Data Narrative

Analysing the data of different institutes

Ruchit Jagodara
Student of CSE department
Roll no. 22110102
IIT Gandhinagar
Gandhinagar,Gujarat
ruchit.jagodara@iitgn.ac.in

***Abstract*— This data narrative aims to provide the data analysis done by Ruchit Jagodara on various institution of USA. This data will give you the information about salary of professors and average score of students who are taking admission into particular institutes.**

A. Overview of the Dataset

Here, the dataset 'aaup.data' contains data total number of professors, associate professor, assistant professor, their average salary and their average compensation given by the institution in which they are working. Also, they have provided the information of institute like its postal code (i.e., state) and type (i.e., I, IIA, IIB).

In another dataset named 'usnews.data', data regarding admission of students in that institution is given. In this dataset, students' average ACT score, average SAT score and average score of different subjects of SAT examination is given for a particular institution. This dataset provides us the information regarding money that students are spending behind different things like room, board, additional fees, book costs, personal spending. This dataset also provides us the information regarding graduation rate, education of professors, donation of alumni for institution, student/faculty ratio, moreover it contains the information regarding the number of applications received and number of accepted applications to take admission in that institution. This also provides us the information that whether that institution is private or public.

B. Scientific Questions/Hypotheses

I. Is there any trend that average salary and average compensation of professors are following with respect to total number of professors? Are these trends similar or not? Verify your answer.
II. Is there any institute that is paying more salary to their associate professor or assistant professor than the average salary of professors of other institute if total number of professor and assistant professor or associate professor are almost same? Is there any trend that these graphs are following and if yes, then at what extent?
III. Which type of institute is paying more salary to their staff? Compare the highest average salary paid in each type of institute.
IV. Which state has highest number of institutions? In which state, professors are given more average salary and average compensation? At which extent these two trends are related?
V. Find the probability that if I choose a public institute, it is type I institute. Also, find the probability if private was selected.
VI. Rejection ratio may depend on the value of institute like where only scholar students are going.
VII. Is a greater number of applications arriving where institute is giving some special functionality?
VIII. Institute where graduation rate is low may get a smaller number of applications.
IX. Professors may be getting more salary if they have PhD qualifications.
X. Behind which thing student have to spend more money? Is that same for public institute and private institute?

C. Answers to the Questions

I. IS THERE ANY TREND THAT AVERAGE SALARY AND AVERAGE COMPENSATION OF PROFESSORS ARE FOLLOWING WITH RESPECT TO TOTAL NUMBER OF PROFESSORS? ARE THESE TRENDS SIMILAR OR NOT? VERIFY YOUR ANSWER.

To find the answer of this question, first of all we will delete all the rows which does not have enough data which we require and to do that we will remove all the rows which contain a '*' in particular column with the help of below syntax,

```
data=data[data[4] != '*']
```

After that we will convert all the values to numeric form because till now they were in string format. We can do this thing with below syntax,

```
y=pd.to_numeric(data[4])
```

After that we will plot graph between number of professors vs average salary or average compensation that they are getting to see if that are following some trend or not, by this we will get the graphs which are presented below.

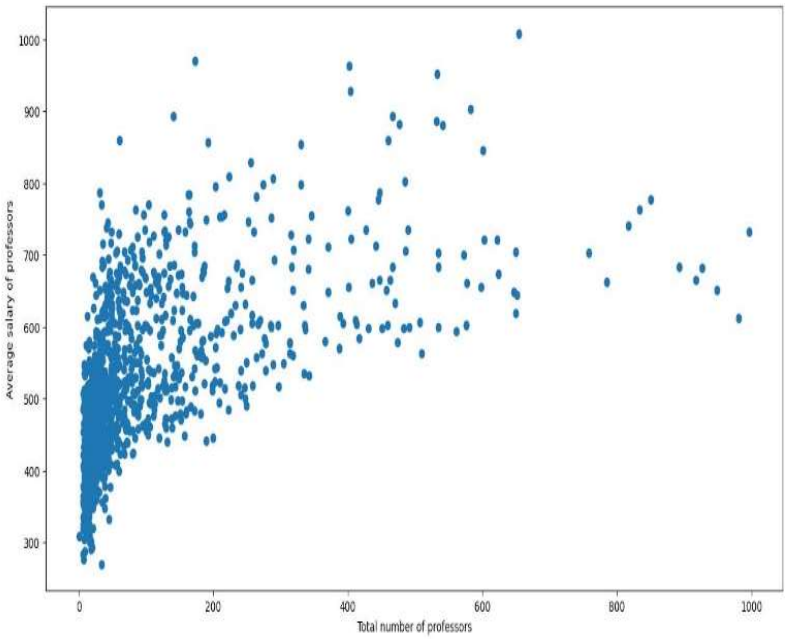

fig. 1 scatter plot between no. of professors and their average salary

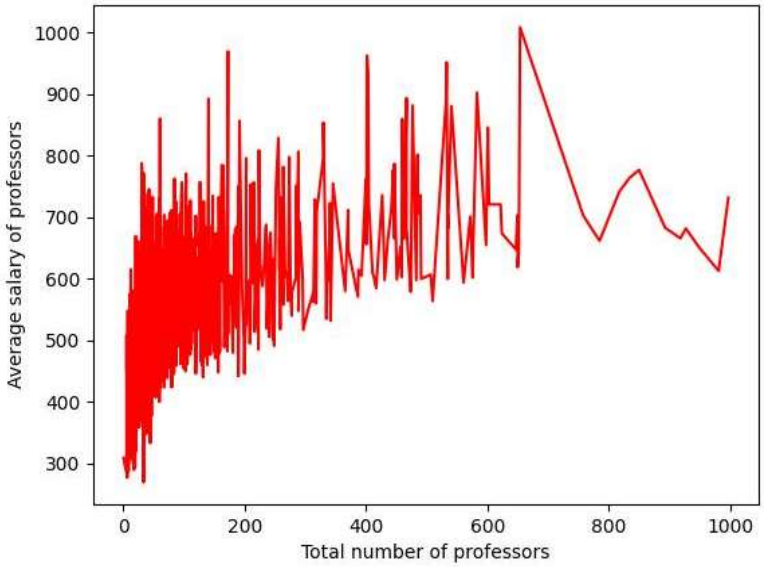

fig. 2 normal plot between no. of professors and their average salary

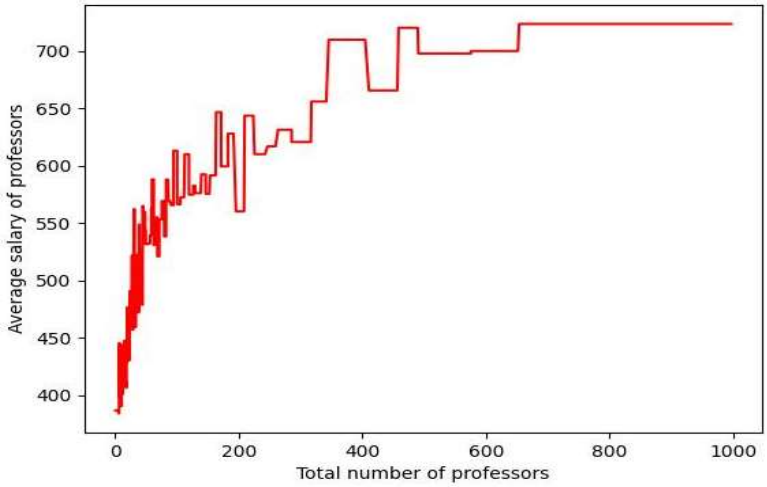

fig. 3 Modified version of fig. 2

Now as this fig. 2 is not readable so we will take average value of nearby points so after doing some operation we will get fig. 3 which is made by taking average of 12 nearby points. This can be done by below syntax,

```
z=[]
i=0

while i<len(x)-1:
    z+=[np.mean(y[i:i+12])]*12
    i+=12
```

Now in fig.3, we can clearly see that the graph is increasing continuously.

Similarly, after doing same operations we will get graph of average compensation of professors.

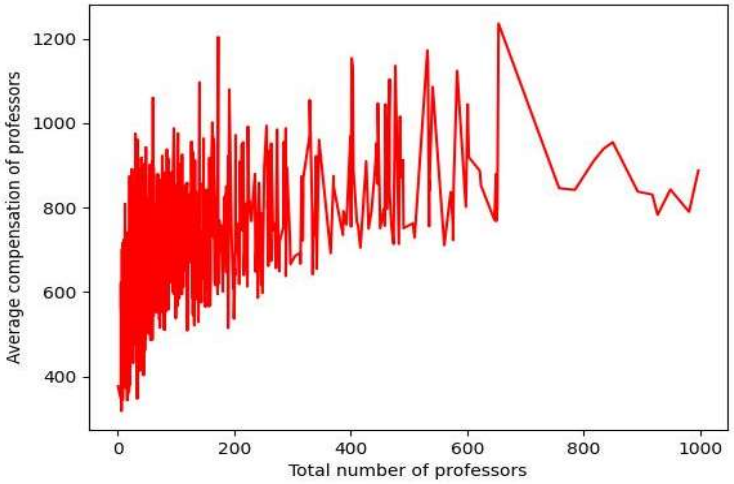

fig. 4 number of professors vs average compensation

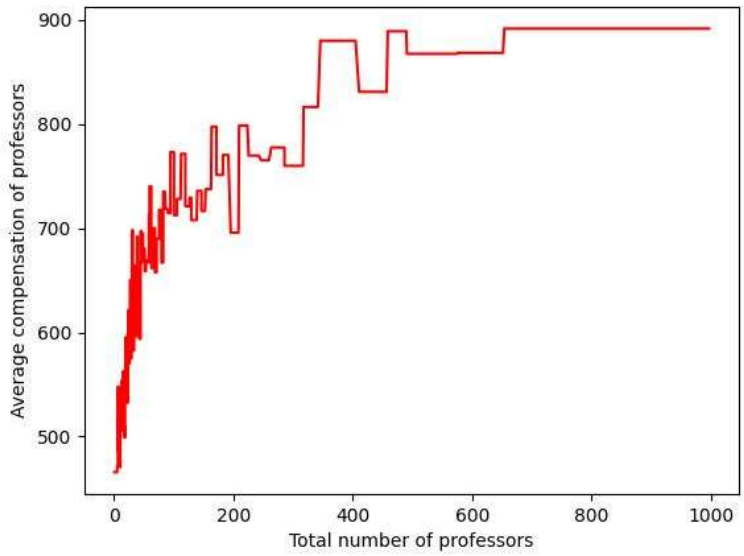

fig. 5 Modified version of fig. 4

Here, as we can see in fig.5, the graph is increasing similarly as of fig. 3 so maybe they are following same trend.

Now, we will calculate correlation factor of these two functions which we got before taking average with the help of numpy.corrcoef() function.

So, by this function we got correlation factor = 0.9971 of these two functions so it indicates that these functions are following the same trend.

II. IS THERE ANY INSTITUTE THAT IS PAYING MORE SALARY TO THEIR ASSOCIATE PROFESSOR OR ASSISTANT PROFESSOR THAN THE AVERAGE SALARY OF PROFESSORS OF OTHER INSTITUTE IF TOTAL NUMBER OF PROFESSOR AND ASSISTANT PROFESSOR OR ASSOCIATE PROFESSOR ARE ALMOST SAME? IS THERE ANY TREND THAT THESE GRAPHS ARE FOLLOWING AND IF YES, THEN AT WHAT EXTENT?

To solve this question, we need to plot the graph of number of associate professors with their average salary and need to compare it with that of professors, similarly we have to do this same thing for assistant professors also.

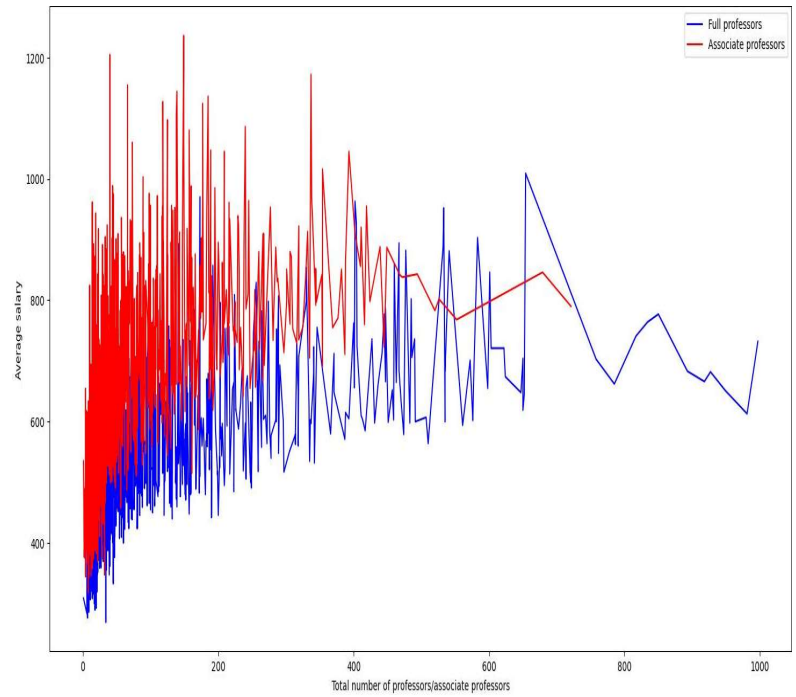

fig. 6 Graph between number of professor/associate professor vs average salary

So, we will get graph like fig. 6 when we are comparing graph between associate professor and professors. Here, we can easily see that if we keep total number of person similar than associate professor have high salary than full professors, we can make this graph very clear by taking average of 12 nearby values and by that we will get graph like fig. 7 which clearly indicates that associate professor has more salary than full professors under some constrains.

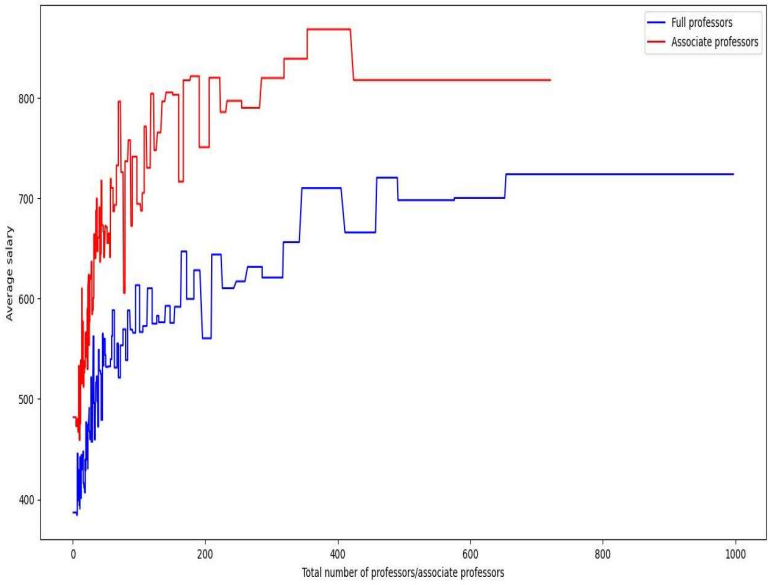

fig. 7 Modified version of fig. 6

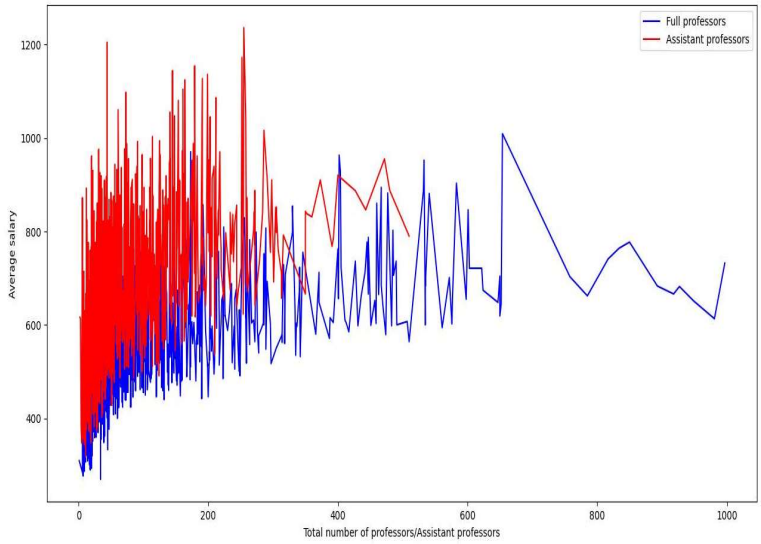

fig. 8 Graph between number of professor/assistant professor vs average salary

Similarly, we get fig. 8 and fig. 9 for assistant professors. We can clearly say that this graphs are almost similar like the fig. 6 and fig. 7 so they maybe follow some common trend.

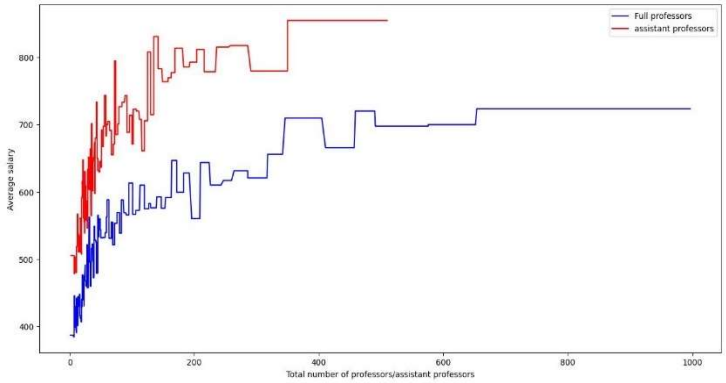

fig. 9 Modified version of fig. 8

Now, we will find correlation factor of that two functions which are related to average salary of associate professor and assistant professor.

Correlation factor is equal to 0.8916, so we can say that these two salary functions are following nearly the same trend.

III. WHICH TYPE OF INSTITUTE IS PAYING MORE MONEY TO THEIR STAFF? COMPARE THE HIGHEST AVERAGE SALARY PAID IN EACH TYPE OF INSTITUTE.

To find the answer of question, we need to make group of institutes by their types so that finally we can calculate the total salary paid and we can plot graph. We can make groups with the help of below syntax.

```
cc=d.groupby([3]).sum().plot(kind='pie',y=17)
```

For this, after making groups we need to multiply average salary with total number of staff to get the value of total salary paid of each of the institutes. After that, we will simply add that values so that we can calculate it with respect to group of that type of institute.

So, by this method we get fig.10.

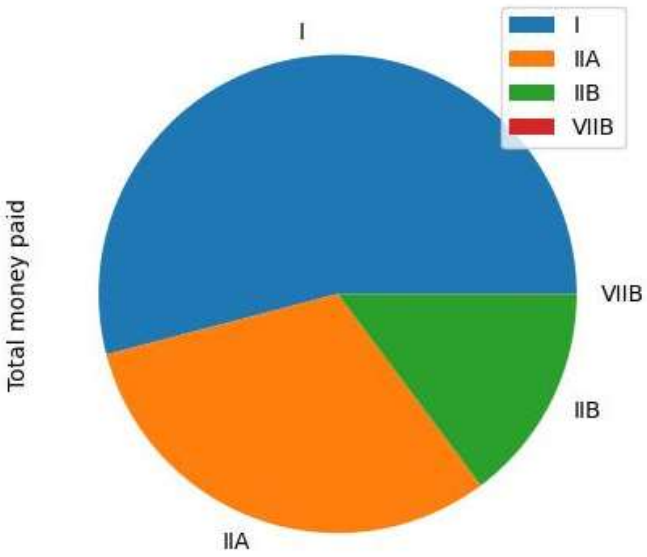

fig. 10 Graph of various types of institutes with respect of total money paid through salary by that group to their staff

In this graph, we can notice that type I which is called most reputed type is spending more money than anyother to their staff through salary.

Now, fig.11, fig.12, fig.13 gives the information regarding the institutes which are giving highest salary to their staff. From fig.11, fig.12 and fig. 13 we can easily say that in paying highest salaey to their staff type I is on the top.

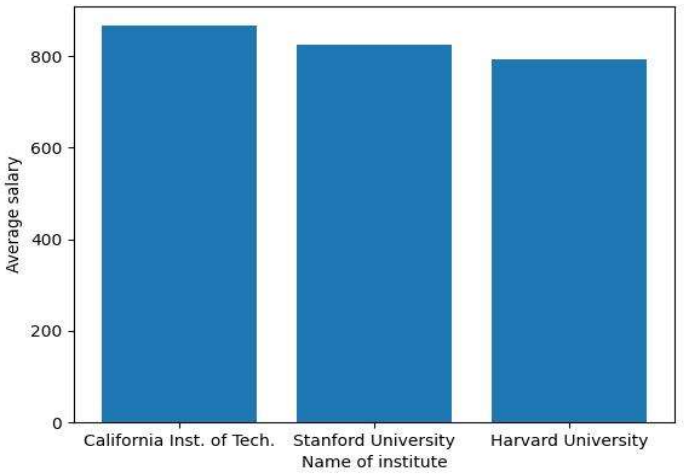

fig. 11 Top 3 Institutes of type I which are paying more salary than anyother institute.

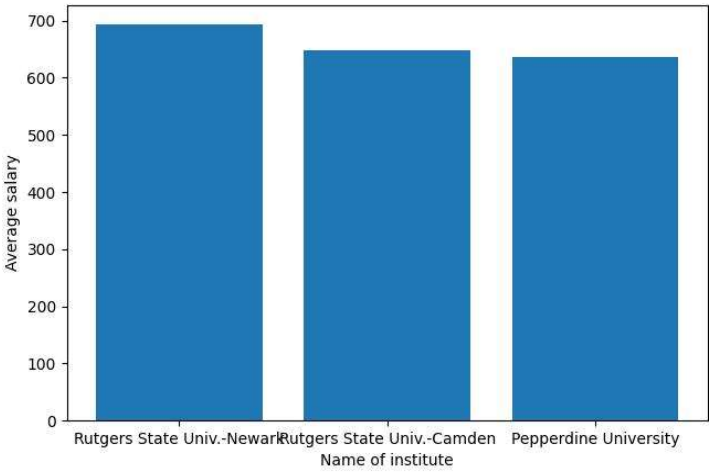

fig. 12 Top 3 Institutes of type IIA which are paying more salary than anyother institute.

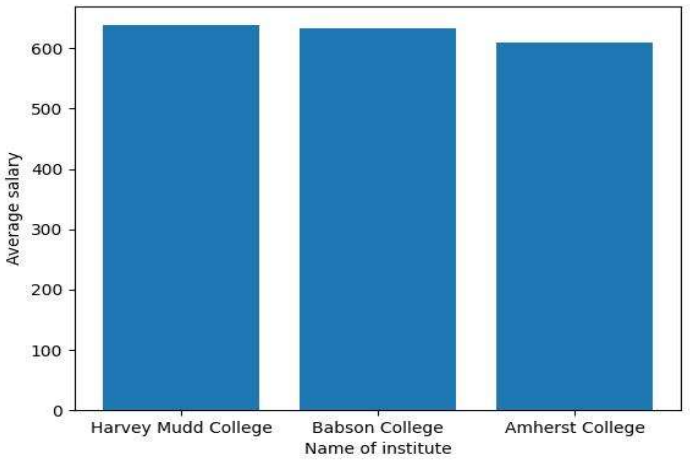

fig. 13 Top 3 Institutes of type IIB which are paying more salary than anyother institute.

IV. WHICH STATE HAS HIGHEST NUMBER OF INSTITUTIONS? IN WHICH STATE, PROFESSORS ARE GIVEN MORE AVERAGE SALARY AND AVERAGE COMPENSATION? AT WHICH EXTENT THESE TWO TRENDS ARE RELATED?

Here, we have given postal codes of each institute which are representing a particular state. To find the total number of institutions per state first of all we will add a column in a original dataframe which contains 1 in all rows. After that, we will make groups with the help of pandas.groupby() function and then we will calculate the total value of that row so that we can get total number of institutes and then we will plot the graph accordingly.

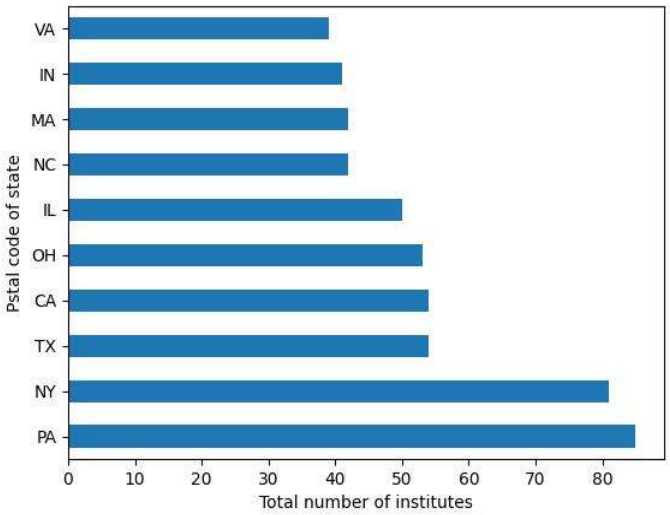

fig. 14 Graph between total number of institutes and postal code of state

As we can see in fig.14 that PA, NY have highest number of institutes as compared to any other state so maybe these two states are more likely to be educational centre of the country.

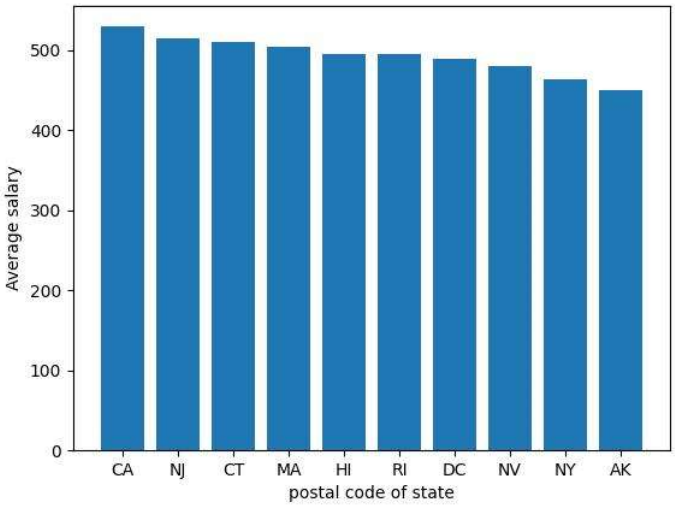

fig. 15 Graph between postal code of state and average salary given to the staff in that state

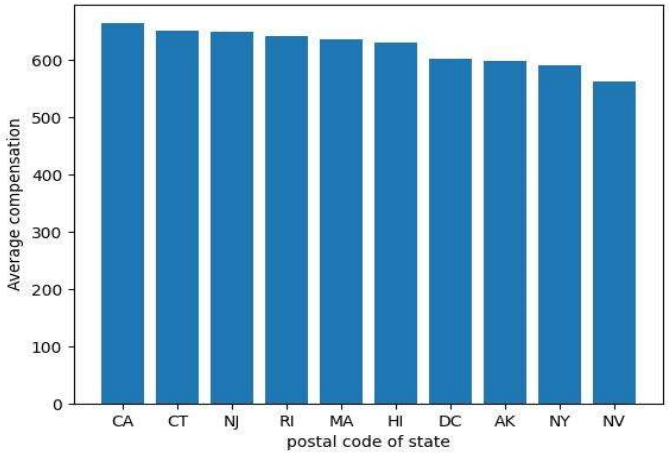

fig. 16 Graph between postal code of state and average compensation given to the staff in that state

Now, we will see that in which state professors are given more salary as compared to other.

So from fig. 15 and fig. 16 we can see that in CA highest salary and highest compensation is given. In this lists, PA and NY are nowhere and this was expected also, because PA and NY are educational centres so there may be more number of professor available than any other state so as availability is more salary will decrease.

Now, if we compare the salary function and compensation function, their correlation factor is 0.8095 so we can easily say that they are correlated.

V. FIND THE PROBABILITY THAT IF I CHOOSE A PUBLIC INSTITUTE, IT IS TYPE I INSTITUTE. ALSO, FIND THE PROBABILITY IF PRIVATE INSTITUTE WAS SELECTED.

To find the probability, firstly we need to find that how many type I institutes are there which are public. So we need to import both datasets in one file only. After that, we will store both the dataframe in two different variables and then in first dataframe, we will add the column of second dataframe which contains the data whether that institute is private or public.

After having this data we can easily make groups of public institutes and then we can find the total number of type I institutes in that. And we can find the probability of type I institute in public institutes. Formula for this probability is shown below.

p = total typeI institutes which are public/total public institutes

So, by this way we get probability p1=0.1446

And similarly, we will find the same probability for private institutes. And we will get probability p2=0.1607

Now we have both the probabilities.

VI. REJECTION RATIO MAY DEPENDS ON THE VALUE OF INSTITUTE LIKE WHERE ONLY SCHOLAR STUDENTS ARE GOING AND WHERE MORE NUMBER OF APPLICATIONAS ARE COMING.

To find the rejection ratio we have subtracted the number of accepted applications from number of received application and after we have divided the value by total number of received application. And as it is obvious that when number of total received application increases rejection ratio increases because total number of seats remains nearly same for most of institutes.

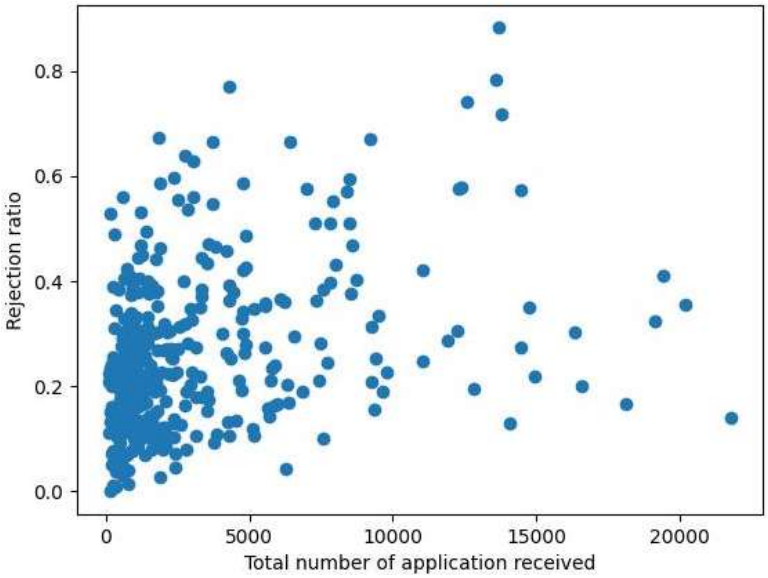

fig. 17 graph between total number of application received vs rejection ratio

Here, in fig. 17 we can easily see that when total number of applications received increases then rejection ratio also increases. But we can see some exceptions between the x-axis value 14000 to 15000.

Institute where number of scholar students are more may reject more applications because there are many brilliant persons who want to take admission in that institute so we will make the graph of number of scholar students means top 10% percent of H.S. classes.

Here, in graph of fig.18 we can see that till the value of 40% on x-axis value of rejection ratio remains nearly same but after that rejection ratio is increasing so much.

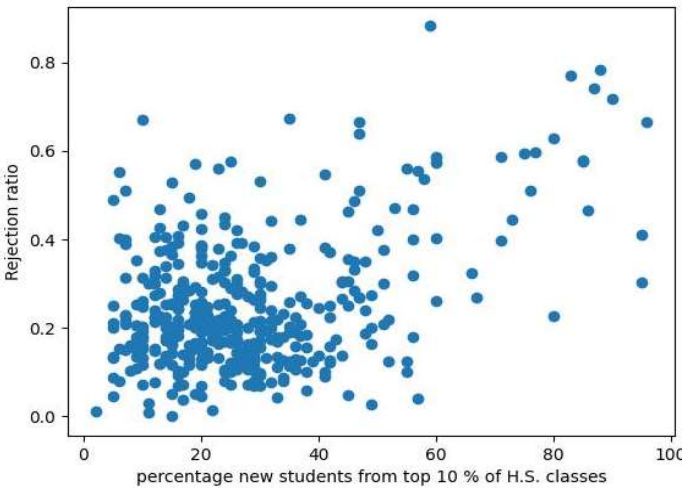

fig. 18 graph between percentage of new students coming from top 10% of H.S. classes vs rejection ratio

Similarly, we can see the trend with the help of average SAT score or average ACT score of students studying in that institution because it also shows that how scholar students are there in that institute.

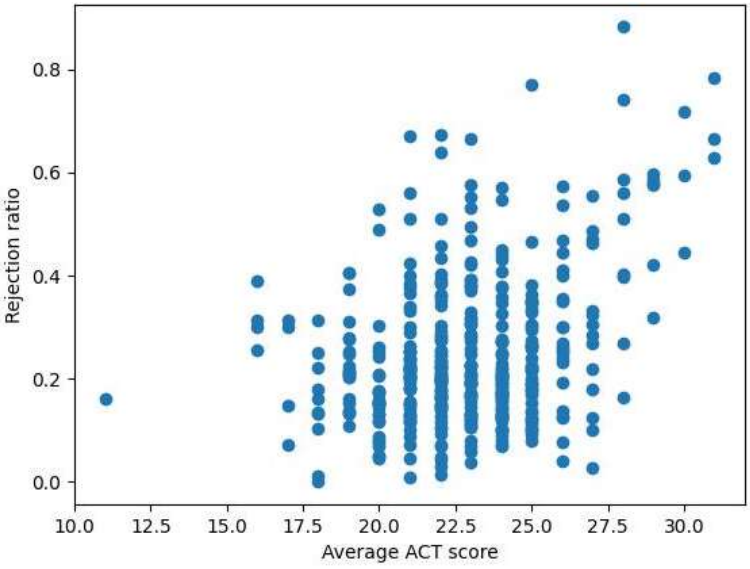

fig. 19 graph between average ACT score and rejection ratio

In fig. 19, we can see that rejection ratio increases but for very high value of ACT score only.

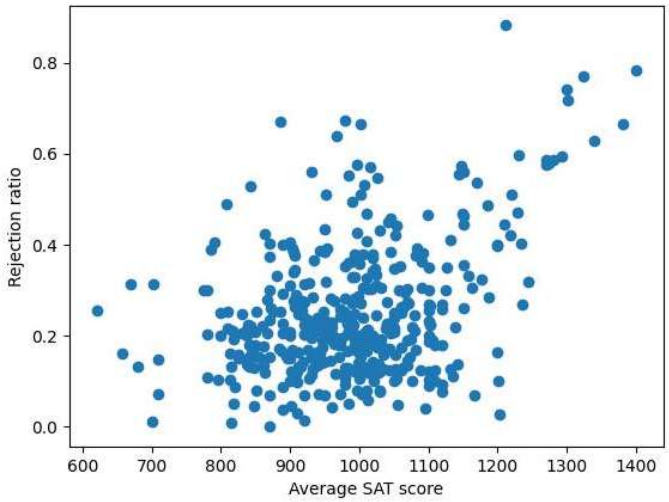

fig. 20 graph between average SAT score and rejection ratio

Similarly, like fig. 13 in fig. 20, we can see that rejection ratio increases but for very high value of SAT score only.

VII. Is a greater number of applications arrive where institute is giving some special functionality?

It is obvious that if some institute are providing more special functions than other institutes than they may get more number of application. So, it is obvious that if we plot a graph between total number of applications received vs student/faculty ratio of that institute.

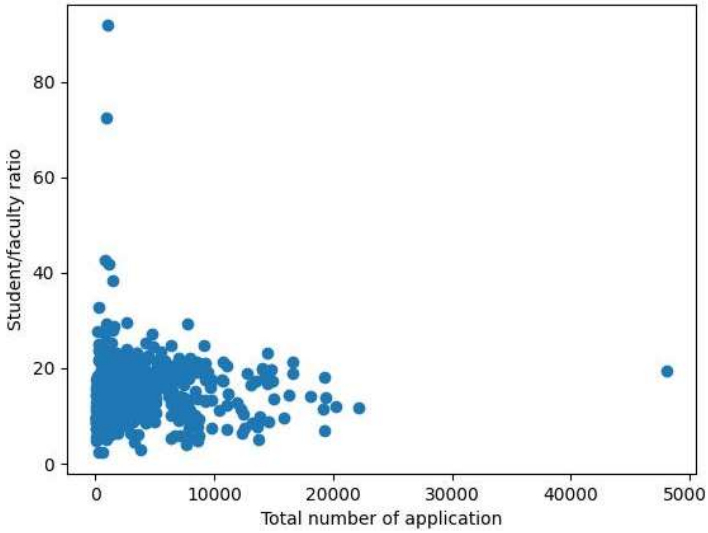

fig. 21 graph between total number of applications received vs student/faculty ratio

Here, in fig. 21 we can see that student/faculty ratio does not depend on each other. And its correlation factor is also equal to 0.060 so we can say that these functions are not correlated.

Similarly, we will plot the scatter graph for Total number of student vs additional fee because if fees are less than students got attracted to that institute because generally people don't want to spend so much amount of money so that they have to take a loan for that.

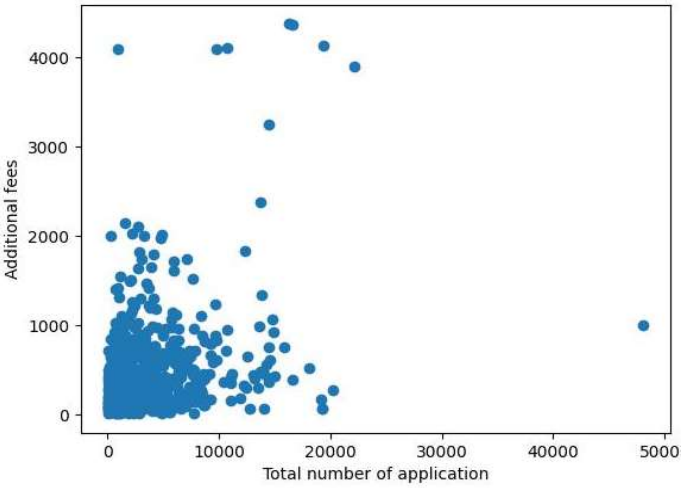

fig. 22 graph between total no. of applications received vs additional fees

In fig. 22, we can not say that total number of applications are depends on the additional fee of the institute. Moreover, correlation factor of these two functions is equal to 0.3531 so those functions are not correlated.

From fig.21 and fig.22 we can see that most of the points are concentrated to one part of the graph only. So, we can say that students generally apply the application in all the institutes.

VIII. Institute where graduation rate is low may get a smaller number of applications.

Institute where graduation rate is very low may not get a greater number of application requests because student generally feel unsecure about their future after taking admission in such kind of institutes.

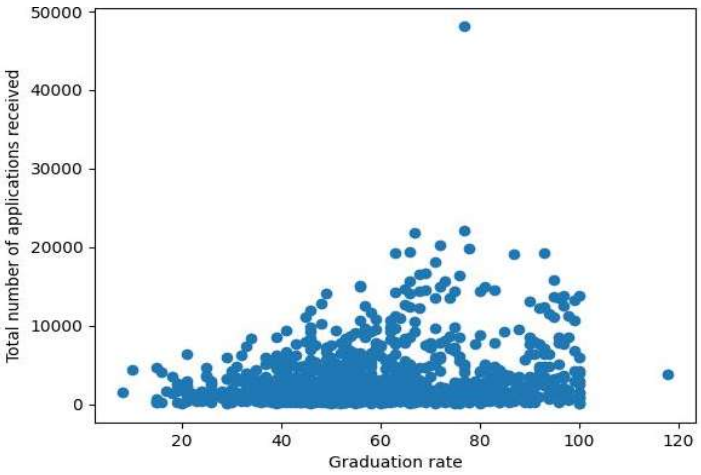

fig.23 graph between no. of applications and graduation rate

In fig.23, we can see that although some points are following our expected trend but still there are many points which are not fitting in our trend. And correlation factor of these two functions is equal to 0.1598 which is very less so we can say that these functions are not correlated.

IX. PROFESSORS MAY BE GETTING MORE SALARY IF THEY HAVE PHD QUALIFICATIONS.

It is generally seen that professors with higher knowledge or higher qualifications is paid more compared to other ones. So, to check this we will plot a graph between percentage of PhD qualified professors and average salary.

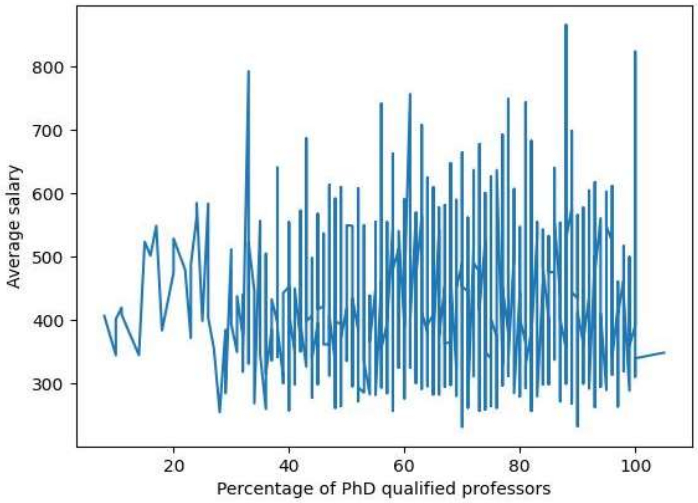

fig. 24 graph between pct. Of PhD qualified professors and average salary

Here, as this graph is not so readable we will take the average value of nearby 11 points and then again plot the graph. So, by doing that, we will get the fig.25.

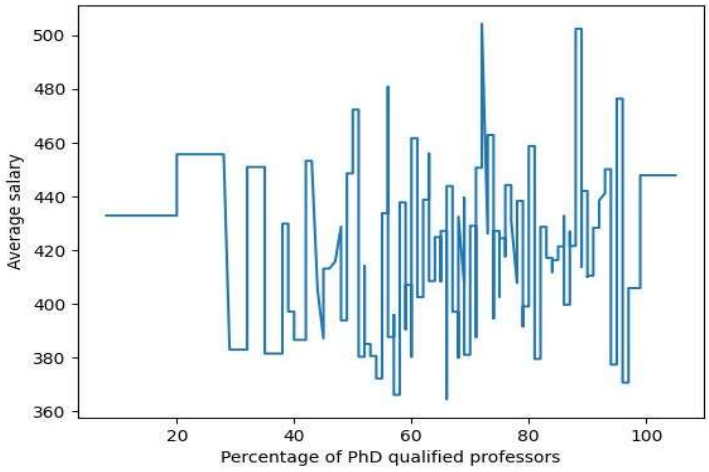

fig.25 Modified version of fig. 24

In fig.25, we can see that the graph is not following a particular trend. So, it is clear that salary given to professors does not depend so much on qualification of professor.

X. BEHIND WHICH THING STUDENT HAVE TO SPEND MORE MONEY? IS THAT SAME FOR PUBLIC INSTITUTE AND PRIVATE INSTITUTE?

To see the overall cost for a student we need to add room and board costs, additional fees, estimated book cost, and estimated personal spending. So, after adding these quantities we created a new column in the dataframe so that we can easily plot that.

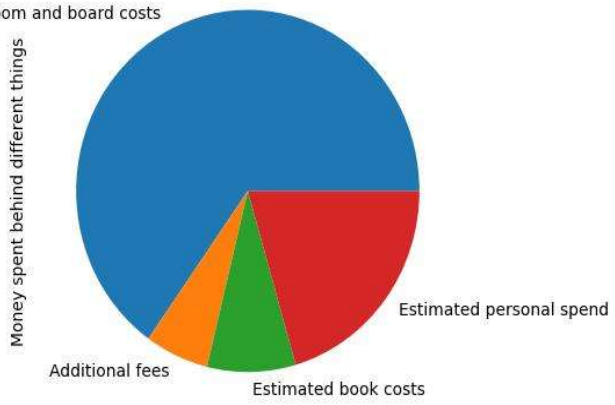

fig. 25 chart of money spent behind different things (for public institute)

Here, fig. 25 shows the money spent behind different things by students who are currently studying in public institute. So, we can see that student have to spent most of money behind room and boards.

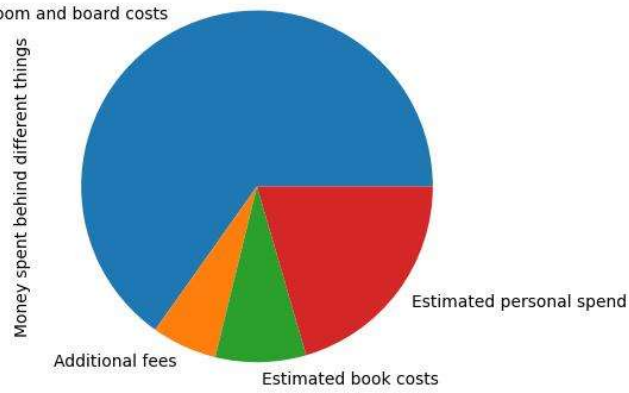

fig. 26 chart of money spent behind different things (for private institute)

In fig.26, we can see that it is exactly similar to fig. 25. So we can see that in private institute also students have to spend more money behind rooms and boards only.

Also, Maximum money spent in public school on rooms and boards is 6776 units, on additional fees is 4374

units, on books is 1125 units and on personal spending is 4288 units. Similarly, Maximum money spent in private school on rooms and boards is 8700 units, on additional fees is 2000 units, on books is 2340 units and on personal spending is 6800 units. So we can say that overall spending done in private institute is much more than public institutes.

D. Details of Libraries and Functions

1. Pandas library is very useful while handling some kind of data especially when we are dealing with csv files.

2. pandas.read_csv() function reads the data of the csv file which we have given as input to the function and we can store the value of that with the help of assigning operator '='.

3. pandas.DataFrame.plot() function is used to plot different graphs. We can also use it as (DataFrame_variable _name).plot() so that we have no need to write the variable name separately in brackets.

4. Matplotlib is a very useful library as it helps to handle the graphs which we are going to plot and it also helps to show the graphs on the screen.

5. NumPy is a library in python that helps us to do a wide range of operations on multidimensional arrays.

6. pd.groupby() function is used to group data in a pandas dataframe based on multiple columns. It creates a 'GroupBy' object, which can be used to apply aggregate functions to the groups.

7. np.corrcoef() is a function provided by NumPy library in python that calculates correlation coefficient matrix for a given set of input arrays.

E. Unanswerable Questions

1. Why the hypotheses 'Institute where graduation rate is low may get a smaller number of applications' is wrong?
2. Why the hypotheses 'Professors may be getting more salary if they have PhD qualifications' is wrong?

F. Summary of the Observations

1. It is generally shown that as number of full professors increases their average salary and compensation is also increasing and they follow the same order.

2. If we compare the average salary and compensation with respect to total number of people then we can see that generally associate professor and assistant professor have more salary than the full professors.

3. Generally, it is seen that type I institute is paying more salary and it is also noticed that type I institute has the top position in giving the highest average salary.

4. It is shown that the state which have a greater number of institutions has no effect on salary or compensation of professors.

5. Rejection ratio depends on the average SAT or ACT score of students and also depends on the percentage of students who are coming from top 10% of H.S. classes. But its effect on rejection ratio is very low.

6. Total number of applications received does not depend on the student/faculty ratio or popularity of the institute, all institutes receive nearly the same number of applications.

7. Students who are studying in private institutes have to spend more money as compared to public institutes.

G. References

1. McKinney, Wes. *Python for data analysis: Data wrangling wit Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

2. Tosi, Sandro. *Matplotlib for Python developers*. Packt Publishing Ltd, 2009.

H. Acknowledgements

I acknowledge that this whole report is written by Ruchit Jagodara, CSE student at IITGN, roll no 22110102 and reserves all rights..