ES 114: Data Narrative 2

Jaidev Sanjay Khalane (22110103)

B.Tech 2022
Computer Science and Engineering
Indian Institute of Technology
Gandhinagar, India
jaidev.khalane@iitgn.ac.in

Abstract— **This report is about the analysis of the data given on the website http://lib.stat.cmu.edu/datasets/colleges/ about the various parameters of American schools, as given in the overview part. Various scientific questions and hypotheses were also created in order to effectively study the data. The software used is Python 3 with modules such as NumPy, Pandas and matplotlib**

I. Overview Of The Dataset

In this dataset, data about various aspects of American schools has been given. The first data set gives data about the school's state, the school's type, and the school's expenditure towards salaries and compensations of faculty and staff at various levels. The dataset also presents the number of faculty and staff members in the schools. The second data set presents the data about average SAT scores in various subjects, such as math and verbal ability required by the students in order to be admitted to the schools. It also mentions the average ACT scores that are required. This dataset also provides data on other parameters such as student-to-faculty ratio, student expenditure, the total number of undergraduate students etc. On the whole, this dataset provides a very wide spectrum of parameters that are analysed using various libraries of python such as Pandas, MatPlotLib, Numpy and Scipy.

II. Scientific Questions about the Dataset

AAUP

Q1. Which state in America has the minimum number of schools? How can we visualise the relative number of schools in various states of America against the names of the states?

Q2. What is the relative proportion of the number of universities against the number of colleges and institutions? How many schools fall under the various categorisations of group I, IIA, IIB, and VIIB? How can we visualise this data?

Q3. What is the mean, variance and standard deviation of the average salaries of full professors, associate professors and assistant professors? How can we visualise it with a histogram? How would a smooth plot of the probability distribution of the salary of the faculties against the probability of finding the faculty with that salary look like, and what can we analyse from that?

Q4. What is the mean, variance and standard deviation of the average compensation of professors, associate professors and assistant professors? How can we visualise it with a histogram? How would a smooth plot of the probability distribution of the compensation of the faculties against the probability of finding the faculty with that compensation look like, and what can we analyse from that?

Q5. Find the mean, variance and standard deviation in the number of full professors, assistant professors, associate professors and instructors at the schools. What would a histogram of the number of faculty members at various schools look like? How would the probability distribution smooth curve of the number of faculties at various ranks look for all the schools, and what can we analyse from that?

Q6. What is the mean, variance and standard deviation of the expenditure of the schools towards salary and compensation? What would a histogram of various schools in this aspect look like? What would the probability distribution curve of getting a school having an expenditure of a particular value look like, and what can we analyse from that?

Q7. What is the mean, variance and standard deviation of the total expenditure of the school? What would a histogram of various schools look like? What would the probability distribution curve of getting a school having an expenditure of a particular value look like, and what can we analyse from that?

- ***Additional Questions at the end***

USNEWS

Q1. How many schools in the survey are privately owned, and how many of them are publicly owned? What is the ratio of the number of private schools to the number of public schools? How can we visualise the relative number of public and private schools?

Q2. What is the mean, variance and standard deviation in the average SAT score in math and verbal abilities accepted by the schools? What would the visualisation of the average SAT score in math and verbal abilities look like against the number of schools at that score? What would the distribution of the probability of getting a school at a particular average SAT math and verbal score look like?

Q3. What is the mean, variance and standard deviation in the average SAT score accepted by the schools? What would the visualisation of the average SAT score and the number of schools at that score look like? What would the distribution of the probability of getting a school at a particular score look like?

Q4. What is the mean, variance and standard deviation in the average ACT score accepted by the schools? What would the visualisation of the average ACT score and the number of schools at that score look like? What would the distribution of the probability of getting a school at a particular score look like?

Q5. What is the mean, variance and standard deviation of the acceptance rate in various schools? How can we visualise the trend of acceptance rate of various schools, and how would the probability distribution curve of getting a school having a particular rate of acceptance look like? What are the five schools having the least rate of acceptance? (Most difficult schools to get admission into.)

Q6. What is the mean, variance and standard deviation of the number of undergraduate students in the schools? How can we visualise the trend of the number of undergraduate students in the schools? What is the probability of getting a school with a particular number of undergraduate students, and what would its graph look like? What are the five schools having the most and the least number of students at the undergraduate level?

Q7. What is the mean, variance and standard deviation of the total expenditure of the undergraduate students at the schools? How can we visualise the trend of the number of schools having a total expenditure within a certain range? What are the top 5 schools having the most expenditure and the five schools having the least expenditure with respect to the students? What would a probability distribution of getting a school with a particular expenditure for undergraduate students look like?

Q8. What is the mean, variance and standard deviation of the total graduation rate of the schools? What would a histogram of the number of schools having a graduation rate within a particular range look like? What would the probability distribution of getting a school having a particular rate of graduation look like? What are the five schools having the largest rate of graduation and the five schools having the least rate of graduation?

Q9. What is the mean, variance and standard deviation of the student-by-faculty ratio at various schools in America? What would the trend of the number of schools having the student-by-faculty ratio within a particular range look like? What would a probability distribution of getting a school having a particular student-by-faculty ratio look like? What are the top five schools having the largest amount of student-by-faculty ratio and the top five schools having the least amount of student-by-faculty ratio?

- ***Additional Questions at the end***

III. Details of Libraries used

The libraries that are used include matplotlib, NumPy, and Pandas.

• matplotlib.pyplot: This library was used for applications such as plotting in the form of a line graph, bar graph, histogram, pie chart etc. and other applications such as legend, the label of axes, title, grid, etc.

• pandas: pandas were used for reading the data from .csv files, creating data frames, iterating through them, and deleting the wrongly formatted data points using the function df.drop. The kernel density estimation KDE was also used to plot smooth probability distributions.

• NumPy: I used the NumPy library for creating NumPy arrays for faster computing and storing data, for calculating mean, for getting the unique elements out of the NumPy arrays, and by enabling the counts, I also got the number of times the unique element has occurred in the array and its index. It was also used to calculate the mean, standard deviation and variance of the data.

IV. Answers to the Questions and Summaries Drawn

AAUP

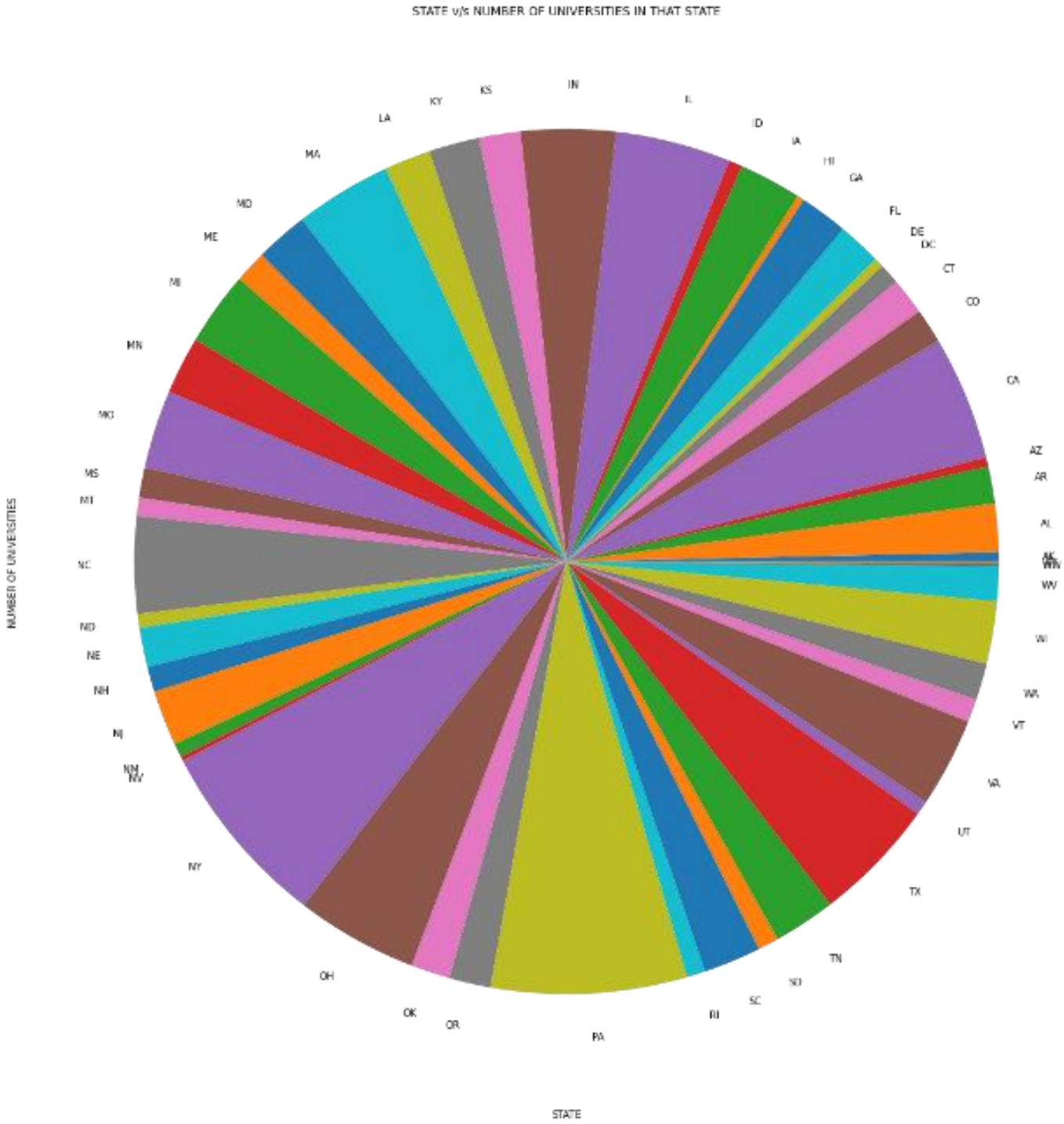
Q1. Which state in America has the minimum number of schools? How can we visualise the relative number of schools in various states of America against the names of the states?

A1. Approach: First of all, the file was opened as a data frame using the Pandas function for .csv files. Then the column of states was extracted as NumPy array and then using the numpy.unique() function, the number of occurrences of unique elements was counted which was later plotted in the form of a bar graph and a pie chart. Then using the following code, the minimum of that array was extracted

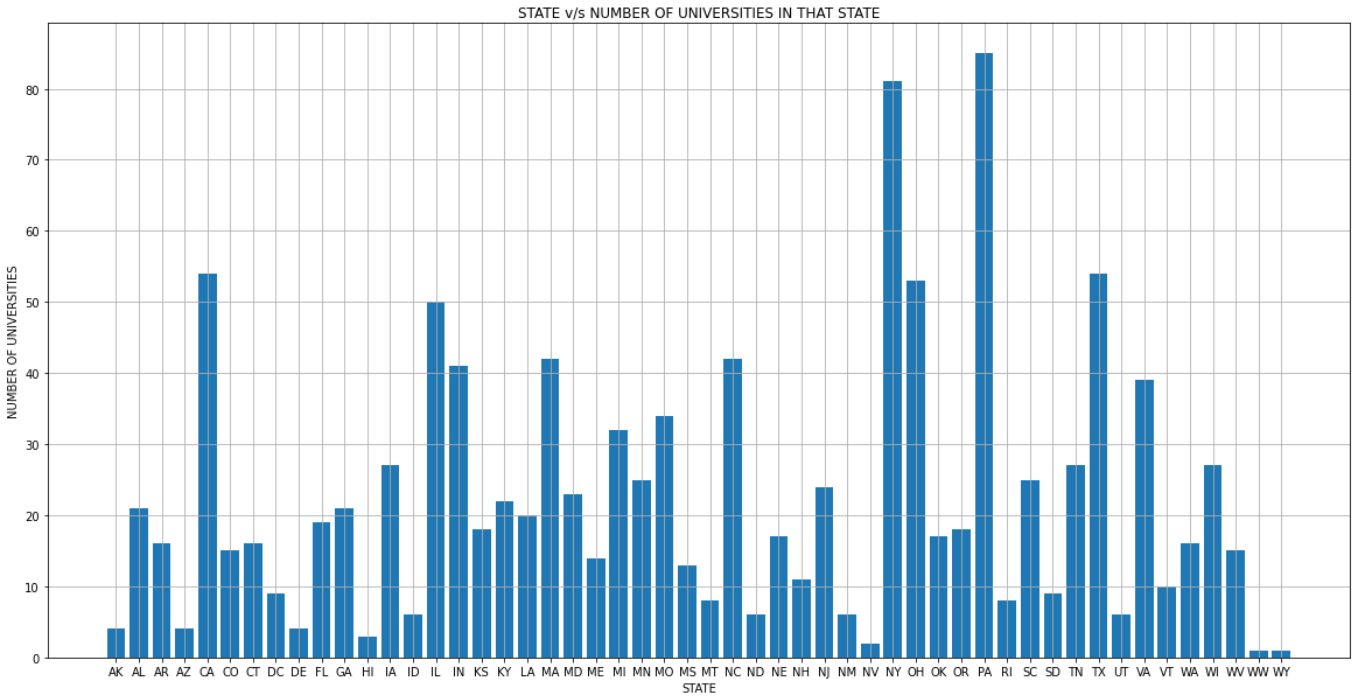
```
min=0
ind=0
```

```
for i in range(len(counts)):
  if counts[i]<min:
    ind=i
print(statenames[ind])
```

Results and Observations:



WY= Wyoming State



We can observe from the graphs that most of the states have almost more than 10 and less than 40 schools, but there are a few states, like New York, California, Texas, etc. which have more than 50 schools. We can also see from the pie chart that these States have very large contribution to the total number of schools. The state with the lowest number of schools is Wyoming followed by Alaska and other states.

Q2. What is the relative proportion of the number of universities against the number of colleges and institutes? How many schools fall under the various categorisations of group I, IIA, IIB, and VIIB? How can we visualise this data?

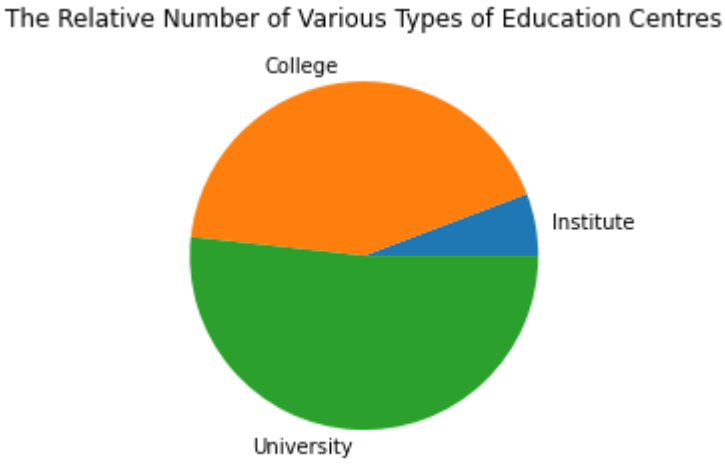
A2. Approach: First of all, the file was opened as a data frame using Pandas and then the column containing the names of the college was converted into an array. Then, there were three variables created named University counter college counter and Institute counter so that if the name of the school had the word college the college counter was increased by one. Similar thing was done with university counter and Institute counter. Then these counters were plotted as given below with their respective labels in the form of a pie chart. The number of universities, colleges and institutes was also printed.

```
for item in college_names:
  if "Univ" in item:
    univ_counter+=1
  elif "College" in item:
    college_counter+=1
  else:
    institute_counter+=1
```

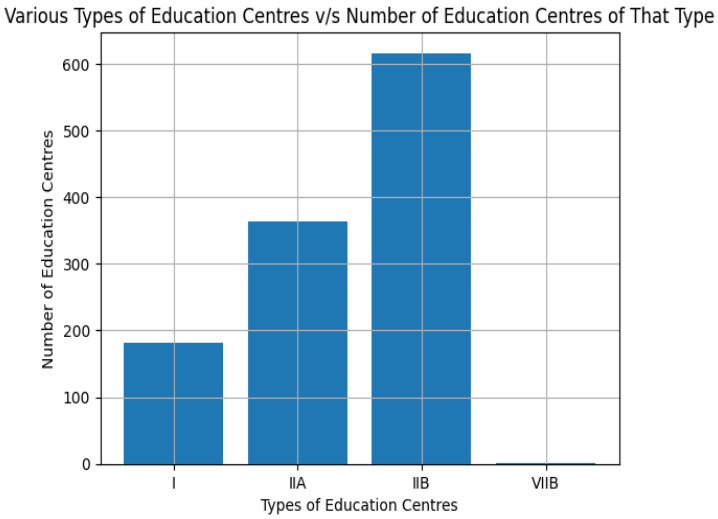
Result and Observation:

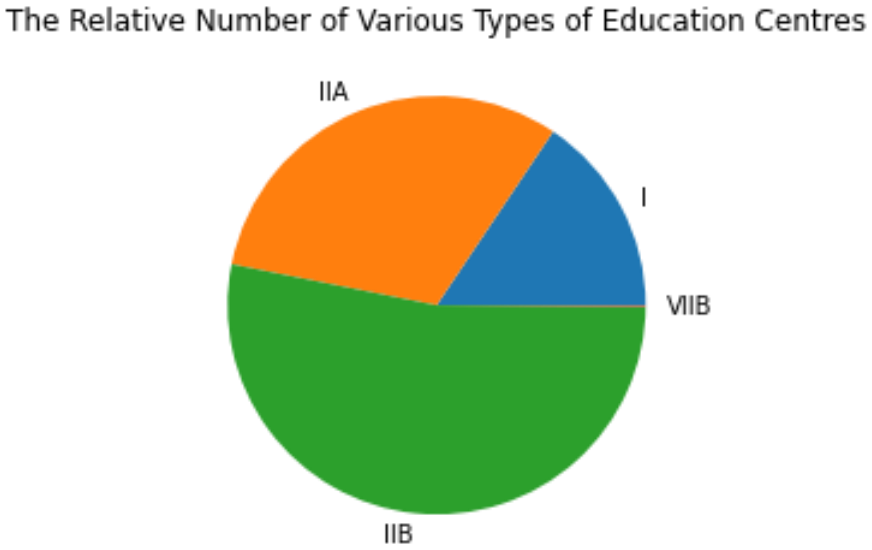
```
Number of Universities: 600
Number of College: 493
Number of Institute: 67
```

As we can observe from the pie chart that majority of the schools are of University type, very few of them are of Institute type and rest of them are colleges.



Part 2: Approach: The approach to this part of the question was similar to that of the first question, the only difference being that the column that was taken as axis was the column containing the type of the schools in terms of Roman numbers.
Result and Observation: We can observe from the results that most of the schools are of type IIB, IIA and some are of type I. Very few schools are of type VIIB.



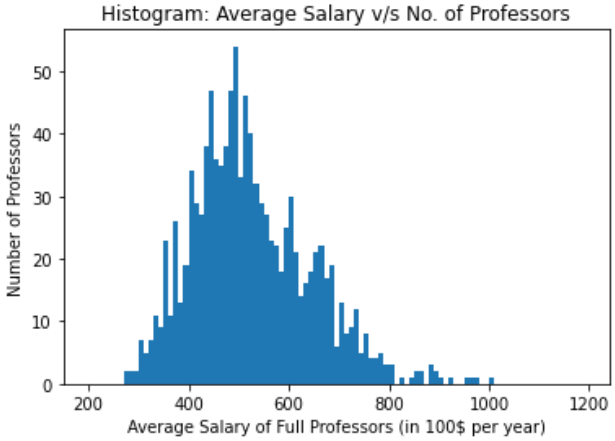


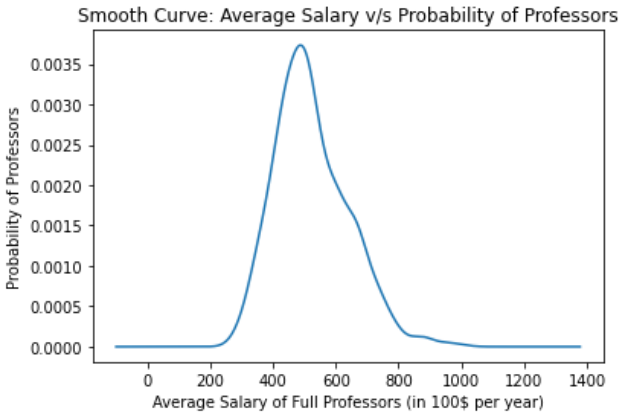
Q3. What is the mean, variance and standard deviation of the average salaries of full professors, associate professors and assistant professors? How can we visualise it with a histogram? How would a smooth plot of the probability distribution of the salary of the professors against the probability of finding the faculty with that salary look like, and what can we analyse from that?

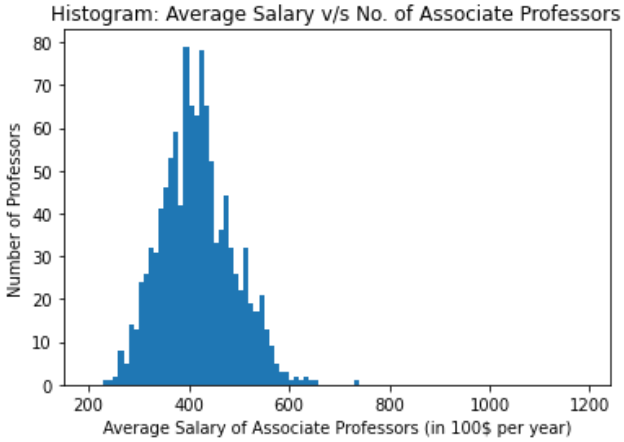
A3. Approach: First of all, the data set was opened as a data frame using Pandas. Then all the rows were iterated in the columns containing the salaries of professors and the columns having invalid data were dropped by using the Pandas function of df.drop. Then the column of salaries was converted into an array which was then mapped as integer entries and then using the numpy functions of mean, variance and standard deviation, the respective values were obtained. Then using the matplolib library, the histogram of salaries was plotted which was then converted into a pandas series and using kernel density estimation (kde), the smooth curve of the probability distribution number of professors versus salary was plotted. The same approach was used for associate professors, assistant professors and all professors also.

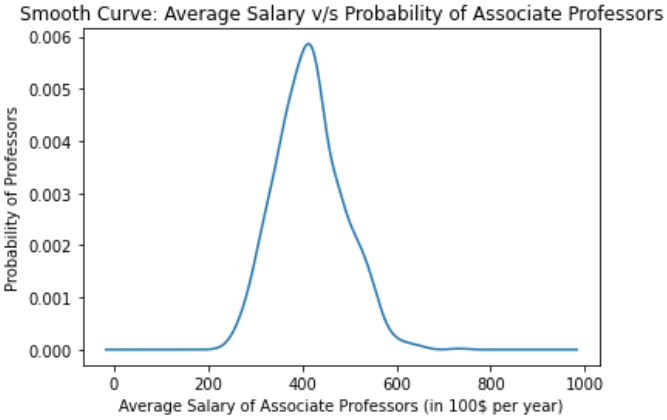
Results and Observations:

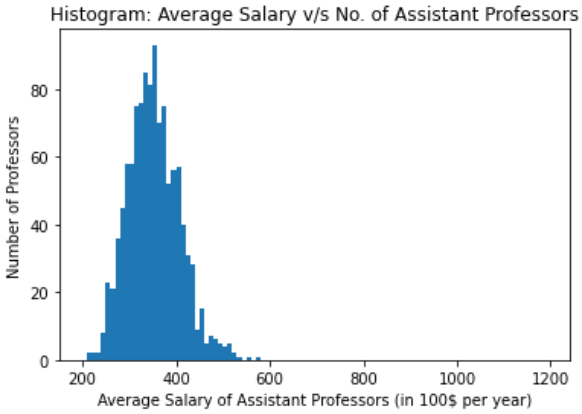
```
Mean Salary of Full Professors: 524.2087912087912
Variance Salary of Full Professors: 13993.976552751277
Standard deviation Salary of Full Professors: 118.2961392132105
Mean Salary of Associate Professors: 416.4261565836299
Variance Salary of Associate Professors: 5117.294369213916
Standard deviation Salary of Associate Professors: 71.53526661174834
Mean Salary of Assistant Professors: 351.9181338028169
Variance Salary of Assistant Professors: 3001.2195303201256
Standard deviation Salary of Assistant Professors: 54.78338735711881
Mean Salary of All Professors: 420.40344827586205
Variance Salary of All Professors: 8515.568263971463
Standard deviation Salary of All Professors: 92.27983671404856
All the values are in terms of 100$
```

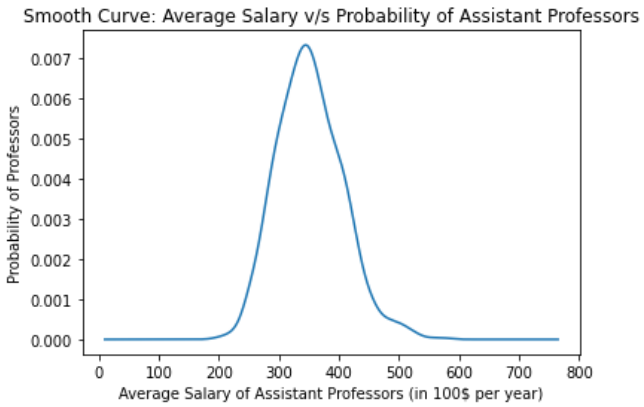


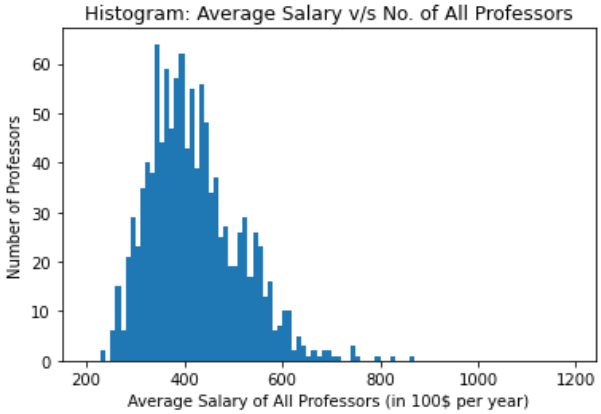


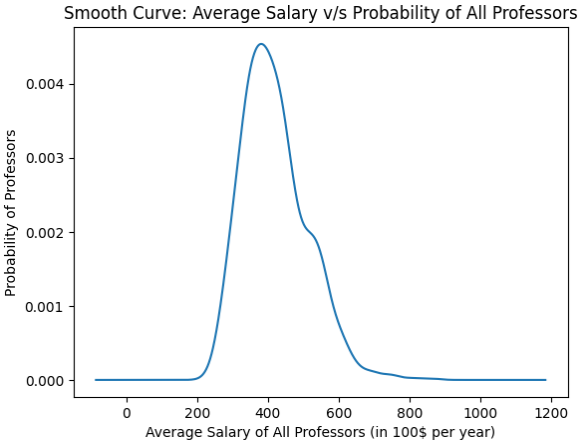












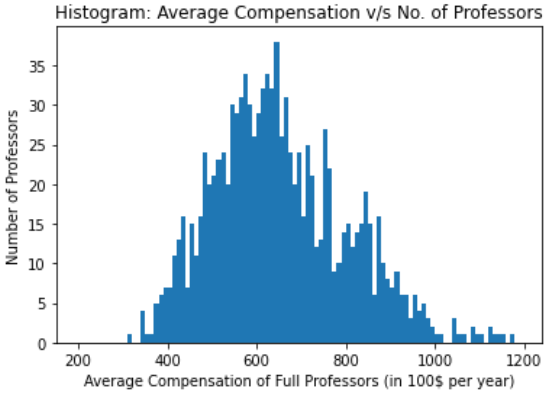
We can observe that the salaries of full professors is much greater than the salaries of associate professors and the salaries of assistant professors. The graphs, however, look similar to a normal distribution type of a function having a peak at a certain value and then decreasing on both sides. We can also see that the variance in the salaries of full professors is the largest, then followed by the variance in the salaries of associate professors, followed by the variance in the salaries of assistant professors. From this, we can say that the distribution of salaries of full professors is more widely distributed compared to the distribution of salaries of associate professors and assistant professors in the schools of America. This can be visually confirmed from the graphs and the histograms of the distributions also.

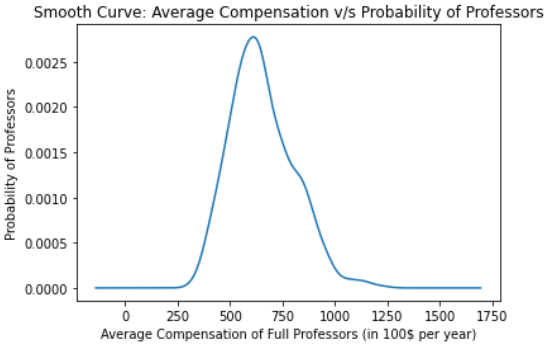
Q4. What is the mean, variance and standard deviation of the average compensation of professors, associate professors and assistant professors? How can we visualise it with a histogram? How would a smooth plot of the probability distribution of the compensation of the professors against the probability of finding the faculty with that compensation look like, and what can we analyse from that?

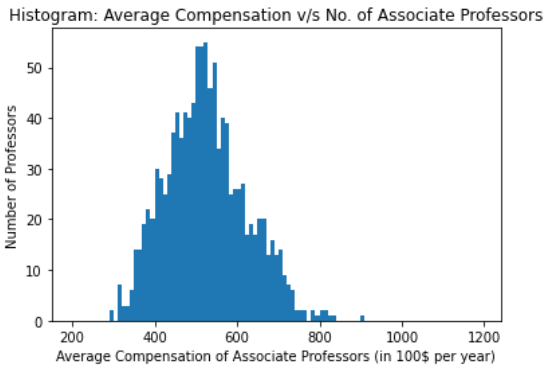
A4. Approach: Same as A3.
Results and Observations:

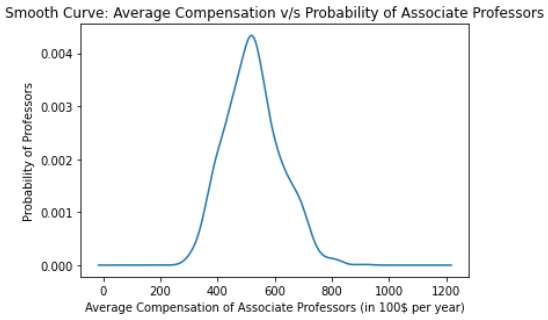
```
Mean Compensation of Full Professors: 653.5668498168499
Variance Compensation of Full Professors: 23022.652124508582
Standard deviation Compensation of Full Professors: 151.73217234492026
Mean Compensation of Associate Professors: 523.8274021352313
Variance Compensation of Associate Professors: 9423.922167272449
Standard deviation Compensation of Associate Professors:
97.07688791505653
Mean Compensation of Assistant Professors: 442.08714788732397
Variance Compensation of Assistant Professors: 5694.127088344327
```

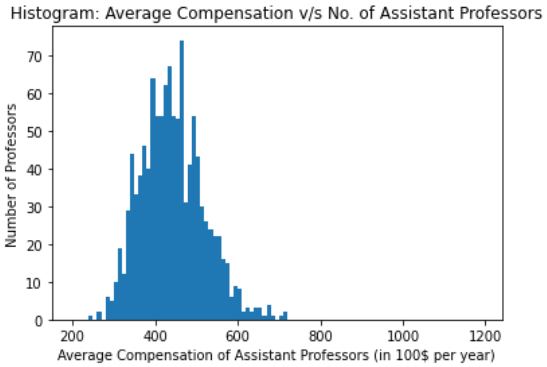
```
Standard deviation Compensation of Assistant Professors:
75.45944002140705
Mean Compensation of All Professors: 526.7275862068966
Variance Compensation of All Professors: 14565.818894173603
Standard deviation Compensation of All Professors: 120.68893443134546
All the values are in terms of 100$
```

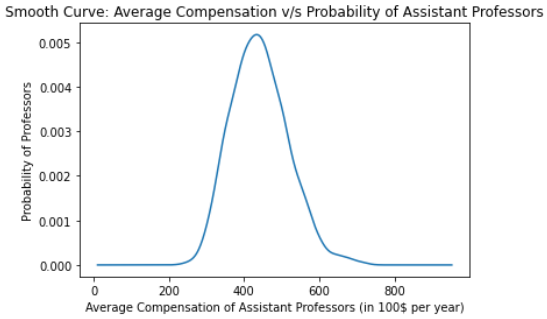


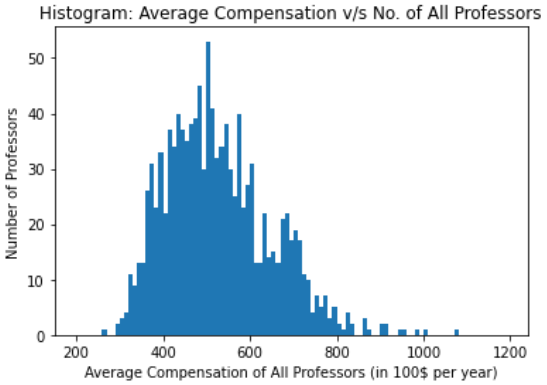


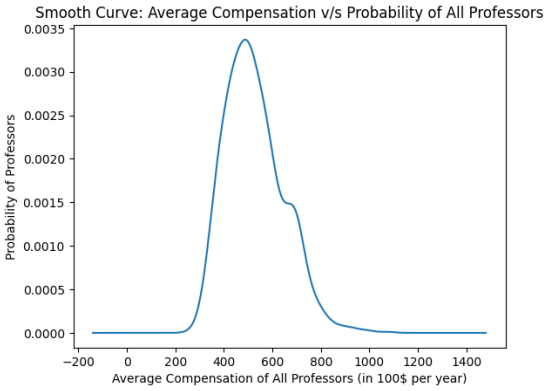






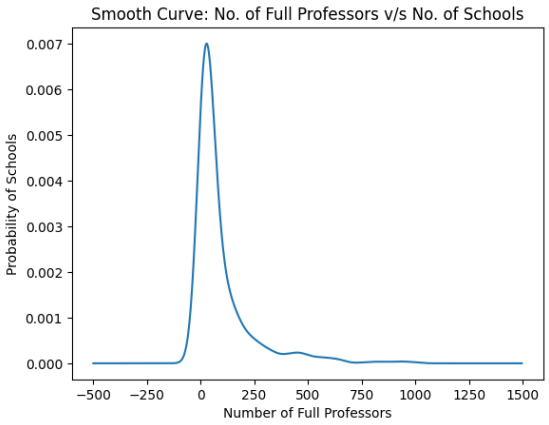






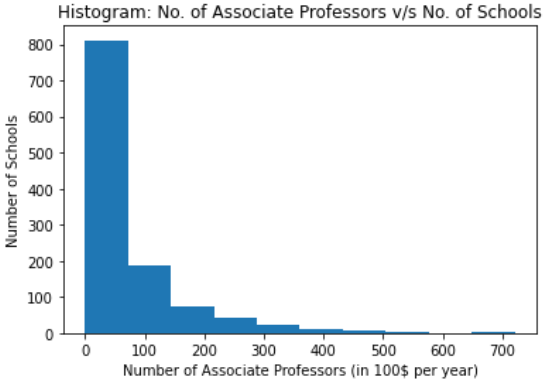
We can observe from the results that the compensation of full professors, in general is greater

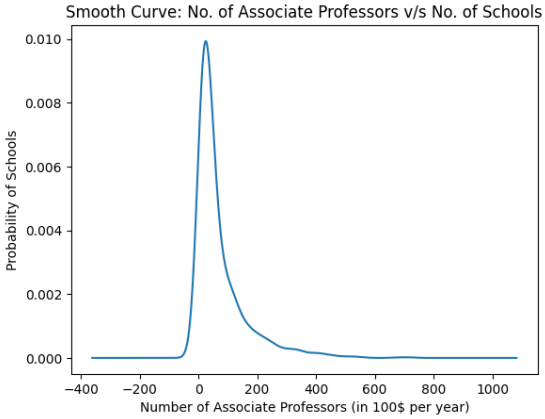
than the compensation of associate professors and assistant professors. But the variance in the compensations of full professors is much greater than the variance in the compensations of assistant and associate professors. We can analyse from the graph also that the peak of the graphs of the smooth function of the probability distribution and the histogram are very close to the mean value of the distribution and the wider graphs have more variance as compared to the narrow graphs. The distribution of the probability of most of the curves in this case is very similar to the Gaussian distribution.

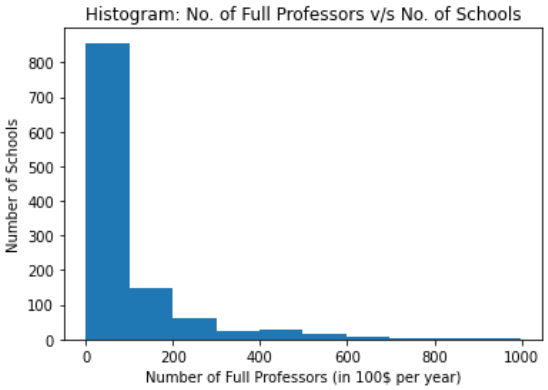


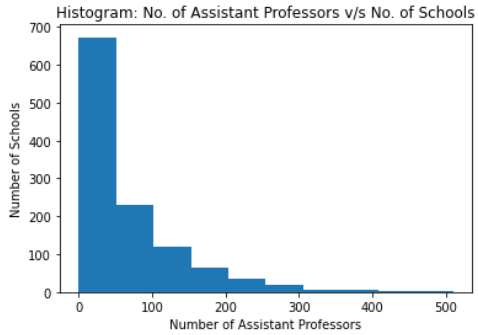
Q5. Find the mean, variance and standard deviation in the number of full professors, assistant professors, associate professors and instructors at the schools. What would a histogram of the number of faculty members at various schools look like? How would the probability distribution smooth curve of the number of faculties at various ranks look for all the schools, and what can we analyse from that?

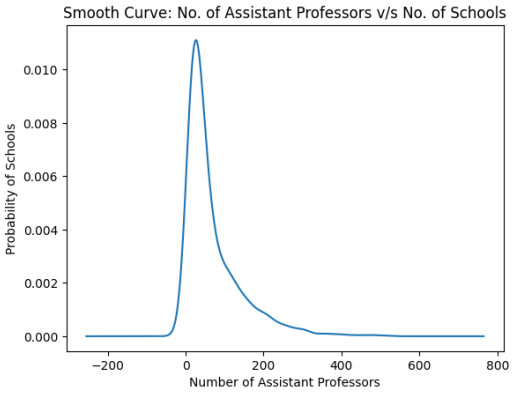
A5. Approach: First of all, the data file was opened as a data frame using the Pandas .read-csv function. Then, all the rows containing the number of faculty members were analysed and the rows containing invalid format of data were dropped using df.drop of Pandas. Then the number of faculty members' column was formed into a numpy array and then mapped as an integer list and sorted. Then using the numpy library, mean, variance and standard deviation were calculated. Then the matplotlib was used to plot the histogram of the number of faculty members. The numpy array was later converted into a pandas series, which was plotted using matplotlib and kernel density estimation (kde) to give a smooth curve of probability of getting a school with a particular number of faculty members.
Results and observation:

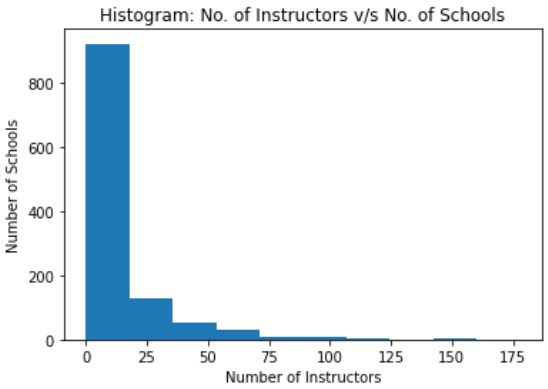


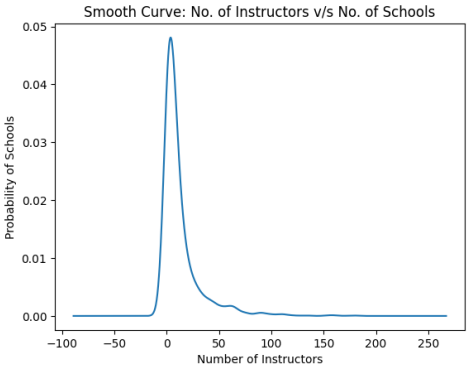













```
Mean Number of Full Professors: 95.17327586206896
Variance Number of Full Professors: 20324.620837544593
Standard deviation Number of Full Professors: 142.56444450684256
Mean Number of Associate Professors: 72.43793103448276
Variance Number of Associate Professors: 7992.704768133176
Standard deviation Number of Associate Professors: 89.40192821261282
Mean Number of Assistant Professors: 68.68620689655172
Variance Number of Assistant Professors: 5297.211878715815
Standard deviation Number of Assistant Professors: 72.78194747817494
Mean Number of Instructors: 12.743103448275862
Variance Number of Instructors: 380.7340041617123
Standard deviation Number of Instructors: 19.51240641647545
```

From the data analysis, we can observe that, on an average, schools have almost 95 full professors, 72 associate professors, 68 assistant professors and 12 instructors. The variance in the number of full professors is the greatest followed by the variance in the number of associate professors, assistant professor and the number of instructors. We can see from the probability distribution also that the peak is almost at the point of mean and the expanse of the probability distribution curve gives an idea about the variance, that is, greater the expanse more will be the variance. We can also see from the histogram that a few schools may also have has many as 1000 full professor, almost 700 associate professor and almost 500 assistant professors and as many as 150 instructors.

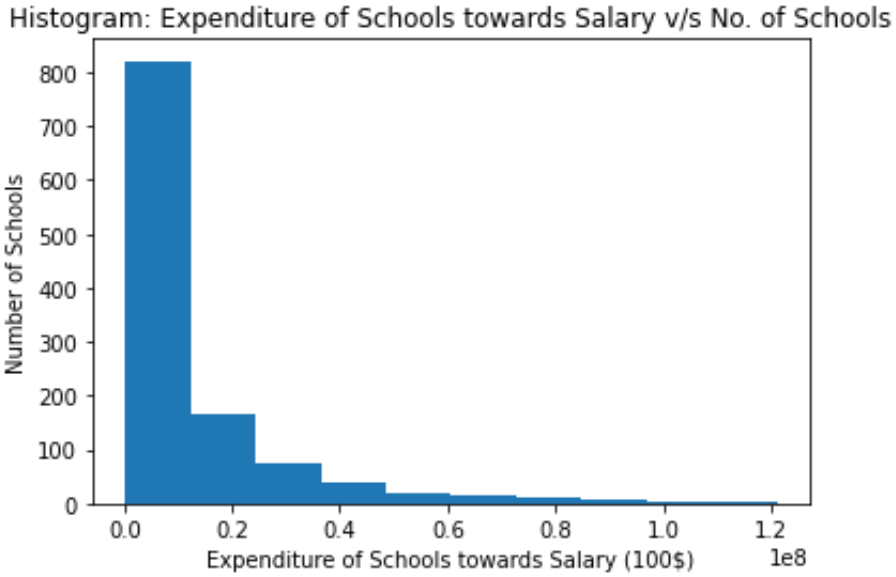
Q6. What is the mean, variance and standard deviation of the expenditure of the schools towards salary and compensation? What would a histogram of various schools in this aspect look like? What would the probability distribution curve of the schools having an expenditure of a particular value look like, and what can we analyse from that?

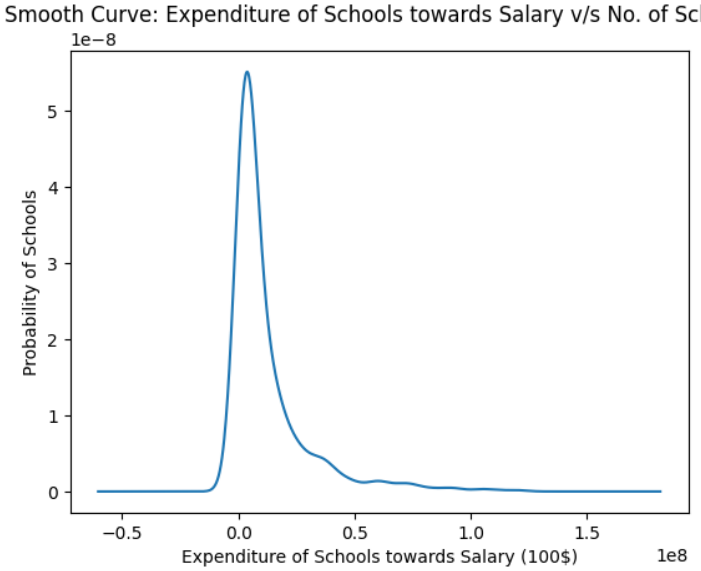
A6. Approach: First of all, the data set was accessed and opened as a data frame using the Pandas library. Then, all the rows were iterated under the columns of expenditure towards salary and the number of faculty members and the wrongly formatted rows were dropped using df.drop function of Pandas.

Then, the number of faculty members and the total salary were multiplied and added into a new array so as to form an array which contains the total expenditure of the school towards the salary. Then, using the numpy functions, mean, variance and standard deviation were calculated. Then using the matplotlib, a histogram of the expenditure of the school towards salaries and the number of school was plotted. Then the probability distribution was plotted using the kde function of Pandas in order to give us the probability distribution of getting a school with a particular expenditure towards salary. A similar approach was used for total compensation also.

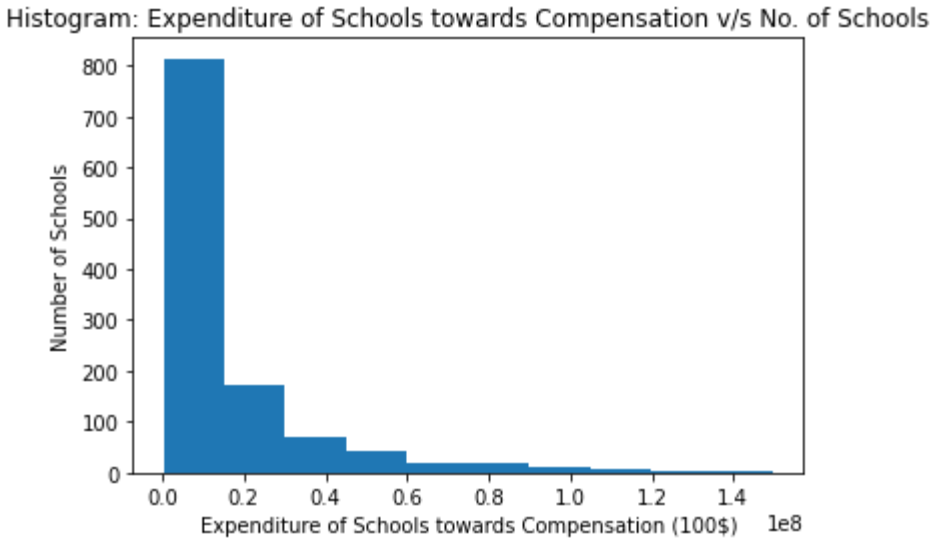
Results and Observations:

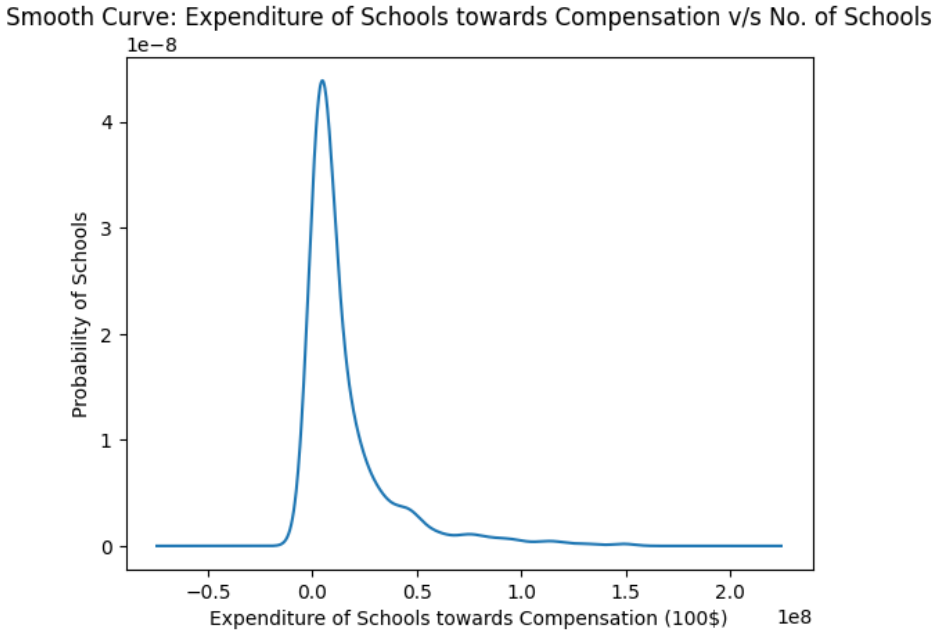
```
Mean Expenditure of Schools towards Salary (100$): 12479640.948275862
Variance Expenditure of Schools towards Salary (100$): 308802896436262.9
Standard deviation Expenditure of Schools towards Salary (100$):
17572788.52192397
```






```
Mean Expenditure of Schools towards Compensation (100$):
15642040.344827587
Variance Expenditure of Schools towards Compensation (100$):
483507476236958.44
Standard deviation Expenditure of Schools towards Compensation (100$):
21988803.42894898
```





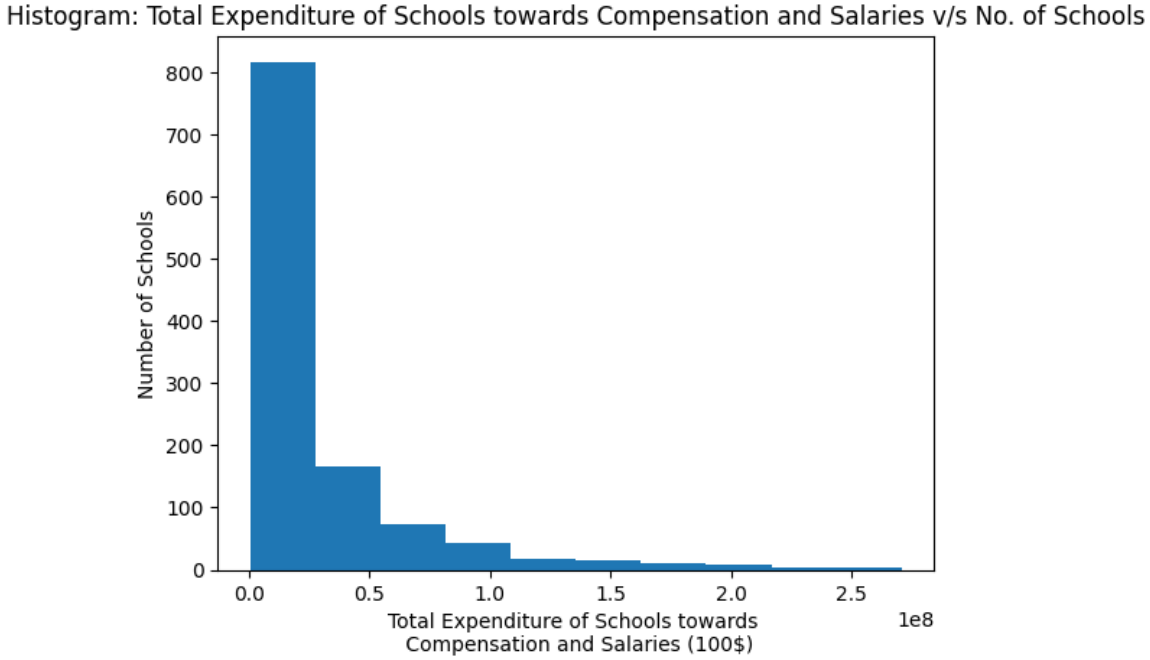
As we can see from the values of mean expenditure towards compensation and mean expenditure towards the salary, we can see that the schools spend more amount of money towards compensation compared to the money spent towards salary. Also, we can see that the variance in the total composition is also greater than the variance of total salary expenditure. This can be visually confirmed from the graphs also.

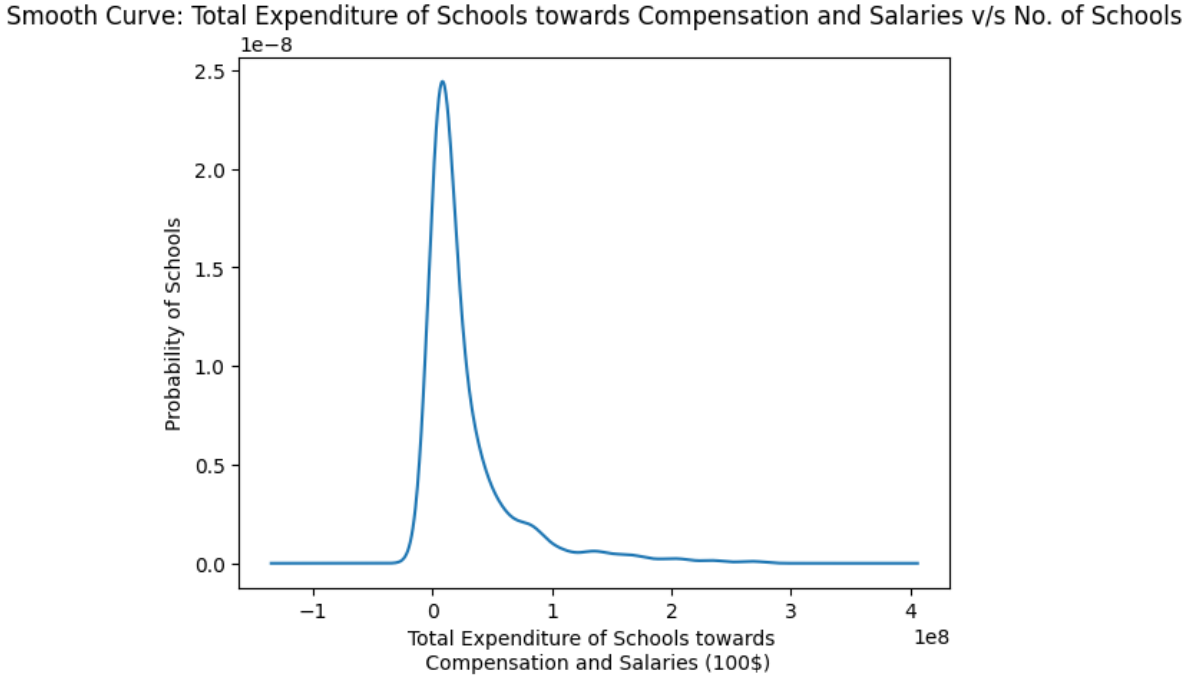
Q7. What is the mean, variance and standard deviation of the total expenditure of the school? How would a histogram of various schools look like? How would the probability distribution curve of the schools having an expenditure of a particular value look like?

A7. Approach: In order to get the total expenditure of the school towards salaries and compensations in total, first of all, the data set was opened as a data frame using the Pandas library. Then, all the rows were iterated in order to find wrongly formatted cells in the columns of number of total faculty members, average salary and average composition and if any wrongly formatted row was found, it was dropped using df.drop of the Pandas. Then, a new array was created in which the sum of the average salary and compensations was multiplied by the number of faculty members in order to get the total expenditure towards the salaries and compensations. This array was then sorted and plotted as histogram using the matplotlib library and then using the numpy library, the mean, variance and standard deviation of the total expenditure of the schools calculated. Using the Pandas library, kernel density estimation was also plotted in order to give probability distribution of

getting a school with a particular expenditure towards salaries and compensations.

Results and Observations:






```
Mean Total Expenditure of Schools towards Compensation and Salaries
(100$): 28121681.29310345
Variance Total Expenditure of Schools towards Compensation and Salaries
(100$): 1564563536164520.5
Standard deviation Total Expenditure of Schools towards Compensation
and Salaries (100$): 39554564.03709337
```

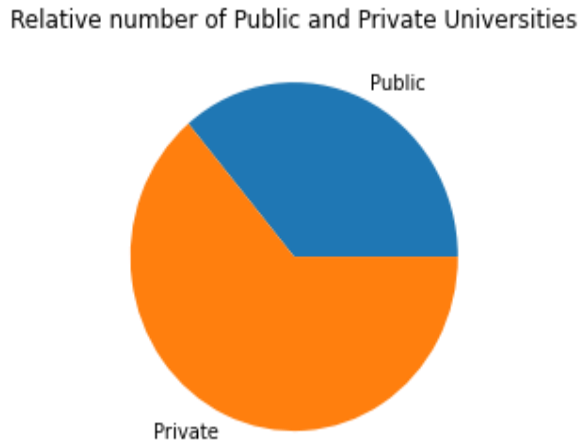
We can see from the graph that the number of schools having very large expenditure is very less compared to the number of schools having a small expenditure. From the histogram and the probability density curve, we can also see that as the expenditure increases, the number of schools having that particular expenditure decreases significantly. Moreover, we can see the mean, standard deviation and variance from the values given above

US NEWS

Q1. How many schools in the survey are privately owned, and how many of them are publicly owned? What is the ratio of the number of private schools to the number of public schools? How can we visualise the relative number of public and private schools?

A1. Approach: To get the data on the number of public and private schools, the data file was opened as a data frame using the Pandas library (.read_csv) function. Then the column containing the data about the public and private nature of the schools was converted into a numpy array such that if the entry of the array was 2, then the school would be public; otherwise, the school would be private. Then the number of unique elements and the count of occurrences was found by using numpy.unique, which returns the type of the school with the count which was printed. In order to get the ratio of the number of private and public schools, the respective counts were divided by each other and in order to visualise the relative number of public and private schools, they were plotted as a pie chart using the matplotlib library.

Results and Observations:




```
Number of private schools= 831
Number of public schools= 471
ratio of the number of private
and public schools=
1.7643312101910829
```

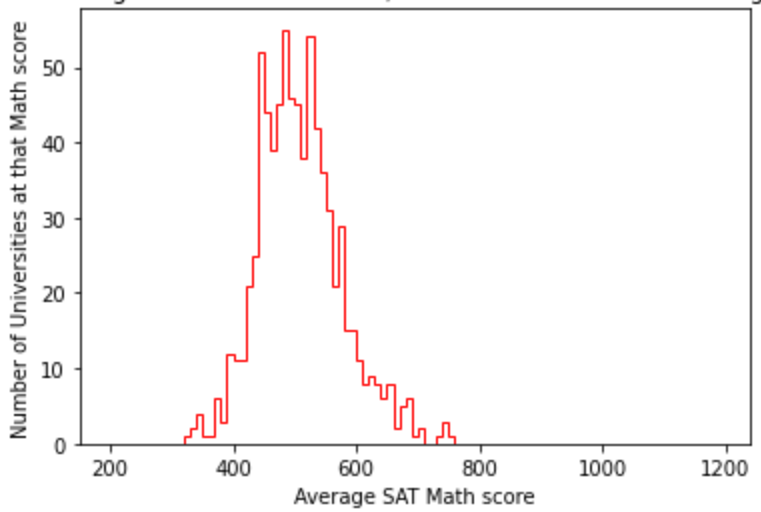
We can observe from the data that there are more private schools compared to public schools, and for every public school, there are 1.76 private schools, as the ratio of the number of private schools to the ratio of number of public schools is almost 1.76.

Q2. What is the mean, variance and standard deviation in the average SAT score in math and verbal abilities accepted by the schools? What would the visualisation of the average SAT score in math and verbal abilities look like against the number of schools at that score? What would the distribution of the probability of getting a school at a particular average SAT math and verbal score look like?

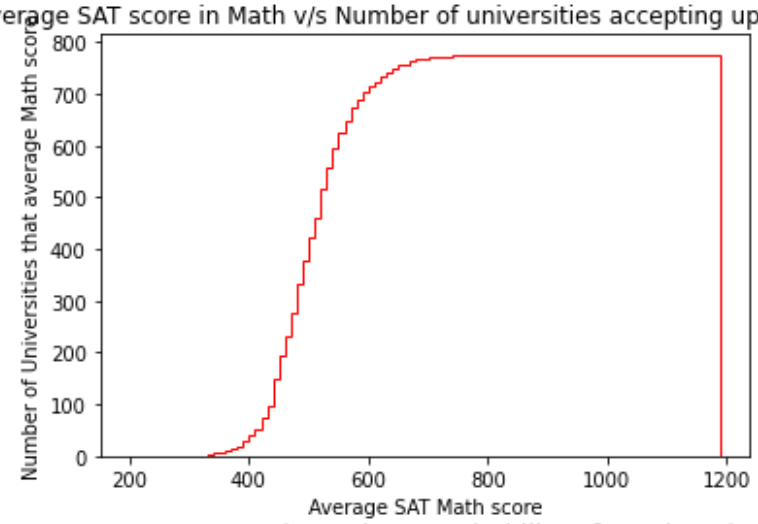
A2. Approach: First of all, the file containing the data was opened as a data frame using the Pandas library. Then, all the rows were iterated in the column of the SAT score of mathematics in order to find wrongly formatted rows, and if a row was found to be wrongly formatted, it was removed using df.drop function of the Pandas library. Then this column was converted into a numpy array which was sorted and mapped as integers. Using the numpy library, the mean, variance and standard deviation of the array was calculated. Then, using the matplotlib library, histograms of various kind were plotted such that the first one was a normal histogram with the histogram type of step; whereas the second one was a cumulative type histogram. Then, the array was converted into a Pandas series and using the kde (kernel density estimation) function of Pandas, the probability distribution of getting a school with a particular SAT score in maths was plotted. A similar procedure was extended for SAT score in verbal abilities also.

Results and Observations:

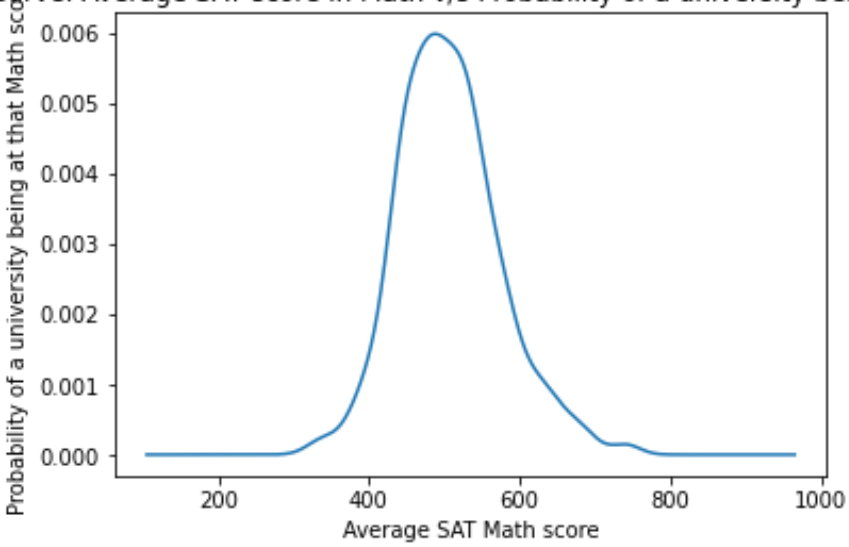
Histogram: Average SAT score in Math v/s Number of universities being at that score



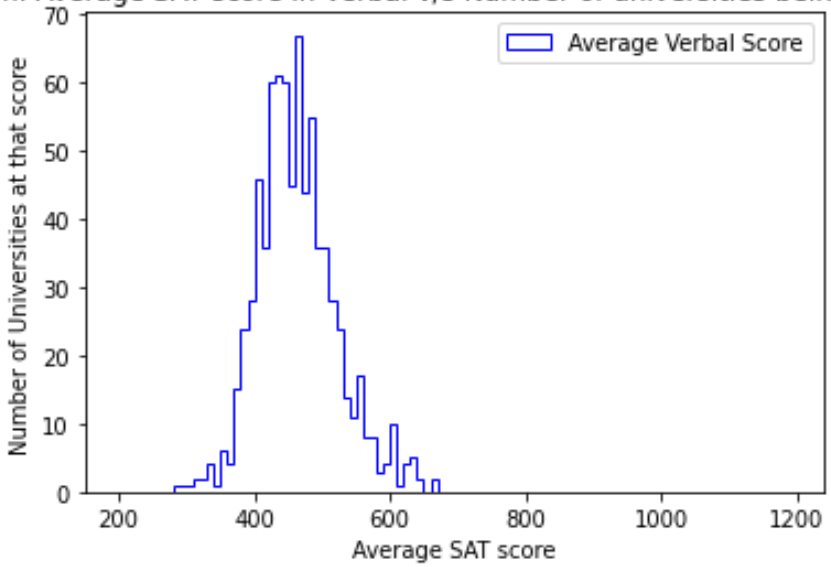
Histogram: Average SAT score in Math v/s Number of universities accepting upto that average score



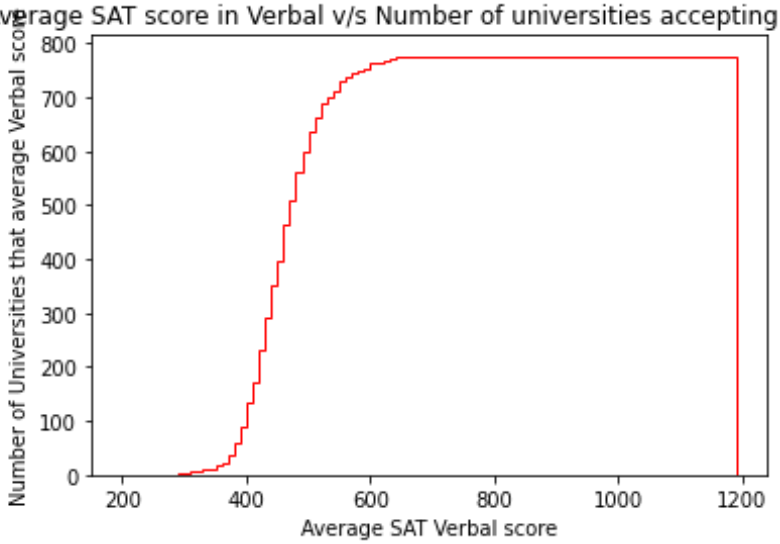
Smooth Curve: Average SAT score in Math v/s Probability of a university being at that score



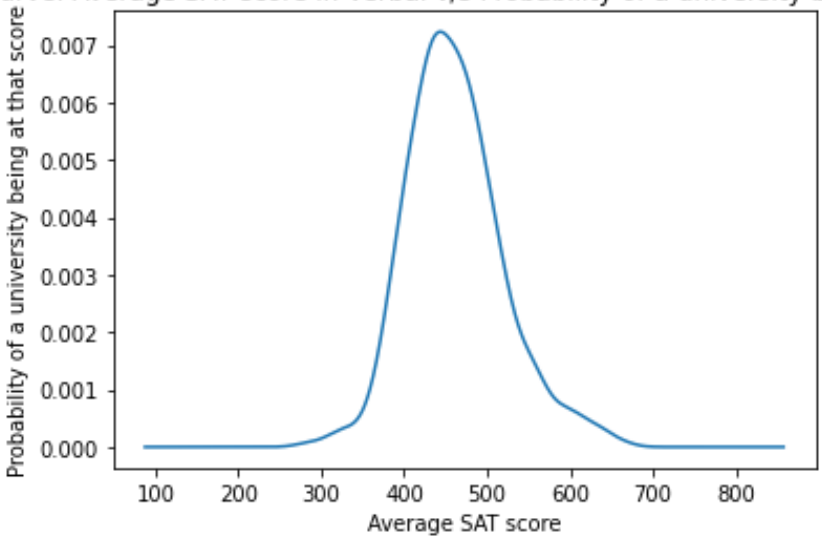
Histogram: Average SAT score in Verbal v/s Number of universities being at that score



Histogram: Average SAT score in Verbal v/s Number of universities accepting that average score



Smooth Curve: Average SAT score in Verbal v/s Probability of a university being at that score




```
Mean SAT Math score of all universities: 506.8595360824742
```

```
Variance in the average SAT Math scores of all universities:
4599.517641021894
Standard deviation in the average SAT Math scores of all universities:
67.8197437404617
```

```
Mean SAT Verbal score of all universities: 461.19716494845363
Variance in the average SAT Verbal scores of all universities:
3398.1479816532046
Standard deviation in the average SAT Verbal scores of all universities:
58.29363585892721
```

From this result, we can observe that the average math score of all the schools was 506, and the standard deviation in SAT math scores was 67. We can also see from the cumulative histogram that below the score of 300, no schools have average scores below a score of 300 in maths, and there are almost no schools with a score average of more than 800. Similarly, we can observe in the case of verbal abilities that the mean score of verbal ability of all schools was at 461 which is less than the math score and the average score of verbal ability of the schools was always greater than 300 marks and was always less than about 700 to 600. From the probability curves, we can also see that the distribution is similar in appearance to the normal Gaussian distribution having a peak at the mean and the variance corresponds to the spread of the graph.

Q3. What is the mean, variance and standard deviation in the average SAT score accepted by the schools? What would the visualisation of the average SAT score and the number of schools at that score look like? What would the distribution of the probability of getting a school at a particular score look like?

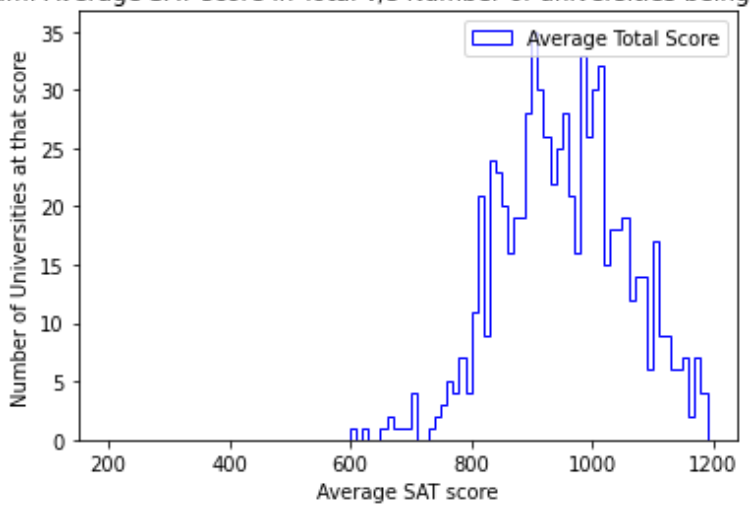
Q4. What is the mean, variance and standard deviation in the average ACT score accepted by the schools? What would the visualisation of the average ACT score and the number of schools at that score look like? What would the distribution of the probability of getting a school at a particular score look like?

A3+A4. Approach: the approach of this question is also similar to the approach of the previous question, the only difference being that we are considering different columns of total SAT score and total ACT score required to join the school.

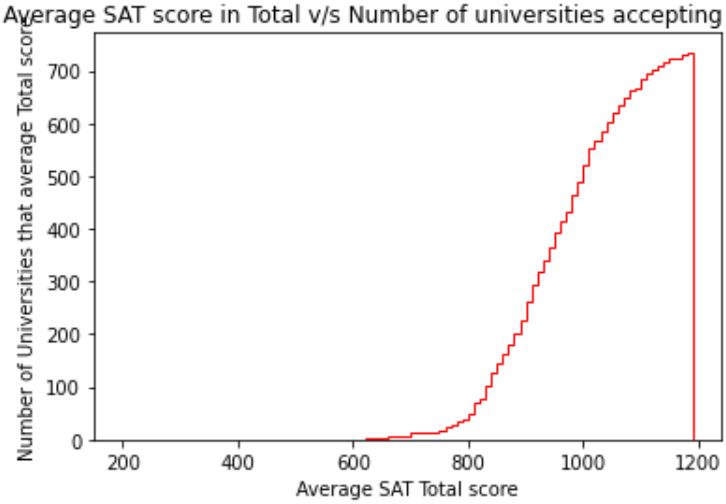
Results and Observations:

```
Mean SAT Total score of all universities: 967.9730077120822
Variance in the average SAT Total scores of all universities:
15271.375878100198
Standard deviation in the average SAT Total scores of all universities:
123.57740844547679
Mean ACT Total score of all universities: 22.1234221598878
Variance in the average ACT Total scores of all universities:
6.649563604389725
Standard deviation in the average ACT Total scores of all universities:
2.5786747767777394
```

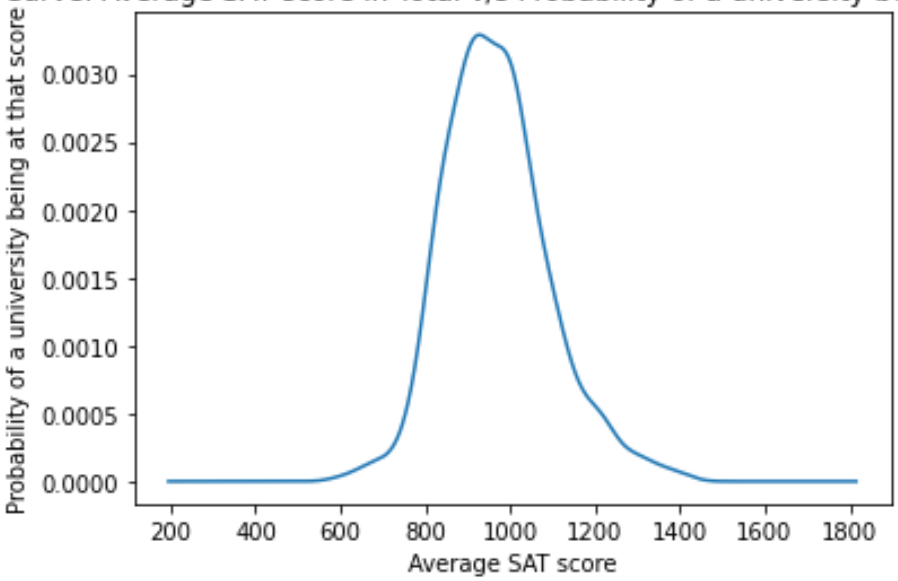
Histogram: Average SAT score in Total v/s Number of universities being at that score



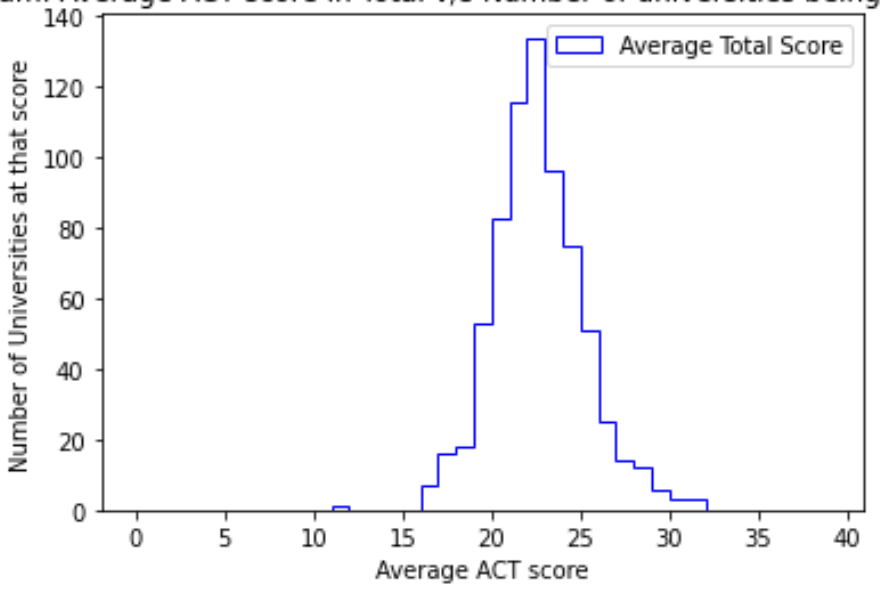
Histogram: Average SAT score in Total v/s Number of universities accepting that average score

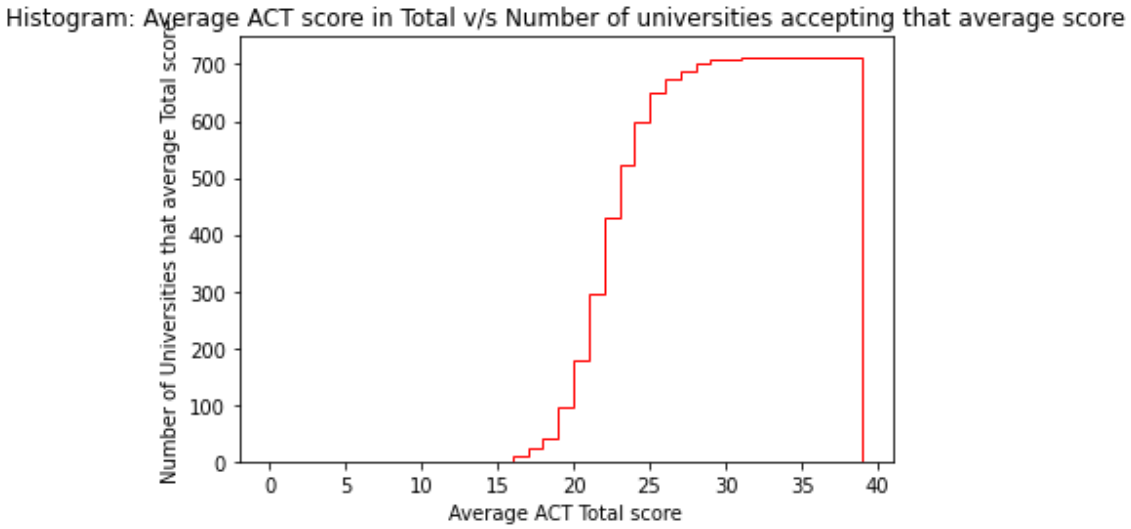


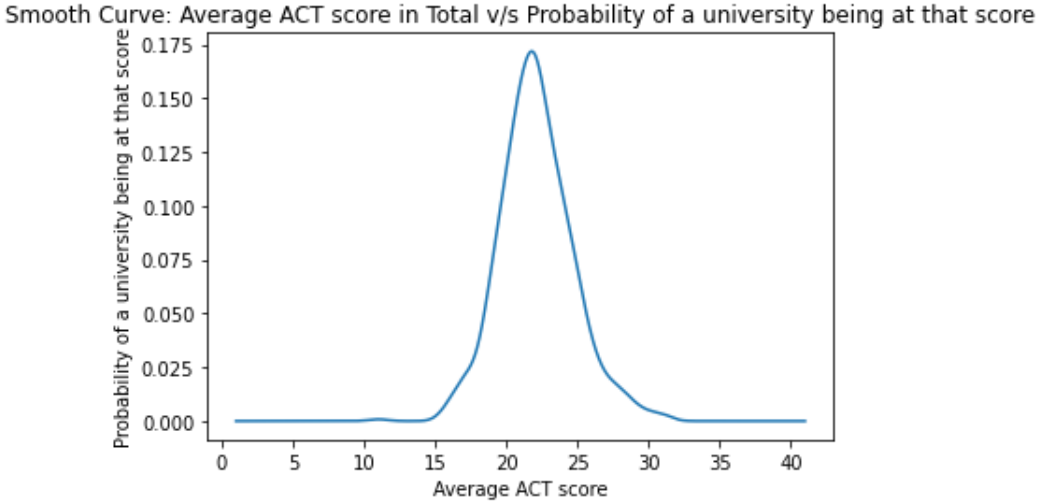
Smooth Curve: Average SAT score in Total v/s Probability of a university being at that score



Histogram: Average ACT score in Total v/s Number of universities being at that score







We can observe that the average score of SAT of all the schools is almost 968, and the average ACT score of all the schools is 22. So, we can see from the cumulative histogram that the minimum average SAT score of the schools is greater than 600, and the highest average SAT score of any school is almost 1200. We can also see from the distribution of scores that the minimum average ACT score of the schools is just greater than 10, and the highest average is between 30 to 35. So, from this data, we can say that if anyone has a score of greater than 600, then it must be enough for that individual to cross the average score of the school with the lowest cut-off score, and if anyone has ACT score of greater than 15, then it must be enough for that individual to cross the average score of the school with the lowest cut-off score. We can also see from the probability distributions that, the peak almost corresponds to the average scores, and the width of the graph corresponds to the variance; that is, greater the variance, more will be the expanse.

Q5. What is the mean, variance and standard deviation of the acceptance rate in various schools? How can we visualise the trend of acceptance rate of various schools and how would the probability distribution curve of getting a school having a particular rate of acceptance look like? What are the five schools having the least rate of acceptance? Most difficult schools to get admission into.

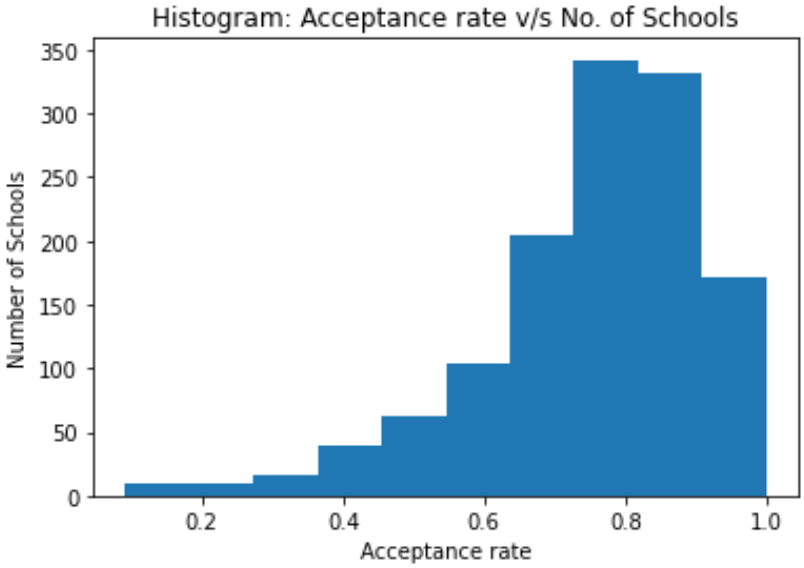
A5. Approach: First of all, the data file was opened as a data frame using pandas .read_csv function. Then the rows were iterated under the columns of the number of applications and the number of accepted applications to search for wrongly formatted rows, and if a row was found to be wrongly formatted, it was removed using df.drop function. Then an empty array was created and the ratio of the number of accepted applications divided by total number of application was appended in that to give us an array which contains the acceptance rate of the students. Then the array was again formed into a data frame with the indices of the names of the schools. Then the first five entries of the sorted data frame were presented by using .head() function. Using the numpy library, the mean, standard deviation and variance of the acceptance rate were

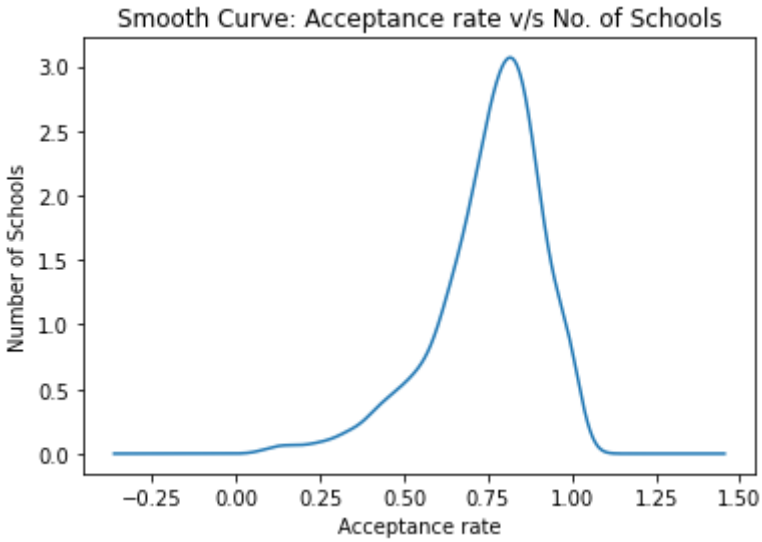
also found out. Using the matplotlib library, the acceptance rate histogram was plotted against the number of schools. Then the numpy array was converted to a pandas series and using kde of the Pandas, it was plotted into a smooth probability distribution curve which provides the probability of getting a school with a particular rate of acceptance.

Results and Observations:

```
Name of the School                        Rate
United States Coast Guard Academy    0.091390
College of the Ozarks                0.103745
United States Military Academy       0.117493
United States Naval Academy          0.126974
Cooper Union                         0.131253
Mean Acceptance rate: 0.7547856665370596
Variance Acceptance rate: 0.02545799011873675
Standard deviation Acceptance rate: 0.15955560196601293
```

Top 5 schools with least acceptance rate





We can observe from results that the mean acceptance rate is around 0.75; that is, out of 4 candidates on an average, the schools accept 3 candidates. The variance in the acceptance rate is small at 0.025 and the standard deviation is almost 0.15. This can be confirmed from the probability distribution curve as well as the histogram, that is, the peak of the probability distribution curve is at around 0.75 (mean) and the variance corresponds to the span of the curve. We can also see from the graph that no school has acceptance rate beyond 1 and no school has acceptance rate less than zero, so this also

satisfies to the conditions of a probability distribution random variable and therefore, we can treat the acceptance rate as the probability of a student to be admitted in a school. We can also see from the analysis that the least rate of acceptance is of 'United States Coast Guard Academy' with 0.09; that is, out of 100 applications, the academy accepts only 9 applications.

Q6. What is the mean, variance and standard deviation of the number of undergraduate students in the schools? How can we visualise the trend of the number of undergraduate students in the schools? What is the probability of getting a school with a particular number of undergraduate students, and what would its graph look like? What are the five schools having the most and the least number of students at the undergraduate level?

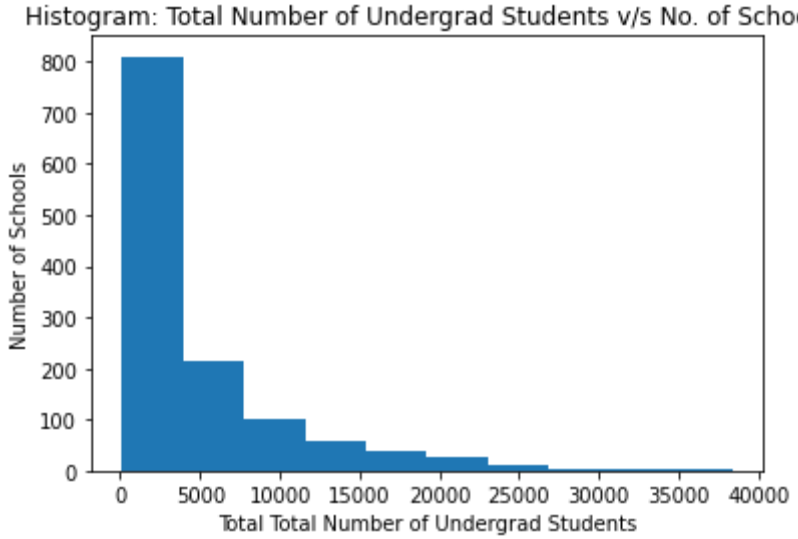
A6. Approach: First of all, the data file was opened as a data frame using the Pandas .read_csv function. Then, the rows of the data frame were iterated through the columns of the number of full-time undergraduate students and the number of part-time undergraduate students. Then, an array was constructed in which the sum of these two columns was appended and then, a new data frame was formed which was indexed with the names of the schools which was later sorted. Then, using the .head() and .tail() functions of pandas, the top 5 and the last 5 schools in terms of the number of undergraduate students was found out. The numpy was also used to calculate the mean, variance and standard deviation. Then, the array was also plotted as a histogram using the matplotlib library and using the kernal density estimation, it was plotted as a probability distribution of getting a school with a particular number of students.

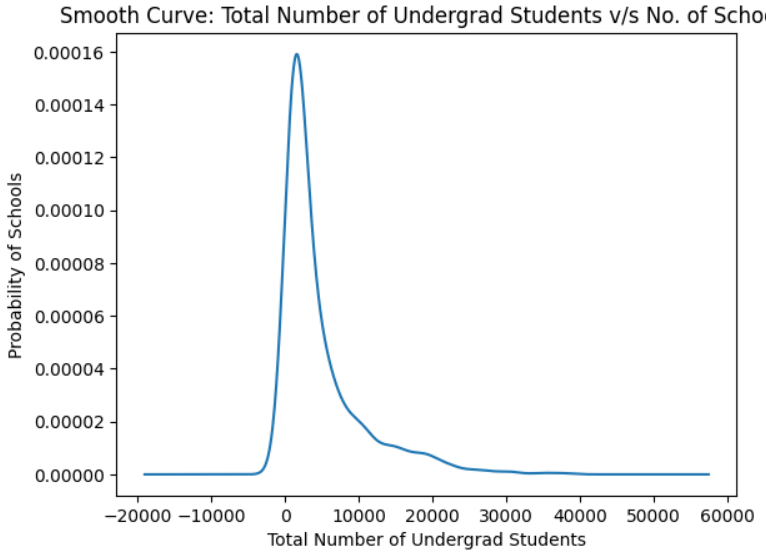
Results and Observations:

```
University of Judaism          100
Christendom College            142
Art Academy of Cincinnati      230
King's College                 264
Marlboro College               271

Pennsylvania State Univ. Main Campus  30963
Texas A&M Univ. at College Station    34441
University of Texas at Austin         35206
Ohio State University at Columbus     37044
University of Minnesota Twin Cities   38338
Mean of Total Number of Undergrad Students: 4807.537431048069
Variance of Total Number of Undergrad Students: 32418853.180041
Standard deviation of Total Number of Undergrad Students: 5693.755630
```

Top 5 schools with least number of undergraduate students

Top 5 schools with most number of undergraduate students





As we can see from the results, the school with the lowest number of students is the University of Judaism, with just a hundred students, and the school with the largest number of students is the University of Minnesota twin cities, with more than 38000 students. The mean of the total number of undergraduate students is at 4800. As we can see from the histogram, majority of the schools have less than 5000 undergraduate students and the number of schools decreases as the number of undergraduate students increases. We can also see from the probability distribution curve that the peak of the probability distribution curve corresponds to the mean number of students, the standard deviation in the number of students is at 5600 which is much greater than the mean of the number of undergraduate students. So, from here, we can also see that the standard deviation is greater than the mean.

Q7. What is the mean, variance and standard deviation of the total expenditure of the undergraduate students at the schools? How can we visualise the trend of the number of schools having a total expenditure within a certain range? What are the top 5 schools having the most expenditure and the five schools having the least expenditure with respect to the students? What would a probability distribution of getting a school with a particular expenditure for undergraduate students look like?

A7. Approach: The data file was first opened as a data frame using the Pandas library and the various kinds of expenditures columns were iterated through rows in order to find the wrongly formatted rows and delete them using df.drop() function of Pandas. Then, a new array was created in which the sum of all the expenditures was appended and a data frame was created in which the data was taken from the array with the index of the names of the schools which was then sorted in the ascending order by the expenditure column and then using the .head() and .tail() functions of Pandas data frame the top 5 and the last 5 schools in terms of total expenditure of students were printed. Using numpy library, the mean, variance and standard deviation of the array was also found out. Using matplotlib library, the array was plotted into a histogram in order to provide the number of schools within a particular range of student expenditure and then using the Pandas kernel density estimation function (kde), the graph of the probability distribution of getting a school with a particular student expenditure was plotted.

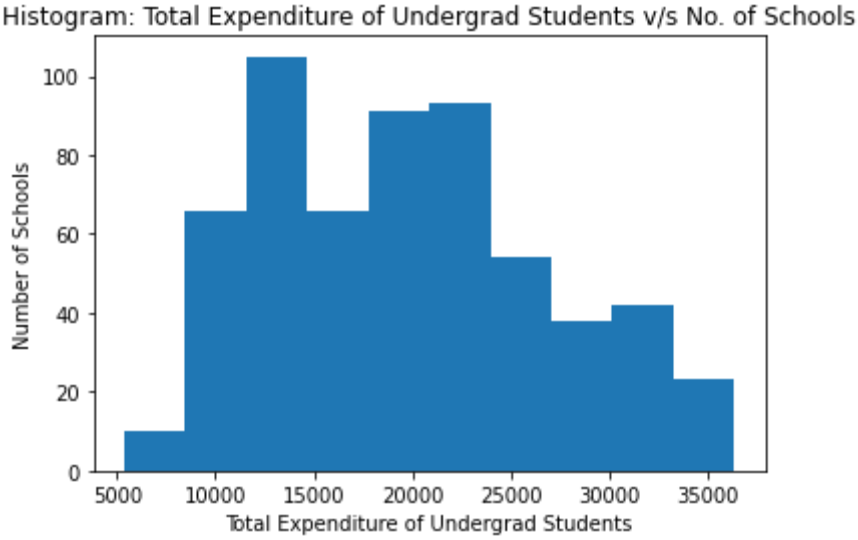
Results and Observations:

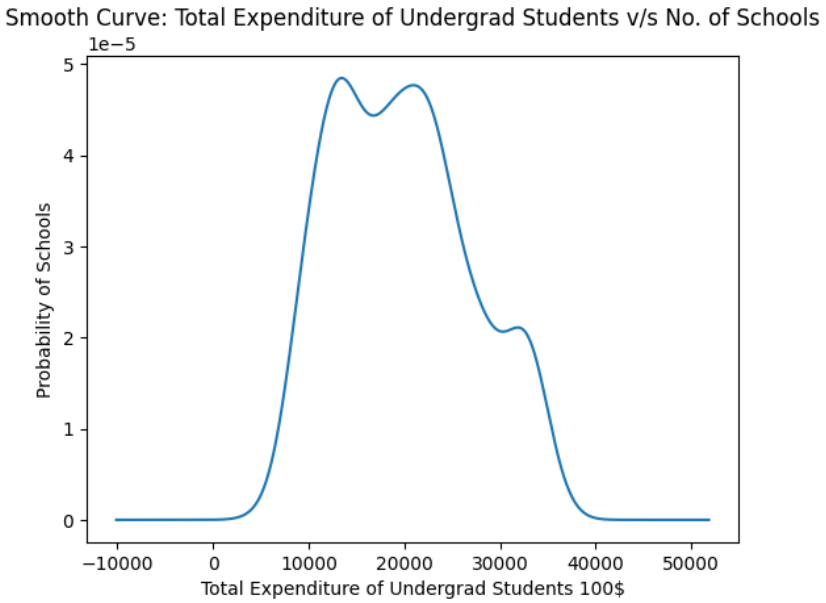
```
Grambling State University                5368
Oklahoma Panhandle State University       6919
Fayetteville State University             7592
University of Sci. and Arts of Oklahoma   7820
Western Carolina University               7857

Massachusetts Institute of Technology  34975
Harvard University                     35060
Georgetown University                  35088
University of Pennsylvania             35440
Barnard College                        36344
Mean of Total Expenditure of Undergrad Students: 19779.700680272108
```

Top 5 schools with the least expenditure to students

Top 5 schools with the greatest expenditure to students

```
Variance of Total Expenditure of Undergrad Students: 50789824.40020362
Standard deviation of Total Expenditure of Undergrad Students
7126.698001192672
```





As we can see from the results that the average expenditure of the students in a few schools such as University of Pennsylvania, Bernard college, etc. is up to 7 times greater than the expenditure in school such as Grambling State University, Okhlahoma State University, etc. We can also see from the results that the mean expenditure is around 20000$.

Q8. What is the mean, variance and standard deviation of the total graduation rate of the schools? What would a trend of the number of schools having a graduation rate within a particular range look like? What would the probability distribution of getting a school having a particular rate of graduation look like? What are the five schools having the largest rate of graduation and the five schools having the least rate of graduation?

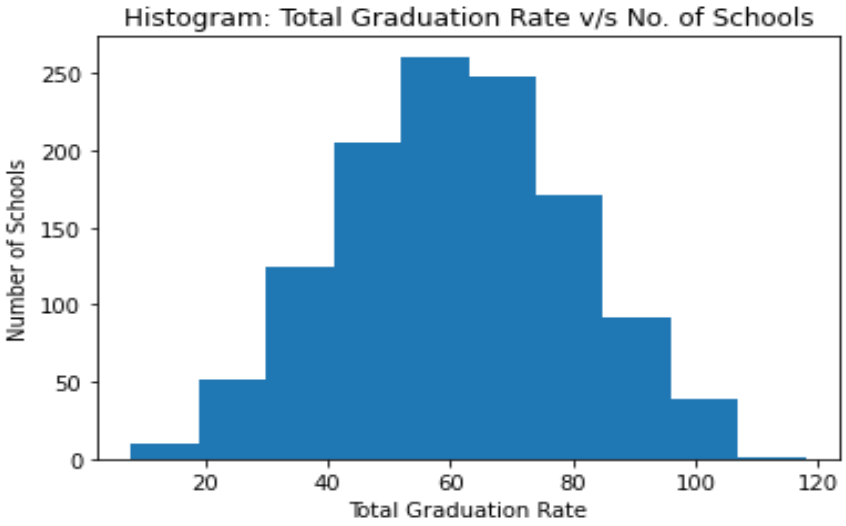
A8. Approach: First of all, data file was converted into a data frame using the Pandas library. Then, the rows were iterated under the column of graduation rate in order to eliminate wrongly formatted rows and the rows with wrong format were removed by using df.drop function. Then an empty array was created in which the graduation rate was appended and a new data frame was constructed out of this array with the indexing of the names of the schools. Using the .head() and .tail() functions of Pandas, the top five and the last five items of the dataframe were printed. Using numpy library, the mean, variance and standard deviation of the graduation rate was also found out and the matplotlib library was used to plot the histogram of graduation rate and a smooth probability distribution curve also plotted in order to give the probability of getting a college having a particular graduation rate using kde of pandas.

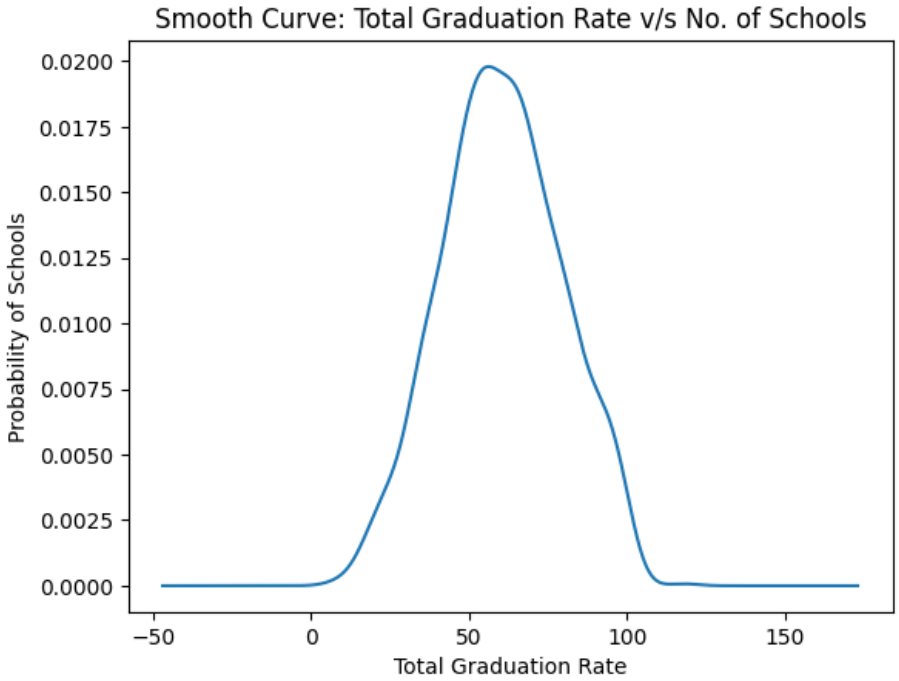
Results and Observations:

```
Rate
University of Houston - Downtown      8
Texas Southern University            10
Alabama State University             15
Montreat-Anderson College            15
Miles College                        15
                        rate
Harvard University       100
Harvey Mudd College      100
Goddard College          100
University of Richmond   100
Cazenovia College        118
Mean of Total Graduation Rate: 60.44305901911887
Variance of Total Graduation Rate: 355.08133794082266
Standard deviation of Total Graduation Rate 18.843602042625044
```

Top 5 schools with the least graduation rate

Top 5 schools with the most graduation rate





As we can observe from the results that some schools such as University of Houston, Texas Southern University, etc. have graduation rates as low as 8 to 10%; whereas a few schools such as Harvard College, University of Richmond, etc. have 100% graduation rate and colleges such as

Cazenovia College have greater than 100% graduation rate. We can see the mean graduation rate of 60% and the standard deviation in the graduation rate of 18% which can be visualised from the probability distribution curve. The histogram also has a peak at around 60% and from the probability distribution curve we can see that there is also probability of getting a college with greater than 100% graduation rate. The distribution of probability of getting college at a particular graduation rate is similar to a normal distribution in terms of appearance.

Q9. What is the mean, variance and standard deviation of the student-by-faculty ratio at various schools in America? What would the trend of the number of schools having the student-by-faculty ratio within a particular range look like? What would a probability distribution of getting a school having a particular student-by-faculty ratio look like? What are the top five schools having the largest amount of student-by-faculty ratio and the top five schools having the least amount of student-by-faculty ratio?

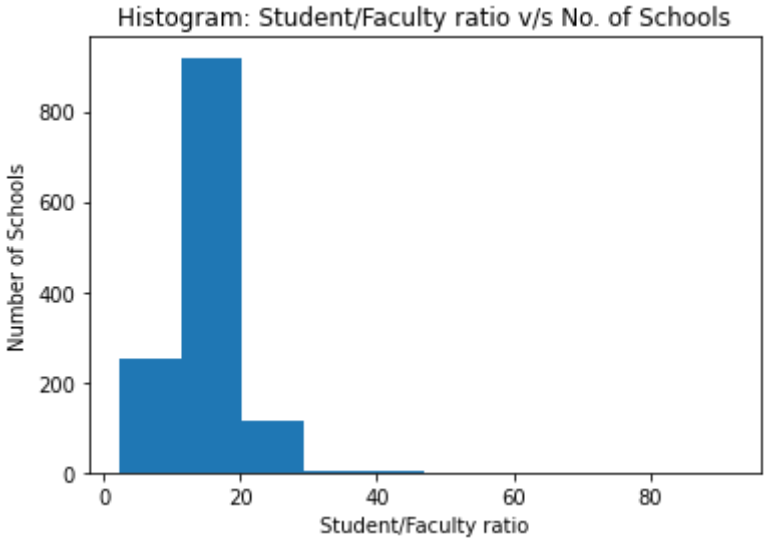
A9. Approach: First of all, the data file was opened as a data frame using Pandas, and then the rows were iterated under the column of student-by-faculty ratio in order to search for wrongly formatted rows. If a row was found to be wrongly formatted, it was dropped using Pandas df.drop function. A new array was created, in which this column was added and it was later formed into a data frame with indexing of the names of the schools. Using the Pandas .head() and .tail() functions, the top 5 and the last 5 schools in terms of student-by-faculty ratio were printed. The numpy library was used to find the mean, variance and standard deviation of the student-by-faculty ratio. A histogram was also plotted in order to visualise the trend in the student-by-faculty ratio against the number of schools within that range and the kernel density estimation was used in order to plot the smooth curve of probability distribution showing the probability of getting a school with a particular student by faculty ratio.

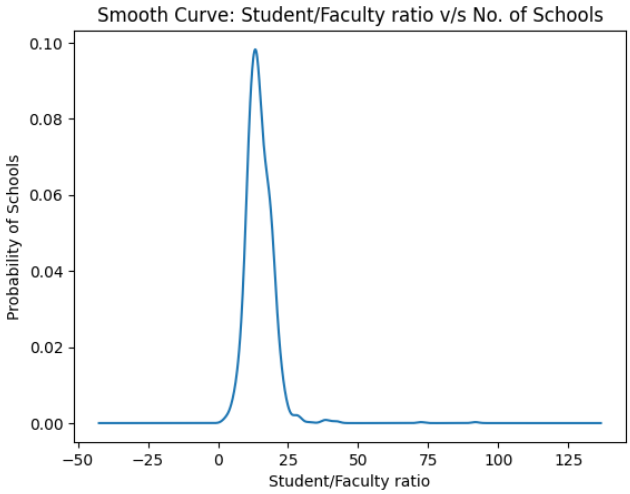
Results and Observations:

```
Prescott College                    2.3
University of Charleston            2.5
Case Western Reserve University     2.9
Johns Hopkins University            3.3
Washington University               3.9

                                    rate
Indiana Wesleyan University         39.8
Langston University                 41.7
University of Arkansas at Monticello  42.6
St. Leo College                     72.4
Northwood University                91.8
Mean of Total Student/Faculty ratio: 14.861046959199385
Variance of Student/Faculty ratio: 26.891985354992674
Standard deviation of Student/Faculty ratio 5.1857482926760605
```

Top 5 schools with the least student-by-faculty ratio

Top 5 schools with the most student-by-faculty ratio





We can analyse from the results that a few schools, such as Prescott College, University of Charleston, etc., have the student-by-faculty ratio as low as 2.3, whereas school such as Northwood University have the student by faculty ratio as large as 91. The mean of the student by faculty ratio was of 14 and the standard deviation was of almost 5. This is shown by the histogram and the probability distribution curve also. From the histogram, we can see that the majority of the school have the student-by-faculty ratio within the range of 10 to 20. Some schools have that within the range of 2 to 10 and a few of them have a ratio greater than 20. We can also see from the probability distribution curve that the peak corresponds to the mean value of the probability distribution, and because of having a low value of variance, the spread of the curve is also less in this case.

ADDITIONAL QUESTIONS

(AAUP dataset)

Q8. What is the relative ratio of expenditure of the schools towards salaries of Faculty members at various levels? How can we visualise this ratio?

Q9. What is the relative ratio of the average spending of various types of schools towards salaries? How can we visualise that relative ratio?

Q10. What is the relative ratio of the average salaries of faculties at various types of schools? How can we visualise that relative ratio?

(USNEWS dataset)

Hypothesis: Do schools with high average scores in SAT provide proportionately larger resources to the students?

Q10. How can we visualise the relation between the parameters of Instructional expenditure per student against the Average Combined SAT score? What can we deduce from this?

Q11. How can we visualise the relation between the parameters of the Student/faculty ratio against the Average Combined SAT score? What can we deduce from this?

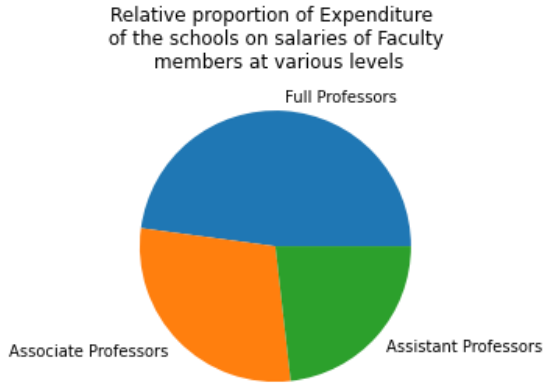
Q12. How can we visualise the relation between the parameters of acceptance rate against the Average Combined SAT score? What can we deduce from this?

ANSWERS TO ADDITIONAL QUESTIONS

AAUP

Q8. What is the relative ratio of expenditure of the schools towards salaries of Faculty members at various levels? How can we visualize this ratio?

8A. Approach: The average values of the salaries of the faculties at various levels were multiplied by the average number of faculty members to get the average expenditure spent by the school towards salaries of various faculty levels, which was plotted as a pie chart and the values were also divided by each other to get the relative ratio of the salaries.
Results and Observations:




```
Spendings towards the salaries of Full Professors : Associate Professors : Assistant Professors :: 2.0639926384005642:1.2479375966850133:1
```

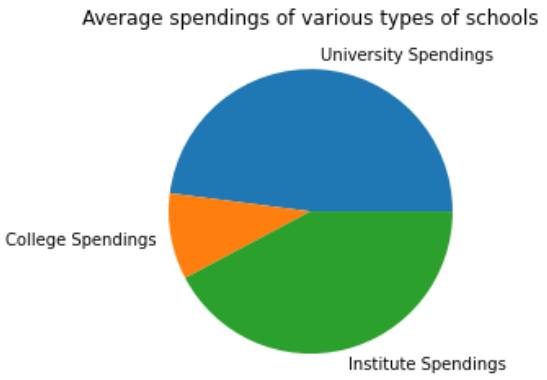
The observations that we can draw from this graph and data is that, the schools spend twice as much money towards the salaries of full Professors compared to that of the Associate and Assistant Professors. Moreover, the schools spend more money towards the salaries of associate professors than that towards assistant professors in total.

Q9. What is the relative ratio of the average spending of various types of schools towards salaries? How can we visualise that relative ratio?

9A. Approach: After opening the dataset as a dataframe using Pandas, it was searched for the wrongly formatted rows, and the wrongly formatted rows were eliminated from the dataframe using the df.drop() function of Pandas. Then some empty arrays were constructed which would contain the total expenditures of Universities, Colleges and Institutes. The expenditure values were added by iterating through the rows and if the name of the school was matching with the initial letters of the word "University", then the total expenditure as calculated in the previous questions was added to the respective array. Using matplotlib, a pie chart was plotted for better visualisation of the relative ratio of the expenditure of Universities v/s Colleges v/s Institutes and the ratios were also calculated and printed.

Results and Observations:

```
University Spendings : College Spendings : Institute Spendings :: 4.889 : 1 : 4.307
```



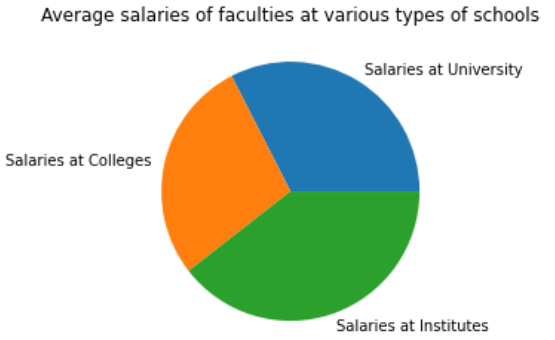
We can observe that the universities and institutes spend 4 times the money spent by colleges in general towards salaries. Universities also spend more money compared to institutes.

Q10. What is the relative ratio of the average salaries of faculties at various types of schools? How can we visualise that relative ratio?

10A. Approach: The approach of the problem was similar to the previous one except for the fact that rather than appending the total expenditures, the average salaries of all professors was added to the respective lists.

Results and Observations:

```
Salaries at University : Salaries at Colleges : Salaries at Institutes :: 1.163 : 1 : 1.411
```



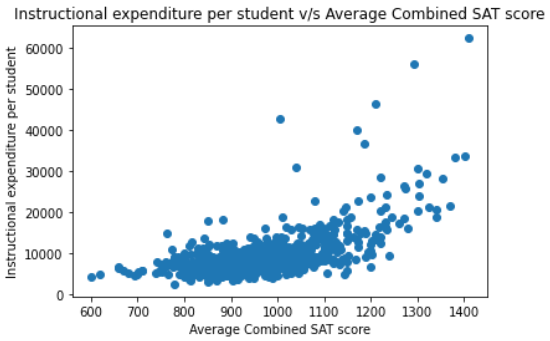
We can observe from this data that the professors at institutes receive the highest salaries on an average, followed by the professors at universities and professors at colleges.

USNEWS

Q10. How can we visualise the relation between the parameters of Instructional expenditure per student against the Average Combined SAT? What can we deduce from this?

A10. Approach: First of all, the dataset was converted to a dataframe using Pandas and then the columns of "INSTRUCTIONAL EXPENDITURE PER STUDENT" AND "THE AVERAGE COMBINED SAT SCORES" were cleaned from wrongly formatted data using the df.drop() function of Pandas. Then the two columns were plotted as a scatter plot for visualization using matplotlib.

Results and Observations:

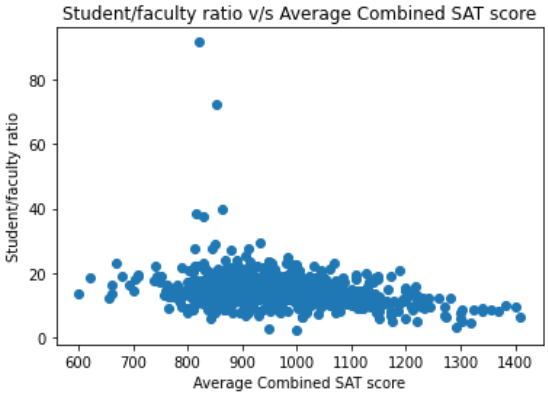


In general, we can see that the Expenditure increases with the increase in SAT score with some exceptions.

Q11. How can we visualise the relation between the parameters of Student/faculty ratio against the Average Combined SAT? What can we deduce from this?

A11. Approach: First of all, the dataset was converted to a dataframe using Pandas and then the columns of "STUDENT BY FACULTY RATIO" AND "THE AVERAGE COMBINED SAT SCORES" were cleaned from wrongly formatted data using the df.drop() function of Pandas. Then the two columns were plotted as a scatter plot for visualization using matplotlib.

Results and Observations:

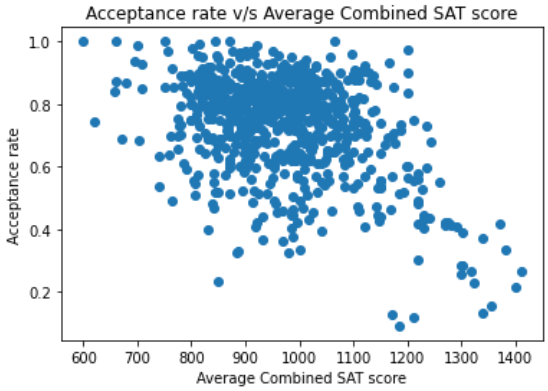


In general, we can see that the student by faculty ratio remains almost constant with the increase in the value of SAT score in general which indicates that there is not much relation between them.

Q12. How can we visualise the relation between the parameters of acceptance rate against the Average Combined SAT? What can we deduce from this?

A12. Approach: First of all, the dataset was converted to a dataframe using Pandas and then the columns of "ACCEPTANCE RATE " AND "THE AVERAGE COMBINED SAT SCORES" were cleaned from wrongly formatted data using the df.drop() function of Pandas. Then the two columns were plotted as a scatter plot for visualization using matplotlib.

Results and Observations:



In general, we can see that the acceptance rate remains almost randomly distributed with the increase in the value of SAT score, which indicates that there is not much relation between them. But we can also see that at lower SAT scores, the acceptance rate is almost 1 and at very high SAT scores, the acceptance rate is very less. This indicates the relative ranking of colleges on the basis of the acceptance rate. The colleges with higher SAT scores are bound to be more selective.

Conclusion of the hypothesis: We can say that the high SAT score of the schools does not always correspond to higher educational facilities for the students.

UNIQUENESS OF APPROACHES USED

The approaches used are unique in the way that the libraries of Pandas, NumPy and matplotlib, along with the prior knowledge of Python, were used in a complementary manner along with the concepts of this course to analyze the various parameters effectively. Various concepts of the course, such as basics of probability, probability distribution, cumulative distribution functions, smoothening the curve using kernel density estimation, mean, variance, standard deviation, correlations, joint random variables, etc., were also used by invoking various functions of the modules. On the whole, this analysis gives the various unseen perspectives of the dataset. The fact that I believe was the most interesting from the AAUP dataset was the expenditure of various types of schools and how colleges lag behind Institutes and universities in terms of the expenditure towards salaries and compensations and which type of schools offer more salary. From the USNEWS dataset, the parameters such as selection rate, graduation rate etc were very interesting to understand and analyse. Rather than going for the function of correlations from pandas, I indirectly tried to visualize that from the scatter plots.

SUMMARY

The 22 questions that were created are more than sufficient to analyse the data from all the perspectives and hence the aim to analyse the data with detailed approach, results and observations has been achieved by incorporating various concepts covered in the class as well as the labs. In the beginning, we have analysed the number of schools in various American states and visualized that with the help of a bar graph as well as a pie chart and found that the state with the least number of schools was Wyoming. It was seen that the schools were not distributed uniformly across all the states and there were some states having 20 times as many schools compared to some other states. We then analysed the relative number of various types of schools that are divided under the categories such as Universities, Colleges, and Institutes, as well as the categories denoted by roman numbers. We also found out the total number of schools under each of those categories. The visualization was provided through a pie chart. We also calculated the mean, variance and standard deviation of the average salaries of full professors, associate professors and assistant professors. We then visualised the trends with a histogram and plotted a smooth curve of the probability distribution of the salary of the professors against the probability of finding the faculty with that salary. We found the curves similar to those of normal distribution with slight deviations. A similar thing was repeated for the parameters such as compensation amount and the number of faculty members. We then also calculated the mean, variance and standard deviation of the expenditure of the schools towards salary and compensation by taking into account various parameters into account simultaneously. This data from all the parameters simultaneously taken was plotted as a histogram to facilitate visualisation. Then the probability distribution curve of the schools having an expenditure of a particular value was also plotted. Then the relative proportions of the spendings of the schools was also analysed in terms of spendings towards the salaries of faculties at various levels was also analysed and visualized with the help of a pie chart. We also analysed and visualized in a similar way the questions such as the average spending of various types of schools (Universities, Colleges, Institutes) and which type of school pays more salaries on an average to the faculty members at various levels.

For the USNEWS dataset, first of all, the schools were analysed on the category of ownership (private and public) and the total number of private and public schools as well as their ratios were also calculated along with a pie chart for the visualization. We then calculated and analysed the mean, variance and standard deviation in the average SAT score in math, verbal ability, total SAT score and ACT score accepted by the schools. We also visualised these parameters with the help of a histogram and also analysed the distribution of probability of getting a school with a score equal to a particular value. We then calculated and analysed the mean, variance and standard deviation of various parameters such as acceptance rates, number of undergraduate students, total expenditure of the undergraduate students at the schools, graduation rate, student by faculty ratio in various schools and then visualised the trend of those parameters for various schools. We also analysed and plotted the probability distribution curve of getting a school having a particular value for those parameters. We also looked for the top five and last five schools with respect to those parameters. We then created a hypothesis that the schools with higher cut-off scores would provide more resources to the students and later proved the hypothesis wrong with the following parameters. We tried to analyse if there was any relation between the average score of a school in SAT and the facilities offered by the schools such as higher instructional expenditure towards the students, better student-to-faculty ratio and also tried to find out if there was any relation between the average SAT score and the acceptance rates. All these things were visualised using scatter plots and negligible relation between these parameters was found.

REFERENCES

[1] “NumPy User Guide.” NumPy user guide - NumPy v1.24 Manual. Accessed March 10, 2023. https://numpy.org/doc/stable/user/index.html

[2] “Matplotlib - Visualization with Python v3.7.0.” Matplotlib. Accessed March 10, 2023. https://matplotlib.org/

[3] “Pandas User Guide.” User Guide - pandas 1.5.3 documentation. Accessed March 10, 2023. https://pandas.pydata.org/docs/user_guide/index.html#user-guide

[4] “Pandas - Fixing Wrong Data.” Pandas - Cleaning Data. Accessed March 10, 2023. https://www.w3schools.com/python/pandas/pandas_cleaning_wrong_data.asp