Analysing Data on Colleges in the US

Shambhavi Agrawal

22110240
Civil Engineering
IIT Gandhinagar
Gandhinagar, India
shambhavi.agrawal@iitgn.ac.in

***Abstract*—This paper analyses the two datasets containing information about various colleges in the US. It answers scientific questions and proves/disproves hypotheses about the dataset using python codes, plots, graphs etc. A few unanswerable questions are included as well.**

I. Overview Of The Dataset

The datasets includes a list of public and private colleges in the US, along with the postal code and Federal ID Number. It contains information about the average SAT and ACT scores. It has scores specified for the different parts of the SAT exam, which includes Maths and Verbal scores and the quartiles of these. It also gives information about the number of applications received and accepted in each college. It tells about the number of new students enrolled and the percentage who were the toppers in high school. Further, it gives financial data like the in-state and out-of-state tuition fees, average salaries and compensation provided to the faculty by their ranks, instructional expenditure per student, room and board costs, additional fees, book costs, estimated personal spending etc. It establishes the credibility of the faculty by giving information about the percentage of the faculty who have Ph.D.s, and terminal degrees. It includes the student-faculty ratio, the number of faculty by rank, the percentage of alumni who donate and most importantly, the graduation rate. This dataset is useful for extracting much information about colleges, the admission system and the pattern it involves and can help analyse the education system of the US.

II. Scientific Questions And Hypotheses

A. *How do the salaries of faculty members vary by academic rank, such as full professors, associate professors, and assistant professors?*

The rank and the salary of faculty members depends on their qualification, teaching experience and many other parameters. Of course, the higher the rank of the faculty member, the higher the salary. Is there a uniform increase in salary when one goes from assistant professor to associate professor and from associate professor to full professor? One would expect that there will be a higher increase in salary from an associate professor to a full professor than for an assistant professor to an associate professor. Is there a higher height increase in a graph showing this relation? Can it prove this hypothesis? Also, does this increase vary for different colleges, or is the increase somewhat similar?

B. *What is the acceptance rate for different colleges?*

The acceptance rate of students in a college depends on a variety of parameters such as quality of education, fees, facilities provided and many more. Usually, prestigious colleges have a lower acceptance rate than other colleges. In other words, if the number of applications received by a college is more and the number of applications it accepts is very less than the college can be considered a good college. Can we identify these types of colleges through the graph plotted between these quantities. On the other hand, one might say that a good university might accommodate many students and hence accept more applications. Is this true?

C. *What is the trend in the number of students rejecting their acceptance into a college?*

Many students receive acceptance letters from colleges, but not all finally enrol in the respective college. This is due to multiple acceptance offers and various other reasons. Can the number of students rejecting their acceptance into a college tell us something about the rank of the college? Alternatively, do good colleges have lesser rejection from students than the not-so-good ones? Can such colleges be identified from this relation?

D. *What is the number of students failing to graduate from college?*

All students that are accepted into a college do not necessarily graduate. In fact, a large percentage of students fail to graduate. What is the rate of failure in graduation? Is it affected by the total number of students? Does more number of students mean that a lesser percentage of students will graduate due to less individual attention?

E. *What is the financial status of students studying in college? How much of their money is spent in housing and boarding and how much is their average personal expenditure?*

Students studying in college come from a range of financial backgrounds. Some can afford a higher personal expenditure while others struggle to pay the boarding cost apart from other costs. Is there a relation between the boarding cost and the estimated personal expenditure of students studying in a particular college? Does higher boarding costs mean students have less money to spend on their personal needs? Also, if

students from a particular college can spend more money on board as well as for their personal expenditure, can we have a higher probability of being right in saying that the college is private? Since usually, students from a good financial background can afford to study in a private college?

F. What is the difference in the fees for private and public colleges? Also, how does the fee for these two types vary for in-state and out-of-state students?

Using the data, can we prove the fact that public colleges have lesser fees than private colleges? Also, public colleges have different fee structures for in-state and out-of-state students. Out-of-students have to pay a higher fee than in-state students since public colleges are also funded by state tax. Is this relation visible in a graph visualising the fee structure? Additionally, private colleges have even higher fees than the out-of-state fees for public colleges. Further, the in-state and out-of-state fee for private colleges is usually the same since it is state-funded [1]. Can these hypotheses be proved by plotting graphs?

G. What is the relation between the SAT and ACT scores accepted by different colleges? Does this relation vary for public and private colleges?

Good colleges require high SAT or ACT scores to get accepted. Nowadays, to get accepted into most colleges, a student needs either ACT or SAT score. Do colleges require a similar percentage of marks for both SAT and ACT exams? What is the trend of ACT and SAT scores in public and private colleges? Do private colleges require higher SAT and ACT scores than public colleges, since most of the prestigious colleges in the US are private colleges? Can we observe patterns in both the scores? Which colleges require high SAT and ACT scores? These colleges can be estimated to be good colleges.

H. Is there a higher graduation rate in colleges that spend more on their students? Is there a difference for public and private colleges?

Do students perform better when they get more resources and facilities? In other words, if more money is spent on individual students, does that lead to a higher graduation rate? Do these things vary for public and private colleges? Do private colleges spend more on their students due their high fees structure?

I. What is the number of public and private colleges in various states across the US? Can we conclude anything about the number of public colleges and the tax spent on education in each state?

Do some states have a disproportionately large number of colleges? Does more number of public colleges mean a high number of private colleges as well? It seems right that the state that spends a higher amount of its tax on education would have more number of public colleges. Can this hypothesis be proved?

J. Does having top students with good scores and more faculty per student in a college affect its graduation rate?

Will a college accepting smart students and more faculty to give individual attention affect the performance of the students? In other words, does the graduation rate of a college have a relation with the average SAT scores of the students and the number of faculty per student? Is it true that students who performed well in SAT will perform well in college as well? Also, does individual attention given by faculty matter in the performance of a student?

K. How satisfied are students after graduating from a particular college? What is the contribution of faculty members in this relation? What is the contribution of the expenditure done per student?

When students are happy and enjoy their college life, they tend to give more donations to the college after they become part of the alumni of that college. Does this happiness depend upon the qualification of the faculty in the college? The better the faculty, the more enjoyable the learning process for the students it will be. Also, it can be assumed that when a college spends more money on students, the students feel more connected to the college. Hence, a large percentage of these students will donate when they become alumni of the college.

L. Is the salary and compensation given to faculty members in accordance with the distribution of faculty members by their rank?

Is the salary distribution of a college uniform? Or is it concentrated on just one rank of faculty member? Is salary and compensation given according to the number of faculty members of a particular rank present in the college?

III. Libraries And Functions

To analyse the given data and answer scientific questions and hypotheses, a few python libraries need to be installed and then imported to help write the code and plot graphs. A few of the python libraries used to assess the dataset are Matplotlib, Pandas, and NumPy. Matplotlib is extremely useful for plotting graphs and curves to visualise the data. It has many features such as labelling the axes, and giving title, colour, and size. Various graphs like bar graphs, double bar graphs, line graphs, scatter, histograms, 3D plots, 2D plots with three variables, pie charts etc. can be plotted using Matplotlib. The Axes3D module helps plot 3D graphs. Pandas is an extremely useful tool for data analysis. It has many functions that can help organise and view the data from different perspectives. It is used to read and change the dataset format into the desired one. Functions like DataFrame and Series. The mean value (mean), standard deviation (std), variance (var), maximum (max), and minimum (min) of a dataset can be easily calculated using the respective functions written inside the parenthesis. It can help plot various types of graphs. NumPy helps create arrays, matrices and perform basic mathematical operations. A few other libraries that would be useful in analysis would be scipy for optimizing data, scikit-learn for advanced data analysis, seaborn for plotting data etc.

IV. Analysing DAta and Answering Questions/Hypotheses

A. *Variation in increase in salary of faculty members of different rank*

1) *Procedure*

First the ratio of increases in salary going from assistant professor to associate professor and from associate professor to full professor is calculated by converting the datatype into integers. The two ratios are then plotted as a scatter graph. Also the mean of all the average salaries of professors of a specific rank from all colleges is calculated and plotted as a bar graph. The following code snippet can be used:

```
diff_one = (aaup['Average salary - asso
ciate professors'].astype(int) - aaup['
Average salary - assistant professors']
.astype(int))/(aaup['Average salary - a
ssistant professors'].astype(int))
diff_two = (aaup['Average salary - full
 professors'].astype(int) - aaup['Avera
ge salary - associate professors'].asty
pe(int))/(aaup['Average salary - associ
ate professors'].astype(int))
plt.figure(figsize = (6, 6))
plt.plot(aaup['Average salary - assista
nt professors'].astype(int), diff_one,
'.')
plt.plot(aaup['Average salary - associa
te professors'].astype(int), diff_two,
'.')
assistmeans = aaup['Average salary - as
sistant professors'].astype(int).mean()
assocmeans = aaup['Average salary - ass
ociate professors'].astype(int).mean()
profmeans = aaup['Average salary - full
 professors'].astype(int).mean()
meansalary = pd.Series([assistmeans, as
socmeans, profmeans], index = ['Assist.
 Prof.', 'Assoc. Prof.', 'Full Prof.'])
meansalary.plot(kind = 'bar', xlabel =
'Rank of faculty', ylabel = 'Mean Salar
y', title = 'Mean of salaries of facult
y by rank')
```

2) *Observation and Results*

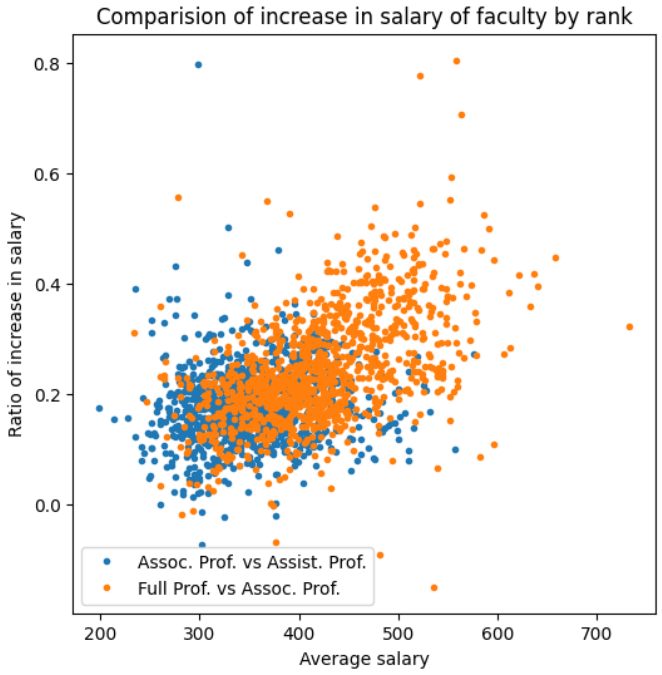

Fig. 1. Graph showing comparison of increase in salary of faculty by rank

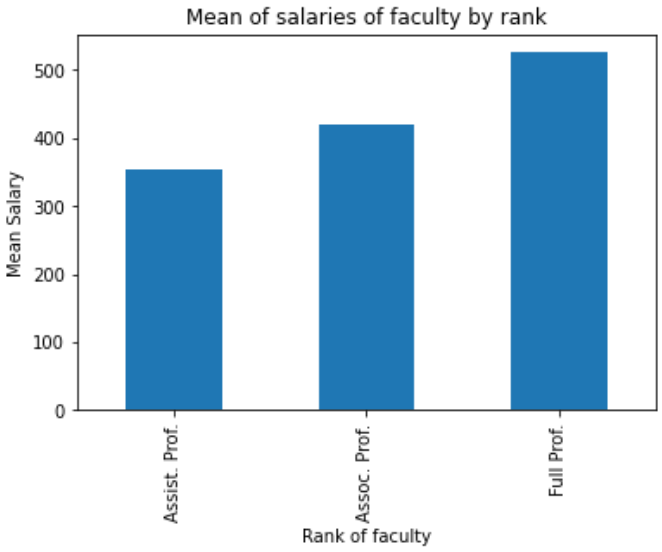

Fig. 2. Bar graph showing mean of average salaries of faculty members by their rank

In the scatter plot plotted, it is visible that the patch of orange dots is shifted a little to the upper right side with respect to the blue patch of dots. This show that the relative increase in salary from associate professor to full professor is higher for some colleges than the relative increase in salary from assistant professor to associate professor. Although interestingly, a lot of the blue and orange dots coincide or are not very far apart which shows that many colleges don't have much of a difference in the relative increase of the salaries of the professors.

From the bar graph, we can generalise that on average most of the colleges give a higher increase in salary from an

associate professor to a full professor than from an assistant professor to an associate professor. Thus, our hypothesis for the increase in salary is proved from these graphs.

B. *Acceptance rate for different colleges*

1) *Procedure*

A scatter graph is plotted between the number of applications received vs the number of applicants accepted by different colleges. We also find the details of the college with the highest number of applications received and the highest number of applications accepted. Calculate the rate of acceptance and plot the kde for the same. The following code snippet can be used:

```
plt.plot(usnews['Number of applications
 received'].astype(int), usnews['Number
 of applicants accepted'].astype(int),
'.')
plt.xlabel('Number of applications rece
ived')
plt.ylabel('Number of applicants accept
ed')
plt.title('Acceptance rate')
print(usnews[usnews['Number of applicat
ions received'].astype(int)>40000])
rate = (usnews['Number of applicants ac
cepted'].astype(int))/(usnews['Number o
f applications received'].astype(int))
rate.plot(kind='kde')
```

2) *Observation and Results*

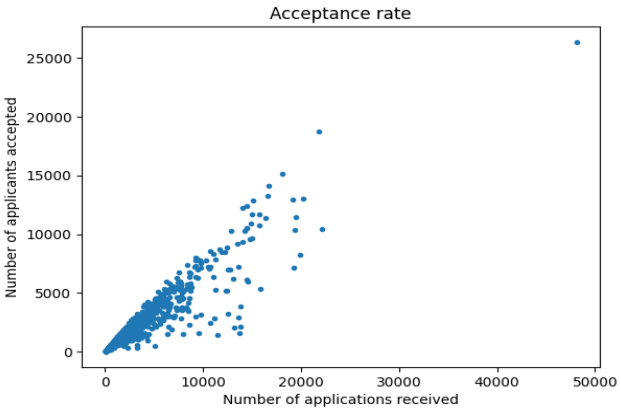

Fig. 3. Graph depicting the number of applications received vs the number of applicants accepted

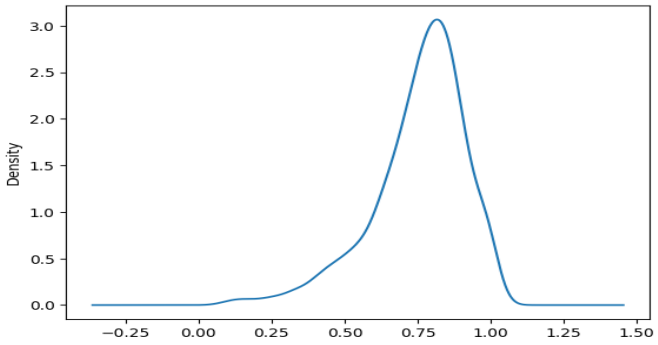

Fig. 4. kde graph depicting acceptance

It is visible that some colleges have an almost 100% acceptance rate since their points can be plotted to form a line similar to y=x. The points near the x axis of the scatter plot show colleges which have very less acceptance rate. Also, most of the points are concentrated in the range of 0-10000 for the number of applications received. An interesting point in this graph is one for which the number of applications received and accepted is very high. The name of the college is Rutgers at New Brunswick which is a public college.

Also, from the kde plotted it can be seen that the peak of the graph shows that the majority of the colleges have an acceptance rate of nearly 75%.

C. *Students rejecting acceptance into college*

1) *Procedure*

We plot the graph between the number of applications accepted vs the number of new students enrolled. Next, we find and mark the extreme points which include a very high number of applicants accepted or a very high number of students enrolled. Next we calculate the rejection rate by computing the number of new students enrolled divided by the number of applicants accepted and subtracting this ratio by 1. We then plot the kde of this function. The following code snippet can be used:

```
y1 = usnews[usnews['Number of new stude
nts enrolled'].astype(int)>7000]
x1 = usnews[usnews['Number of applicant
s accepted'].astype(int)>15000]
plt.plot(usnews['Number of applicants a
ccepted'].astype(int), usnews['Number o
f new students enrolled'].astype(int),
'.')
plt.plot(x1['Number of applicants accep
ted'].astype(int), x1['Number of new st
udents enrolled'].astype(int), 'r.')
plt.plot(y1['Number of applicants accep
ted'].astype(int), y1['Number of new st
udents enrolled'].astype(int), 'g.')
```

```
r_rate =1 - ((usnews['Number of new students enrolled'].astype(int))/(usnews['Number of applicants accepted'].astype(int)))
r_rate.plot(kind = 'kde')
```

2) Observation and Results

tot

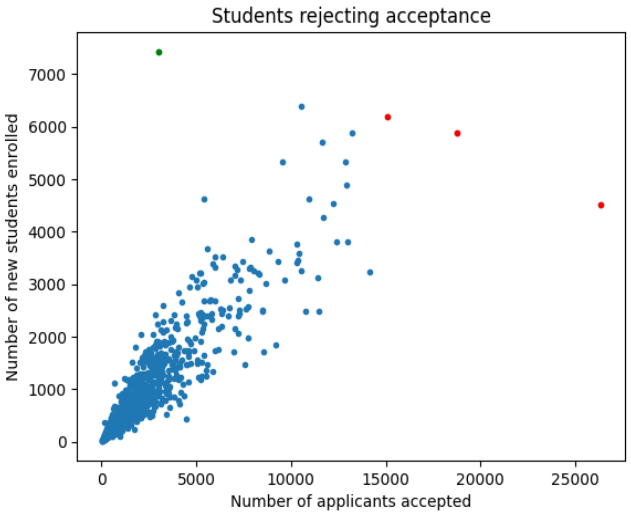

Fig. 5. Graph of number of applicants accepted vs the number of new students enrolled

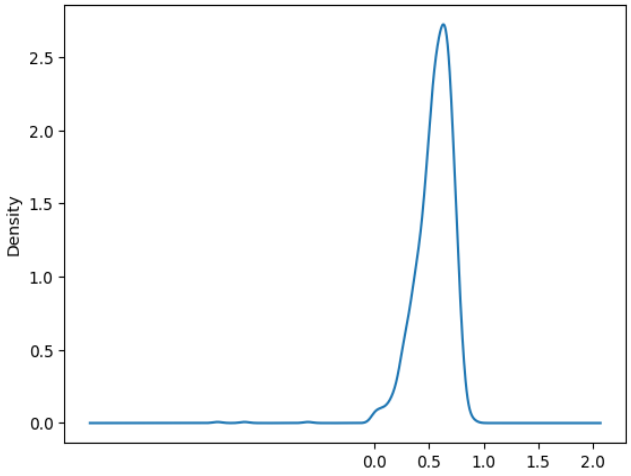

Fig. 6. kde of the rate of rejection of acceptance into college by students

For a lot of points in the scatter plot, the number of new students enrolled is much less than even 1/5th of the number of applicants accepted. The points marked in red show the colleges that accept many students, but a huge part of these students did not enrol, showing that maybe these colleges are not very prestigious. In the kde plot the rejection rate for a majority of the colleges is nearly 0.75, which is pretty high.

D. Analysing failure of students to graduate

1) Procedure

First the total number of undergraduates is calculated by taking the sum of the number of fulltime undergraduates and the number of half time undergraduates. Next, we plot the scatter plot for the total number of undergraduates vs the graduation rate. The following code snippet can be used:

```
total_ug = usnews['Number of fulltime undergraduates'].astype(int) + usnews['Number of parttime undergraduates'].astype(int)
plt.plot(total_ug, usnews['Graduation rate'].astype(int), '.')
plt.xlabel('Total number of undergraduates')
plt.ylabel('Graduation rate')
plt.title('Graduation of students')
fail = total_ug*(100-usnews['Graduation rate'].astype(int))
fail.plot(kind = 'kde')
```

2) Observation and Results

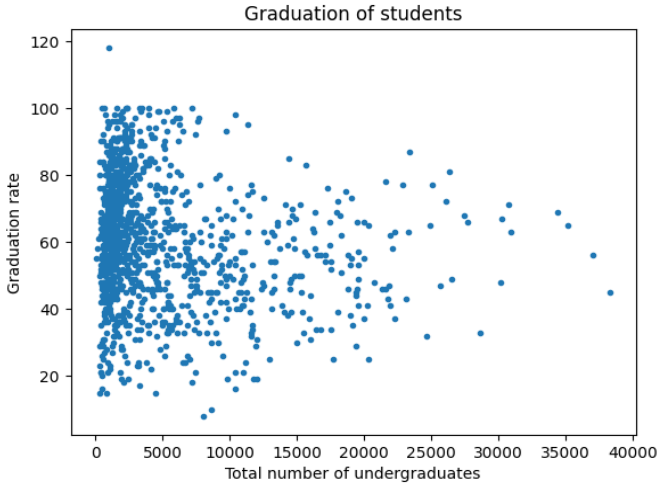

Fig. 7. Graph depicting the graduation rate vs the total number of undergraduates

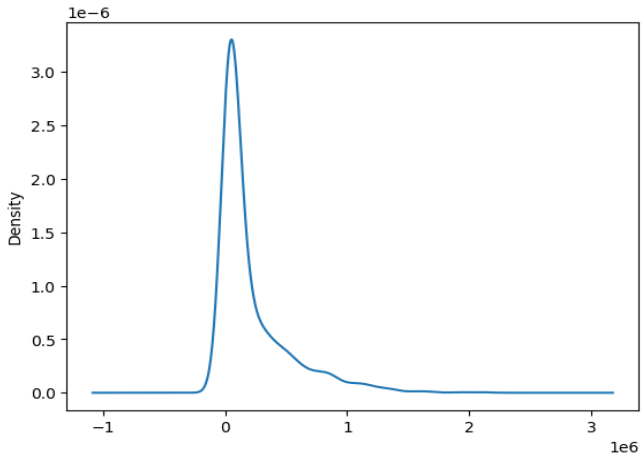

Fig. 8. kde of the number of students failing to graduate

From the scatter plot it can be observed that for less number of undergraduate students in the range 0-10000, the graduation rate does not show any significant pattern i.e. it shows a large variation in values. Interestingly for higher number of undergraduates the graduation rate is somewhat between 40%-80%, which shows that maybe due to less individual attention or less money spent per individual etc., the graduation rate is not near 100%. Another interesting feature to note is that the graduation rate is very scattered and is not very high for most of the colleges showing a serious problem in the education of students. Thus our hypothesis is somewhat true for higher number of students but does not give any definite result for less number of students.

E. *Analysing the financial status of students studying in different colleges*

1) *Procedure*

First, find the total room and board costs by taking the sum of these costs. Then plot a scatter plot between the total boarding cost vs the estimated personal spending of a student from a particular college. Also, find and mark the point which shows very high personal spending. The following code snippet can be used:

```
totalcost = usnews['Room and board costs'].astype(int) + usnews['Room costs'].astype(int) + usnews['Board costs'].astype(int)
rich = usnews[usnews['Estimated personal spending'].astype(int)>5000]
plt.plot( totalcost, usnews['Estimated personal spending'].astype(int), '.')
plt.plot(rich['Room and board costs'].astype(int) + rich['Room costs'].astype(int) + rich['Board costs'].astype(int),
 rich['Estimated personal spending'].astype(int), 'r.')
```

2) *Observation and Results*

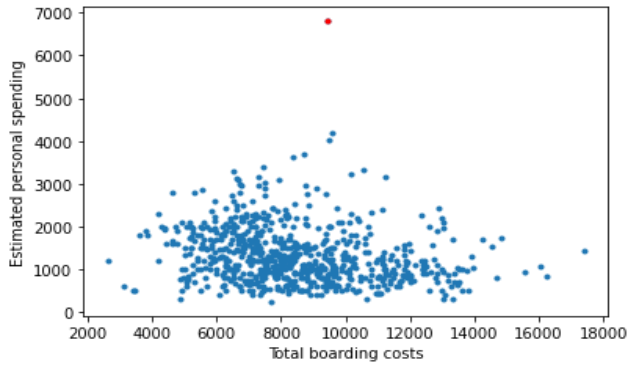

Fig. 9. Graph plotted between total boarding costs vs the estimated personal spending

We can observe from the scatter plot that a higher boarding cost does not necessarily mean that personal spending reduces. This shows that not all students are very rich. Even if some pay high boarding cost, they cut down on personal spending. Although some points show low boarding costs, we can observe a little higher personal spending, which shows they do not have to save as much as the students paying high boarding costs. Hence, though for some points, our hypothesis is correct, we cannot generalise it for the whole dataset. Interestingly there is one point marked in red in the plot that shows significantly high personal spending with median boarding cost. The point refers to Saint Louis University, which is a private college. This proves our proposed hypothesis that such students with good financial backgrounds would be studying in a private college.

F. *Comparison of fees for private and public colleges and in-state and out-of-state.*

1) *Procedure*

First, divide the colleges into private and public. Next, plot the in-state vs out-of-state tuition fee for private and public colleges. Add legend to the plot. The following code snippet can be used:

```
public = usnews[usnews['Public/Private']==1]
private = usnews[usnews['Public/Private']==2]
plt.plot(public['In-state tuition'].astype(int), public['Out-of-state tuition'].astype(int), '.')
plt.plot(private['In-state tuition'].astype(int), private['Out-of-state tuition'].astype(int), '.')
plt.xlabel('In-state tuition')
plt.ylabel('Out-of-state tuition')
plt.title('Comparision of In-state and Out-of-state tuition fee')
plt.legend(['Public', 'Private'])
```

2) Observation and Results

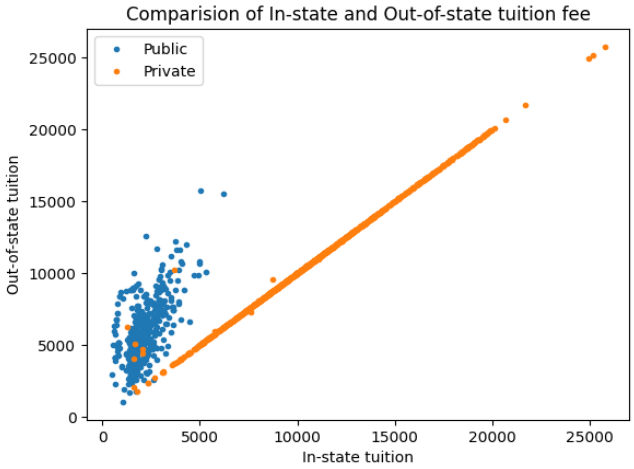

Fig. 10. Graph showing the relation between In-state and Out-of-state tuition fee for public and private colleges

From the graph, we can observe that for public colleges, the out-of-state tuition fee is much higher than the in-state tuition fee, thus proving our hypothesis that due to state funding, public colleges charge a higher fee to out-of-state students. Interestingly, we can see that the curve for the private colleges resembles y=x line, showing that the in-state and out-of-state tuition fee is the same, thus proving our hypothesis. Additionally, except for a few private colleges, the fee is much higher than public colleges and in fact even higher than the out-of-state tuition fee, which again proves the proposed hypothesis.

G. Analysis of SAT and ACT scores with respect to private and public colleges.

1) Procedure

First, we identify the private and public colleges and make separate data frames for these two. Next, plot the required SAT and ACT scores for the different colleges. Find and mark the colleges which need high SAT as well as ACT scores in the graph. The following code snippet cam be used:

```
good = private[private['Average ACT sco
re'].astype(int)>=30.0]
good_two = private[private['Average Com
bined SAT score'].astype(int)>1250]
plt.plot(public['Average Combined SAT s
core'].astype(int), public['Average ACT
 score'].astype(float), '.')
plt.plot(private['Average Combined SAT
score'].astype(int), private['Average A
CT score'].astype(float), '.')
plt.plot(good['Average Combined SAT sco
re'].astype(int), good['Average ACT sco
re'].astype(int), 'g.')
plt.plot(good_two['Average Combined SAT
 score'].astype(int), good_two['Average
 ACT score'].astype(int), 'g.')
from IPython.display import display
display(good_two)
```

2) Observations and Results

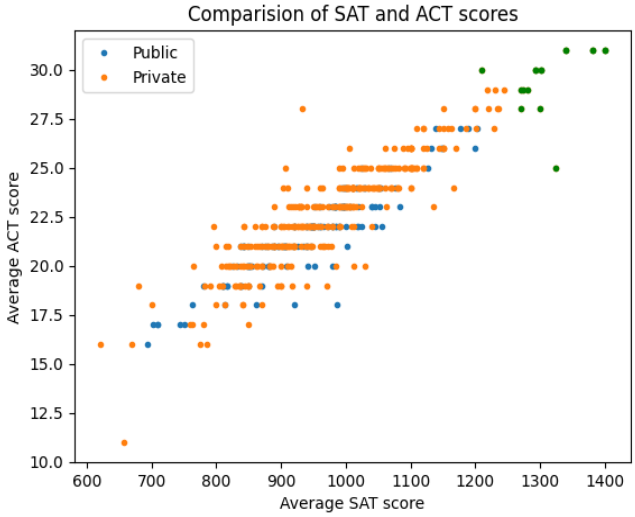

Fig. 11. Graph depicting the relation between average combined SAT scores and ACT scores for public and private colleges

index	FICE	College name	Postal code	Public/Private	Average Math SAT score	Average Verbal SAT score	Average Combined SAT score	Average ACT score
75	1170	Claremont McKenna College	CA	2	670	600	1270	28
78	1173	Pomona College	CA	2	700	640	1340	31
151	1424	Wesleyan University	CT	2	660	620	1280	29
307	1739	Northwestern University	IL	2	670	600	1270	29
431	2115	Amherst College	MA	2	685	639	1324	25
454	2178	Massachusetts Institute of Technology	MA	2	742	639	1381	31
494	2077	Johns Hopkins University	MD	2	686	606	1292	30
653	2920	Duke University	NC	2	687	615	1302	30
1023	3378	University of Pennsylvania	PA	2	680	594	1274	29
1039	3401	Brown University	RI	2	680	620	1300	28

Fig. 12. Dataframe depicting the good colleges based on high SAT and ACT scores

The graph is seen to have small straight horizontal lines, which show that the colleges require particular discrete ACT scores i.e. the values are not in a continuous range. Also, if we perform the linear regression of these data points, then we will get a straight line with some slope. This shows that most of the colleges that require a particular SAT score approximately have similar average ACT scores. Also, as the average SAT score increases, the average ACT score increases as well, showing that if a college requires a high SAT score, then it requires a high ACT score too. Not surprisingly, it can be observed that the highest SAT and ACT score required by public colleges does not go as high as it goes for private colleges. This proves our hypothesis that most of the prestigious colleges in the US are private colleges. The points marked in green represent the colleges that need high SAT and ACT scores. From the above table

we can recognize a few prestigious private colleges that require high scores to get into. For example, Stanford University, Massachusetts Institute of Technology, University of Pennsylvania, Brown university etc.

H. *Analysing the relationship between the instructional expenditure per student and the graduation rate and comparing this for public and private colleges.*

1) *Procedure*

Identify the public and private colleges and make separate data frames for these two. Next, plot the instructional expenditure per student vs the graduation rate of different public and private colleges. The following code snippet can be used:

```
print(public['Instructional expenditure per student'].astype(int).max())
plt.plot(public['Instructional expenditure per student'].astype(int), public['Graduation rate'].astype(int), '.')
plt.plot(private['Instructional expenditure per student'].astype(int), private['Graduation rate'].astype(int), '.')
plt.xlabel('Instructionl expenditure per student')
plt.ylabel('Graduation rate')
plt.title('Instructional expenditure vs Graduation rate')
plt.legend(['Public', 'Private'])
```

2) *Observation and Results*

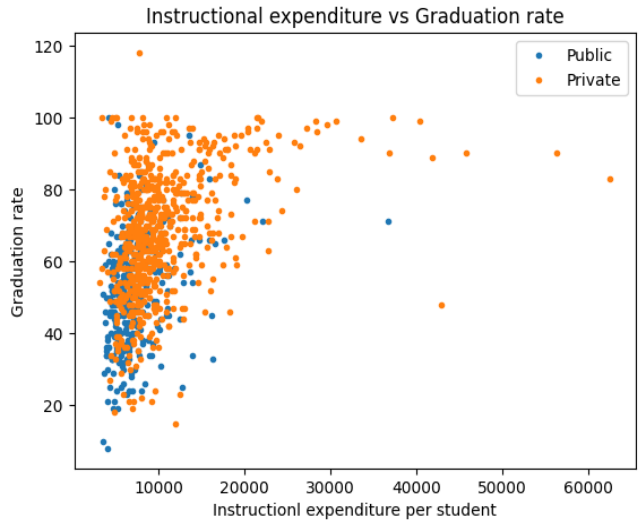

Fig. 13. Graph depicting relation between the instructional expenditure per student vs the graduation rate

From the graph we can observe that graduation rate takes a range of values for similar instructional expenditure i.e. there is no direct relation between these two parameters, thus proving our hypothesis wrong that the more the resources spent on a student, the higher the graduation rate of a college. An interesting observation from the graph is that some private colleges have very high instructional expenditure per student which proves a part of our hypothesis that since private colleges have high fees they spend more resources on an individual student. Also, the maximum instructional expenditure per student amongst the public colleges is 36704, though majority of them limited to somewhere near 20000.

I. *Distribution of public and private colleges state-wise in the US and its relation with the amount of tax spent on education in the state*

1) *Procedure*

We first distribute the dataset to public and private colleges and make separate data frames. Next, these data frames are grouped according to the state postal code. A double bar graph is then plotted with the state in the x-axis and the number of colleges in the y-axis. The following code snippet can be used:

```
public_inst = public.groupby('Postal code').size()
private_inst = private.groupby('Postal code').size()

df = pd.DataFrame({"public":public_inst, "private":private_inst})
ax = df.plot.bar(figsize = [10, 4], color=["SkyBlue", "IndianRed"], rot = 90, title = "Distribution of Institutions", ylabel = 'Number of institutions')
plt.show()
```

2) *Observation and Results*

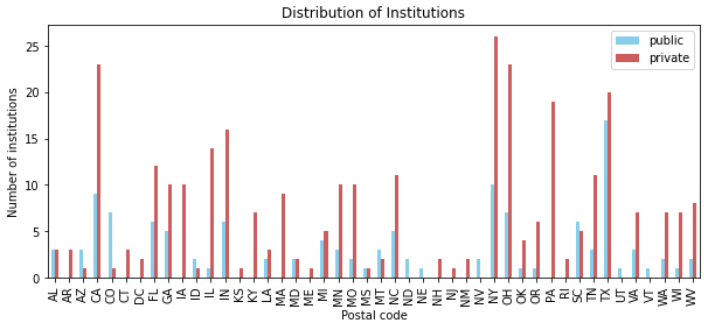

Fig. 14. Double bar graph depicting the number of public and private colleges in different states of the US

We can observe that there are a high number of private colleges across the US. The majority of the states have more number of private colleges than the number of public colleges. Also, it cannot be said for all states, but it can be observed that a state with high number of public colleges has a high number of private colleges. Thus proving part of

our hypothesis true. An interesting point to note here is that Texas has the highest number of public colleges and then comes California and New York. It is known that the states spending a large amount of money on education includes California and New York, but Texas is not in the top of this list. Hence part of our hypothesis that states spending more on education would have a high number of public colleges, is true for states like California, New York, Florida etc. but shows a contradiction for states like Texas.

J.Analysis of dependence of graduation rate on factors like good students and individual attention of faculty per student

1) Procedure

Import the Axes3D module to plot 3d graphs. Distribute and make two different data frames, one for public colleges and the other for private colleges. Scatter plot a 3D graph by setting the projection of axes to 3d and then specifying the data frames corresponding to the axes. Also, plot a 2D scatter graph with three variables, one for public and the other for private colleges. The following code snippet can be used:

```
fig = plt.figure(figsize=(6, 6))
threedee = plt.axes(projection='3d')
threedee.scatter(public['Average Combin
ed SAT score'].astype(float), public['S
tudent/faculty ratio'].astype(float), p
ublic['Graduation rate'].astype(float),
 color='SkyBlue')
threedee.scatter(private['Average Combi
ned SAT score'].astype(float), private[
'Student/faculty ratio'].astype(float),
 private['Graduation rate'].astype(floa
t), color='pink')
fig, ax = plt.subplots()
graph = ax.scatter(public['Average Comb
ined SAT score'].astype(float), public[
'Student/faculty ratio'].astype(float),
 c=public['Graduation rate'].astype(flo
at))
var_three = fig.colorbar(graph)
ax.set_xlabel('Average Combined SAT sco
re')
ax.set_ylabel('Student/faculty ratio')
ax.set_title('Public colleges')
var_three.ax.set_ylabel('Graduation rat
e')
```

2) Observation and Results

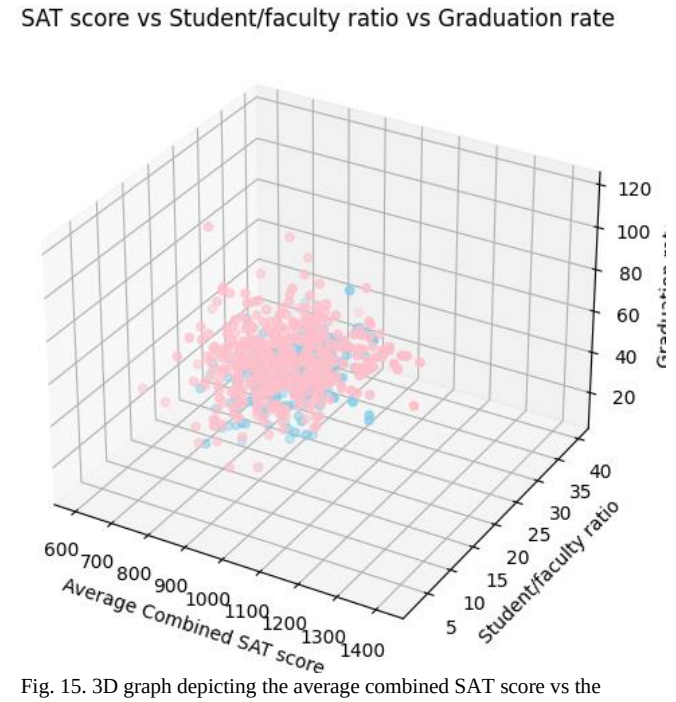

Fig. 15. 3D graph depicting the average combined SAT score vs the student/faculty ratio vs the graduation rate

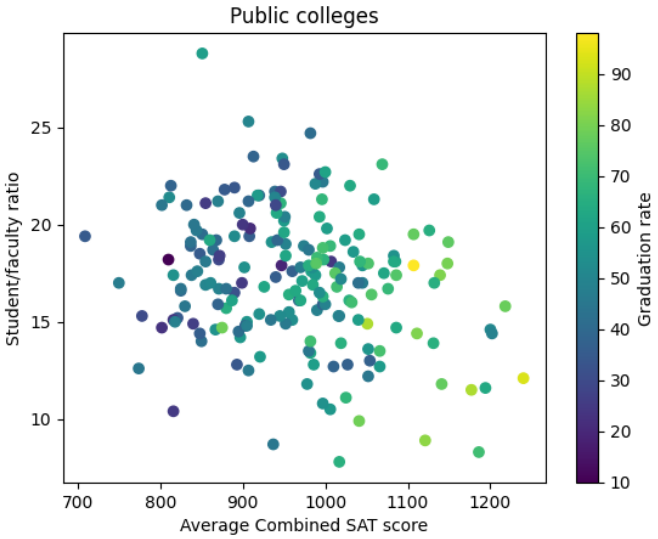

Fig. 16. Graph depicting SAT score and student/faculty ratio with graduation rate for public colleges.

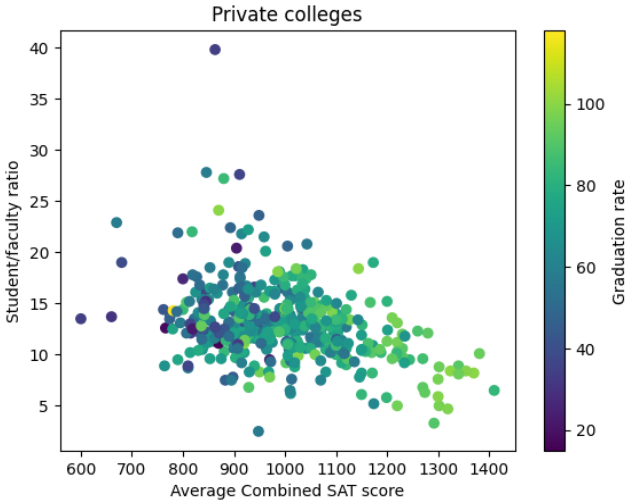

Fig. 17. Graph depicting SAT score and student/faculty ratio with graduation rate for private colleges.

From the above graphs, we can observe that when the average combined SAT score increases, the graduation rate also increases, which means that there is a direct correlation between the smartness of students and the graduation rate in the college. This proves part of our hypothesis that students who perform well in SAT also perform well in college. If we observe the graph, we do not see any significant relation between the student/faculty ratio and the SAT score or the graduation rate. It seems that the individual attention received by a student does not give a significant change in the graduation rate, nor is it a parameter for students having high SAT scores to join colleges with a good student/faculty ratio.

K. *Analysing the satisfaction and connection of the students towards the college after they become alumni of the college.*

1) *Procedure*

Distribute the data into two data frames, one for public and the other for private colleges. Next, we find the number of faculty members with high qualifications by taking the sum of faculty members with Ph.D. and terminal degrees. We then plot a 3D graph with the axes having percentage of faculty with high qualification, instructional expenditure per student and the percentage of alumni who donate. Also, plot two separate 2D graphs with these three variables for both public and private colleges. The following code snippet can be used:

```
faculty_pub = public['Pct. of faculty w
ith terminal degree'].astype(float) + p
ublic['Pct. of faculty with Ph.D.s'].as
type(float)
faculty_priv = private['Pct. of faculty
 with terminal degree'].astype(float) +
 private['Pct. of faculty with Ph.D.s']
.astype(float)
fig = plt.figure(figsize=(6, 6))
threedee = plt.axes(projection='3d')
threedee.scatter(public['Pct.alumni who
 donate'].astype(float), public['Instru
ctional expenditure per student'].astyp
e(float), faculty_pub)
threedee.scatter(private['Pct.alumni wh
o donate'].astype(float), private['Inst
ructional expenditure per student'].ast
ype(float), faculty_priv)
fig, ax = plt.subplots()
graph = ax.scatter(public['Instructiona
l expenditure per student'].astype(floa
t), faculty_pub, c=public['Pct.alumni w
ho donate'].astype(float))
var_three = fig.colorbar(graph)
```

2) *Observation and Results*

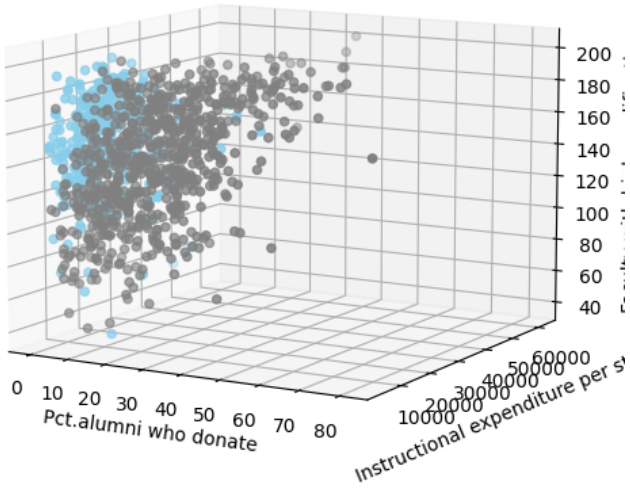

Fig. 18. 3D graph depicting relation between the percentage of alumni who donate vs the instructional expenditure per student vs the percentage of faculty with good qualification

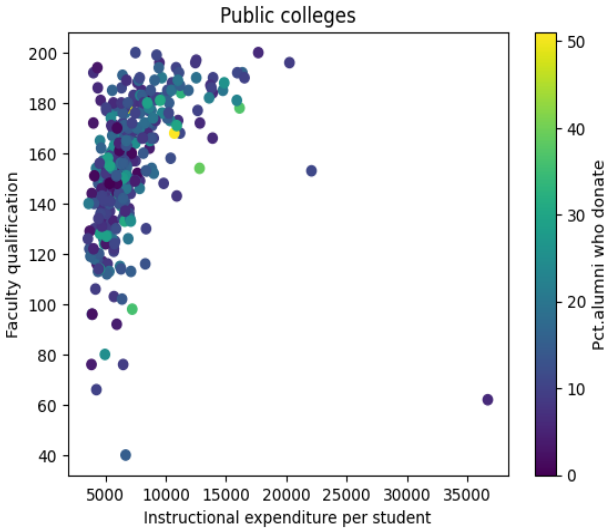

Fig. 19. Graph depicting instructional expenditure per student vs the faculty qualification with the percentage of alumni who donate for public colleges

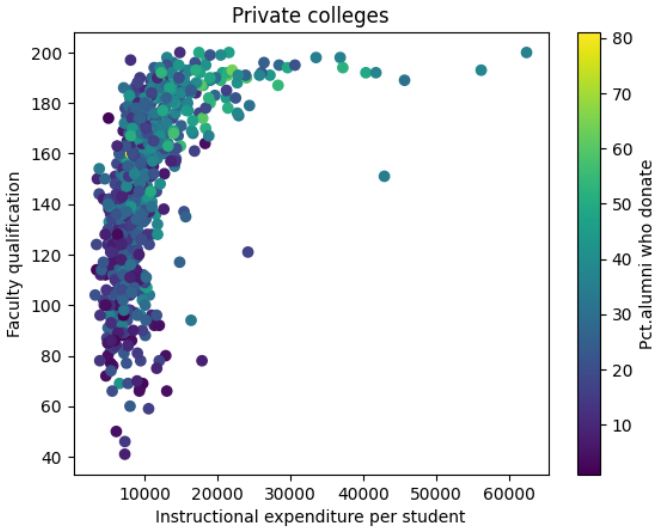

Fig. 20. Graph depicting instructional expenditure per student vs the faculty qualification with the percentage of alumni who donate for public colleges

In the above graphs, it can be observed that for public colleges, there is not much significant relation between the percentage of alumni who donate and the faculty qualification and the expenditure per student. For private colleges, although there is not a clear distinction but as the percentage of qualified faculty increases, the percentage of alumni who donate also increase. This percentage also slightly increases as the instructional expenditure per student increases. Hence we cannot establish a clear connection between the three parameters, which disproves our hypothesis of students being happy due to high percentage of qualified faculty and high expenditure per student.

L. Analysing the salary and compensation of faculty members by their distribution rank-wise.

1) Procedure

First, calculate the mean value of the number of faculty members, the average salary of them and the average compensation provided based on the rank. Then pie charts of each of these data are subplotted. The following code snippet can be used:

```
mean_sal = [profmeans, assocmeans, assi
stmeans]
mean_n = [profmean_n, assocmean_n, assi
stmean_n]
mean_c = [profmean_c, assocmean_c, assi
stmean_c]

labels = 'Full Professors', 'Associate
Professors', 'Assistant Professor'

plt.figure(figsize = (20, 6))
plt.subplot(131)
plt.pie(mean_n, labels=labels, autopct=
'%1.1f%%', startangle=90)
plt.axis("equal")
plt.title('Number of faculty')
```

2) Observation and Results

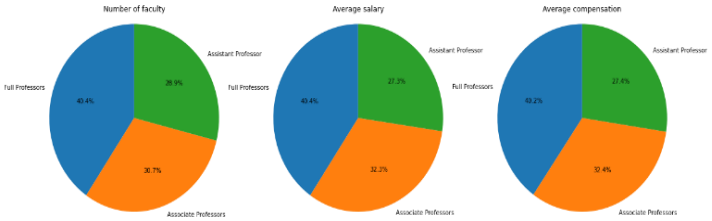

Fig. 21. Pie charts representing the average salary, average compensation and the number of faculty member by their rank

From the pie charts above, we can observe that the salary and compensation distribution is in accordance with the number of faculty members of each rank. That is the salary and compensation given is not concentrated towards any particular rank of the faculty member. We can see that the number of full professors is 40.4%, the average salary for them is 40.4% and the average compensation is 40.2%. All these values are very similar to each other. Similarly, for assistant professors, the values of all these parameters revolve around 28%. For associate professors, the value is approximately 31%.

V. Summary Of The Observations

Using the data about the colleges in the US, we analysed the relation between the average salaries amongst the faculty

members and proved our hypothesis about the increase in salary as we go higher in the rank. Some of the data proved our hypotheses true, like the fees for in-state and out-of-state for private and public colleges, while others like the relation of the instructional expenditure per student with the graduation rate when plotted proved our hypothesis to be false. This shows that trends keep on changing with time and perhaps for some of the hypothesis, more data was required.

VI. Unanswered Questions

1) Which college has the highest percentage of new students from the top 10%of their high school class, and is there a relationship between this percentage and the tuition rate?

2) How does a college's location (based on the state postal code) affect its average SAT scores and tuition rates?

3) How does the ratio of full-time to part-time undergraduates at a college relate to the percentage of faculty with Ph.D.'s?

4) What is the average acceptance rate of universities in different regions of the United States?

5) What is the distribution of faculty salaries across different regions of the United States, and are there any notable differences or outliers?

Acknowledgment

The author would like to thank Prof. Shanmuganathan Raman and his entire team for their support and guidance in this project.

References

[1] Swati, Mamatha, Rajendra Das, and Sameer Kamat. "Difference between in-State vs out-of-State Tuition." MBA Crystal Ball, March 17, 2017. https://www.mbacrystalball.com/blog/2017/03/17/in-state-vs-out-of-state-tuition-differences/.

[2] "Pandas Documentation#." pandas documentation - pandas 1.5.3 documentation. Accessed March 30, 2023. https://pandas.pydata.org/docs/.

[3] "Matplotlib 3.7.1 Documentation#." Matplotlib documentation - Matplotlib 3.7.1 documentation. Accessed March 30, 2023. https://matplotlib.org/stable/index.html.

[4] NumPy documentation. Accessed March 30, 2023. https://numpy.org/doc/.