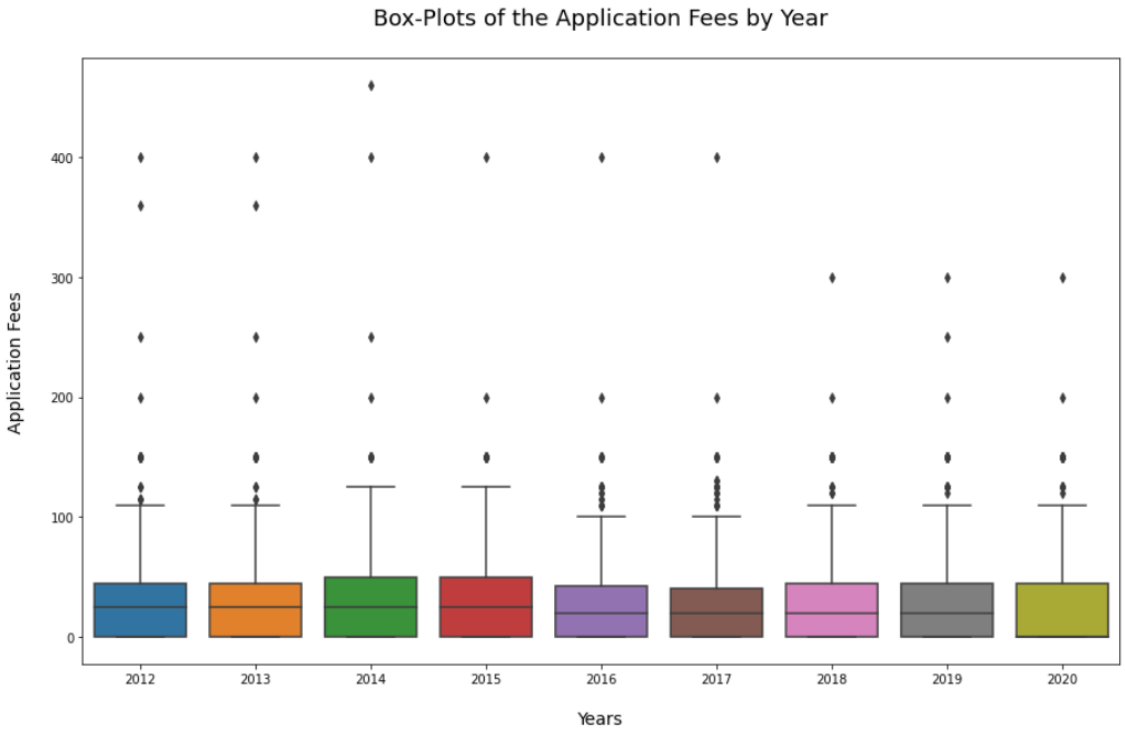
2012-2020 Data overall

Given these results, I proceeded to work with the whole data set for years 2012-2020.

Before proceeding with all of the data I decide to perform a little more data cleaning to adjust the models and try to improve the algorithms performance. I decided to eliminate from the data, HEIs located outside the continental US and Service Schools (i.e. primarily military schools)

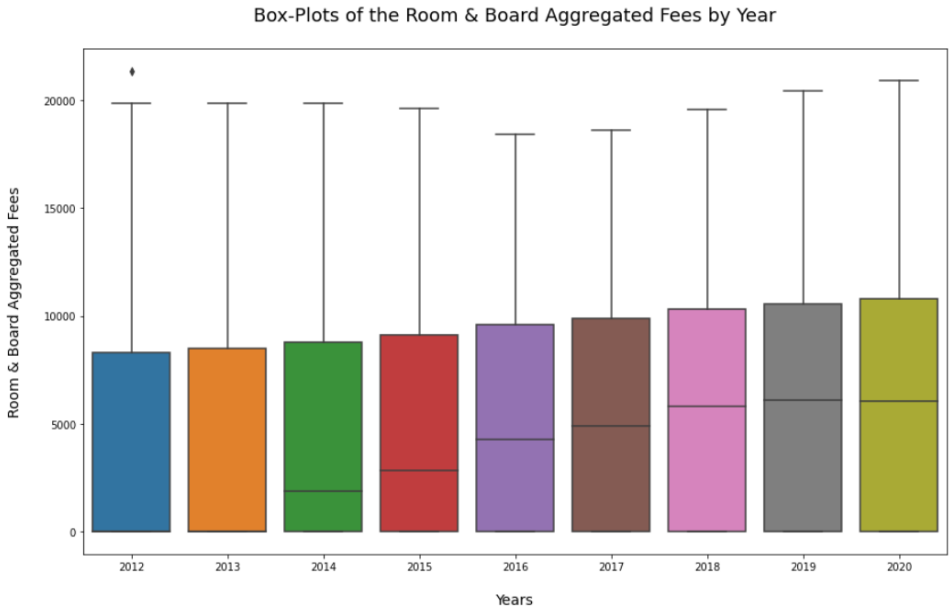
Then, I took a look at the data distribution per year of the numerical features in the dataset

Application fees have remained fairly constant throughout the years, although this feature still includes a few outliers and there has been a tendency of the fees to increase over the years.

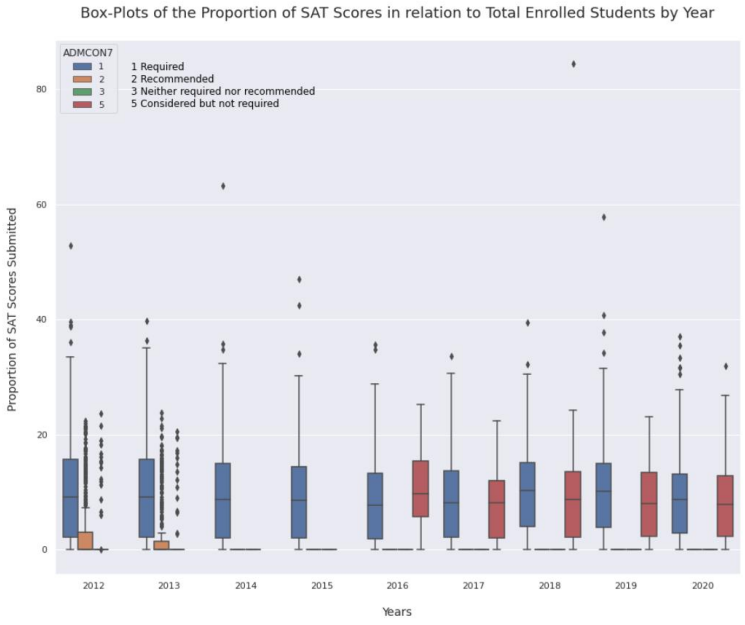


As we know, means are pulled in the direction of outliers, thus the data seems to indicate that overall and throughout the years fees have been higher than it might appear at first glance. However, there was an attempt around 2016 to keep these additional fees in check, as the contraction in median application fees suggests. This measure, may have also been one aimed at trying to encourage and enhance enrollment (or at least applications) by minority students and generally students less affluent.

There has been a gradual and more significant increase in the cost of the median Room & Board as the boxplots below show and confirm what we already saw in our graphs. Median values have shown an increase across the board and our distribution of costs are heavily skewed right.

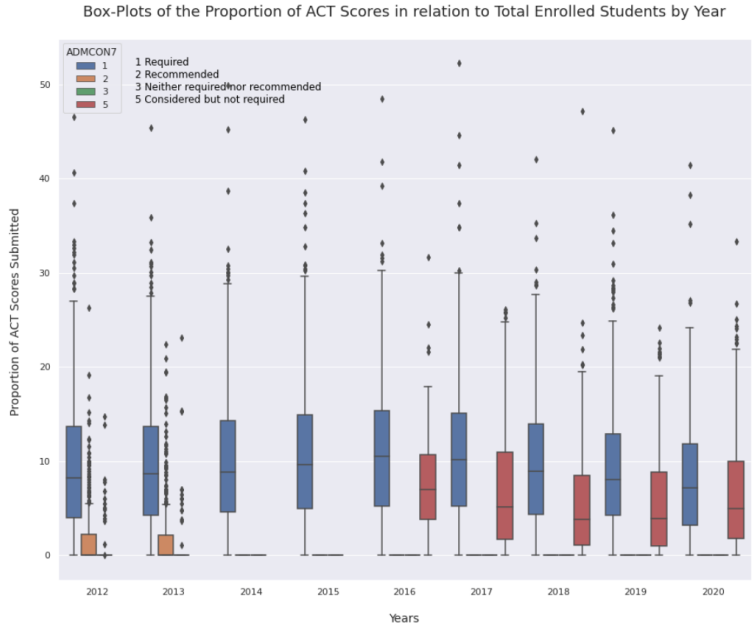


Taking a look at the SAT Scores, it is important to analyze this data not taking into account non-reported scores due to the Institutional Policy, i.e. filtering out the institutions who do not require students to report them. From the box-plots below it is apparent that there has been a change in policy that occurred starting after 2013. Less and less institutions require SAT score reporting and, after 2015, the number of institutions considering but not requiring SAT scores has become definitely significant.

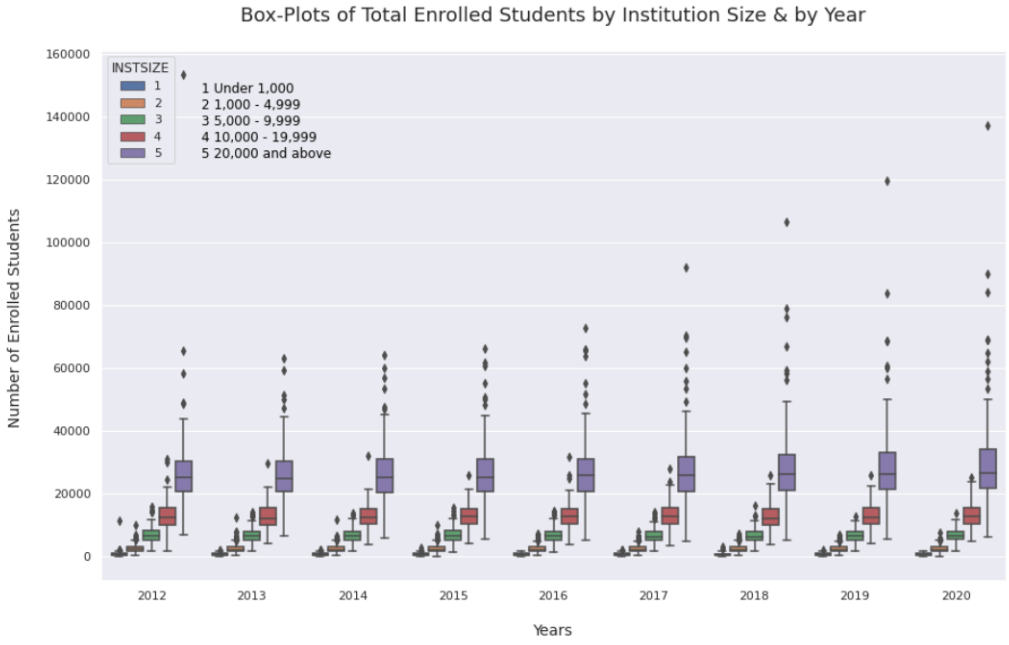


We must however be careful about drawing conclusions upon this data as it may be reflecting, at least in part, a change in reporting preferences. Interestingly, in spite of the fact that some of the larger HEIs across the country are those recently pushing for not requiring standardized tests at all, the number of institutions adopting such policies is still too low to be significant.

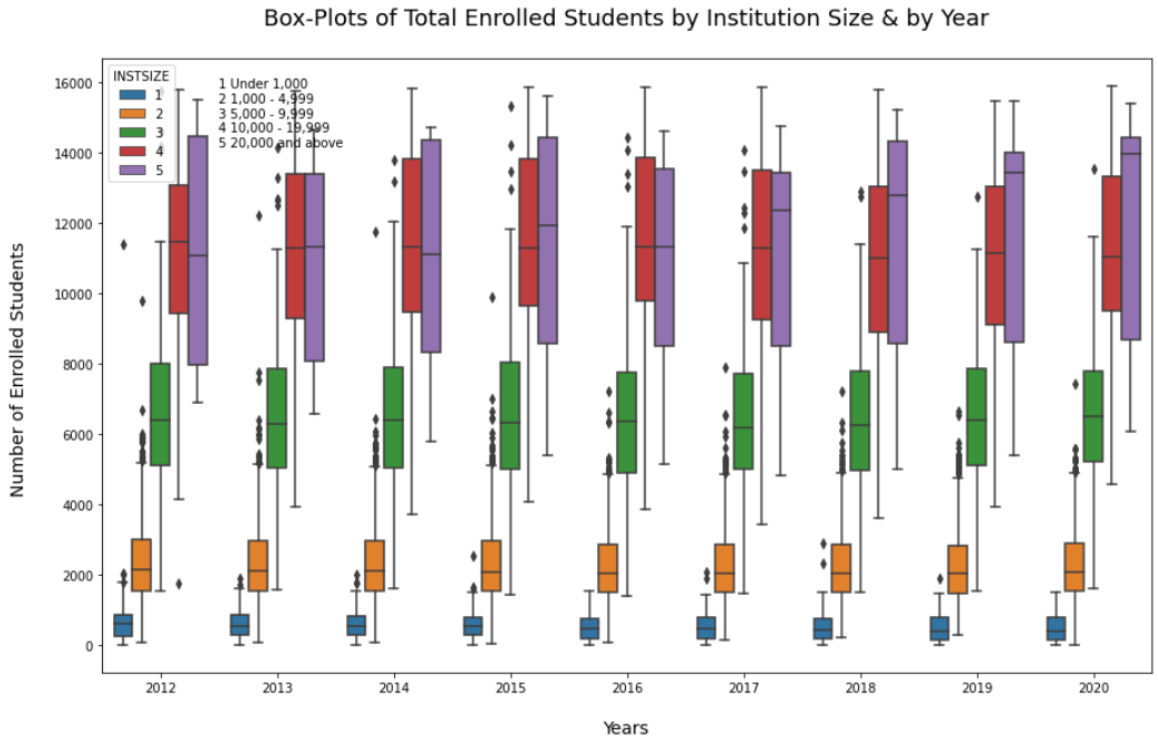
The same seems to be true for ACT Scores as well as the following figure shows:



The distribution by year of Total Enrolled students by HEIs' size does not present wild changes over the period, but there is still a presence of outliers.



Eliminating Outlier data in terms of Total Student Enrollment the distributions still present fairly stable picture over the years, with the greatest fluctuations present in the larger HEIs (codes 4 and 5, enrolling more than 10,000 students each year).

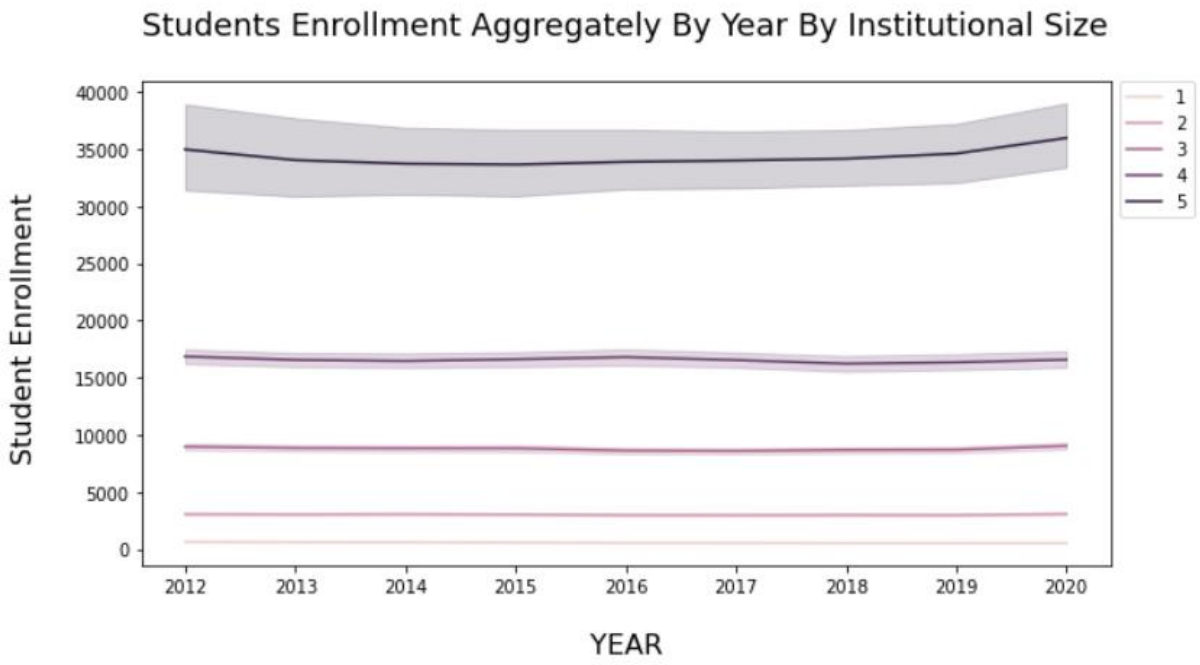


The data regarding student enrollment within the largest institutions then, seems to contradict, at least for this cluster, talks of contraction ins tudent enrollments. In fact, the medial value of enrolled students has actually gone up for the largest institutions. While the second largest institutions may have been those that experienced the greatest fluctuations during these years.

Taking the whole data set together, we must remember that we are dealing with a Time Series of data points. This makes our data richer but we also have to keep this element into consideration when applying our ML models.

Before proceeding in deciding how to split the data to apply it to our ML models, I chose to dive into taking a closer look at the Time Series and how my numerical features, and enrollment numbers in particular, were affected by the change in time.

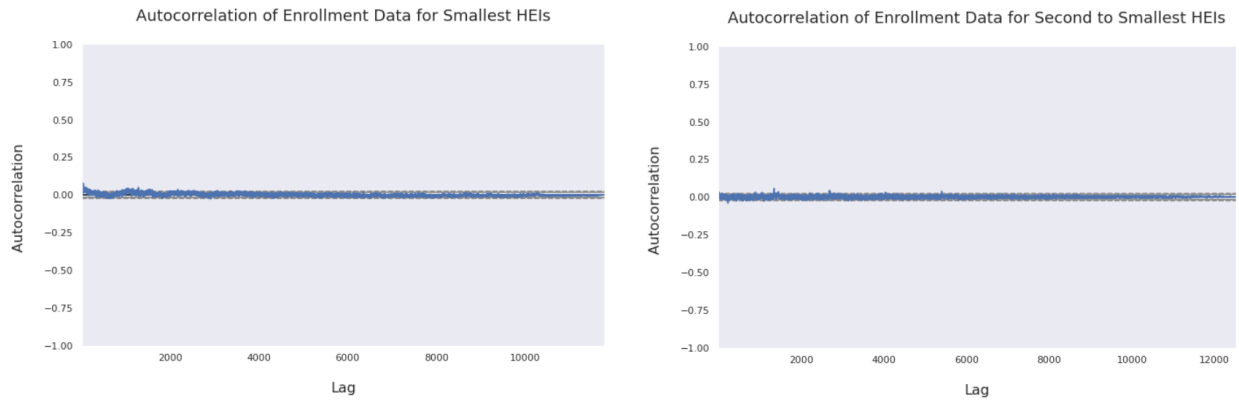
Given the nature of the data, the results showed that the data tends were fairly stationary over time. This is not too surprising, since enrollment data is recorded on a yearly basis and is not affected by seasonality in a significant manner (or at least any seasonality would be mitigated by both the fact that the data is recorded yearly and that changes in population trends tend to occur over long periods of time). As the graph below suggests with the shadow area around our trend lines, the greatest fluctuations in overall enrollment levels by size occur within the second to largest and largest HEIs. Overall, however, all institutional groups aggregately seem to maintain fairly constant enrollment levels.

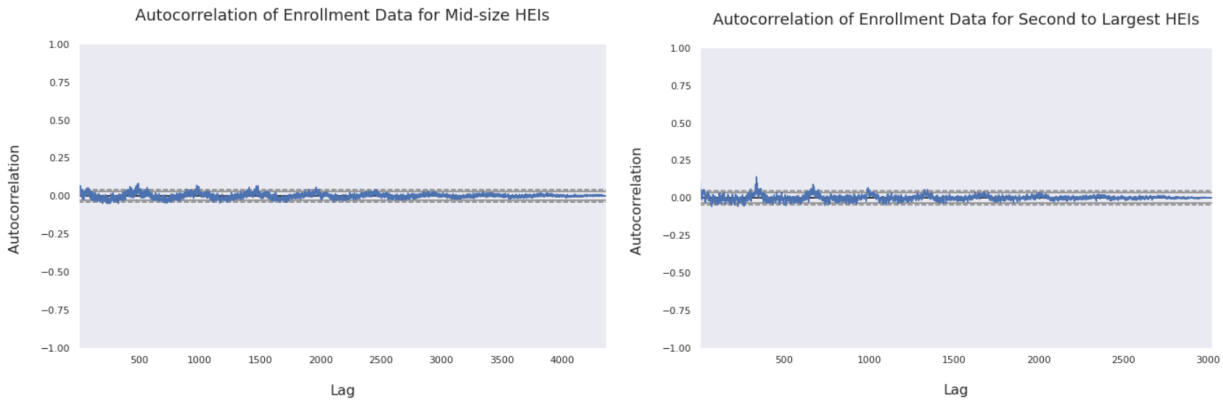


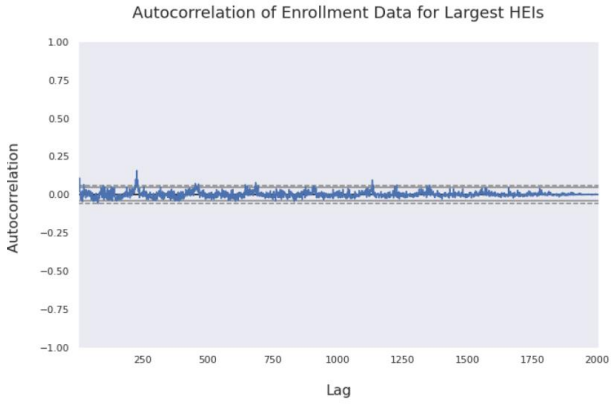
The second important aspect I checked was the presence of correlation between student enrollment values over time. I performed the Augmented Dickey-Fuller Test and the results glaringly confirmed our hypothesis of non-correlation, with a p-value of practically zero for each of the subsets of data regarding enrollment by HEIs size. This result is a bit more surprising since I would expect there to be some level of correlation between values from one year to the next.

As institutional size increases, there appears to be an increase correlation between enrollment levels and a more visible cyclical nature of enrollments. However, this information is not supported by the scatter plot display of enrollment levels compared to the previous' period.

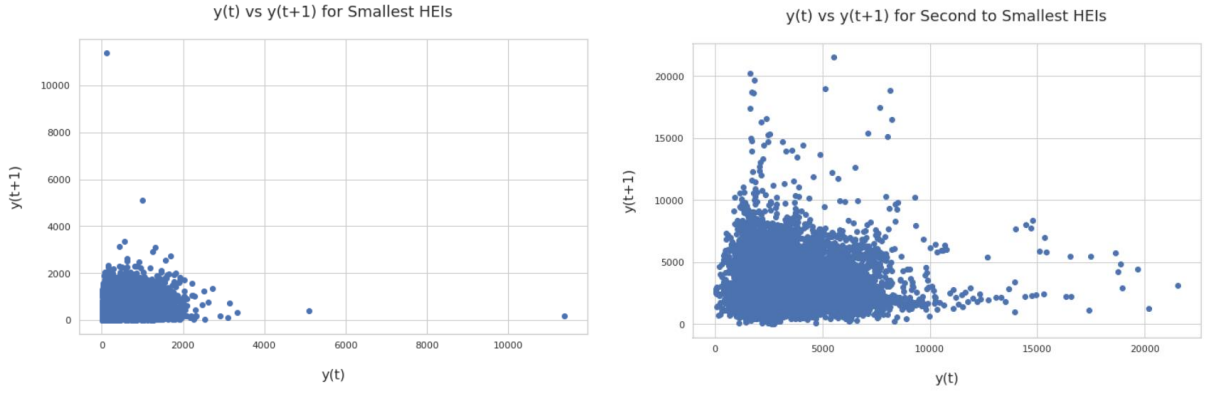
This would require further investigation on our part but it may be due to the lack of uniformity within these groups of HEIs.

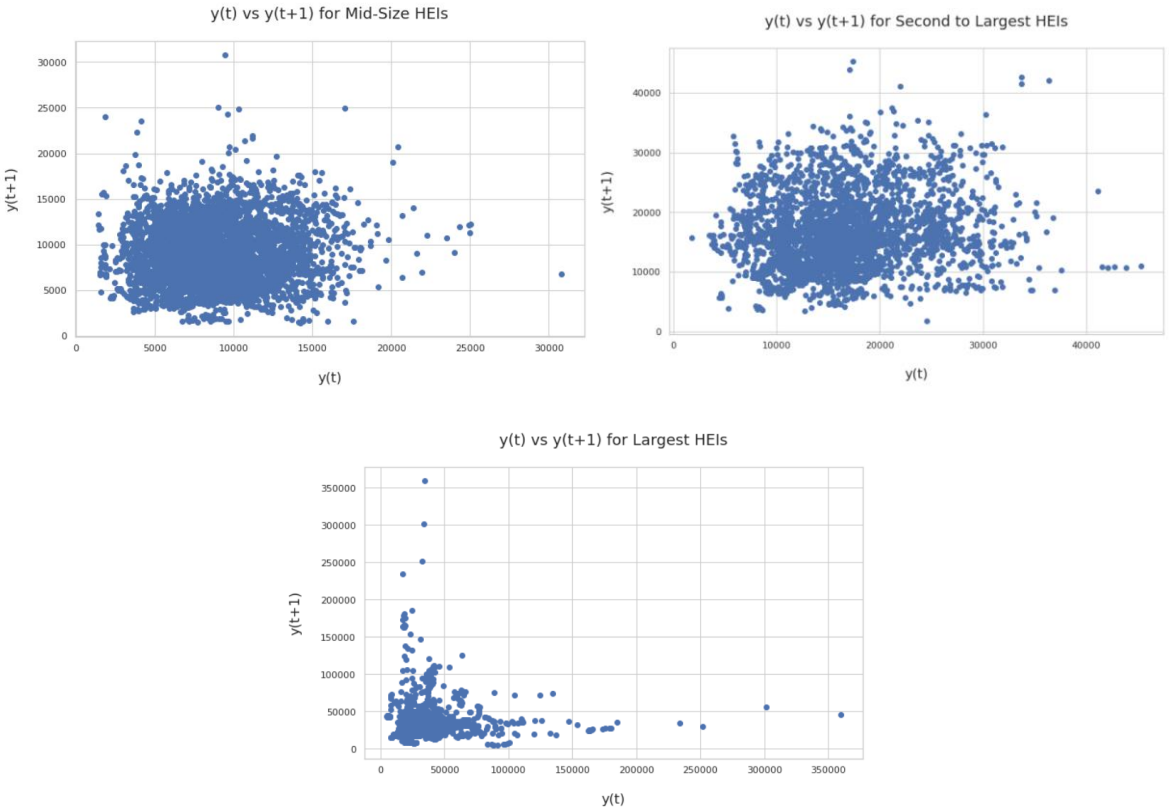






These results are also easily read from the scatter plot showing the relationship between Student Enrollment levels between lag periods which shows no clear relationship between the data



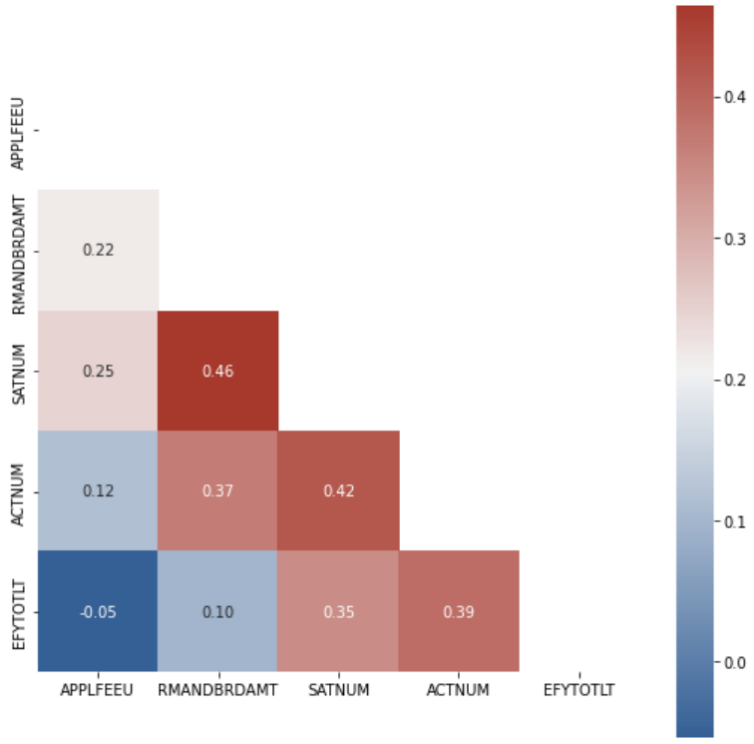


These results ultimately challenge us in trying to find a model that can be used to explain or highlight features affecting student enrollments over time. However, it also allows us to take a closer look at the influence each individual feature has on our target variable (Student Enrollment) without worrying too much about the fact that our data is a time series. It would be interesting to pursue an autoregressive analysis as well, as a next step to try to disclose the forces at play in student enrollment over time. I postpone this to future next steps for this work.

This said, in order to justify the relatively poor results obtained with our regression model over 2020 data and to explain the results we might achieve on the overall data set I decided to take a step back and check for the correlation existing between the numerical variables, independently of time and aggregately. I fully expected there to be little to no correlation between the variables again and in fact these results were confirmed.

The lack of strong correlations between the numerical variables part of our data set ultimately affect the ease with which we can achieve high accuracy with out models. It is reasonable in fact to suspect that low variable interaction as measured by correlation will reduce the influence each variable, and therefore the influence each change in such variable, can potentially have on our target.

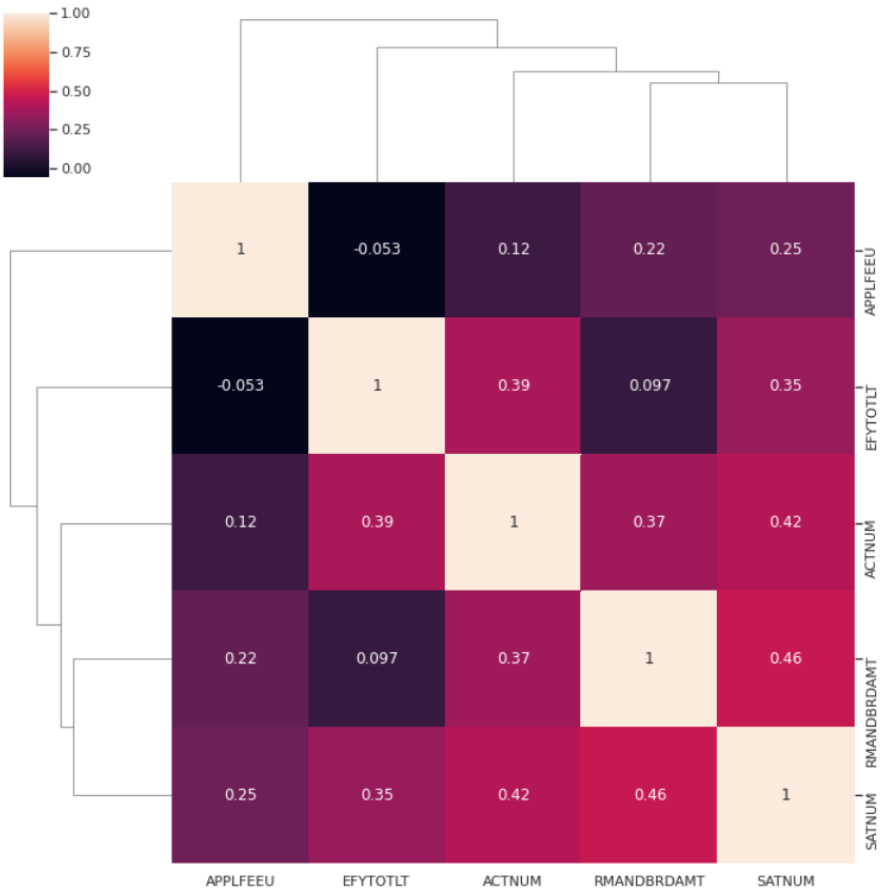
Below is a first graph of the correlation existing between our numerical features alone.



The purpose of checking these correlations was to understand how the variables affect each other statically - Pearson's correlation can give us this information and it is depicted above. As we can see, student enrollment levels are limitedly correlated with Number of SAT (35%) and ACT (39%) submitted scores - but here it is likely that the correlation coefficient is highly influenced by the sheer size of the institutions across the continental USA still requiring such submissions. While application fees (5%) and aggregate costs of room and board (10%) seem to represent coefficients that may be more reliable in terms of their values and implied effects. Surprisingly however, even the cost of Room and Board does not seem to be particularly correlated with total number of students enrolled - this may be linked to the fact that here, I are analyzing only HEIs not adopting open admission policies and only HEIs attracting undergraduate students.

Overall, these value confirm that the variables are fairly independent from each other and do not affect each other greatly – consequently they represent limitedly useful tools to leverage increases in enrollment by HEIs. (Even though we must keep in mind that the existence of correlation between these variables does not translate into a direct causal effect.)

In an attempt to depict an even clearer picture, I chose to pull together a second graph – a cluster map - to capture how the various "seasonality" (here, changes over time) of the variables affects the others. Again it is apparent that there is little influence between the numerical values part of our data-set. However, our data seems to indicate that HEIs who receive the most number of reported SAT scores also receive greater numbers of ACT Scores and tend to be more expensive institutions. They tend to allure greater number of enrollees, but this may be an indication of the fact that they can also represent the more exclusive HEIs or the HEIs with greater demand of services. Finally, application fees still seem to matter the least in determining final enrollment decisions and affect the other numerical variables here included the least.



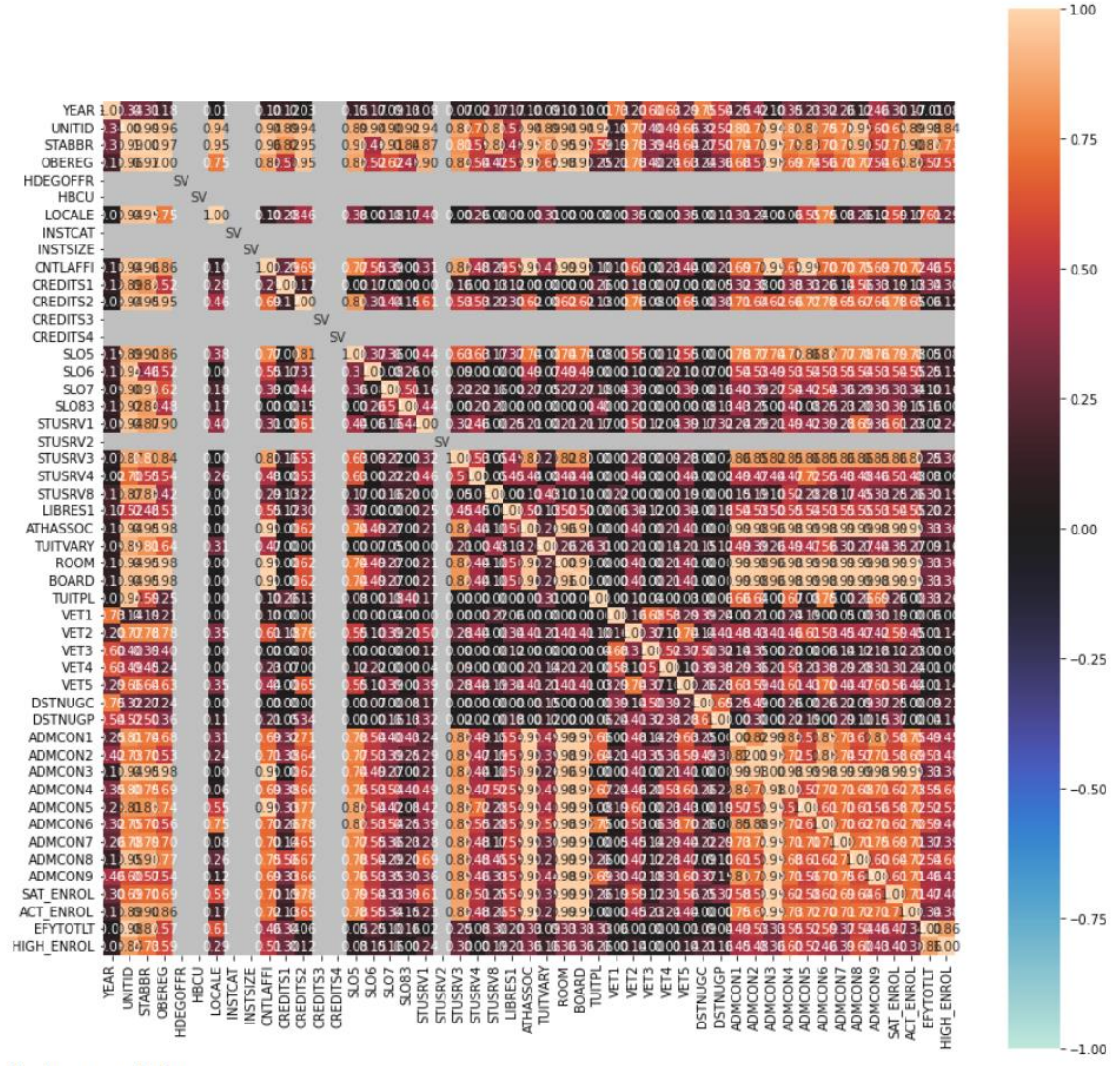
Having checked more closely the relationship existing between numerical features I took a closer look at the associations existing between categorical features as well. A look at the data aggregately over all of the years and over all of the features is not particularly revealing. (see graph in the appendix) thus I decided to check features' association but correcting based on institutional size. The information emerging from the analysis of the association of categorical features was more interesting and definitely starts highlighting differences existing between sub-groups of HEIs.

First of all, this analysis led to the discovery of an interesting peculiarity for the largest HEIs (those enrolling over 20,000 students per year).Namely, the fact this group, while numerous, is never able to meet a target high enrollment of 85% capacity, unless compared and rescaled based on the second largest HEIs size cluster. This seems like an interesting discovery on many levels and would beg the question of trying to identify how this cluster of HEIs was separated out – i.e. when and how the decision was made to consider this group as a separate group overall, instead of being considered as outliers pertaining to the HEIs coded as in group 4. Clearly the implications and consequences of the separation of this group for our analysis and for policy decisions, including financial support, can be significant.

Largest HEIs

Continuing my closer look at the largest HEIs, I noticed the following characteristics:
- The institutions are a minority within the group (only 94 in total)
- None are HBCUs
- All offer both Undergraduate and Graduate programs (which makes sense)
- Most are Private organizations (86 vs 8) with the ones Not for profit constituting the larger portion (72 vs 14) – interesting find
- All accept Advance Placement credits (at least on paper)
- None accept credit for Life experiences, placing enphsis on the fact that these institutions are primarily geared towards academics
- All have academic counseling offices

Furthermore, the features mostly associated with student enrollment seem to be in traditional academically related features(such as HS GPA, Letters of recommendation etc.). Finally, the institutions' location also seems to impact enrollment.

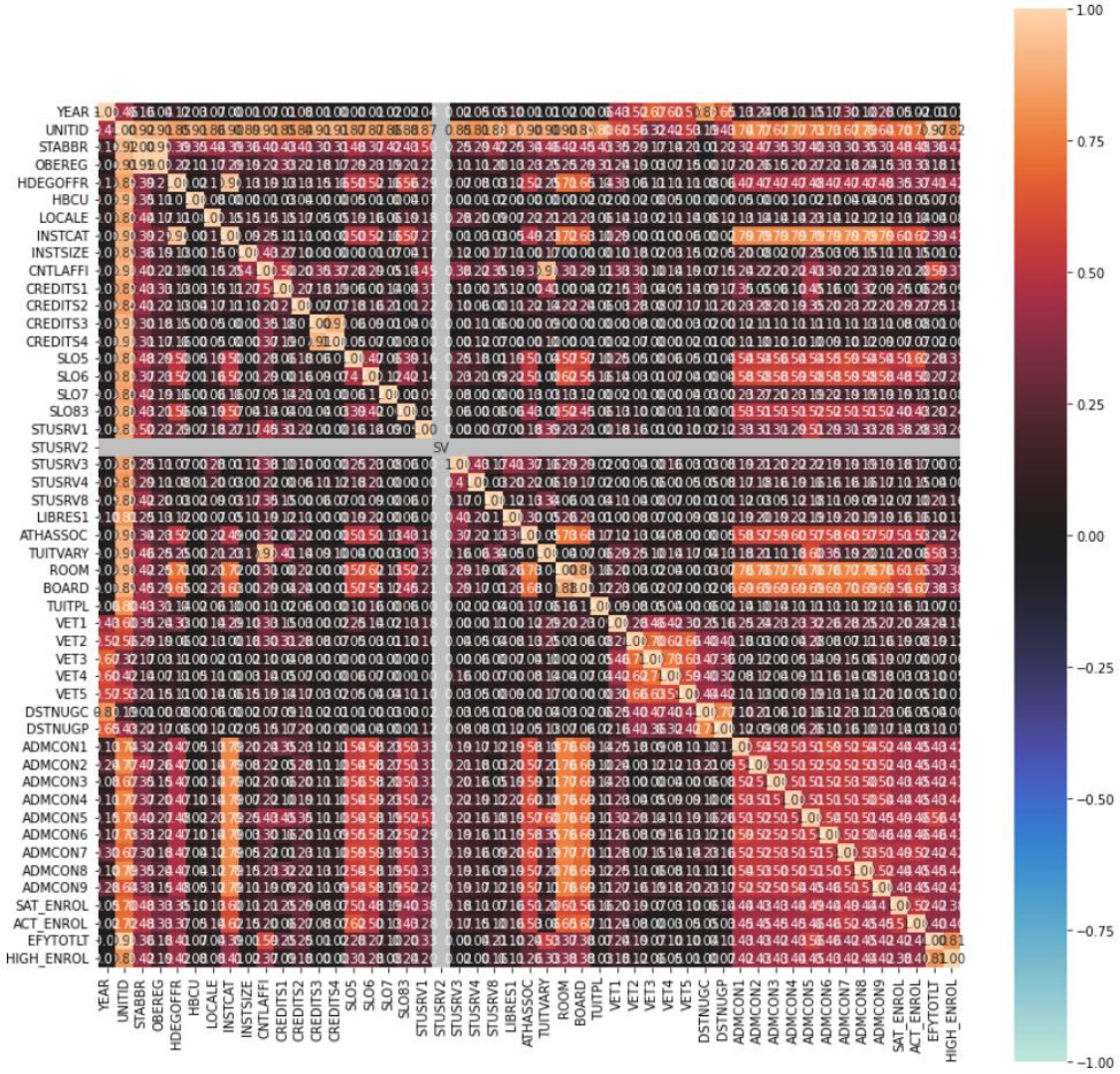


Second to Largest HEIs

The features mostly associated with enrollment for this group align somewhat with those for the 2nd largest group, however, notable differences are:
- Dual Enrollment services which play a significant role for this group
- Being part of Athletic Associations (not surprisingly), and
- the presence of tuition plan assistance programs and distance education (contrary to the 2nd largest group)

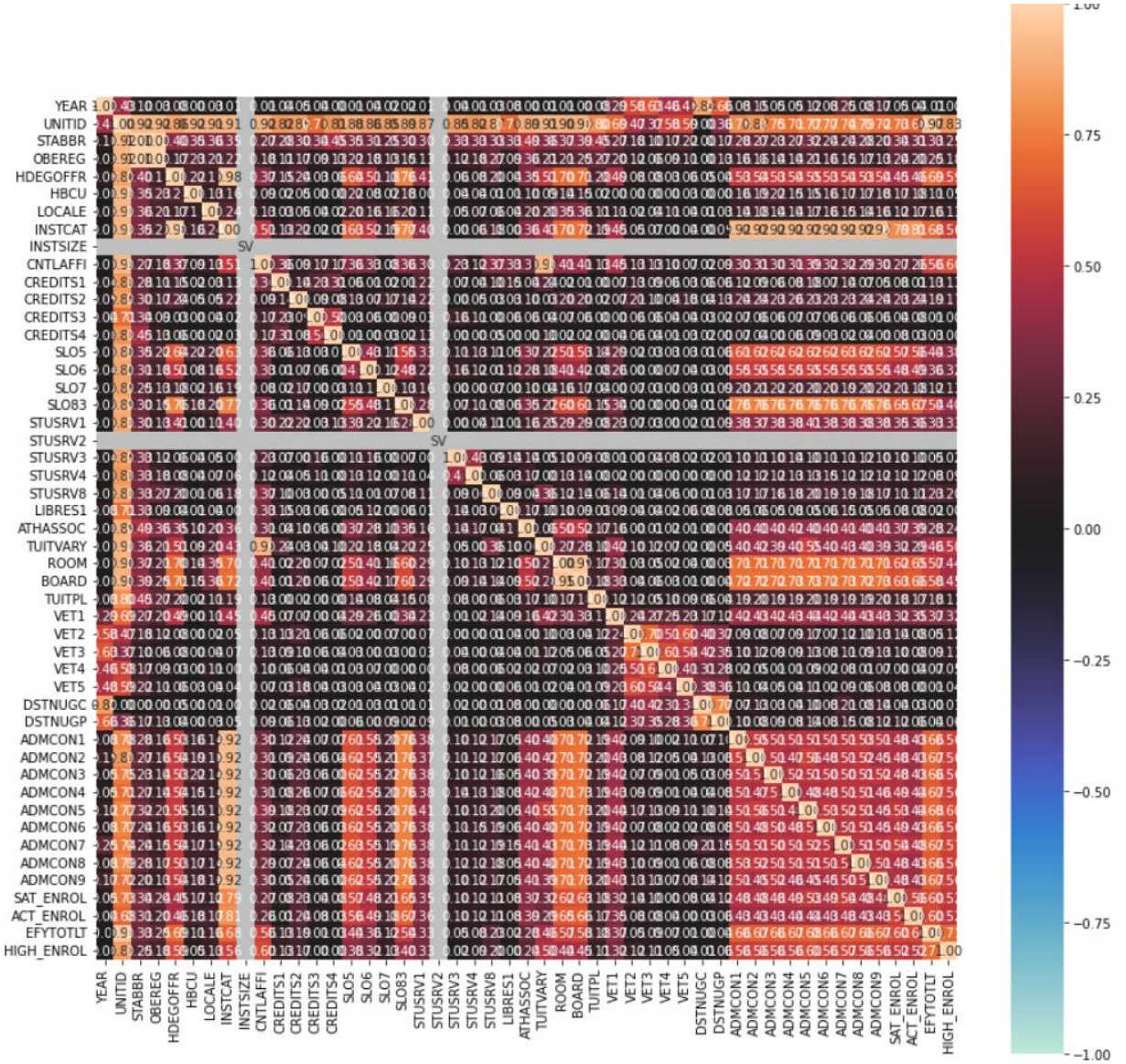
Among the features that are less impactful, surprisingly, we find Teacher certification programs, Tuition Assistance plans, as well as Veterans-oriented programs.

Career counseling services and career placement services do not distinguish this group from the Largest HEIs as do typical features that matter in High School (here captured by the ADMCON series, underscoring again that these institutions are likely to be academically focused.



Mid-sized HEIs

HEIs pertaining to the middle cluster in terms of size indicate that enrollment is associated more closely with the type of institutions (offering only undergraduate programs or not). Here too enrollment levels are associated with typically academic-related features.
Services aimed at returning students (as the STUSRV series indicates) have different effects (some positive, as services to support students with children, while others, such the offering of career counseling services are non-existent – interestingly and surprisingly). HEIs' location remains important but not to the same extent, while services aimed at military personnel do.
We start being able to see a slight but significant change in the type of HEIs that are classified in this cluster and also potential for a diversification of services offered to attract (or fight for) students who might otherwise enroll in larger institutions.



Smallest and Second to Smallest HEIs

The largest two clusters are those including the smallest HEIs in terms of size. Here, the clusters are fairly equally numbered however they include a wide array of different types of institutions.

Comparing the two clusters we see that significant associations emerge looking at Institutional Categories (degree granting vs non-degree granting) and type of control the HEIs is under (private vs public), especially for the Second-to-Smallest HEIs. The presence of Study abroad programs, interestingly, seems to be a factor attracting student enrollment in second to smallest HEIs, while teacher-certification programs have greater appeal for students enrolling in the smallest HEIs, together with Employment Services and Placement Services. Furthermore, and generally speaking, traditional features associated with High School student performance are not as closely knit to enrollment for these clusters of HEIs. (see graphs in the appendix)

These initial results are not that surprising as these two clusters of HEIs could be attracting less traditional students, students with a more limited array of choices open to them, generally younger in age and less experienced. The HEIs seem to represent less-traditional education organizations addressing a much wider need of educational outcomes.

This preliminary analysis of the categorical features and their association with enrollment is useful prior to implementing our ML algorithm as we can try to assess which features might be those we want to focus on and generally get a sense of the homogeneity of the data we are working with. Again we must factor in the lack of uniformity in our data. Furthermore, some of the associations emerging here could represent interesting extensions of my research.

Splitting our Time Series

Prior to proceeding with the application of machine learning algorithms to the overall dataset there is one more preparatory step I must take: choosing how to best split the data to take into consideration the fact that we have yearly recordings of our features and thus are dealing with Tine Series, even if our data is fairly stationary in nature.

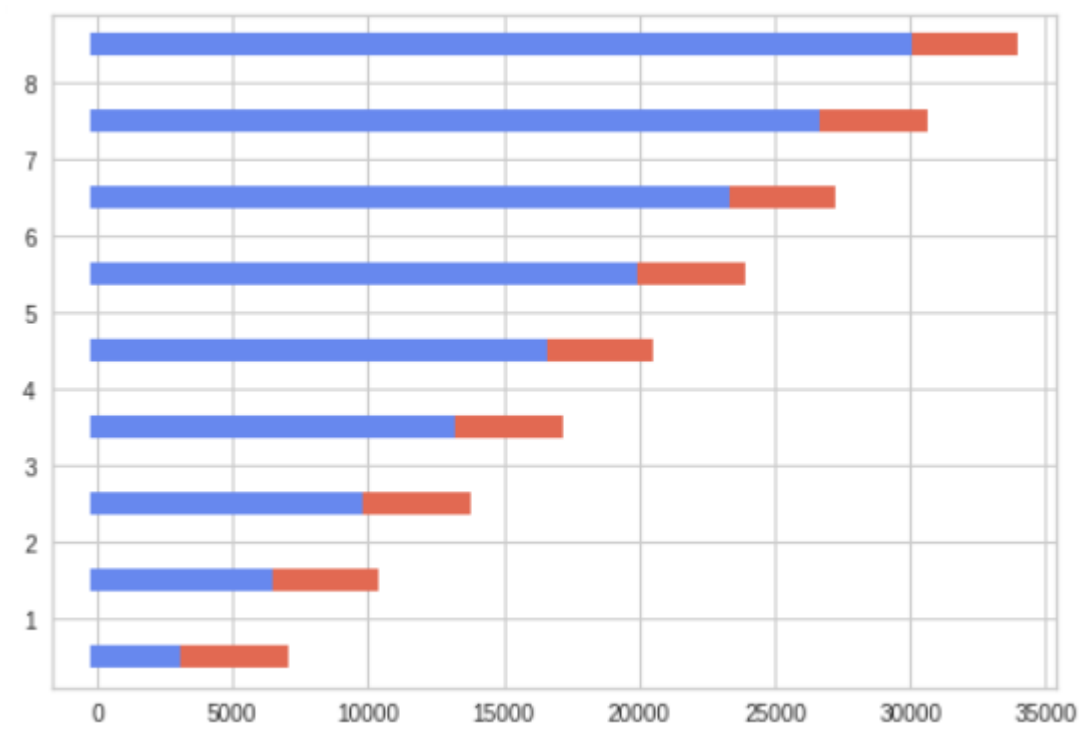
A simplistic approach would be to just split the data irrespective of the year it was recorded on but this would not allow us to use our data properly or at least fully, especially if we intend to use the data for forecasting purposes or decision making.

A possible approach would be to use multiple splits of the data between training and testing data sets such that we would train multiple models on training datasets of varying size, while maintaining our testing data set of identical size for comparative purposes.This approach is usually applied on larger datasets and sometimes referred to as cross-validation approach on the data. However, it does not allow to retain some of the information emerging from the intrinsic time-series nature of our dataset.

A different approach is using a sliding split method known as a Walk Forward Testing Method where the Testing data is always composed of data coming from the dataset one-time-stamp ahead of the data used to train the model. I chose this last method as it seems to me the most suitable for the yearly collected dataset I have to work with.

Below is a representation of the process used to split the time-series data. The Testing data set is always the same size, although randomly selected, while our training data set increases at every split. Overall I chose 9 splits since our data covers nine years and is fairly stationary in nature and fairly independent of the previous years' data. Note that this approach ensures that the testing set remains constant in size throughout the process and it also represents the data that is closest in time to the present, within each sub-split.

Walk forward method of selecting Time-Series data

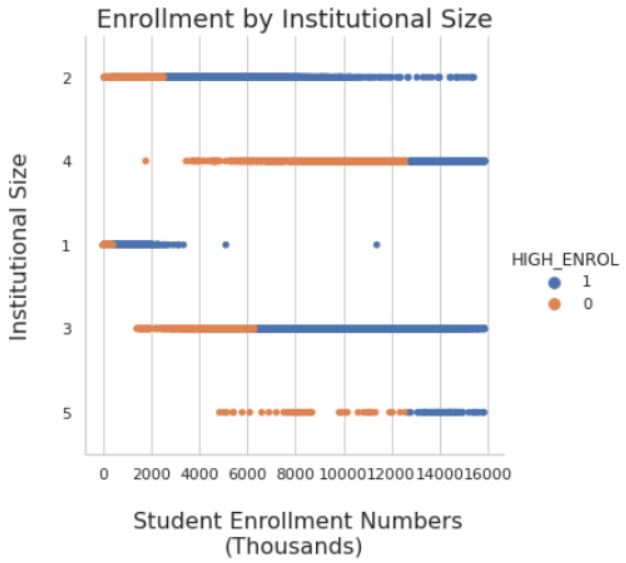


Given the analysis on existing associations among the variables, and prior to running our Machine Learning algorithms I take a last look at the institutions' size in relation to high versus low enrollment levels. As we can see from the graph below, the data referring to the second to largest and largest Institutions actually overlaps when looking at the HEIs from the perspective of meeting high enrollment ratios. Here we have defined high enrollment as enrollment at or above 85% of the yearly total (code 1 - blue) versus low enrollment (code 0 - orange).

The two clusters including the largest HEIs overlap and are skewed towards the lower ratio's side, showing homogeneity between them, and highlighting the fact that this may be the cluster

of HEIs who have been suffering the most in terms of contraction of overall enrollments over the years analyzed. (as also suggested by our previous analysis)

While the temptation to aggregate the Second to largest (code4) and largest HEIs (code 5) together is strong, I chose not to do this across the board, to follow upon the analytical decisions that were made by IPDES a priori. However, as far as feature analysis is concerned, I did merge the two largest groups of HEIs in spite of their heterogenous nature.



XGBoost Regression

Given the best results achieved with XGBoost regression on our 2020 data, I stick to this algorithm to analyze the overall dataset over the 2012-2020 time frame. Taking a look at how the XGBoost Regression model performs I finally start seeing an improvement in our performance in terms of overall predictions however the lack of uniformity of our data, the significant overlap that exists between institutional classes, and the limited effect our selected features have on enrollment overall, limit somewhat our results.

I choose to exclude quantitative variables from our data set, except for our target variable (student enrollment) given the limited correlations these showed and based upon the fact that their influence was captured indirectly by other categorical features. For example, the presence of lodging and boarding services is captured in our data set by two separate categorical features; SAT and ACT scores' submission, was captured by a new and ad-hoc feature created to register the level of submission corrected by institutional size; the only feature whose information is lost is application fee, but its influence in the decision making process of whether to enroll was minimal, as we have already seen.

Let's take a look at how the algorithm performs across institutional groups in terms of size.

Largest HEIs

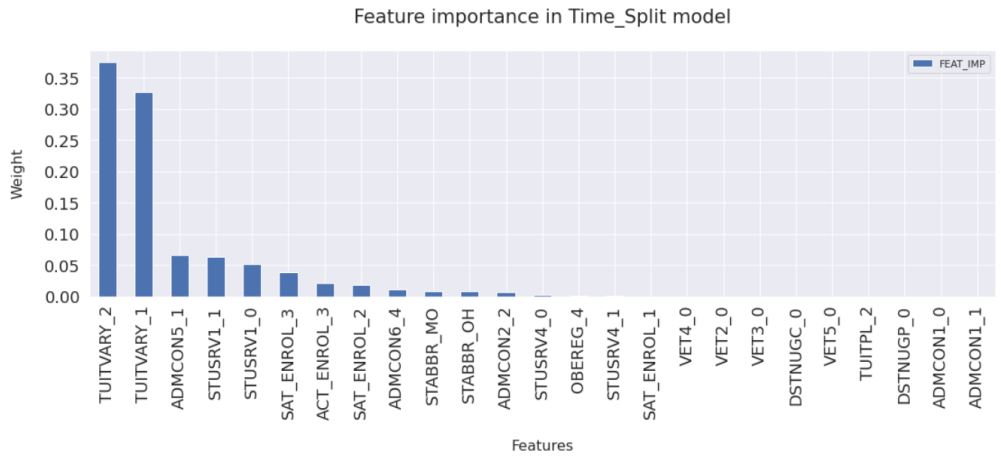
Our Root Mean Square Error (RMSE) is good measure of how well our model is performing overall, ie how well our model can predict actual student enrollment based on the categorical features we used. For our model, the RMSE tells us directly how many students we are off - on average - in our yearly prediction of Enrollment Numbers. The presence of outliers is surely affecting our predictions, however, our model's RMSE stays around about roughly a value of 2000, which is not too bad given these are the largest HEIs.

Comparing results among the splits, the model's training accuracy for the largest HEIs ranges between (50.24%, 59.25%) and the testing accuracy between (32%, 62.75%), which is not great but not awful either, especially considering the wide array of sizes this group includes. Our predictions for this group of HEIs tend to be pulled in the direction of the outliers and thus tend to be higher as confirmed by our scatter plot below.

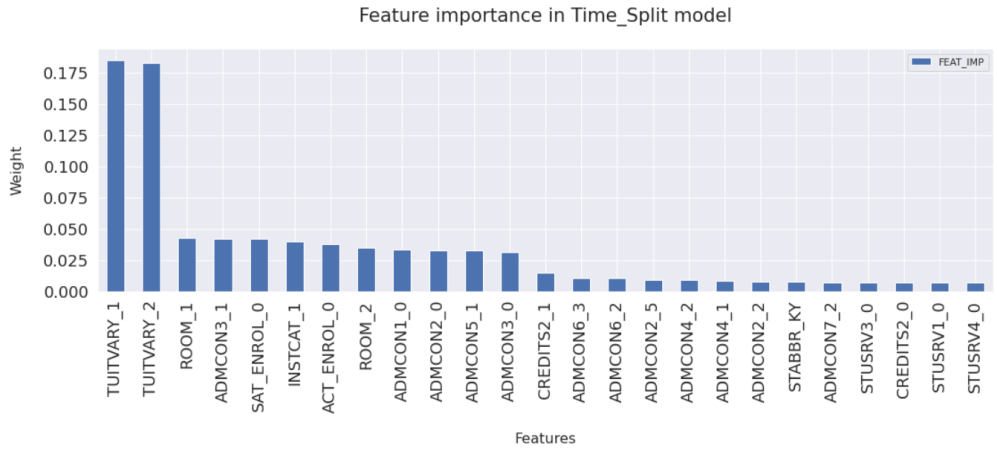
Taking a look at features' importance, what seems to emerge is that for these large institution, the presence or not of differences in the tuition charge for in-district, in-state, out-of-state students, matters, cancels itself out. This could be indicating that the organization's recognition – due to its size – may be playing a larger role than differences in charges across the group.

Offering boarding is important for these HEIs; The HEIs seem to be competing more consistently among each other through the traditional high school features students are concerned about when first applying to HEIs: School Records and High School Recommendations still matter the most. SAT scores effectively are influencing enrollment less consistently when we compare oldest splits against more recent ones, apparently confirming a shift in policy overall.

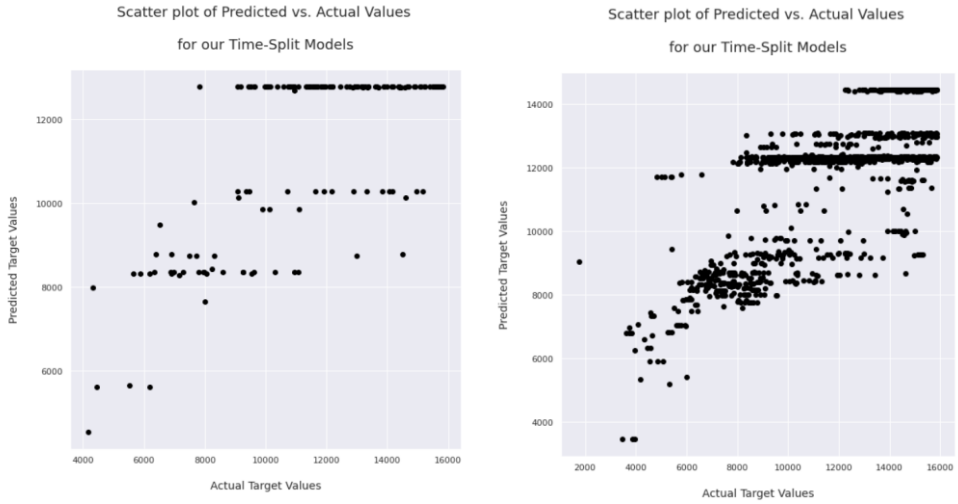
First Split (smallest amount of data) for the Largest HEIs



Largest amount of data (largest split) for the Largest HEIs



Scatter plot of predictions versus actual for the Largest HEIs (smallest versus largest split)



Middle-sized HEIs

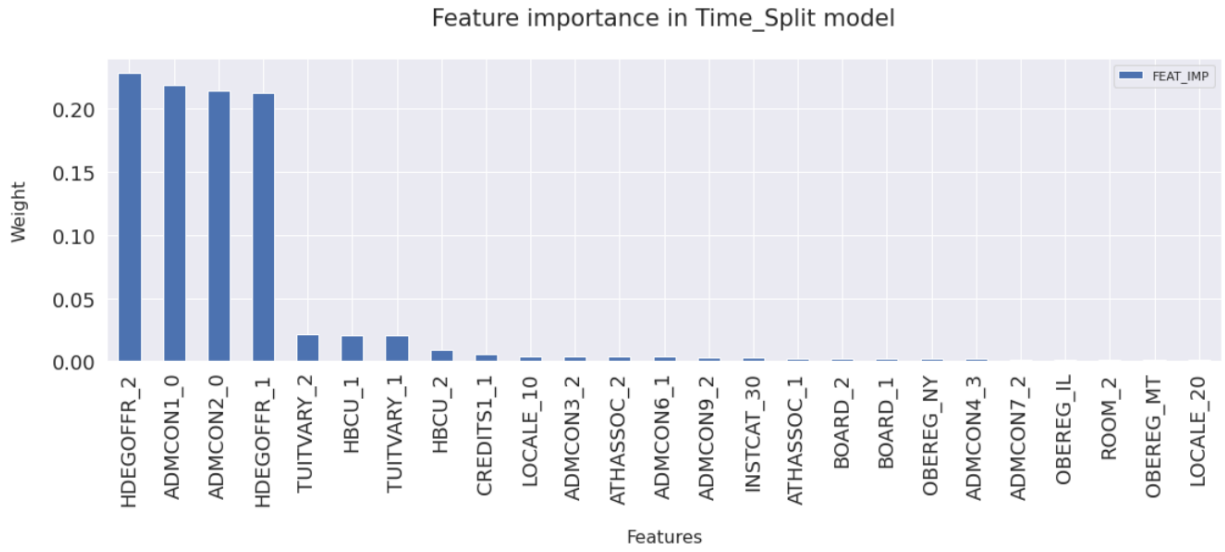
Our Root Mean Square Error (RMSE) for this size of HEIs ranges between about (1854 and 2074), which is not too bad but indicates a fair amount of variability within the data set. The model's training accuracy ranges between (56.77% and 62.7%) and the testing accuracy between (54.19% and 60.54%), which is a significant improvement from our previous group of HEIs. The models seems to do a better job at predicting enrollment levels based on the features, and taking a look at the scatter plot of predicted versus actual data points we can see that the cloud of points does tend to align itself more closely around our 45 degree line (indicating better

performance of our model) however, it also shows the variability of the data by not displaying a particularly tight cloud.

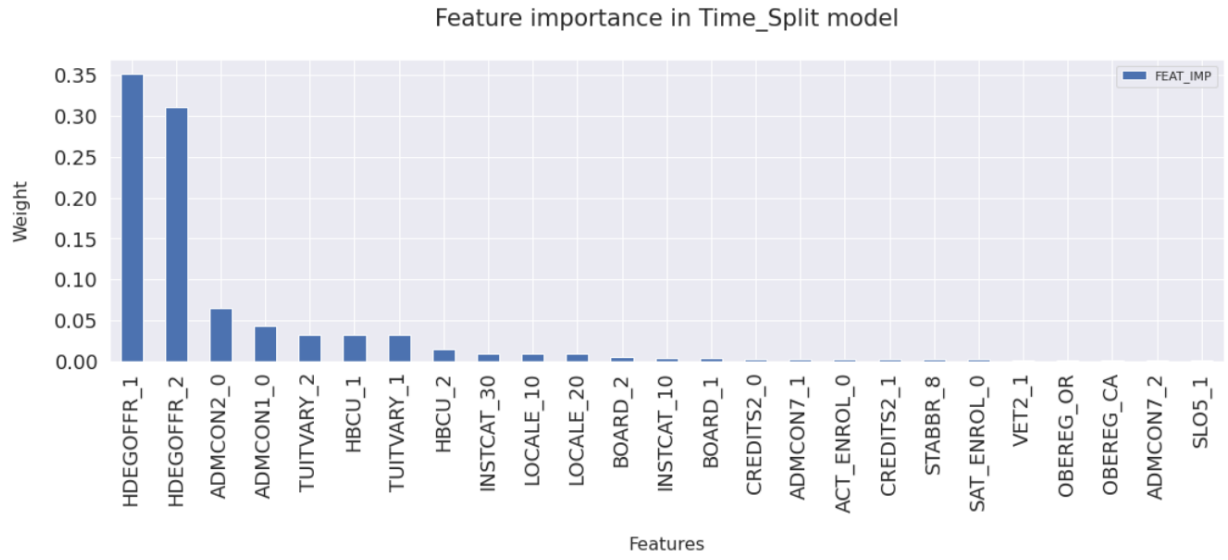
Taking a look at these mid-sized institution, the features' importance is quite different. The presence and offering of graduate level programs versus the lack of the presence of graduate level programs both appear to affect enrollment significantly, almost cancelling each other out, as does the presence of different tuition plans for in State and out of State students. This indicates that variety of institutions present in this subgroup. Most of these HEIs appear to be chosen based on their location (urban and suburban), they do not offer boarding services, and do not use school rankings or high school GPAs as a leverage to increase enrollment.

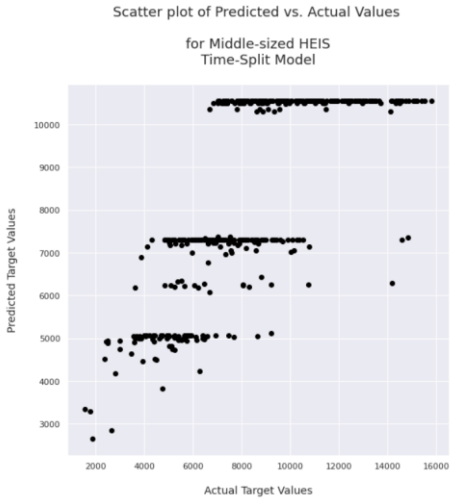
Being an HBCU seems to be an important feature, which could imply higher minority students' enrollment in this group a characteristic that could be reinforced by their location in urban and suburban settings. Finally, for this group admission test scores still seem to play a role. However, this feature appears both as a positive and a negative, suggesting that there may be a shift starting to occur on the role of these tests on enrollment.

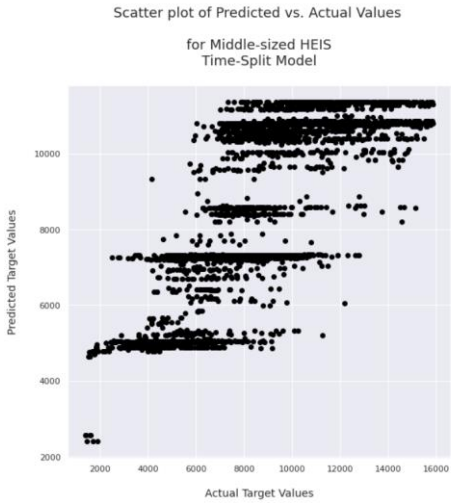
First Split (smallest amount of data) for the Middle sized HEIs



Last Split (largest amount of data) for the Middle sized HEIs







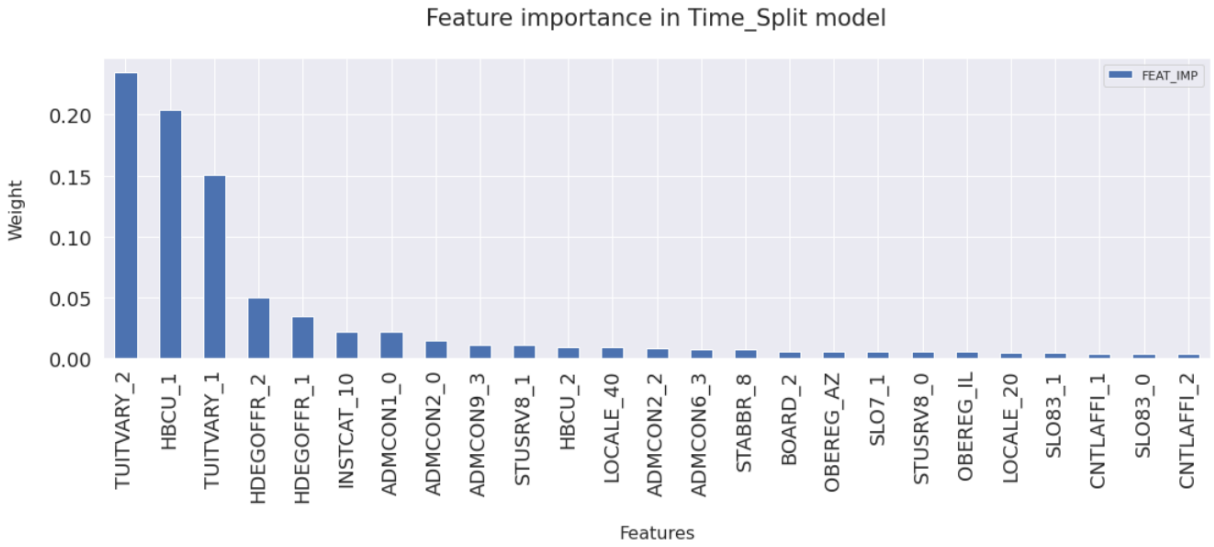
Second to Smallest-sized HEIs

Our Root Mean Square Error (RMSE) for this size of HEIs ranges between about (1289 and 1395), which is not too bad. The data points are more closely clustered together although there is still variability within the group. The model's training accuracy ranges between (41.28% and 47.84%) and the testing accuracy between (35.56% and 46.82%). Both models perform worst than for the previous groups of HEIs. This lack of improvement may be connected to the large variety of institutions with differing characteristics present in the group.

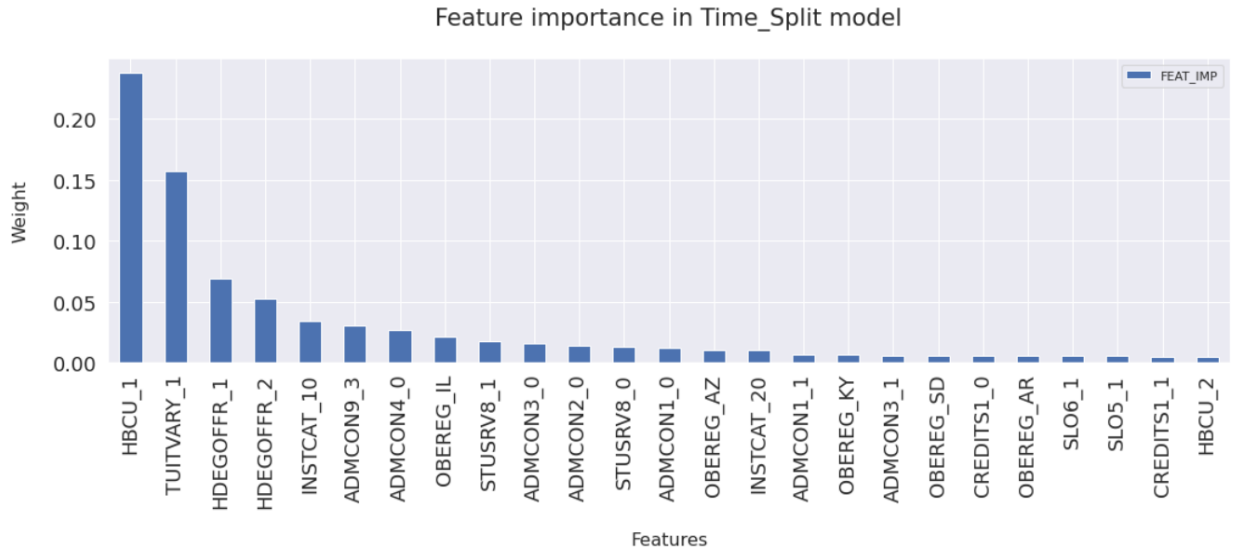
Judging from our scatter plot of predicted versus actual data points we see that predictions seem to be consistently underestimated. The presence of outlier data and variability of the HEIs affects our results. These results could also indicate a more significant lack of influence on the part of traditionally reported features (and obviously those we have selected for our analysis) on enrollment decision choices on the part of students.

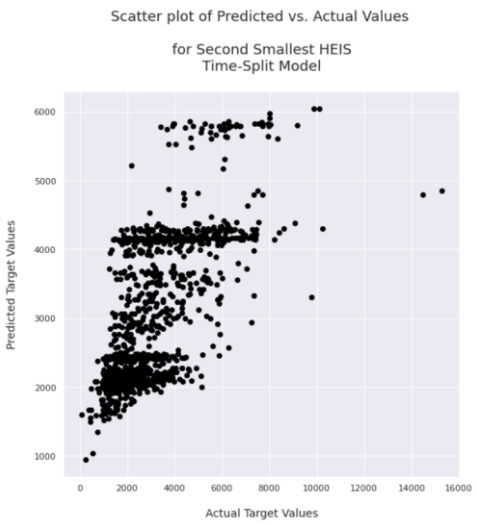
The features' importance indicates that students enrolling in this group of HEIs tend to be close to the schools they join. The feature HBCU indicates that this is an attractive feature for enrollment, which could indicate higher numbers of minority students enrolling in this type of institution. The presence of both graduate and undergraduate programs seems to matter less than they do for larger institutions. However, given the fact that the presence of graduate programs tend to affect the size of the HEIs, this feature still influences the levels of enrollment probably indirectly through confounding variables, such as name recognition of the HEIs. Furthermore, these institutions seem to accept alternative tests to SATs and ACTs, which could also indicate the presence of minorities in higher numbers. Interestingly here the HEIs location seems to be relatively important, as it emerges as among the top features affecting high enrollment numbers. This is an interesting find that would need further attention as it could indicate the relative importance given to education in these States/areas or the relative greater numerosity of this type of institution in certain areas of the continental USA.

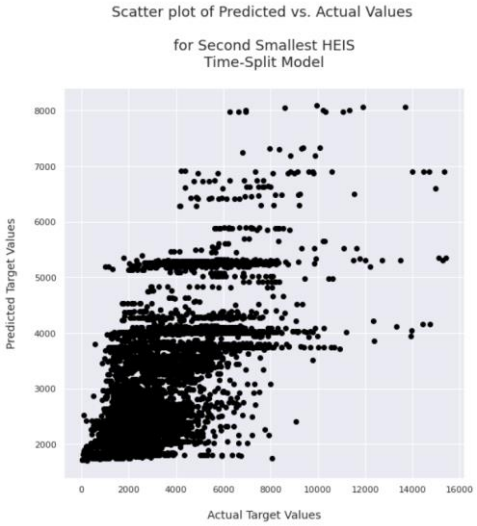
Smallest split of the data pertaining to second to smallest sized HEIs



Largest split of the data pertaining to second to smallest sized HEIs







Smallest HEIs

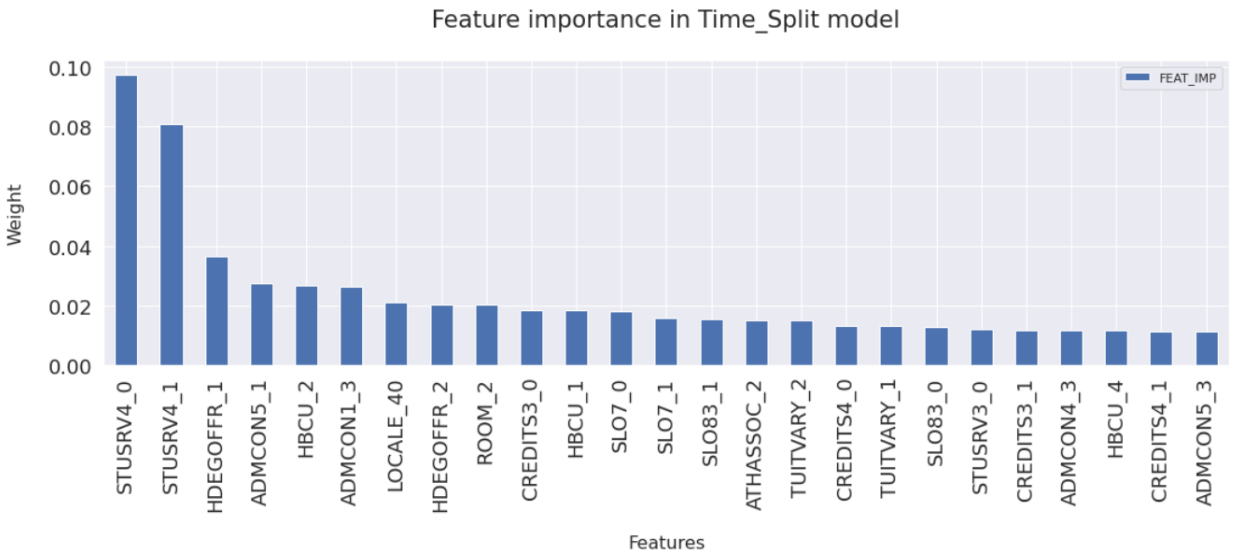
Finally, our Root Mean Square Error (RMSE) for the smallest HEIs of all, ranges between about (330 and 495). This group has the smallest RMSE however, given that they also represent the group with the lowest enrollment levels, this reduction may be misleading. The data points are more closely clustered together and there is a lot less variability within the group. The model's training accuracy however ranges between (24.3% and 38.3%) and the testing accuracy between (12.36% and 28.7%), which is the worst performance yet and indicate that for this group, the predictions tend to underestimate actual enrollment levels more consistently across the board.

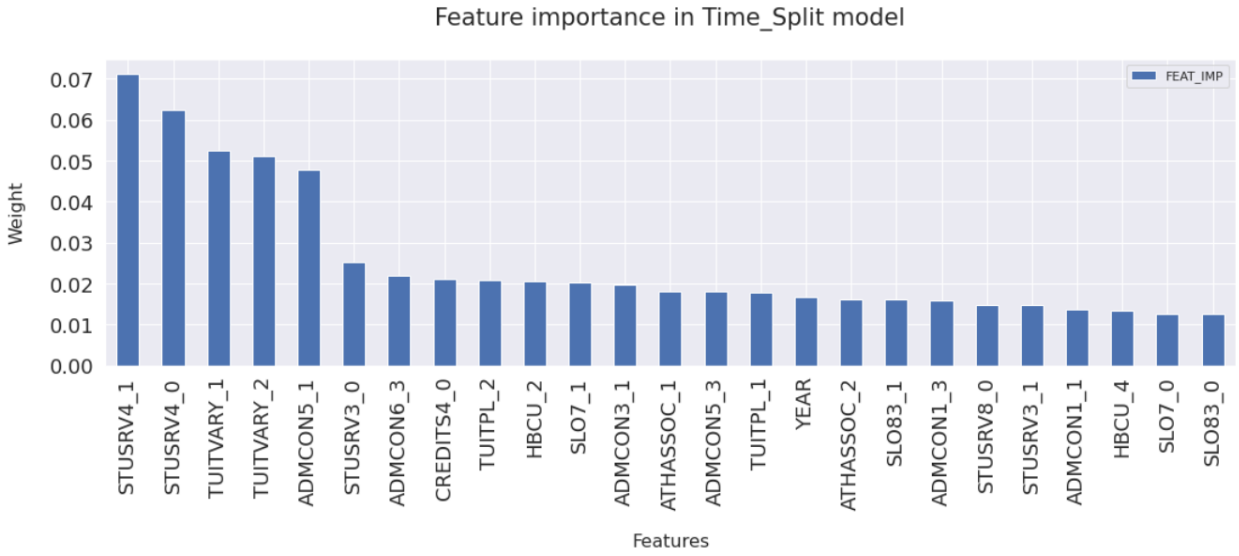
One of the reasons for our model's poor performance for this group of HEIs may be linked to the fact that the features whose influence we are analyzing maybe inadequate to explain student enrollment for this type of institution. In fact, reading these results in parallel to the fact that this group of institutions also seem not to be as traditionally aligned with academics, could help explain the relatively poor results.

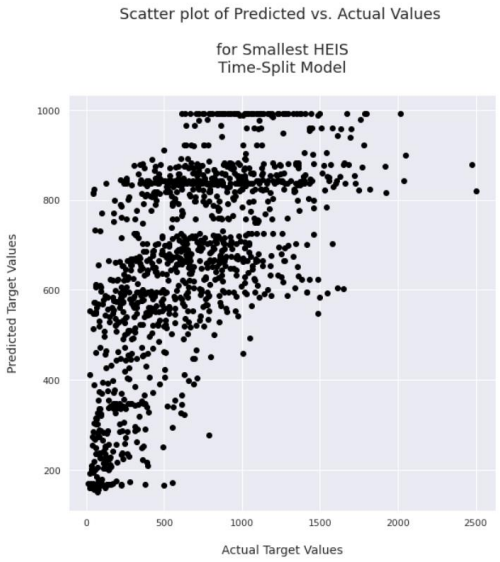
In any case, features' importance indicates that students enrolling in this group of HEIs are influenced by the presence of student services that tend to be geared towards an older population or a working population, such as placement services, the acceptance of alternative forms of competency, and the offer of weekend and evening classes. All features that support us looking at this group as non-traditionally academic-oriented.

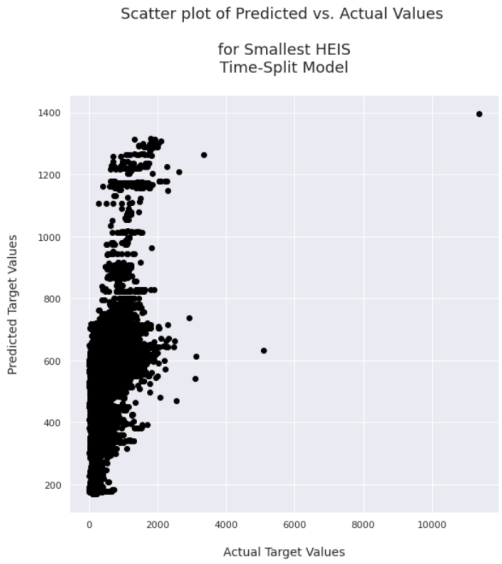
Finally, and surprisingly the year appears as a feature of relevance for our predictions. This would seem to indicate that there is a lot more variability in the enrollment levels across time for this group of HEIs, making predictions harder and thus model performance worse.

This may be due to the intrinsic nature of the group of HEIs , however, it also may be due to inconsistencies in the data collection which lead to questions of data reliability and correctness.







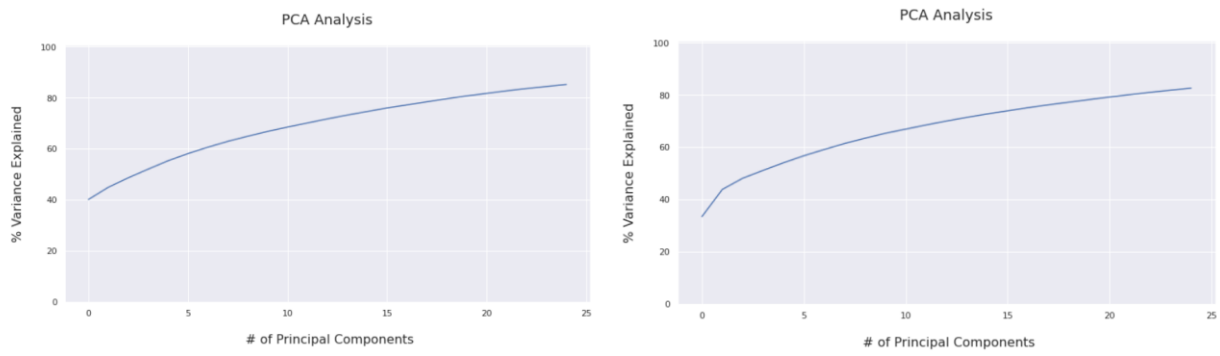


Pricipal Component Analysis on Mid-sized HEIs

As a final push to try to identify better results and also confirm results across different ML algorithms, I applied PCA analysis to the Middel HEIs grouping.

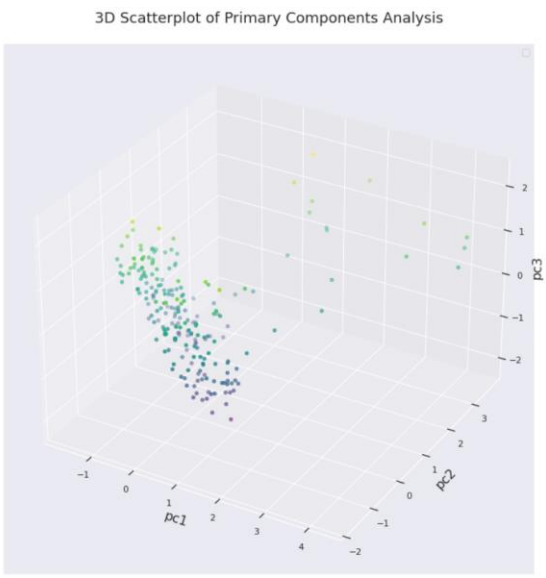
Only taking all of the 25 principal components used to build our regression model do we reach a level of explained variance similar to what we have achieved from our data before – see below.

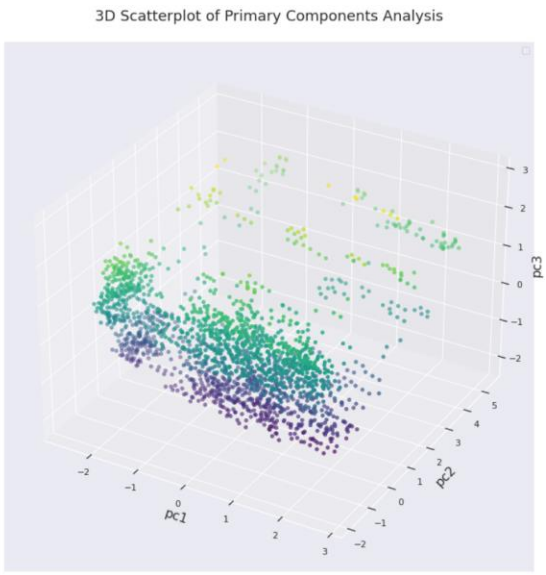
Explained variance by our PCA components on Middle Sized HEIs all together (split n1 – left - and n 9 – right - compared)



However, taking a look at the clustering of these data points based on the 3 first PCAs, we can clearly see that there are two clusters of data that emerge and become more apparent with each split of the time series (split n 1 - left - vs split n9 – right - below) and that within each of the two clusters there seem to be three more uniform aggregations that emerge.

This could be supporting our previous results, as indicated from our feature importance results achieved via XGBoosting Regression. In fact when we subdivide out of our Middle-sized HEIs throup of those offering only undergraduate programs from the rest, we can clearly see that the clusters go down to one (see image below), supporting our results so far. Our principal component analysys results go down but this only indicates how relatively important the single features are in explaining our target, so that would be expected. and also the data within clusters is not homogenous.





Conclusions

Publicly disclosed information on the part of HEIs is extensive in nature however it is also somewhat inaccurate and keeping it up to date is challenging, especially for the wide number of smaller HEIs making up the majority of institutions present over the US territory. The consequence in terms of data analysis is significant because, as we all know, analytical conclusions are only as good as the data that is used to pursue them. This said, IPDES data does allow to see new and interesting trends emerging from the characteristics of our HEIs and it allows us to offer at least preliminary answers to our research questions.

First and foremost, our analysis does support the conclusion that different features characterizing HEIs have different impact on enrollment levels. As mentioned, segmenting the institutions based on size, however, seems necessary to better capture which features have the most impact on which type of institution.

While this may appear as an obvious result, it is important to keep in mind and underline, to support policy decision makers and the public's better understanding of the impact of interventions in higher education. Each cluster of HEI based on size appears to address diverse educational needs or at least "education consumers' needs"; each cluster is characterized by different long-term objectives; and each cluster is controlled by mostly different entities. Interventions "for higher education" must keep this in mind.

Across all HEI sizes there have been changes in the influence different feature seem to have in terms of enrollment levels across the most recent years. Some of these changes are a direct consequence of changes in policy decisions made at the Federal level (an example of this is the greater emphasis placed on opening HEIs up to greater equity in education through interventions aimed at supporting enrollment by non-white students). However, not all of the changes seem to be truly implemented with an objective of long term public benefit in mind, rather, they may be

more reactive in nature and follow changes in enrollment patterns - both in terms of numbers and in terms of student racial composition- instead. This notwithstanding there seems to be a transformation occurring within HEIs that is reflecting a broader societal change occurring within the continental USA. Greater variety in the elements that appear to influence enrollment is occurring. The long-term question is: will these measures be sufficient to support the continuous growth or at least the permanence of all segments of HEIs.

The largest HEIs appear to be challenged in terms of how to maintain enrollment levels high, but so are the segments of the smallest HEIs who suddenly find themselves competing for students who in the past were not enrolling in more traditional academic institutions.

HEIs within each cluster compete with the other institutions within the group to attract student enrollments but as data shows, this competition is spilling over across groups. IPDES' different features data collection help identify some of the differences existing between clusters of HEIs however, as we saw the lines within the groups are not as clearly defined as they were in the past. This information can be usefully applied to help HEIs identify changes that may help them target their enrollment strategies better. - Certainly additional information could enhance our understanding of the leverages open to HEIs in this respect.

Overall, our first two research questions (a) do reported features affect student enrollment; and b) with an eye towards IDEA, do different HEIs fare differently) have been addressed and answered affirmatively. Regarding our third question, further research is recommended and additional data – beyond the data shared by IPDES – would be necessary to reach greater clarity. However, definitely there seems to be an increasingly smaller role being played by standardized tests in student enrollments, but not as much in traditionally academic institutions. The cause of this shift seems to be driven by the need to boost enrollments but it also may be driven by shifts in policy priorities started back with the Obama Administration and geared towards boosting enrollment by students from different social-economic backgrounds.

References:

Data:

All of the variable choices are spread across various data sheets publicly available and can be found https://nces.ed.gov/ipeds/datacenter/DataFiles.aspx?gotoReportId=7&fromIpeds=true

National Student Research Clearinghouse Research Center (NSCRC) https://nscresearchcenter.org/stay-informed/

Articles:

Ashwin Raj Unlocking the True Power of Support Vector Regression. Using Support Vector Machine for Regression Problems Oct 3, 2020 https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0

Two Decades of Change in Federal and State Higher Education Funding. Recent trends across levels of government. Issue Brief October 15, 2019. Projects: Fiscal Federalis. Phillip Oliff Project Director Student Loan Research https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2019/10/two-decades-of-change-in-federal-and-state-higher-education-funding

Gagnon-Bartsch J.A., Sales A.C., Wu E., Botelho A. F., Erickson J.A., Miratrix L.W., Heffernan N.T. (Submitted on 7 May 2021), Precise Unbiased Estimation in Randomized Experiments using Auxiliary Observational Data, Cornell University. https://arxiv.org/abs/2105.03529

Jon Marcus, The Hechinger Report Most Americans don't realize state funding for higher ed fell by billions. Education Feb 26, 2019 12:20 PM EDT https://www.pbs.org/newshour/education/most-americans-dont-realize-state-funding-for-higher-ed-fell-by-billions

State Higher Education Funding Cuts Have Pushed Costs to Students, Worsened Inequality October 24, 2019 | By Michael Mitchell, Michael Leachman and Matt Saenz https://www.cbpp.org/research/state-budget-and-tax/state-higher-education-funding-cuts-have-pushed-costs-to-students

Zhang Q. (2019). A Class of Association Measures for Categorical Variables Based on Weighted Minkowski Distance. *Entropy*, *21*(10), 990. https://doi.org/10.3390/e21100990

Web links:

https://towardsdatascience.com/the-python-glob-module-47d82f4cbd2d#:~:text=glob%20(short%20for%20global)%20is,pattern%20by%20using%20wildcard%20characters

https://machinelearningmastery.com/autoregression-models-time-series-forecasting-python/#:~:text=Autoregression%20is%20a%20time%20series,range%20of%20time%20series%20problems.

https://towardsdatascience.com/detecting-stationarity-in-time-series-data-d29e0a21e638

https://www.statology.org/dickey-fuller-test-python/

https://machinelearningmastery.com/time-series-data-stationary-python/

https://people.duke.edu/~rnau/411arim.htm

https://datascience.stackexchange.com/questions/54138/how-can-time-series-analysis-be-done-with-categorical-variables

https://datascience.stackexchange.com/questions/57341/regression-methods

https://towardsdatascience.com/advanced-time-series-analysis-in-python-decomposition-autocorrelation-115aa64f475e

https://machinelearningmastery.com/feature-selection-with-categorical-data/

https://github.com/shakedzy/dython

https://towardsdatascience.com/the-search-for-categorical-correlation-a1cf7f1888c9

https://stats.stackexchange.com/questions/443878/when-to-use-theils-u-and-cramers-v-the-danger-of-symmetrical-data

https://towardsdatascience.com/an-introduction-to-logistic-regression-for-categorical-data-analysis-7cabc551546c

https://medium.com/@curryrowan/simplified-logistic-regression-classification-with-categorical-variables-in-python-1ce50c4b137

NOTE: additional web links are included in the notebooks associated with this research

Appendix

Variable lists pre-selection

Purely for exemplary purposes, below is a list of the variables included in the dictionary files for a specific year for each data file that was used for the project. The lists are to be considered expemplary because, as already stated, there was some discrepancy across files and across years in terms of the actual variables included in each final file. All discrepancies were corrected by hand.

Institutional Characteristics 'ic' dictionary file 2020

	varname	imputationvar	varTitle
1	UNITID		Unique identification number of the institution
2	PEO1ISTR		Occupational
3	PEO2ISTR		Academic
4	PEO3ISTR		Continuing professional
5	PEO4ISTR		Recreational or avocational
6	PEO5ISTR		Adult basic remedial or high school equivalent
7	PEO6ISTR		Secondary (high school)
8	CNTLAFFI		Institutional control or affiliation

9	PUBPRIME		Primary public control
10	PUBSECON		Secondary public control
11	RELAFFIL		Religious affiliation
12	LEVEL1		Certificate of less than 1 year
13	LEVEL1A		Certificate of less than 12 weeks
14	LEVEL1B		Certificate of at least 12 weeks, but less than 1 year
15	LEVEL2		Certifiicate of at least 1 year, but less than 2 years
16	LEVEL3		Associate's degree
17	LEVEL4		Certificate of at least 2 years, but less than 4 years
18	LEVEL5		Bachelor's degree
19	LEVEL6		Postbaccalaureate certificate
20	LEVEL7		Master's degree
21	LEVEL8		Post-master's certificate
22	LEVEL12		Other degree
23	LEVEL17		Doctor's degree - research/scholarship
24	LEVEL18		Doctor's degree - professional practice
25	LEVEL19		Doctor's degree - other
26	CALSYS		Calendar system
27	FT_UG		Full-time undergraduate students are enrolled
28	FT_FTUG		Full time first-time degree/certificate-seeking undergraduate students enrolled
29	FTGDNIDP		Full-time graduate (not including doctor's professional practice) students are enrolled
30	PT_UG		Part-time undergraduate students are enrolled
31	PT_FTUG		Part time first-time degree/certificate-seeking undergraduate students enrolled
32	PTGDNIDP		Part-time graduate (not including doctor's professional practice) students are enrolled
33	DOCPP		Doctor's professional practice students are enrolled
34	DOCPPSP		Doctor's professional practice students are enrolled in programs formerly designated as first-professional
35	OPENADMP		Open admission policy
36	VET1		Yellow Ribbon Program (officially known as Post-9/11 GI Bill, Yellow Ribbon Program)
37	VET2		Credit for military training
38	VET3		Dedicated point of contact for support services for veterans, military servicemembers, and their families
39	VET4		Recognized student veteran organization
40	VET5		Member of Servicemembers Opportunity Colleges
41	VET9		Services and programs are not available to veterans, military servicemembers, or their families?
42	CREDITS1		Dual enrollment
43	CREDITS2		Credit for life experiences
44	CREDITS3		Advanced placement (AP) credits

45	CREDITS4		Institution does not accept dual, credit for life, or AP credits
46	**SLO5**		ROTC
47	SLO51		ROTC - Army
48	SLO52		ROTC - Navy
49	SLO53		ROTC - Air Force
50	**SLO6**		Study abroad
51	SLO7		Weekend/evening college
52	**SLO8**		Teacher certification (below the postsecondary level)
53	SLO81		Teacher certification: Students can complete their preparation in certain areas of specialization
54	SLO82		Teacher certification: Students must complete their preparation at another institution for certain areas of specialization
55	**SLO83**		Teacher certification: Approved by the state for initial certifcation or licensure of teachers.
56	SLO9		None of the above special learning opportunities are offered
57	YRSCOLL		Years of college-level work required
58	STUSRV1		Remedial services
59	STUSRV2		Academic/career counseling service
60	STUSRV3		Employment services for students
61	STUSRV4		Placement services for completers
62	STUSRV8		On-campus day care for students' children
63	STUSRV9		None of the above selected services are offered
64	LIBRES1		Library resources/services: Physical facilities
65	LIBRES2		Library resources/services: Organized collection of printed materials
66	LIBRES3		Library resources/services: Access to digital/electronic resources
67	LIBRES4		Library resources/services: Staff trained to provide and interpret library materials
68	LIBRES5		Library resources/services: Established library hours
69	LIBRES6		Library resources/services: Access to library collections that are shared with other institutions
70	LIBRES9		Library resources/services not provided
71	TUITPL		Any alternative tuition plans offered by institution
72	TUITPL1		Tuition guaranteed plan
73	TUITPL2		Prepaid tuition plan
74	TUITPL3		Tuition payment plan
75	TUITPL4		Other alternative tuition plan
76	DSTNUGC		Undergraduate level distance education courses offered
77	DSTNUGP		Undergraduate level distance education programs offered
78	DSTNUGN		Undergraduate level distance education not offered
79	DSTNGC		Graduate level distance education courses offered
80	DSTNGP		Graduate level distance education programs offered

81	DSTNGN		Graduate level distance education not offered
82	DISTCRS		Distance education courses offered
83	DISTPGS		Distance education programs offered
84	DSTNCED1		Undergraduate level programs or courses are offered via distance education
85	DSTNCED2		Graduate level programs or courses are offered via distance education
86	DSTNCED3		Does not offer distance education opportunities
87	DISTNCED		All programs offered completely via distance education
88	DISAB		Percent indicator of undergraduates formally registered as students with disabilities
89	DISABPCT	XDISABPC	Percent of undergraduates, who are formally registered as students with disabilities, when percentage is more than 3 percent
90	ALLONCAM		Full-time, first-time degree/certificate-seeking students required to live on campus
91	TUITVARY		Tuition charge varies for in-district, in-state, out-of-state students
92	ROOM		Institution provide on-campus housing
93	ROOMCAP	XROOMCAP	Total dormitory capacity
94	BOARD		Institution provides board or meal plan
95	MEALSWK	XMEALSWK	Number of meals per week in board charge
96	ROOMAMT	XROOMAMT	Typical room charge for academic year
97	BOARDAMT	XBORDAMT	Typical board charge for academic year
98	RMBRDAMT	XRMBDAMT	Combined charge for room and board
99	APPLFEEU	XAPPFEEU	Undergraduate application fee
100	APPLFEEG	XAPPFEEG	Graduate application fee
101	ATHASSOC		Member of National Athletic Association
102	ASSOC1		Member of National Collegiate Athletic Association (NCAA)
103	ASSOC2		Member of National Association of Intercollegiate Athletics (NAIA)
104	ASSOC3		Member of National Junior College Athletic Association (NJCAA)
105	ASSOC4		Member of National Small College Athletic Association (NSCAA)
106	ASSOC5		Member of National Christian College Athletic Association (NCCAA)
107	ASSOC6		Member of other national athletic association not listed above
108	SPORT1		NCAA/NAIA member for football
109	CONFNO1		NCAA/NAIA conference number football
110	SPORT2		NCAA/NAIA member for basketball
111	CONFNO2		NCAA/NAIA conference number basketball
112	SPORT3		NCAA/NAIA member for baseball
113	CONFNO3		NCAA/NAIA conference number baseball

114	SPORT4		NCAA/NAIA member for cross country/track
115	CONFNO4		NCAA/NAIA conference number cross country/track

Institutional Characteristics hd dictionary file 2020

	varname	**imputationvar**	**varTitle**
1	UNITID		Unique identification number of the institution
2	INSTNM		Institution (entity) name
3	IALIAS		Institution name alias
4	ADDR		Street address or post office box
5	CITY		City location of institution
6	STABBR		State abbreviation
7	ZIP		ZIP code
8	FIPS		FIPS state code
9	OBEREG		Bureau of Economic Analysis (BEA) regions
10	CHFNM		Name of chief administrator
11	CHFTITLE		Title of chief administrator
12	GENTELE		General information telephone number
13	EIN		Employer Identification Number
14	DUNS		Dun and Bradstreet numbers
15	OPEID		Office of Postsecondary Education (OPE) ID Number
16	OPEFLAG		OPE Title IV eligibility indicator code
17	WEBADDR		Institution's internet website address
18	ADMINURL		Admissions office web address
19	FAIDURL		Financial aid office web address
20	APPLURL		Online application web address
21	NPRICURL		Net price calculator web address
22	VETURL		Veterans and Military Servicemembers tuition policies web address
23	ATHURL		Student-Right-to-Know student athlete graduation rate web address
24	DISAURL		Disability Services Web Address
25	SECTOR		Sector of institution
26	ICLEVEL		Level of institution
27	CONTROL		Control of institution
28	HLOFFER		Highest level of offering
29	UGOFFER		Undergraduate offering
30	GROFFER		Graduate offering
31	HDEGOFR1		Highest degree offered
32	DEGGRANT		Degree-granting status
33	HBCU		Historically Black College or University

34	HOSPITAL		Institution has hospital
35	MEDICAL		Institution grants a medical degree
36	TRIBAL		Tribal college
37	LOCALE		Degree of urbanization (Urban-centric locale)
38	OPENPUBL		Institution open to the general public
39	ACT		Status of institution
40	NEWID		UNITID for merged schools
41	DEATHYR		Year institution was deleted from IPEDS
42	CLOSEDAT		Date institution closed
43	CYACTIVE		Institution is active in current year
44	POSTSEC		Primarily postsecondary indicator
45	PSEFLAG		Postsecondary institution indicator
46	PSET4FLG		Postsecondary and Title IV institution indicator
47	RPTMTH		Reporting method for student charges, graduation rates, retention rates and student financial aid
48	INSTCAT		Institutional category
49	C18BASIC		Carnegie Classification 2018: Basic
50	C18IPUG		Carnegie Classification 2018: Undergraduate Instructional Program
51	C18IPGRD		Carnegie Classification 2018: Graduate Instructional Program
52	C18UGPRF		Carnegie Classification 2018: Undergraduate Profile
53	C18ENPRF		Carnegie Classification 2018: Enrollment Profile
54	C18SZSET		Carnegie Classification 2018: Size and Setting
55	C15BASIC		Carnegie Classification 2015: Basic
56	CCBASIC		Carnegie Classification 2005/2010: Basic
57	CARNEGIE		Carnegie Classification 2000
58	LANDGRNT		Land Grant Institution
59	INSTSIZE		Institution size category
60	F1SYSTYP		Multi-institution or multi-campus organization
61	F1SYSNAM		Name of multi-institution or multi-campus organization
62	F1SYSCOD		Identification number of multi-institution or multi-campus organization
63	CBSA		Core Based Statistical Area (CBSA)
64	CBSATYPE		CBSA Type Metropolitan or Micropolitan
65	CSA		Combined Statistical Area (CSA)
66	NECTA		New England City and Town Area (NECTA)
67	COUNTYCD		Fips County code
68	COUNTYNM		County name
69	CNGDSTCD		State and 114TH Congressional District ID
70	LONGITUD		Longitude location of institution
71	LATITUDE		Latitude location of institution
72	DFRCGID		Data Feedback Report comparison group created by NCES

73	DFRCUSCG		Data Feedback Report - Institution submitted a custom comparison group

12-Month- Enrollment effy dictionary file 2020

	varname	**imputationvar**	**varTitle**
1	UNITID		Unique identification number of the institution
2	EFFYALEV		Level and degree/certificate-seeking status of student
3	EFFYLEV		Undergraduate or graduate level of student
4	LSTUDY		Original level of study on survey form
5	EFYTOTLT	XEYTOTLT	Grand total
6	EFYTOTLM	XEYTOTLM	Grand total men
7	EFYTOTLW	XEYTOTLW	Grand total women
8	EFYAIANT	XEFYAIAT	American Indian or Alaska Native total
9	EFYAIANM	XEFYAIAM	American Indian or Alaska Native men
10	EFYAIANW	XEFYAIAW	American Indian or Alaska Native women
11	EFYASIAT	XEFYASIT	Asian total
12	EFYASIAM	XEFYASIM	Asian men
13	EFYASIAW	XEFYASIW	Asian women
14	EFYBKAAT	XEFYBKAT	Black or African American total
15	EFYBKAAM	XEFYBKAM	Black or African American men
16	EFYBKAAW	XEFYBKAW	Black or African American women
17	EFYHISPT	XEFYHIST	Hispanic or Latino total
18	EFYHISPM	XEFYHISM	Hispanic or Latino men
19	EFYHISPW	XEFYHISW	Hispanic or Latino women
20	EFYNHPIT	XEFYNHPT	Native Hawaiian or Other Pacific Islander total
21	EFYNHPIM	XEFYNHPM	Native Hawaiian or Other Pacific Islander men
22	EFYNHPIW	XEFYNHPW	Native Hawaiian or Other Pacific Islander women
23	EFYWHITT	XEFYWHIT	White total
24	EFYWHITM	XEFYWHIM	White men
25	EFYWHITW	XEFYWHIW	White women
26	EFY2MORT	XEFY2MOT	Two or more races total
27	EFY2MORM	XEFY2MOM	Two or more races men
28	EFY2MORW	XEFY2MOW	Two or more races women
29	EFYUNKNT	XEYUNKNT	Race/ethnicity unknown total
30	EFYUNKNM	XEYUNKNM	Race/ethnicity unknown men
31	EFYUNKNW	XEYUNKNW	Race/ethnicity unknown women
32	EFYNRALT	XEYNRALT	Nonresident alien total
33	EFYNRALM	XEYNRALM	Nonresident alien men
34	EFYNRALW	XEYNRALW	Nonresident alien women

Admissions and test Scores dictionary file 2020

	varname	imputationvar	varTitle
1	UNITID		Unique identification number of the institution
2	ADMCON1		Secondary school GPA
3	ADMCON2		Secondary school rank
4	ADMCON3		Secondary school record
5	ADMCON4		Completion of college-preparatory program
6	ADMCON5		Recommendations
7	ADMCON6		Formal demonstration of competencies
8	ADMCON7		Admission test scores
9	ADMCON8		TOEFL (Test of English as a Foreign Language
10	ADMCON9		Other Test (Wonderlic, WISC-III, etc.)
11	APPLCN	XAPPLCN	Applicants total
12	APPLCNM	XAPPLCNM	Applicants men
13	APPLCNW	XAPPLCNW	Applicants women
14	ADMSSN	XADMSSN	Admissions total
15	ADMSSNM	XADMSSNM	Admissions men
16	ADMSSNW	XADMSSNW	Admissions women
17	ENRLT	XENRLT	Enrolled total
18	ENRLM	XENRLM	Enrolled men
19	ENRLW	XENRLW	Enrolled women
20	ENRLFT	XENRLFT	Enrolled full time total
21	ENRLFTM	XENRLFTM	Enrolled full time men
22	ENRLFTW	XENRLFTW	Enrolled full time women
23	ENRLPT	XENRLPT	Enrolled part time total
24	ENRLPTM	XENRLPTM	Enrolled part time men
25	ENRLPTW	XENRLPTW	Enrolled part time women
26	SATNUM	XSATNUM	Number of first-time degree/certificate-seeking students submitting SAT scores
27	SATPCT	XSATPCT	Percent of first-time degree/certificate-seeking students submitting SAT scores
28	ACTNUM	XACTNUM	Number of first-time degree/certificate-seeking students submitting ACT scores
29	ACTPCT	XACTPCT	Percent of first-time degree/certificate-seeking students submitting ACT scores
30	SATVR25	XSATVR25	SAT Evidence-Based Reading and Writing 25th percentile score
31	SATVR75	XSATVR75	SAT Evidence-Based Reading and Writing 75th percentile score
32	SATMT25	XSATMT25	SAT Math 25th percentile score
33	SATMT75	XSATMT75	SAT Math 75th percentile score
34	ACTCM25	XACTCM25	ACT Composite 25th percentile score

35	ACTCM75	XACTCM75	ACT Composite 75th percentile score
36	ACTEN25	XACTEN25	ACT English 25th percentile score
37	ACTEN75	XACTEN75	ACT English 75th percentile score
38	ACTMT25	XACTMT25	ACT Math 25th percentile score
39	ACTMT75	XACTMT75	ACT Math 75th percentile score

List of Final Variables Selected

1 'EFYTOTLT' = Our target variable is total enrollment

2 'YEAR' = Our data refers to the years between 2012 and 2020

3 'UNITID' = ID code for each Heigher Education Institution

4 'STABBR' = State Abbreviation

5 'INSTNM' = Institution's name

6 'OBEREG'= Code indicating the Bureau of Economic Analysis Regions the US is divided into (MD is in region 2)
- 0 - US Service schools
- 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 - (Not in the Continental USA)

7 'HDEGOFFR'= Code indicating the level of Degree offered by the Academic-Oriented HEI
- 11 Doctor's degree - research/scholarship and professional practice
- 12 Doctor's degree - research/scholarship
- 13 Doctor's degree - professional practice
- 14 Doctor's degree - other
- 20 Master's degree
- 30 Bachelor's degree
- 40 Associate's degree

8 'GROFFER' = Code indicating whether the HEI offers Graduate Level degrees
- 1 Graduate degree or certificate offering
- 2 No graduate offering

9 'HBCU' = Code indicating whether the HEI is an Historical Black College or University
- 1 - Yes
- 2 - No

10 'LOCALE' = Code indicating the territorial location of the HEI
- 11 City: Large
- 12 City: Midsize
- 13 City: Small
- 21 Suburb: Large
- 22 Suburb: Midsize
- 23 Suburb: Small
- 31 Town: Fringe
- 32 Town: Distant
- 33 Town: Remote
- 41 Rural: Fringe
- 42 Rural: Distant
- 43 Rural: Remote

11 'INSTCAT' = Code indicating the type of institution
- 1 - indicates institutions offering undergraduate programs
- 2 - indicates institutions offering both graduate and undergraduate programs

12 'INSTSIZE'= Code indicating the range of students enrollable by year
- 1 Under 1,000
- 2 1,000 - 4,999
- 3 5,000 - 9,999
- 4 10,000 - 19,999
- 5 20,000 and above

13 'CNTLAFFI'= Code indicating the type of control the HEI is subject to
- 1 Public
- 2 Private for-profit
- 3 Private not-for-profit (no religious affiliation)
- 4 Private not-for-profit (religious affiliation)

14 'OPENADMP' = Code indicating whether the HEI adopts an Open Enrollment Policy or not
- 1 = Yes
- 2 = No

15 'CREDITS1'= Dual enrollment

16 'CREDITS2'= Credit for life experiences

17 'CREDITS3'= Advanced placement (AP) credits

18 'CREDITS4'= Institution does not accept dual, credit for life, or AP credits
- Code 0 = No
- Code 1 = Yes

19 'SLO5' = ROTC

20 'SLO6' = Study abroad

21 'SLO7' = Weekend/evening college

22 'SLO83' = Teacher certification: Approved by the state for initial certifcation or licensure of teachers
Code 0 = No
Code 1 = Yes

23 'STUSRV1 Remedial services

24 STUSRV2 Academic/career counseling service

25 STUSRV3 Employment services for students

26 STUSRV4 Placement services for completers

27 STUSRV8 On-campus day care for students' children
Code 0 = No
Code 1 = Yes

28 'LIBRES1' = Variable indicating whether the HEI has a library
Code 0 = No
Code 1 = Yes

29 'ATHASSOC' = Member of National Athletic Association
1 - Yes
2 - No

30 'APPLFEEU' = Application Fee

31 'TUITVARY' = Tuition charge varies for in-district, in-state, out-of-state students
1 - Yes
2 - No

32 'ROOM' = HEI provides housing
1 - Yes
2 - No

33 'BOARD', = HEI provides meal plan
1 - Yes
2 - No

34 'TUITPL' = Any alternative tuition plans offered by institution
1 - Yes
2 - No

35 VET1 Yellow Ribbon Program (officially known as Post-9/11 GI Bill, Yellow Ribbon Program)

36 VET2 Credit for military training

37 VET3 Dedicated point of contact for support services for veterans, military servicemembers, and their families

38 VET4 Recognized student veteran organization

39 VET5 Member of Servicemembers Opportunity Colleges
Code 0 = No
Code 1 = Yes

NOTE: Data for these VET variables are missing for years 2012 and 2013

40 'DSTNUGC' = Undergraduate Level distance Education Courses Offered

41 DSTNUGP Undergraduate level distance education programs offered
Code 0 = No
Code 1 = Yes

42 'RMANDBRDAMT' = Room and Board costs

43 ADMCON1 Secondary school GPA

44 ADMCON2 Secondary school rank

45 ADMCON3 Secondary school record

46 ADMCON4 Completion of college-preparatory program

47 ADMCON5 Recommendations

48 ADMCON6 Formal demonstration of competencies

49 ADMCON7 Admission test scores

50 ADMCON8 TOEFL (Test of English as a Foreign Language

51 ADMCON9 Other Test (Wonderlic, WISC-III, etc.)
0 Not Applicable
1 Required
2 Recommended
3 Neither required nor recommended
4 Do not know
5 Considered but not required

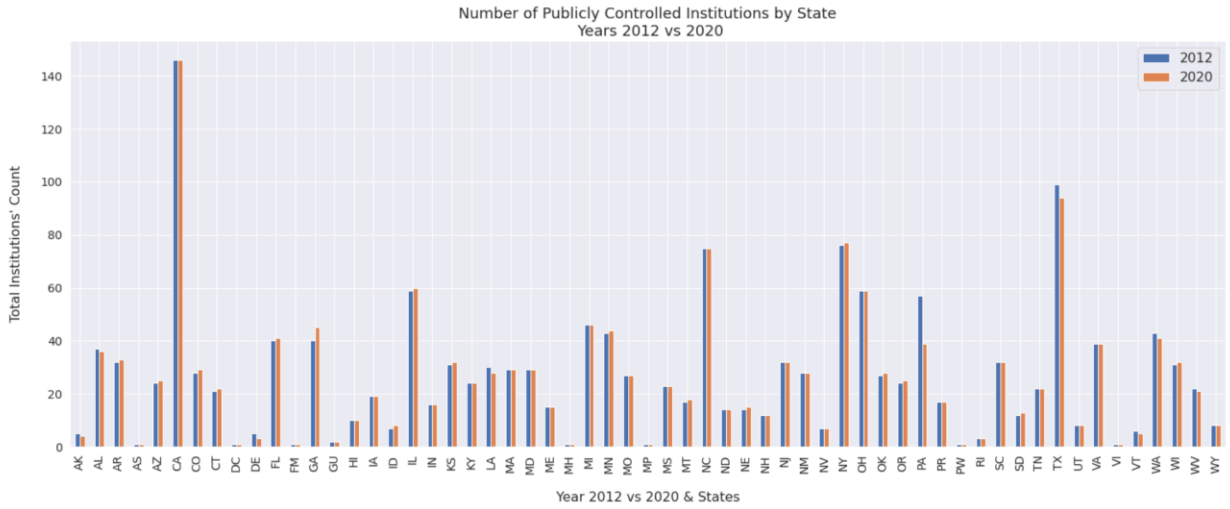
52 SATNUM Number of first-time degree/certificate-seeking students submitting SAT scores

53 ACTNUM Number of first-time degree/certificate-seeking students submitting ACT scores

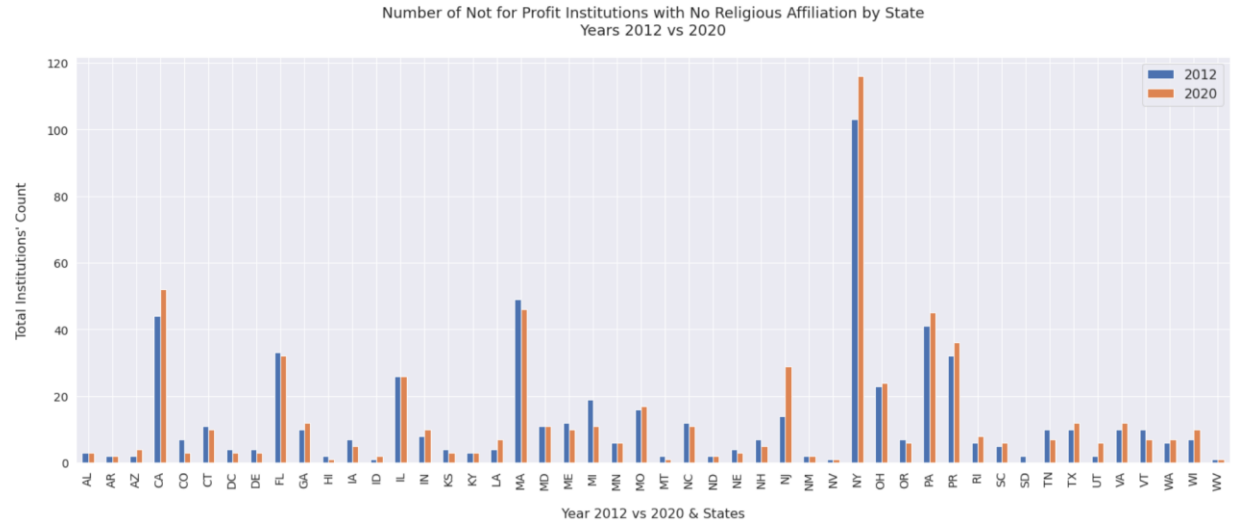
54 EFFYTOT Number of total students aggregately enrolled in undergraduate programs for each year in the Fall of the year

Additional Graphs and Tables

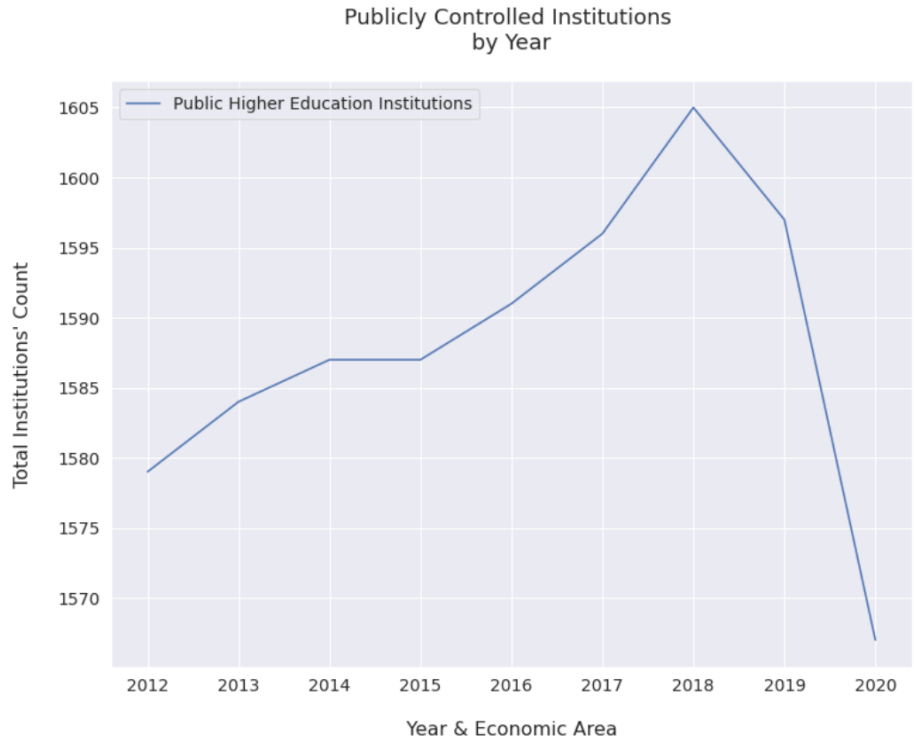
Change in number of Publicly Controlled HEIs between 2012 and 2020



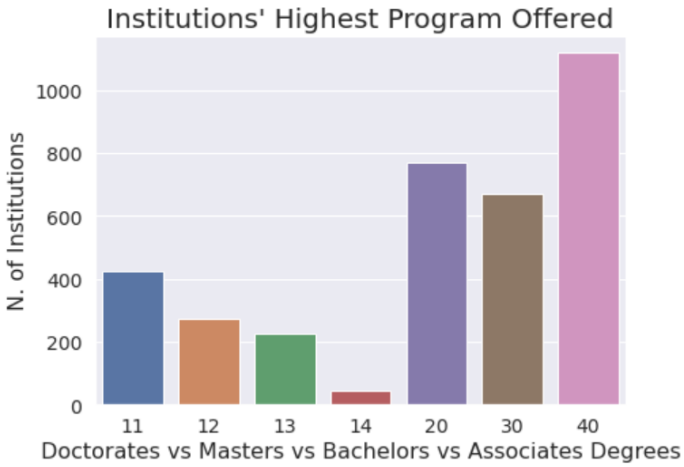
Change in number of Not for Profit HEIs between 2012 and 2020



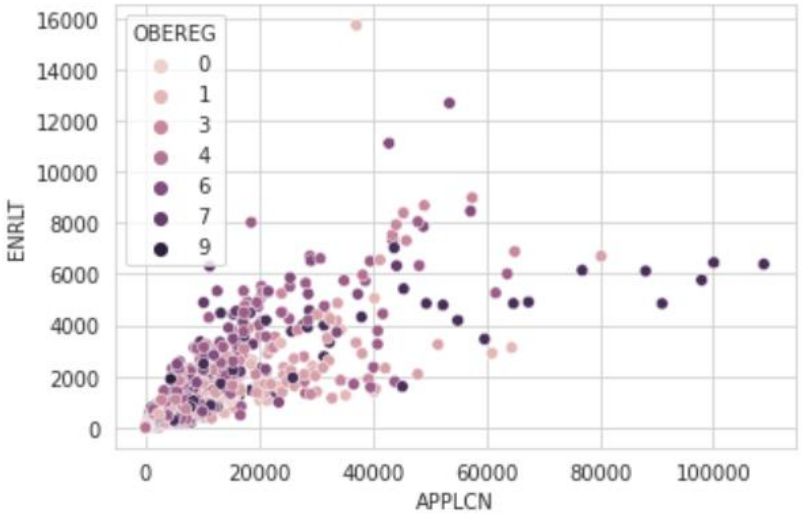
Publicly controlled institutions have also contracted in number, however this is a relatively more recent phenomenon having only started in 2018 – it definitely warrants greater scrutiny but may be due in part to HEIs restructuring and reorganizations given the overall limited numbers involved or in a delay in reporting.



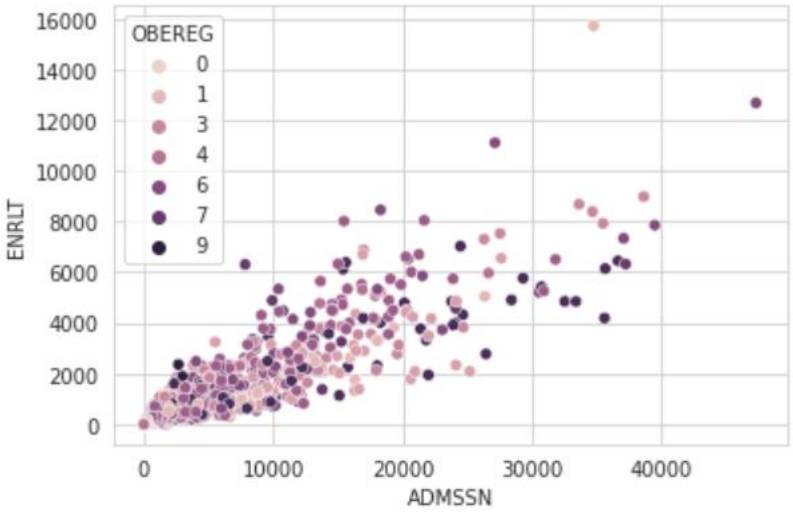
HEIs offering lower level degrees (identified with codes 20,30, and 40 and representing – in that order – Masters, Bachelors and Associate level degrees) compose the majority of our data, whereas Doctorate level degrees (here identified with codes in the teens) aggregately seem to match the numbers of Associate level degrees at least for 2020.
(NOTE: These are not the focus of our analysis)



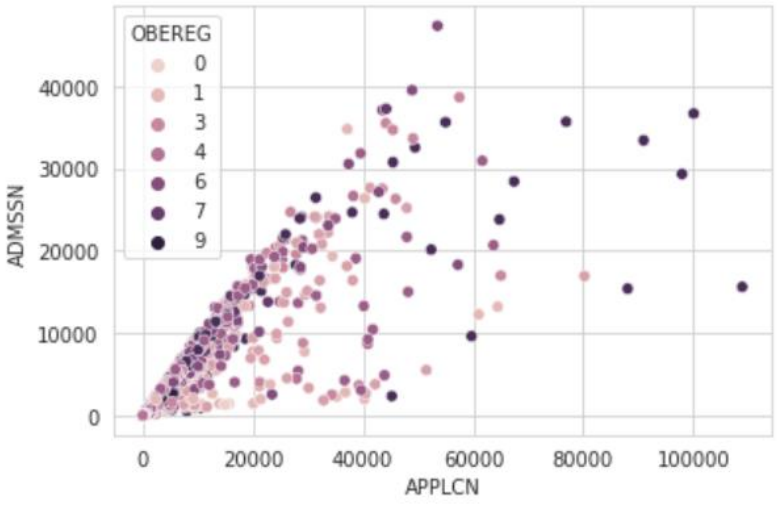
Total Number of Applications versus Total Enrollment by Economic Area (2020 Data only)



Total Number of Admissions versus Total Enrollment by Economic Area (2020 Data only)

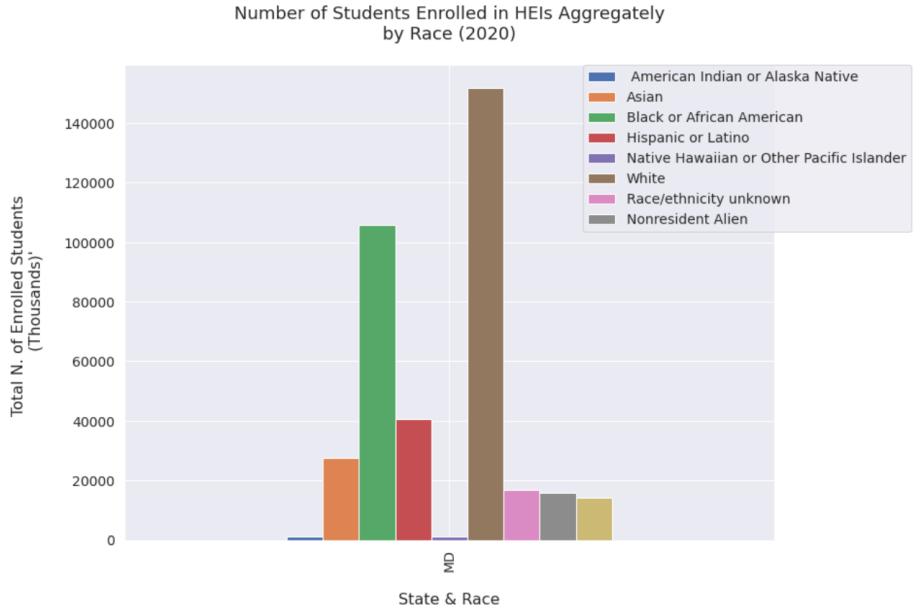


Total Number of Applications versus Total Admissions by Economic Area (2020 Data only)

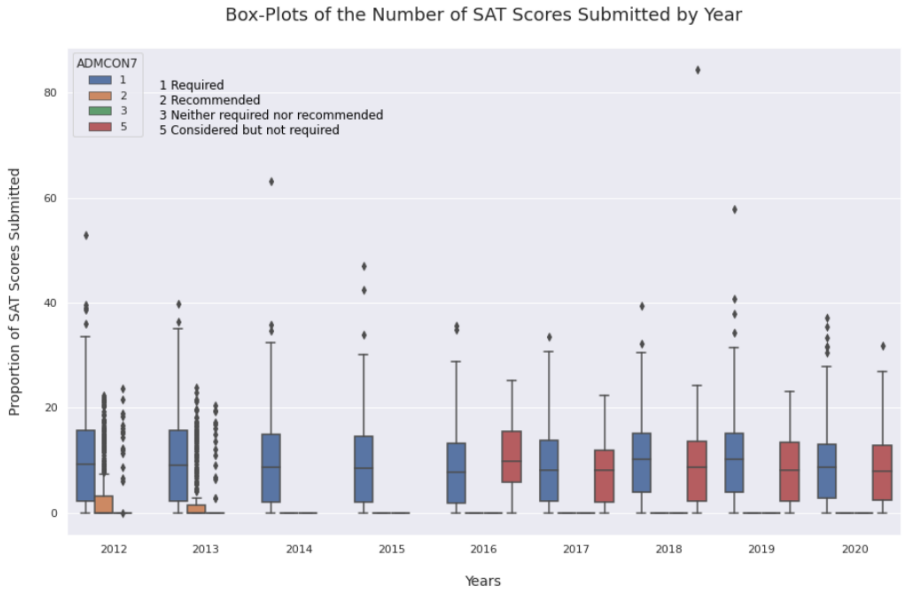


Undergraduate Enrollment in Maryland 2019-2020 data – by Race

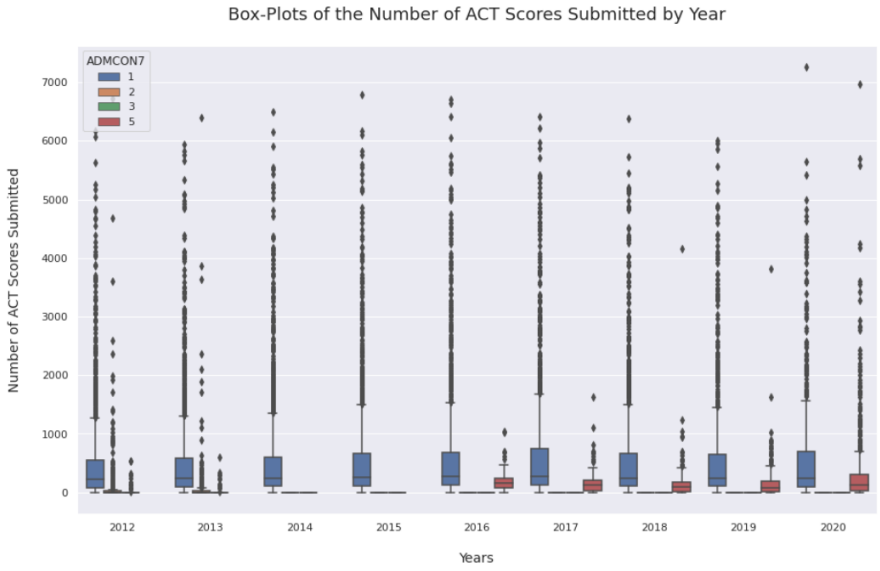
When we focus on student enrollment, aggregately in 2020, the distribution of undergraduate students across the different races still shows a disproportionate presence of White students when compared to each of the other race groups individually, practically consistently across all States. Yet, aggregating the data we can see a more even split between students who identify as White versus non-White students, if not a change in trend with non-White students starting to become more numerous, as Maryland data confirms (below).



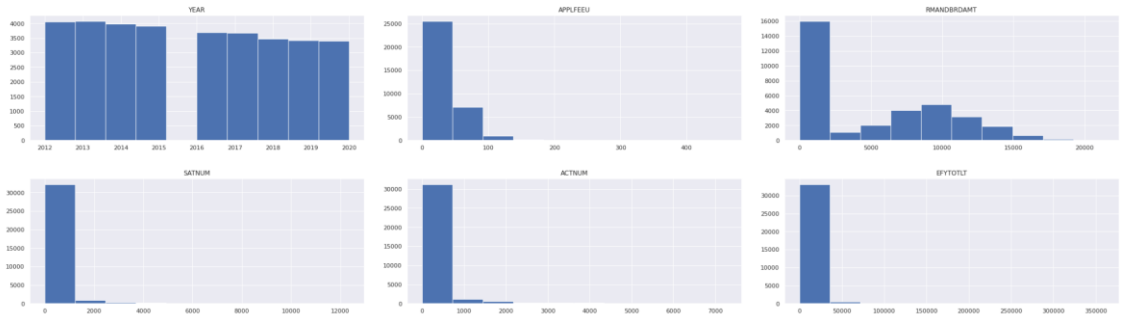
Number of SAT Scores submitted by Year for HEIs having some form of SAT reporting requirement



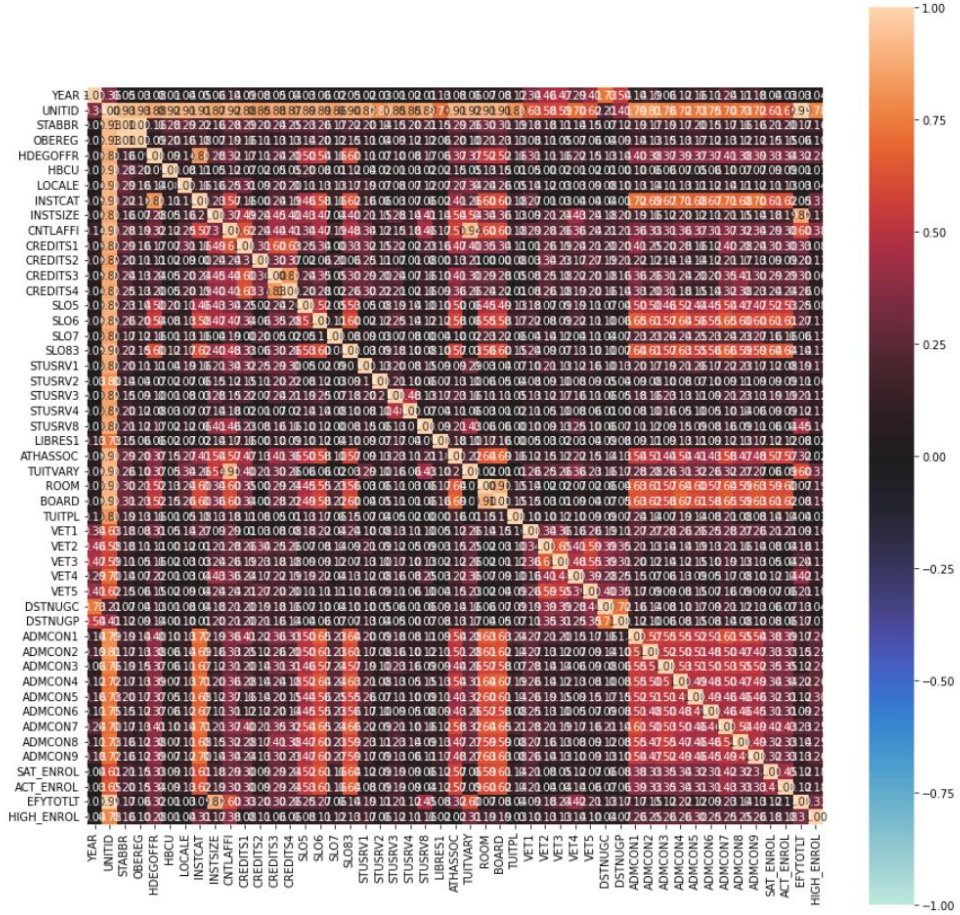
Number of ACT Scores submitted by Year for HEIs having some form of SAT reporting requirement



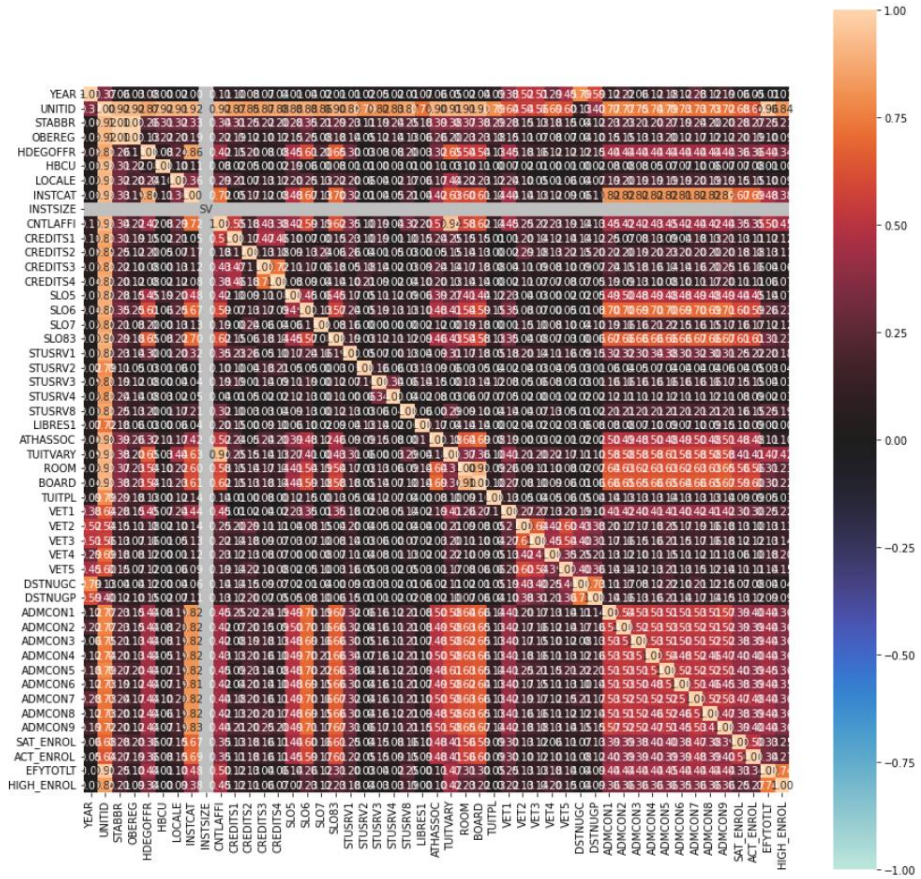
Distribution of the numerical variables for the whole data set



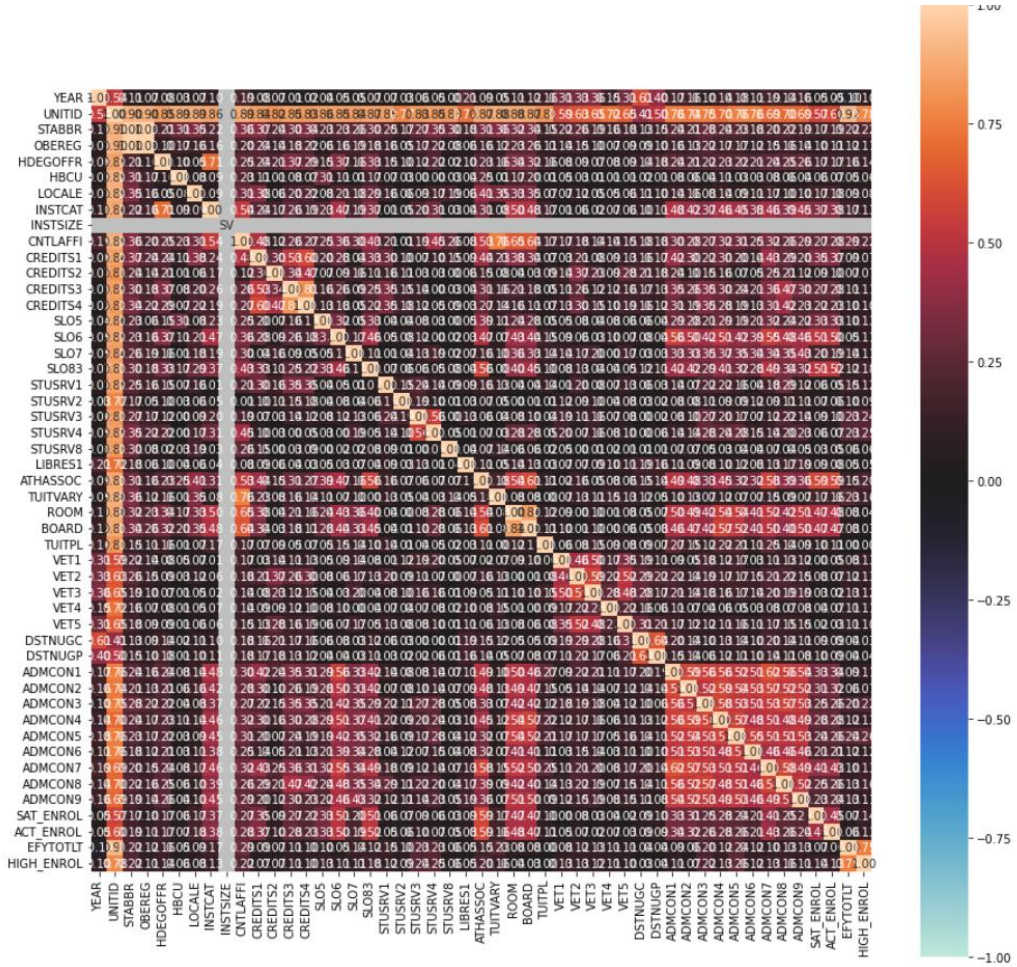
Categorical Features' Association levels aggregately



Associations of HEIs second to smallest in size



Associations of HEIs smallest in size



Pipeline used to process the data prior to using our XGBoost Regressor