Data Narrative 2

Analyzing data of over 1300 institutes in the United States covering a wide range of variables

Akshat Barnwal
Civil Engineering
Indian Institue of Technology
Gandhinagar, India
akshat.barnwal@iitgn.ac.in

***Abstract*—This is a report on the dataset provided by the library: *http://lib.stat.cmu.edu/datasets/colleges/* which consists of two data files, one called AAUP and the other USNEWS. Both these files contain different data about variables such as tuition cost, admission test scores, number of students and professors and so on. Using this data, one can infer various findings and the see how certain factors affect outcomes such as the graduation rate. It is a useful dataset for anyone looking to get into institutes in the USA and wants to do research on them beforehand.**

Keywords—Python, Pandas, Seaborn, libraries, CSV, KDE.

I. Overview of DataSet

The datasets combined, cover information about over 1300 institutes of the USA. The dataset library contains of three kinds of files:

A. *DATA file*

A special *.data* file that contains data, but is unreadable without special software.

B. *CSV files*

The following two CSV files are present in the repository and they are the primary source of where we obtain our data from:

- **aaup_data.csv** – This file contains information on the number of professors, their salaries, and compensations, etc.
- **usnews_data.csv** – This file contains much more data variables such as entrance exam scores, the number of students, tuition and living costs, student-faculty ratio, and the graduation rate.

C. *DOC files*

The doc files are what were used to name the columns of the CSV files. Otherwise, the CSV files did not have column name. And as such, the DOC files played a key role in naming and understanding the data.

II. Scientific Questions/Hypotheses

A. ***Question A:*** *What kind of distribution pattern does the graduation rate follow across all the universities?*

This is a very basic question; however, it lays the foundation of something very natural. The outcome of attending a university is **graduating** from it. And as such, graduation rate is one of the primary variables we can use to gauge how good or bad a university is. It will also be interesting to see how graduation rates vary across the country overall, and without any constraints (we can get into constraints later when discussing specific factors).

B. ***Question B:*** *How do the average ACT and SAT scores of the students attending universities with a high graduation rate compare to the ones attending universities with lower graduation rates?*

This question in a way, is basically asking – what is a good score to get into a good university. And we all know how important this criterion is. Comparing two of the most important exams for higher education gives us a good enough idea of how they correlate to better graduation rates.

C. ***Question C1:*** *What is the In-State and Out-of-State tuitions for different universities?*

Tuition fee is one of the major deciding factors one consideres before joining a university. Comparing the In-State and Out-of-State tuition will help a resident of a particular state decide how much impact joining an university out of their state will have on them financially.

Hypothesis C2: *Tuition fee of Private vs Public universities.*

Private Universities generally tend to have higher fees and tuition compared to Public universities. We will check wether this hypothesis is true or not.

D. ***Question D:*** *Graduation and Acceptance Rate of Public and Private Universities*

This question seeks to answer how graduation rate and acceptance rate are related to each other. It also aims to compare the same for both Public and Private universities.

E. ***Question E:*** *What is the total expense of living for a university student in the different regions of the United States?*

This question aims to help a student understand which places in the US are better sutied to their financial situation.

Given how most students don't have a lot of financial freedom or security, this is an important question that seeks to help them to help decide how much affect living in a certain region in the US would do to their finances.

F. ***Question F:*** *What is the distribution of the universities in the USA according to their type: I, IIA or IIB.*

This interesting question looks at the distribution of I, IIA and IIB type colleges in the United States. This is a classification used by the Carnegie Classification of Institutions of Higher Education to categorize colleges and universities based on various factors including research activity, and programs offered.

- I (or Doctoral Universities) – include institutions that award doctoral degrees (Ph.D., Ed.D, etc.) and they usually have a high level of research activity.
- IIA (or Master's Colleges and Universities) – include institutions that awards primarily master's degrees and have a moderate level of research activity.
- IIB (or Master's Colleges and Universities) – are similar to IIA institutes, but award primarily bachelor's degrees and have a moderate level of research activity. They are also less diverse in the programs they offer compared to IIA institutes.

G. ***Question G:*** *Salary Distribution across the the different university types.*

Given how I, IIA and IIB universities are different in the programs they offer and research activity. It makes sense that the professors working in them also have different salaries. This question aims to figure out this trend in the different types of universities.

H. ***Question H:*** *Number of Working Professors in each type of University.*

Continuing on the previous question, this question aims to see what is the distribution of the number of professors working in each kind of institution.

I. ***Question I:*** *Average Salary vs Average Compensation of Professors working in the different types of universities.*

Compensation and Salary are two similar yet different values for each institute type. This question seeks to find out the difference between the two. And how both the variables vary across different kinds of institutes.

J. ***Question J:*** *Total Expenditure on Professors in the different kinds of universities.*

It would only make sense that given how the universites vary depending on their type, the total expense that they do on their teaching staff (salary + compensation) will vary. This question aims to answer how it varies.

III. Details of Libraries used

A. ***Pandas***

Pandas is a powerful and open-source python library for data manipulation and analysis in Python. It provides a fast, flexible, and expressive data structure for working with large structured data. The primary data structure used is the DataFrame, a two-dimensional table with labelled axes. It also includes Series, which is a one-dimensional labelled array.

Pandas provides a wide array of tools comprising of various functions and methods for manipulating and analysing data, including data cleaning, transformation, and aggregation. It can also be used to plot said data without the aid of the matplotlib library.

As such, it is no surprise that Pandas happens to be the de-facto standard Python library for working with data and is a popular choice for data professionals and researchers alike.

B. ***Matplotlib***

Matplotlib is a popular data visualization library in Python that has a variety of tools for creating visualizations, both static and animated. It is widely used in data analysis to understand data visually.

Matplotlib provides a variety of plots such as line plots, bar plots, histograms, scatter plots, distribution plots and many more. These plots are also highly customisable, allowing users to change everything from axis limits, colours, font styles, titles, labels and much more. The plots are also publication-quality which is a great plus point.

Moreover, Matplotlib is a versatile library and can be used with a variety of data structures such as lists, arrays, as well as Pandas Series and DataFrames.

C. ***Seaborn***

Seaborn is another data visualization library in Python that is built on Matplotlib. Seaborn however, provides a higher-level interface for creating attractive and informative statistical graphics.

Seaborn is primarily designed to work with Pandas DataFrames and provides functions for creating various types of plots, and has much more options than Matplotlib. It also provides functions for statistical data visualization such as regression, distribution, and categorical plots.

This is an advanced library that is really helpful in plotting advanced plots and as an added bonus, the plots are much more aesthetic and pretty.

D. ***Numpy***

Although not used much in this data narrative, NumPy is one of the most essential libraries to use when getting started with data analysis. Useful libraries of the likes of Pandas and Scipy are built upon existing funcitons and methods of the NumPy library.

NumPy, short for Numerical Python, is a powerful library for scientific computing Python. One of the key features of

it is the multidimensional array object, ndarray, and various derived object classes from it.

IV. Answers to the Questions (with appropriate illustrations)

A. ***Question A:***

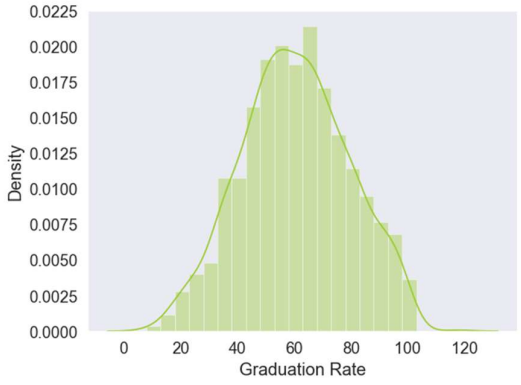

Fig. 1
This histogram shows the density distribution of the graduation rates across all the universities in the United States.

From the given plot in Fig. 1, we get the distribution of the graduation rates across all the given universities in the United States. It is notable to see that the distribution very closely resembles a natural Gaussian (Normal) distribution with the mean at around 60%. This does make sense in the way that most institutes have a decent graduation rate which lies in the middle and there are very few outliers where the rate is either exceptionally low or high.

B. ***Question B:***

From the KDE plot in Fig. 2.1, we can see that an average ACT score of around 20-22 and an average SAT score of 850-900 was obtained by the students joining institutes with graduation rates lower than 40%.

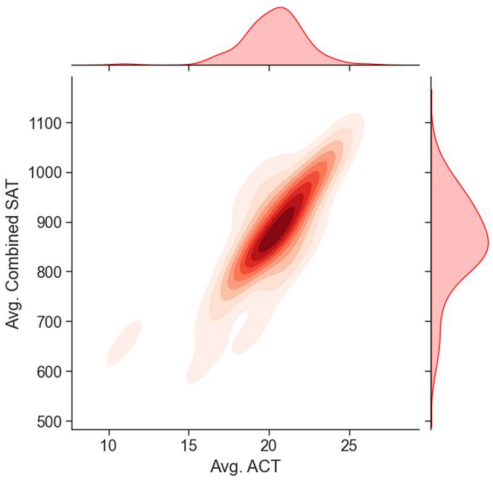

Fig. 2.1
This KDE plot shows the density of average ACT and SAT scores obtained by students who got into universities with graduation rates below 40%.

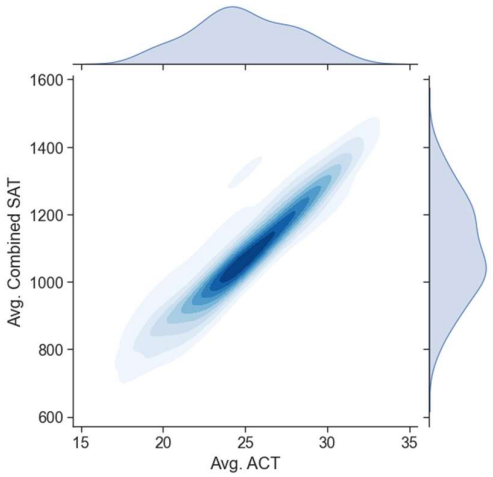

Fig. 2.2
This KDE plot shows the density of average ACT and SAT scores obtained by students who got into universities with graduation rates higher than 80%.

On the other hand, from Fig. 2.2, we can see a clear difference in the students joining institutes with graduation rates above 80% - having average ACT scores well above 25 and SAT scores above 1000.

C. *Question C1 and Hypothesis C2:*

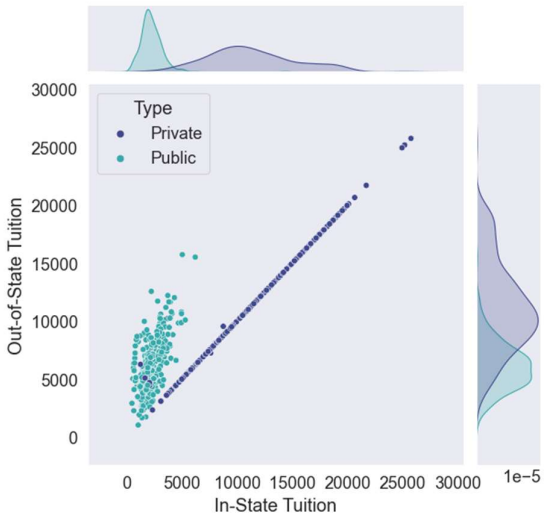

Fig. 3
The trend of In-State and Out-of-State Tuition fee for Public and Private universities.

Plotting a scatterplot of the In-State and Out-of-State Tuition fee shows us one major pattern that was overlooked when designing this question. That there is no concept of In-State and Out-of-State tuition for Private institutions. But other than that, we can clearly see how the tuition for Public universities is crowded in the lower parts of the graph compared their Private counterparts. Another notable pattern, which was to be assumed, is that Out-of-State tuition is quite greater than In-State tuition, sometimes almost double. This goes to show that studying In-State could be a viable benefit to a lot of students who might not have the finances to pursue an expensive education.
As for our hypothesis as well, we can clearly see that there is quite a stark difference in Public vs Private tuition fees. With the Private universities going way beyond $25,000 in some cases and the Public universities lying well within the limit of $15,000 in tuition fees.

D. *Question D:*

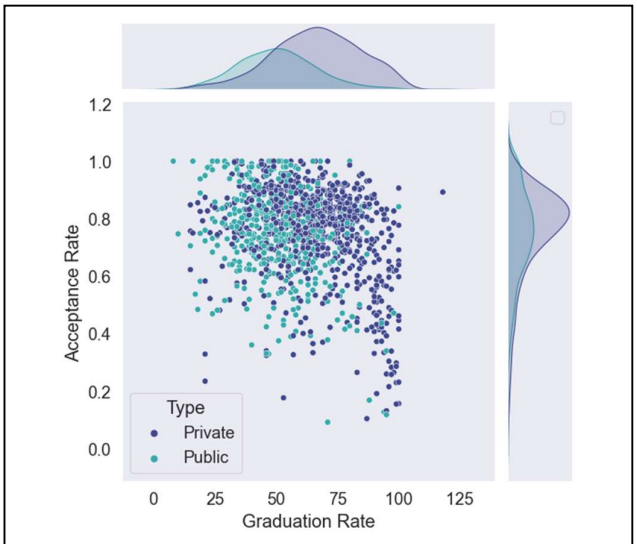

Fig. 4
A scatter-jointplot showing how Public and Private universities compare in their Acceptance and Graduation Rates.

This plot might seem a bit chaotic at first. But taking time to understand it provides valuable insights into the functioning of Private and Public universities.
First of all, looking at the KDE plots on the sides, we can see that the one for Private universities are much denser, indicating a larger intake of students compared to Public universities. Not only that, Private universities also have a skewed distribution towards a higher Graduation Rate indicating that they indeed do perform better in the primary term that we use to rate universities. This aligns with the common expectation that Private universities, although take higher fees, are better universities when it comes to graduation rates and education overall. Another interesting factor to see here is that the plot for the Public universities is much more spread out indicating that there is a much more even distribution in their Acceptance Rates and Graduation Rates, something very different from the Private universities that we talked about just before.

E. *Question E:*

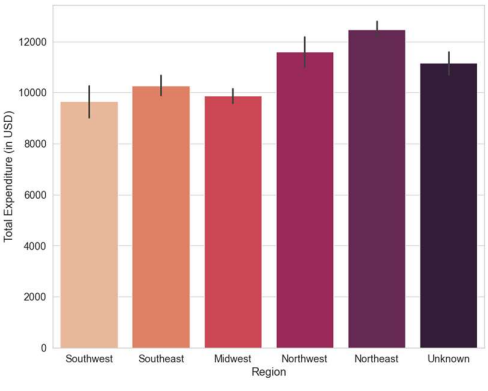

Fig. 5
This bar plot shows the total living cost for a student in the different regions of the United States.

From the plot in Fig. 5, we can see that schools in the Northern region are the most expensive, considering all the factors from tuition fee to living costs. Especially the Northeast which is the most expensive. On the other hand, schools in the Southern states are relatively quiet cheaper for students with the Southwest being the most affordable to place to study. It was a unique approach to group the schools into regions by creating a dictionary with the regions as the keys and the state codes as the values. This was done to create a less cluttered plot as plotting for every state would take up a lot of space and make a mess out of the data. Hence, the data was grouped by region instead of states directly.

F. *Question F:*

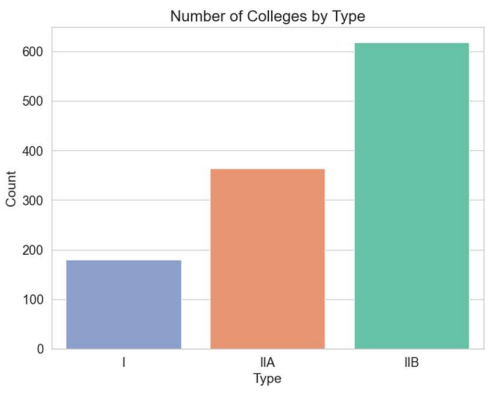

Fig. 6
This bar plot shows the number of each type of college, namely I, IIA and IIB.

The plot in Fig. 6 shows that type I schools are a lot less common than type IIA and IIB. And this should make intuitive sense as they are more specialised and offer doctoral degrees instead of master's and bachelor's. Also, they are a lot more research-focused compared to the other two types. This only adds to them being lesser in number compared to the other types. And with that thinking it only makes sense that type IIB schools are much more common to find.

G. *Question G:*

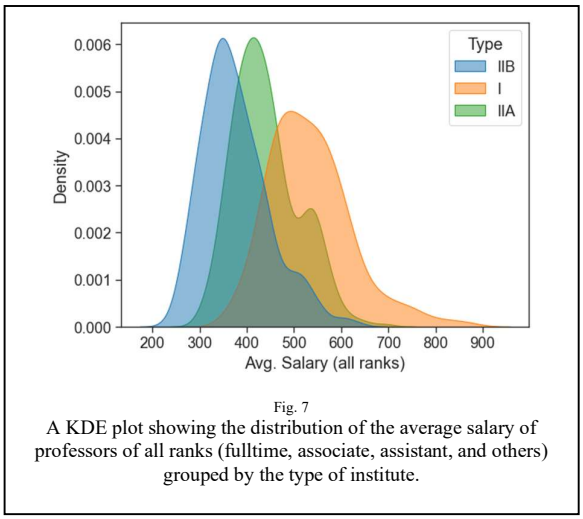

Fig. 7
A KDE plot showing the distribution of the average salary of professors of all ranks (fulltime, associate, assistant, and others) grouped by the type of institute.

From the plot in Fig. 7, it is clearly visible that professors in type I universities are paid significantly higher compared to IIA and IIB universities. This might be due to the fact that teaching in such universities requires much more experience and intellect as they are primarily research-focused and offer specialised degrees. And as for type IIA and IIB universities, the plot shows that a major percentage of professors teaching in them are paid the same amount.

H. *Question H:*

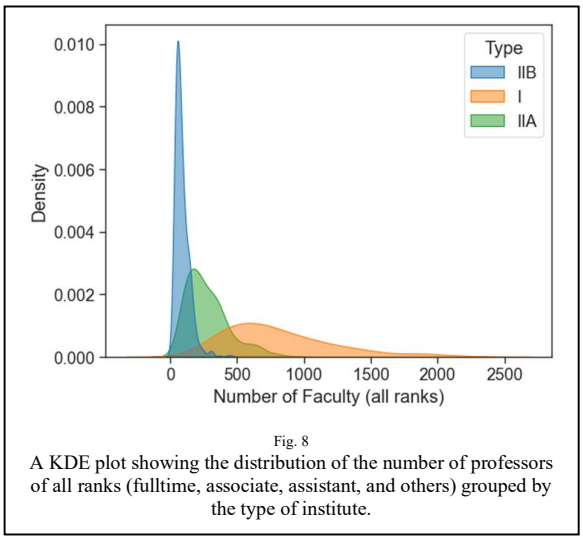

Fig. 8
A KDE plot showing the distribution of the number of professors of all ranks (fulltime, associate, assistant, and others) grouped by the type of institute.

From the KDE plot in Fig. 8, we can see a spike in the plot of type IIB schools. It shows that almost all the type IIB schools have somewhere around 100-200 professors

combined. Whereas the number of professors in type I schools are much more varied and spread out all the way from 500 to over a 1000. This variety appears to emanate from the fact that different type I universities are specialised in different fields and as such have a varying number of teaching staff.

*I. **Question I:***

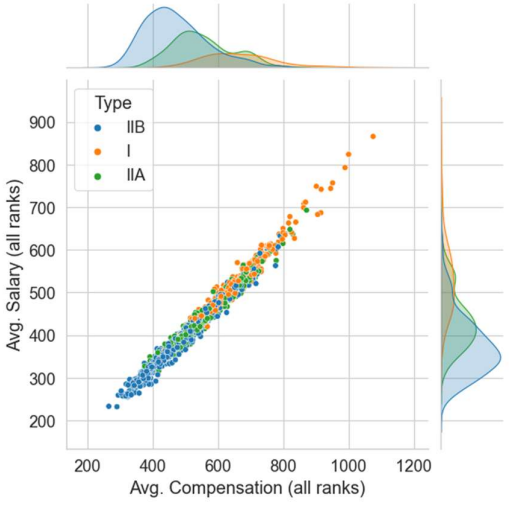

Fig. 9
A scatter/KDE plot showing the trend of Average Salary vs Average Compensation for each type of school.

The plot in Fig. 9 shows how the Average Salary and Average Compensation vary for professors in each type of school. It can be seen that even though they are different terms, the values remain almost the same. Meaning that aside from their salary, the compensation that the professors get in these universities are similar. And the unique thing to notice here is that they remain the same even for different types of universities. Albeit, as discussed in one of the previous questions, both the compensation and salary are highest in type I schools, followed by types IIA and IIB respectively.

*J. **Question J:***

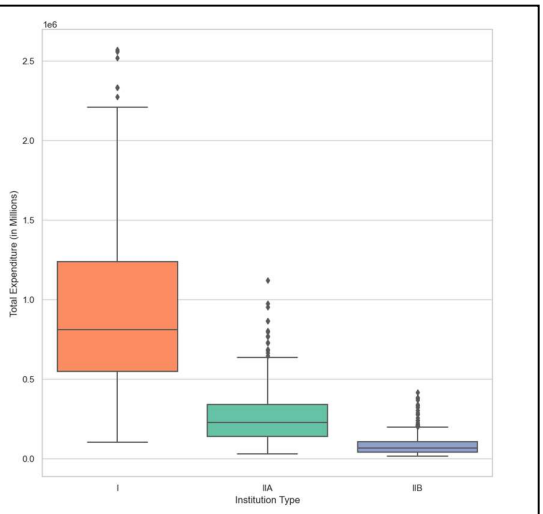

Fig. 10
A boxplot depicting the total expenditure done by institutes on all their professors depending on their type.

In the plot of Fig. 10, we can see again that type I schools spend the most amount of money on their teaching staff (including full-time, associate and assistant professors). And this is further aggravated by the fact that they also usually tend to have more teaching staff. There are little to no outliers for the type I schools but however, the variance is quite strong in their case. On the other hand, in the case of types IIA and IIB schools, the total expenditure is substantially lower but they have a fair amount of outliers that or on the higher side indicating that there are some schools that spend more on their teaching staff.

V. Summary of the Observations

The following are some key observations from the USNEWS dataset:

- Private schools tend to perform better than their Public counterparts. Having lower acceptance rates, a better student-faculty ratio, higher tuition and a lot of better performing students lead to them having consistently higher graduation rates.
- In the case of Public schools, though their In-State tuition fee is quite low, the Out-of-State tuition fee is significantly higher, in some cases, almost double.
- The graduation rate across all the schools is a Gaussian (Normal) distribution with a mean of around 60%.
- The Northeast is the most expensive region of the US to study in and the Southwest is the most affordable.

The following are some key observations from the AAUP dataset:

- Type I institutes can be considered as elite level institutions. They are a lot less in number compared to IIA and IIB type institutes and offer specialised programs and provide primarily doctoral degrees.
- Type I institutes also have higher pay for their teaching staff as well an overall higher number of teaching staff compared to types IIA and IIb of institutions.

VI. Unanswerable Questions

While the given dataset contains a lot of information on various matters for different colleges. There are some limitations to the kinds of question we can answer.

Following are some of the questions which I came across while thinking about the dataset that I believe could not be answered with the current data:

- Is there any direct correlation between the pay of the professors and the living costs for the students?
  This question could have been answered if there was interplay between the two datasets. But as there was not and the required data was in separate datasets, the question was deemed unanswerable.
- The weightage of different factors, such as SAT and ACT scores, in determining outcomes such as Graduation Rate and Acceptance Rate.
  Given my limited knowledge of working with Python libraries such as Pandas and Seaborn, it proved technically difficult to plot the required relationship between them.

VII. References

Following are the websites that I could help from while working this report. Note that these are just the links to the parent website where the help was taken from, I could not reference every single question or doubt that I referred to these websites for.

[1] https://pandas.pydata.org/docs/
[2] https://seaborn.pydata.org/
[3] https://matplotlib.org/stable/index.html
[4] https://stackoverflow.com/
[5] https://www.geeksforgeeks.org/pandas-tutorial/
[6] https://youtu.be/6GUZXDef2U0

VIII. Acknowledgement

I would like to thank Professor Shanmuganathan Raman for providing us this opportunity to write this data narrative. In the labs, we have familiarized ourselves with multiple Python libraries in the form of lab questions. However, working on an assignment like this helped us practice using these tools practically. Also, in this assignment, I got to explore a new Python library Seaborn to make interesting and novel plots to handle and visualize data in much more efficient ways.