**DETERMING THE COLLEGE SCORECARD:**

**Our Objective:**

Determine a Score for each college in the United States, based on various features. College Scorecards make it easier for students to search for a college that may be a good fit for them.

They can use the College Scorecard to find out more about a college's affordability and value so they can make more informed decisions about which college to attend.

**Data Overview:**

The dataset has been retrieved from US Government Site. These data are provided through federal reporting from institutions, data on federal financial aid, and tax information. These data provide insights into the performance of institutions that receive federal financial aid dollars, and the outcomes of the students of those institutions. A complete set of these data for all active Integrated Postsecondary Education Data System (IPEDS) Title IV non-administrative unit institutions are available on the Scorecard data webpage and API1. A subset of these data elements is displayed on the consumer-facing College Scorecard website for the subset of currently operating institutions whose predominant undergraduate awards are:

• Associate degrees

• Bachelor’s degrees

• Certificates (for only those institutions offering and awarding Associate and/or Bachelor’s degrees).

Many data elements are drawn directly from, or derived from, data reported to the IPEDS. 2 Note that some colleges report combined data that pertain to more than one IPEDS institution. The reporting institution (called the “parent”) also indicates an estimated proportion of the data that applies to each of the institutions for which it is reporting (called “child” institutions). In the Scorecard data, these proportions are used to allocate data reported by parent institutions to the child institutions to simplify both consumer and research use of the data.

Data Source: <https://catalog.data.gov/dataset/college-scorecard>

**Project Client:**

For this project, we have assumed that the US government wants to develop an application that can give the college applicants a good idea about how much value they can get on their investment. The score will be assigned based on features such as '**Affordability**' and '**Number of Placements done for the Alumni**' etc.

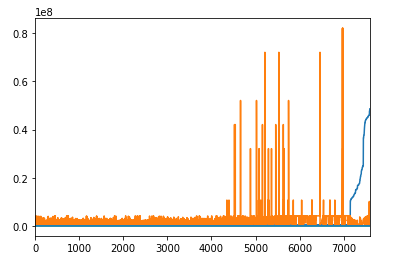
**Initial Assumptions Made:**

### Our dataset contains **295028** Null values. We assume that the data is missing completely at random (i.e., the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random. When data are MCAR, the analysis performed on the data is unbiased; however, data are rarely MCAR).

**Data Preprocessing:**

Our dataset contains 7593 rows where each row represents details of a particular college (in U.S.). Also our dataset contains 123 columns in which we have filtered out 108 dimensions that help us in determining the overall value of that particular college. Our dataset contains 295028 Null values. If I attempt to drop all rows from our dataset then we are left with only 13 dimensions or columns which is a massive loss of information which is unacceptable for any further model that we build on this dataset. Hence, in order to remedy this situation, we use Linear Interpolation of missing data.

Following is the graphical plot signifying the Linear Interpolation of Missing plot:



The spikes in the data represent outliers in our dataset (which we have dealt in the next section).

**Dealing with Outliers in our Dataset:**

After computing the distance (in terms of Standard Deviation) from the mean for every column we find that following are the columns that have outliers:

##### *1: NPT43\_PRIV: Average net price for $48,001-$75,000 family income (private for-profit and nonprofit institutions)*

##### *2: NPT44\_PRIV: Average net price for $75,001-$110,000 family income (private for-profit and nonprofit institutions)*

##### *3: NPT45\_PRIV: Average net price for $110,000+ family income (private for-profit and nonprofit institutions)*

##### *4: NPT4\_PUB: Average net price for Title IV institutions (public institutions)*

##### *5: NPT41\_PUB: Average net price for $0-$30,000 family income (public institutions)*

##### *6: NPT43\_PUB: Average net price for $48,001-$75,000 family income (public institutions)*

##### *7: NPT44\_PUB: Average net price for $75,001-$110,000 family income (public institutions)*

##### *8: NPT45\_PUB: Average net price for $110,000+ family income (public institutions)*

##### *9: NPT41\_PRIV: Average net price for $0-$30,000 family income (private for-profit and nonprofit institutions)*

##### *10: NPT42\_PRIV: Average net price for $30,001-$48,000 family income (private for-profit and nonprofit institutions)*

##### *11: NPT43\_PRIV: Average net price for $48,001-$75,000 family income (private for-profit and nonprofit institutions)*

##### *12: NPT44\_PRIV: Average net price for $75,001-$110,000 family income (private for-profit and nonprofit institutions)*

##### *13: NPT45\_PRIV: Average net price for $110,000+ family income (private for-profit and nonprofit institutions)*

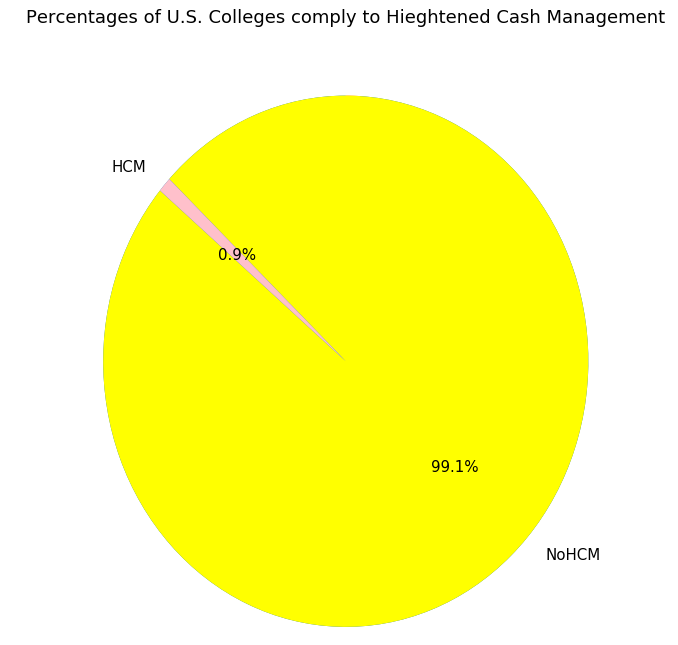
**Why we didn’t drop the outliers?**

What we observe from the dimensions (that have outliers) is that most outliers are because some colleges impose tuition fees that are far above the average mean imposed by other colleges for particular student’s income groups. In my opinion, college ranking should be impacted if the college does not provide any scholarships to low-income students. Hence, we decided that college fees charges for different income groups is a valuable information to retain and therefore we do not drop the outlier.

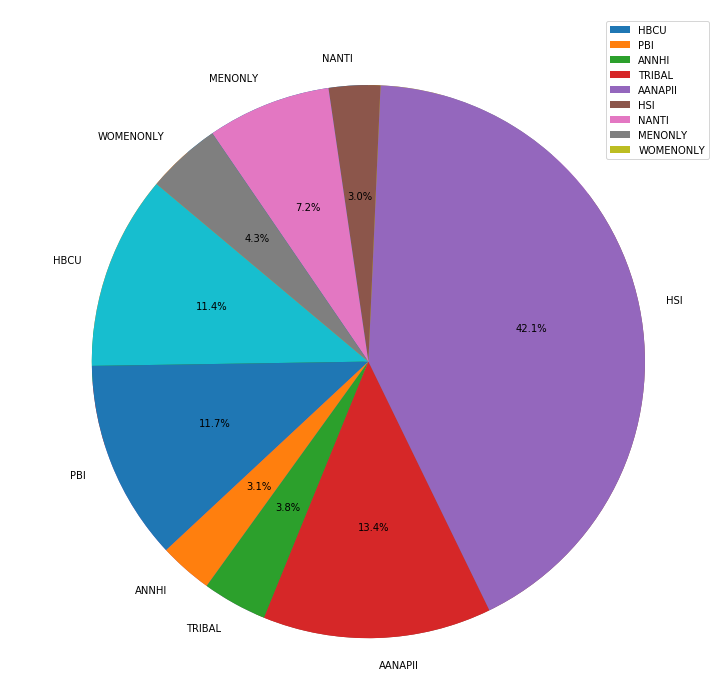
**Data Exploration:**

#### **Finding out what percent of U.S. Colleges comply to Heightened Cash Management**

Below we explore the percent of colleges that follow strict policies to deal with cash:

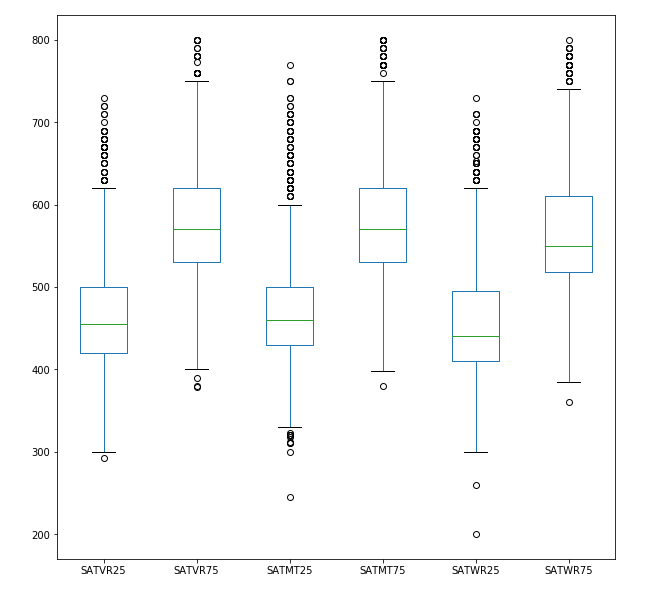


***Displaying the percentage distribution of those colleges whose populations are predominantly ethnic minorities***



##### *We observe that Hispanic Serving Institutions (HSI) are much more in percentage (42.1%) as compared to other minority colleges*

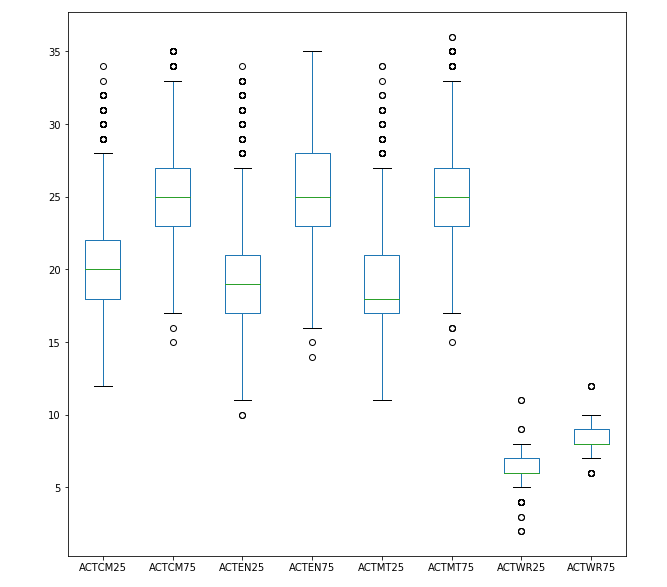
##### *Displaying the distribution of SAT scores of students in the form of box plots*



**Inference from the above box plot:**

* The lower number is for the 25th percentile of students who *enrolled* in (not just applied to) the college. For the school above, 25% of enrolled students received a math score of 590 or lower.
* The upper number is for the 75th percentile of students who enrolled in the college. For the above example, 75% of enrolled students got a math score of 740 or lower.
* For the school above, if you have an SAT math score of 745, you would be in the top 25% of applicants for that one measure. If you have a math score of 320, you are in the bottom 25% of applicants for that measure.
* Being in the bottom 25% is obviously not ideal, and any candidate’s admissions chances will be lessened, but they still have a chance of getting in. Assuming the school has [holistic admissions](https://www.thoughtco.com/what-are-holistic-admissions-788426), factors such as strong [letters of recommendation](https://www.thoughtco.com/letters-of-recommendation-788889), a [winning application essay](https://www.thoughtco.com/tips-for-winning-college-application-essay-788384), and [meaningful extracurricular activities](https://www.thoughtco.com/what-counts-as-an-extracurricular-activity-788878) can all help compensate for less-than-ideal SAT scores.
* Same idea can be applied on SATVR critical thinking.

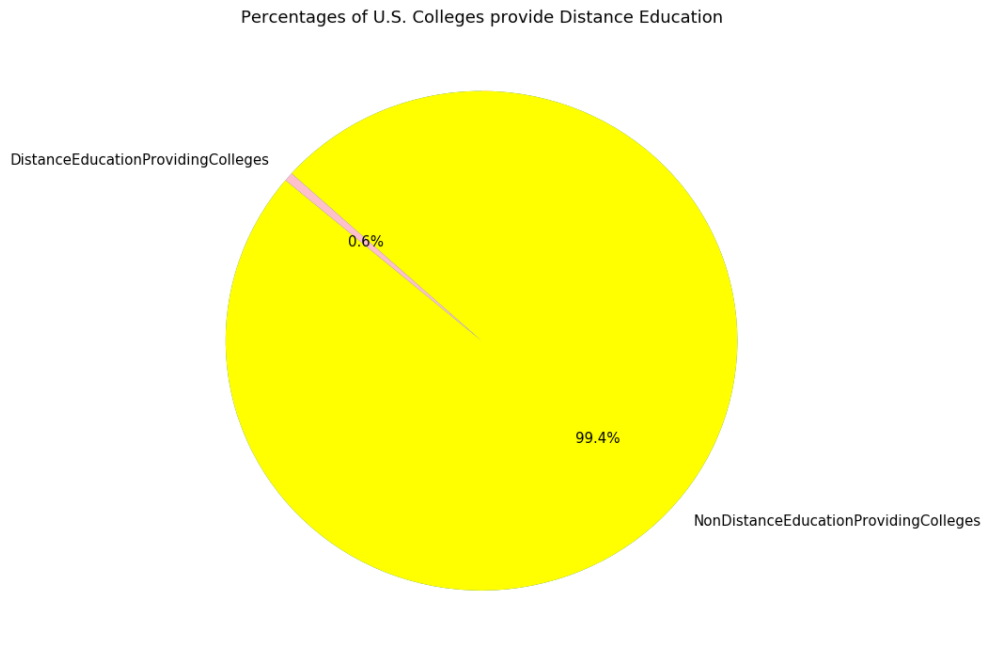
##### *Displaying the distribution of SAT scores of students in the form of box plots*



**Inference from the above box plot:**

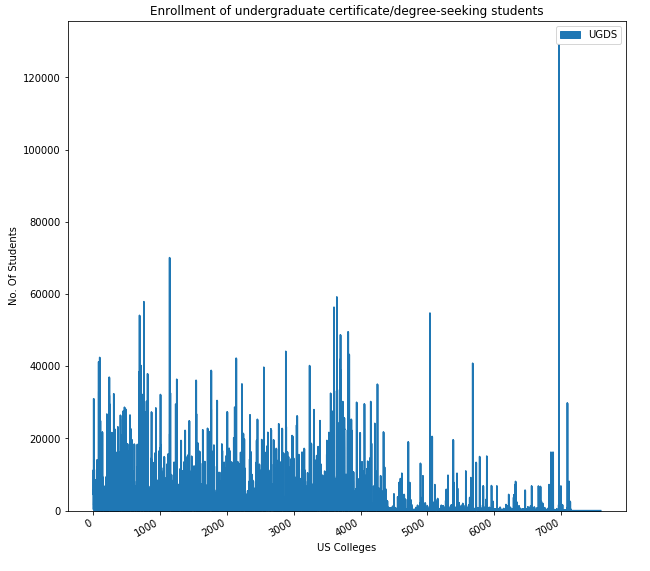
* The lower number is for the 25th percentile of students who *enrolled* in (not just applied to) the college. For the school above, 25% of enrolled students received an English score of 27 or lower.
* The upper number is for the 75th percentile of students who enrolled in the college. For the above example, 75% of enrolled students got a math score of 35 or lower.
* For the school above, if you have an ACT English score of 38, you would be in the top 25% of applicants for that one measure. If you have an ACT English score of 10, you are in the bottom 25% of applicants for that measure.
* Being in the bottom 25% is obviously not ideal, and any candidate’s admissions chances will be lessened, but they still have a chance of getting in. Assuming the school has [holistic admissions](https://www.thoughtco.com/what-are-holistic-admissions-788426), factors such as strong [letters of recommendation](https://www.thoughtco.com/letters-of-recommendation-788889), a [winning application essay](https://www.thoughtco.com/tips-for-winning-college-application-essay-788384), and [meaningful extracurricular activities](https://www.thoughtco.com/what-counts-as-an-extracurricular-activity-788878) can all help compensate for less-than-ideal SAT scores.
* Same idea can be applied on ACTCM, ACTMT and ACTWR.

##### *Displaying what percent of U.S. Colleges provide Distance Education*



As observed, 0.6% of the colleges support Distance Education

##### *Displaying the area distribution of enrollment of undergraduate certificate/degree-seeking students*

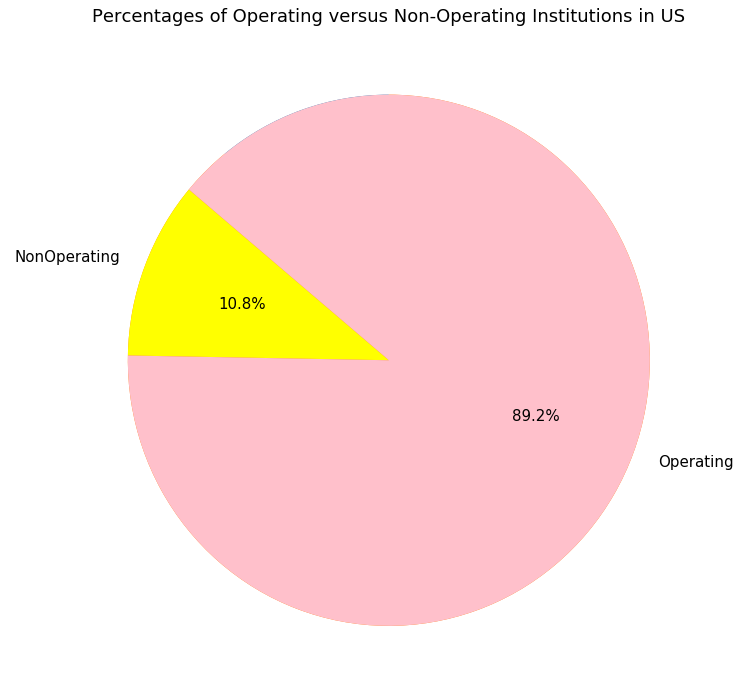


##### In the above area chart, I have plotted the area distribution of undergraduate certificate/degree-seeking students.

#### **Displaying the percentage distribution of various ethnicities for all undergraduate degree seeking students.**

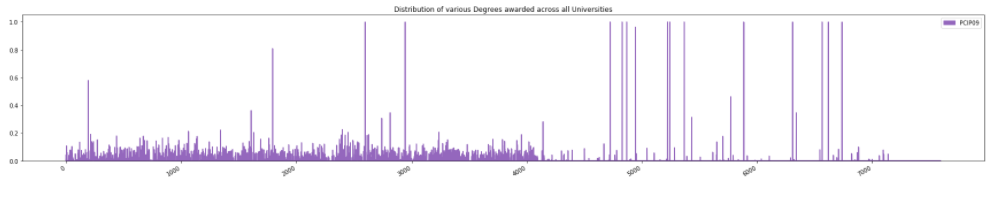
##### *Among undergraduate degree seeking students, white ethnic population is leading with 56.5%*

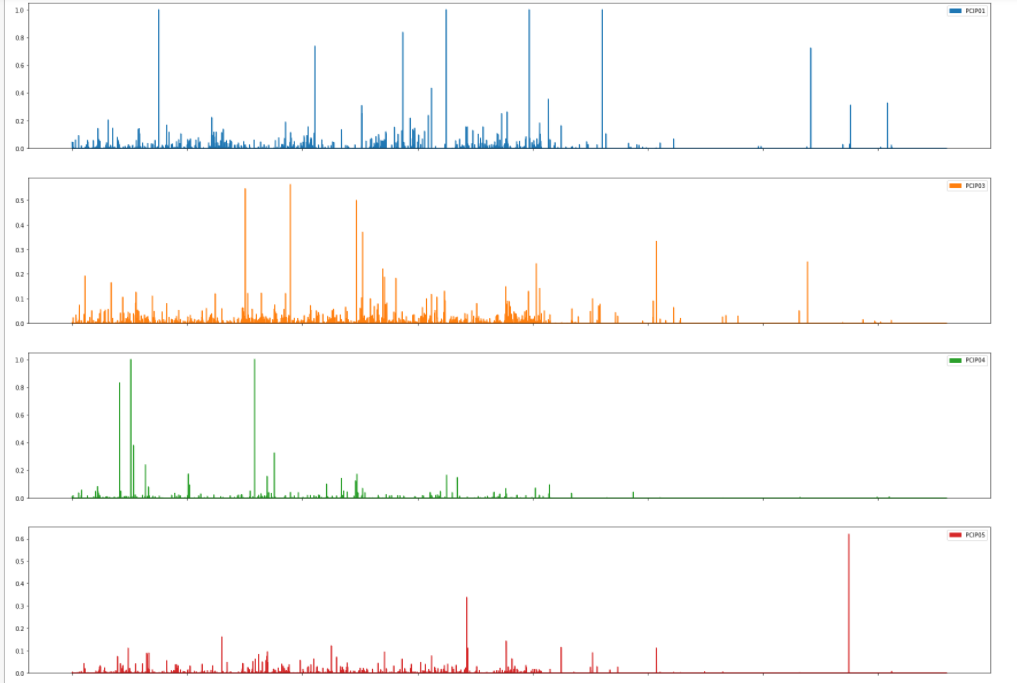
#### **Displaying the percentages of Operating versus Non-Operating Institutions in US**



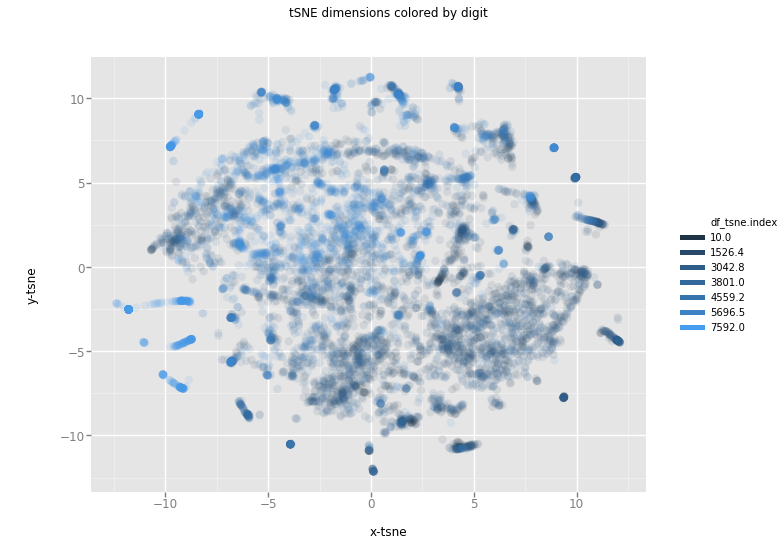
**As observed, 10.8% of US colleges are not operational.**

##### *Area distribution of various courses that are being offered in various fields such as Agriculture, Natural Resources, Ethnic Studies etc.*





**Visualizing the overall Dataset using TSNE Unsupervised Machine Learning Algorithm:**



**The data points scattered around a mean. The only thing that catches the eye are the outliers as highlighted above. The cause of outliers is because some colleges impose tuition fees that are far above the average mean imposed by other colleges for particular student’s income groups. In my opinion, college ranking should be impacted if the college does not provide any scholarships to low-income students.**