Final Project

by

Neha Lad

George Mason University

Title: College Scorecard Analysis

1. Introduction:

Financing a college degree is a complicated process, students need to be aware about the cost, duration for graduation and earnings post graduation. Few years back US government announced the creation of college scorecard project to answer all the above questions.

College scorecard project was basically designed to increase transparency, in admission process. It compares data on basis of cost and value for higher education institutions in US. College scorecard provides data files with data about institutions, cost for education, graduation rate of an institute, employment rate post graduation, amount rate borrowed by students as well as loan default rate. Beside this it also captures data of test scores, salaries after 6 to 10 years after graduation, size of a school etc.

2. Nature of Data Curation:

Who collected the data?

College scorecard data is published by United States Department of Education. Administration created college scorecard to provide reliable and unbiased information about college performance. The Department of Education collaborates with different national institutes like Loan Repayment Program for debts information of students, Center for Labor Market Studies for unemployment rate, US bureau of labor statistics for underemployment etc.

Why did they collect the data?

The main aim for publishing these files was to provide prospective postsecondary students enough data to make informed decisions while selecting college/universities. Beside this it also served the purpose to improve free inquiry, transparency in admission process, accountability by universities.

What is the nature of the data given the purpose of the data collection?

The data is public and was intended for students planning for post secondary studies. It targeted every aspect related to admission from SAT score to family income. From concentration in specific field of studies to earnings after 4 to 10 years after graduation. Hence, the main audience were students and their parents, this data helps them to compare the college ratings gradings considering several parameters.

Given the nature of the data, how can you adjust and leverage the data?

Though there are several advantages of College Scorecard, the drawbacks cannot be neglected. The data accounts only for those students who receiving financial aid, but it does not consider those students who are not eligible or does not qualify for it. The scorecard does mention the marks required to qualify for a college, but college admission process does consider academic and personal background. The scorecard doesn't take into consideration experience of a candidate while applying for admissions, which is also one of the potential flaws.

3. **Questions:**

   The project aims to address below questions:

   1. Does the SAT score impact the admission Rate?

      This question answers the relationship between SAT score and admission rate.
   2. Does the admission rate of an institution affect the salary of its faculty?

      We might be under the impression that more the admission rate, more the faculty salary, however, the analysis results contradicts the understanding.
   3. Which regions award the predominant degree?

      This question answers which region has the highest or lower rates of awarding predominant degrees.
   4. Does on campus or distance education impact faculty salary?

      This question helps to answer the relationship between 2 variables faculty salary and education imparting method.

Is there any privacy, quality, or other issues with this data?

Privacy: Many elements are available only for Title IV recipients, or students who receive federal grants and loans. These data are reported at the individual level to the National Student Loan Data System (NSLDS), which is used to distribute federal aid, and published at the aggregate institutional level. While some institutions report these data at the campus level (8-digit OPE ID), data produced for this site are rolled up to the institution level (6-digit OPE ID). In these cases, IPEDS institutions sharing a common 6-digit OPEID are all assigned the same sum or (student-weighted) average outcome or median outcome for students across all branches of the institution for NSLDS or tax-data derived measures

Quality: For this analysis, the file chosen consists of data from the year fall 2017-18. Some of the patterns that seems to be prevalent this year, need not necessarily exists for other years.

Other Data Issues: Only 32% of the institutes have reported the admission rates. Hence the data is incomplete and so is the analysis, as it does not provide full fledge picture. Values like faculty salary, tuition fees are provided as average values. Hence, they divert us from original picture.

Assumptions:

SAT score is directly related with admission rate. i.e. more the SAT score, higher the admission rate. Beside this, there is one more assumption that more the admission rate more will be the rate of the faculty. Also, we are under the assumption, that faculty imparting education by distance learning are earning more as compared to faculty who are imparting education by both means i.e. on campus and distance learning.

Benefits from Data analysis:

1. Students are the main stakeholders of this analysis. The analysis will help them to understand the relationship between SAT score and admission rate.
2. Beside this, the analysis will help faculty notably, by comparing the variability in average salaries for those who conduct distance learning and on-campus and distance learning both.
3. The analysis also benefits the parents of students, as it gives them the clear picture of the cost for education region wise, school wise. Moreover, the comparison helps them to make a conscious decision for their children's admission.
4. The analysis also serves as an input to researchers to predict the educational system trend for the particular year and perform predictive analysis about progress or degrade for the next year.

4. **Requirements and Resources needed: What software and hardware resources you have used in this project?**

   Software:

   - R: This is used for Statistical analysis and data exploration.
   - Python: Used for data exploration.
   - Tableau Desktop: Tableau software is used for visualizations.
   - Microsoft Excel: Excel used for Pre-Processing on dataset.

   Hardware:

   The complete project is performed on laptop of the following configurations:

   - Device Processor: Intel Core i7, Device Processor Speed: 1.6 GHz , Memory: 16GB

What kinds of pre-processes were needed to make use of the data, and why?

The data is extremely raw data and needed some sort of processing before using the variables. While performing the correlation test the variable needed to eliminate the null values. The categorical variables were given number in order to fetch the counts. The counts were then plotted on graphs so that a meaningful graph could be generated from the data.

What are the advantages and limitations of the target dataset in answering your questions?

Advantages:

The dataset helps to narrow down the approach for admission process based on several parameter like SAT score, institute rating, employment rate, earnings post graduation after 4 years upto 10 yrs etc. It facilitates easy comparison in terms of faculty salary, admission rate SAT score etc.

Disadvantages:

As this target only one year's data, it is not taking into consideration the trends of previous years.

5. **Descriptive Analysis**

Briefly describe the dataset:

The data set is "College Scorecard" and is available in comma separated values file format. It contains information of all different variables and admission rate for the year fall 17-18. The data is published on gov website: https://catalog.data.gov/dataset/college-scorecard

Prepare and describe relevant metadata:

Attributes	Datatype	Sample Values
Name: String	String	George Mason University
Alias: String	String	GMU
Location: String	String	Virginia
URLs: String	String	https://www2.gmu.edu/
Title IV Eligibility Type: Integer	Integer	24
Accrediting Agency: String	String	SACSCOC
Public/Private Nonprofit/Private For-Profit: String	String	Public
Carnegie Classifications: String	String	doctoral Institute
Religious Affiliation: String	String	None
Distance-Only: Boolean	Boolean	N
Institution Revenues/Expenses: Integer	Integer	$100,000,000K
Heightened Cash Monitoring 2: Boolean	Boolean	Y
Academic Areas Offered: Integer	Integer	40
Number of Programs Offered and Six Largest Programs: Integer and String	Integer and String	12
Open Admission Policy: Integer	Integer	22
Admission Rate: Float	Float	47.8
SAT and ACT Scores: Float	Float	87.6
Average Cost of Attendance, Tuition and Fees: Integer	Integer	65
Average Net Price: Integer	Integer	$10,000
Number of Undergraduate Students: Integer	Integer	5600
Undergraduate Student Body by Race and Gender: Float	Float	67.5
Undergraduate Students by Part-Time/Full-Time Status: Float	Float	45.4
Undergraduate Students by Family Income: Float	Float	78.8

Number of Institutions to Which Students Sent FAFSAs: Float	Float	56.7
Percent of Undergraduates Receiving Federal Loans: Float	Float	55.7
Percentage of Pell Students: Float	Float	45.3
Cumulative Median Debt25: Integer	Integer	$1,000
NSLDS Completion and Transfer Rates: Float	Float	15.2
Mean and Median Earnings: Integer	Integer	85,000
Threshold Earnings: Float	Float	60
Cohort Default Rate: Float	Float	56.7
Repayment Rate on Federal Student Loans39: Float	Float	34.7

Descriptive Statistics:

The figure below shows the descriptive summary of college scorecard. In this summary, all values have been given as per statistical analysis for each attributes in the dataset. The range of max and min, median and mean values are used to analysis the specification of Numeric data type

```
In [10]: print df.describe()
```

```
              UNITID         OPEID        OPEID6  SCH_DEG         HCM2  \
count  7.112000e+03  7.112000e+03   7112.000000      0.0  7112.000000
mean   1.867121e+06  1.866750e+06  16755.217660      NaN     0.013920
std    6.958737e+06  3.315240e+06  14589.294487      NaN     0.117168
min    1.006540e+05  1.002000e+05   1002.000000      NaN     0.000000
25%    1.740728e+05  3.229750e+05   3224.750000      NaN     0.000000
50%    2.289995e+05  1.056050e+06  10542.000000      NaN     0.000000
75%    4.506005e+05  3.025825e+06  30106.000000      NaN     0.000000
max    4.900540e+07  8.209884e+07  42698.000000      NaN     1.000000

              MAIN    NUMBRANCH      PREDDEG      HIGHDEG      CONTROL  ...  \
count  7112.000000  7112.000000  7112.000000  7112.000000  7112.000000  ...
mean      0.767717     3.883155     1.834505     2.233830     2.129218  ...
std       0.422319     9.088975     1.057762     1.340925     0.834518  ...
min       0.000000     1.000000     0.000000     0.000000     1.000000  ...
25%       1.000000     1.000000     1.000000     1.000000     1.000000  ...
50%       1.000000     1.000000     2.000000     2.000000     2.000000  ...
75%       1.000000     2.000000     3.000000     4.000000     3.000000  ...
max       1.000000    73.000000     4.000000     4.000000     3.000000  ...

       OMAWDP8_NOTFIRSTTIME  OMENRUP_NOTFIRSTTIME  OMENRYP_FULLTIME  \
count           3813.000000           3813.000000       3940.000000
mean               0.507685              0.299692          0.009377
std                0.226742              0.207287          0.032317
min                0.000000              0.000000          0.000000
25%                0.327900              0.145200          0.000000
50%                0.510500              0.264200          0.002500
75%                0.676500              0.409100          0.011000
max                1.000000              1.000000          0.680400
```

Explore Dataframe:

Data frame structure: The data frame below is generated using R. It shows that in the year fall 17 to 18 of college scorecard there are overall 7112 observations and 1977 variables.

```
> str(data1)
'data.frame':   7112 obs. of  1977 variables:
 $ ï..UNITID                     : int  100654 100663 100690 100706 100724 100751 100760
100812 100830 100858 ...
 $ OPEID                         : int  100200 105200 2503400 105500 100500 105100 100700
 100800 831000 100900 ...
 $ OPEID6                        : int  1002 1052 25034 1055 1005 1051 1007 1008 8310 100
9 ...
 $ INSTNM                        : Factor w/ 6978 levels "A T Still University of Health
Sciences",..: 89 6230 253 6231 94 5951 1139 407 423 422 ...
 $ CITY                          : Factor w/ 2473 levels "Aberdeen","Abilene",..: 1522 18
5 1398 986 1398 2222 28 93 1398 97 ...
 $ STABBR                        : Factor w/ 59 levels "AK","AL","AR",..: 2 2 2 2 2 2 2 2
 2 2 ...
 $ ZIP                           : Factor w/ 6243 levels "00602","00602-0960",..: 2445 24
32 2472 2454 2464 2437 2417 2439 2473 2501 ...
 $ ACCREDAGENCY                  : Factor w/ 44 levels "Accreditation Commission for Educ
ation in Nursing",..: 41 41 41 41 41 41 41 41 41 41 ...
 $ INSTURL                       : Factor w/ 5933 levels "abcollege.edu",..: 1568 5455 17
25 5460 1674 5452 2170 1871 1895 1886 ...
 $ NPCURL                        : Factor w/ 5450 levels "00334bf.netsolhost.com/bcguest/
CET/npcalc.htm",..: 681 2457 5428 374 2622 394 1335 478 2802 1302 ...
 $ SCH_DEG                       : Factor w/ 1 level "NULL": 1 1 1 1 1 1 1 1 1 1 ...
```

Explore Dataset:

The python is used to further analyze the data frame attribute. In the dataset, ACCREDAGENCY highest value count is 1218 i.e. Highest Learning Commission agency has accredited 1218 bodies.

```
import pandas as pd
import numpy as np
```

```
data1=pd.read_csv("C:/Users/Neha/Documents/AIT 580/MERGED2017_18_PP.csv")
```

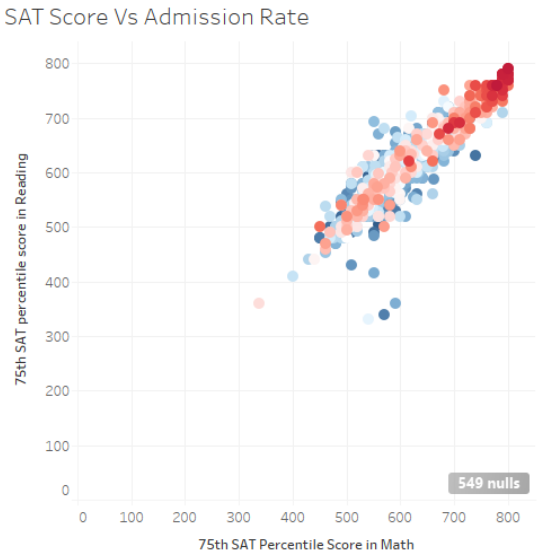
```
total_count1=data1['ACCREDAGENCY'].value_counts('')
```

```
total_count1
```

```
Higher Learning Commissio
n
               1218
```

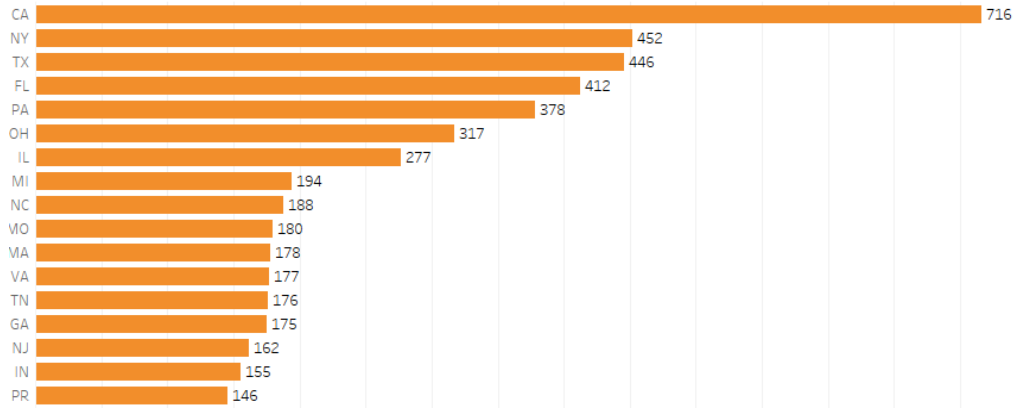
6. Results/Findings:

Scatter Plot: We have assumed that high SAT score results in high admission rate in our assumptions section. However, high SAT score doesn't guarantee high admission rate. The graph clearly depicts this scenario. The dark red points shows that the SAT score is high the admission rate is low.



Bar Graph:

The graph below shows the distribution of institutes across different states in US. Clearly, CA (California) is having highest no of institutes followed by New York, Texas, Florida, Pennsylvania etc



Box Plot:

This box plot is derived from python. Here we have plotted the distanceonly parameter against average faculty salary. The boxplot in orange is showing salary of faculty conducting only Distance classes. From this we can clearly conclude that the faculty conducting both on campus and distance classes are more as compared to faculty only conducting distance classes.

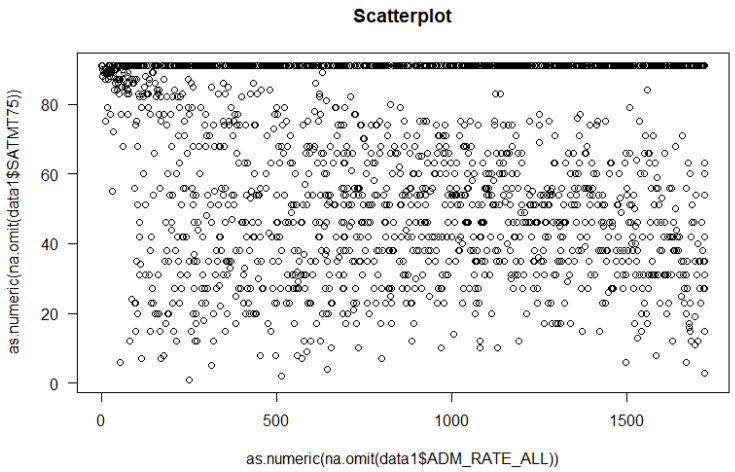
Correlation Analysis:

The Relationship between the Admission Rate and SAT score can be analyzed using correlation test, which is done in the R studio. The admission rate and SAT score are weakly correlated with each other. Therefore, with increase in SAT score the admission rate should reduce slightly. However, these two attributes are not highly correlated with each other. The sample estimated value of correlation is $0.48 < 1$, means not having high correlation. From the fig below it shows both values are not highly correlated with each other.

```
> cor.test(as.numeric(na.omit(data1$ADM_RATE_ALL)),as.numeric(na.omit(data1$SATMT75))
,method="pearson")

        Pearson's product-moment correlation

data:  as.numeric(na.omit(data1$ADM_RATE_ALL)) and as.numeric(na.omit(data1$SATMT75))
t = 46.915, df = 7110, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4682507 0.5037502
sample estimates:
      cor
0.486201
```



Regression Testing:

The linear regression is used to see relationship between Admission rate and SAT score in verbal reading and SAT score in math's. The linear regression test is done in R studio.

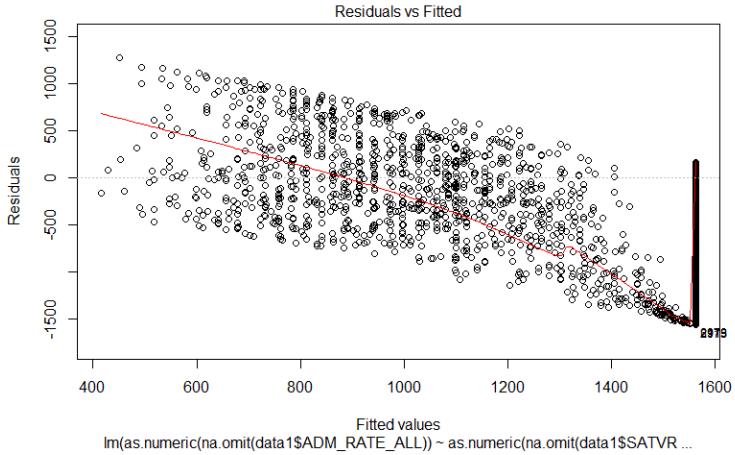
```
Residuals:
    Min      1Q  Median      3Q     Max
-1563.1   149.2   159.9   159.9  1272.3

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      388.414     23.739  16.362  < 2e-16
as.numeric(na.omit(data1$SATVR75))    5.030      1.234   4.075 4.64e-05
as.numeric(na.omit(data1$SATMT75))    7.337      1.336   5.493 4.09e-08

(Intercept)                     ***
as.numeric(na.omit(data1$SATVR75)) ***
as.numeric(na.omit(data1$SATMT75)) ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 410.1 on 7109 degrees of freedom
Multiple R-squared:  0.2382,    Adjusted R-squared:  0.238
F-statistic:  1111 on 2 and 7109 DF,  p-value: < 2.2e-16
```

Regression plot is plotted below. Even though the SAT score is increasing there is increase in admission rate but it is not considerable. However there is a trend that with low admission rate the 75th percentile SAT score is high.



Hypothesis Testing:

The Hypothesis test shows that, the p-value is less than 0.05. We reject the null hypothesis. Therefore, with increase in SAT score need not increase the admission rate and vice versa.

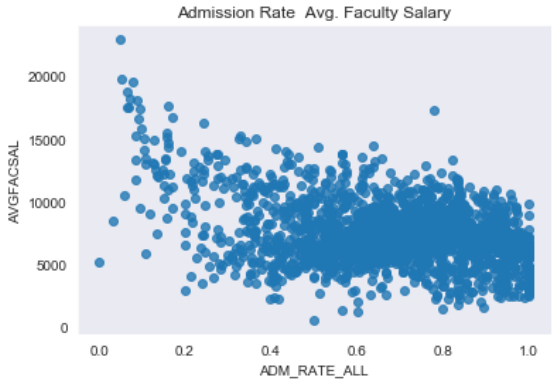
```
> TestResult <- t.test(as.numeric(na.omit(data1$ADM_RATE_ALL)),as.numeric(na.omit(d
ata1$SATMT75)),mu=8,alternative="two.sided",conf.level = 0.95)
> TestResult

        Welch Two Sample t-test

data:  as.numeric(na.omit(data1$ADM_RATE_ALL)) and as.numeric(na.omit(data1$SATMT75
))
t = 247.44, df = 7132, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 8
95 percent confidence interval:
 1376.521 1398.378
sample estimates:
 mean of x  mean of y
1471.15453   83.70487
```

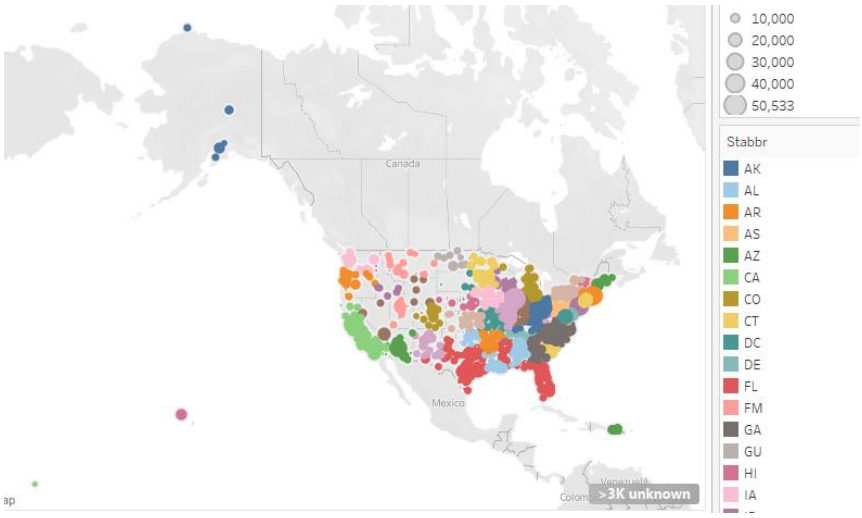
Scatterplot:

In this scatter plot we have plotted admission rate against average faculty salary. It seems that they both are negatively correlated i.e. with increase in admission rate there should be negligible decrease in average faculty salary.



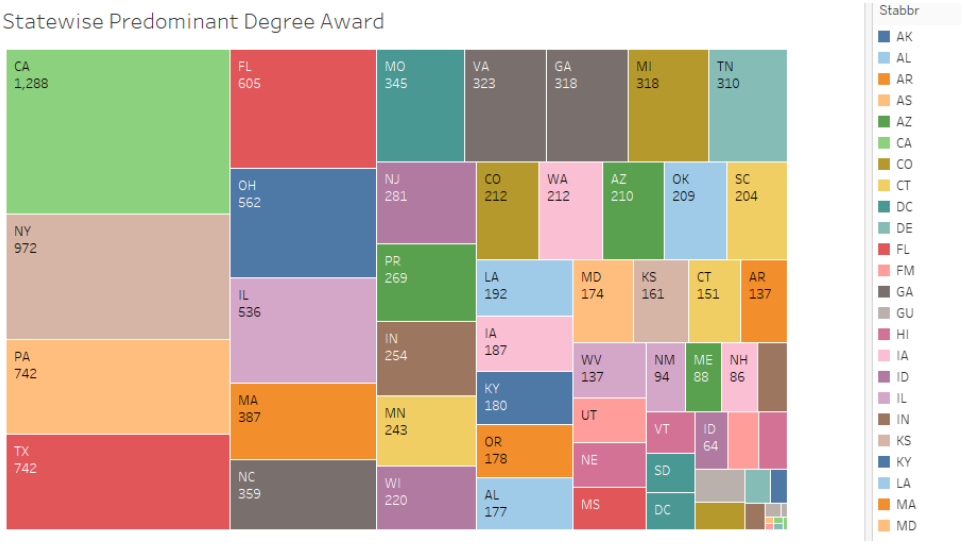
Map Based Visualization:

The below map shows state wise average salary distribution of faculty. The size of dots depicts the maximum salary of faculty. This helps to figure out which region has least and most avg faculty salary.



TreeMap Visualization:

Below tree map shows regions that awards predominant degree. Tree map clearly helps to identify the leading and lagging states who awards predominant degree in all the institutes across the region.



7. Conclusion and Future Works:

With the analysis on some of the variables in college scorecard, here are few interesting observations which answers all the questions we targeted in this project:

- The SAT test scores and Faculty salary have weak negative correlation with admission rate.
- The State of California has maximum no of institutes and tops in terms of average faculty salary and awarding predominant degree.

- With the increase in 75th percentile of SAT score the admission rate is reduced.
- The average salary of faculty conducting online and distance only classes is more than the faculty conducting distance only class.

Advantages:

Due to limited dataset, the trends of particular year i.e. fall 17-18 are clearly tracked.

3 prime analysis were conducted based on SAT scores, Regional institutes and Faculty salary. This targeted all the 3 stakeholders Students, Parents and Faculty.

Limitations:

The data provided for variables like faculty salary and tuition fees were average values.

Only 17% of institutes reported the SAT and ACT scores. Hence the analysis is based only on available data. The trends are subjected to change with change in data.

Future Work:

Analysis can be done at institute level i.e. by categorization public and private. Further, tuition fees variable can be analyzed, and its impact can be studied.

Students receiving federal loans can further be analyzed for employment ratios.

8. References:

1. Data Dictionary | https://collegescorecard.ed.gov/data/documentation/
2. USING FEDERAL DATA TO MEASURE AND IMPROVE THE PERFORMANCE OF U.S. INSTITUTIONS OF HIGHEREDUCATION| https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf
3. Understanding the College Scorecard | Jonathan Rothwell | Monday, September 28, 2015 | https://www.brookings.edu/opinions/understanding-the-college-scorecard/
4. What's Missing From the College Scorecard? | By Carol D'Amico | October 31, 2017| https://www.realcleareducation.com/articles/2017/10/31/whats_missing_from_the_college_scorecard_110219.html

9. **Explain/ Define Terms:**

IPEDS: Integrated Postsecondary Education Data System

NSLDS: National Student Loan Data System

OPE: Office of Postsecondary Education

ED: Education Department of United States

US: United States