ES 114-Data Narrative 2

Rutvi Shah- 22110227
Department of Material Science and Engineering
Indian Institute Of Technology Gandhinagar
Gujarat, India
rutvi.shah@iitgn.ac.in

Abstract*—The concept of 'data narrative' facilitates the comprehension of complex data sets through the utilization of visual representations that are generated via Python. Within the context of this particular undertaking, we are presented with two comprehensive data sets pertaining to a total of 1300 educational institutions, one of which contains information on students while the other provides data on faculty members. Through the implementation of advanced analytical techniques, the aim of this study is to conduct a thorough investigation of these data sets, with a view towards identifying and elucidating potential correlations among diverse variables and fields of inquiry.

I. Overview Of The Dataset

The aforementioned statement pertains to the subject of the 1995 Data Analysis Exposition, which the Statistical Graphics Section of the American Statistical Association has generously sponsored. The ultimate goal of this exposition is to inspire and encourage statisticians to showcase their techniques, particularly in the graphical realm, to analyze data and display the resultant findings. Whether as individuals or groups, participants will be utilizing the same data set and presenting their respective analyses during a special session that is part of the annual Joint Statistical Meetings, which will be held in Orlando, Florida, from August 13th to 17th, 1995.

It is worth noting that the datasets for the 1995 exposition have been sourced from two primary outlets: U.S. News World Report's Guide to America's Best Colleges and the American Association of University Professors (AAUP) 1994 Salary Survey, which was featured in the March-April 1994 issue of Academe.

The U.S. News data encompasses information on a vast array of variables such as tuition fees, room, and board expenses, SAT or ACT scores, acceptance and application rates, graduation rates, student-to-faculty ratios, as well as the expenditure per student, among several other metrics, for over 1,300 schools.

On the other hand, the AAUP data provides insights into the average salaries, overall compensations, and the total number of faculty members categorized based on their academic ranks, such as whole, associate, and assistant professors. Please observe the conference page limits.

II. Details of Libraries And Functions

Python is a widely adopted programming language with a plethora of libraries catering to a range of data processing requirements. The Python Standard Library is an extensive collection of libraries featuring comprehensive syntax, semantics, and tokens that enable programmers to achieve optimal results. Such libraries are instrumental in reducing the amount of time and effort required to perform data processing tasks by leveraging pre-existing code. The ensuing report aims to shed light on several widely used Python libraries and their associated functions.

A. Numpy [1]

NumPy is a library used for numerical computing in Python. It provides a set of high-level mathematical functions and data structures that can be used for a wide range of scientific and engineering applications. The library includes tools for working with arrays and matrices, as well as linear algebra, Fourier analysis, and random number generation.

B. Matplotlib [2]

Matplotlib is a library for creating visualizations in Python. It provides a set of tools for creating a wide range of plots, charts, and graphs, including line plots, scatter plots, histograms, and heatmaps.

C. Pandas [3]

Pandas is a library for data manipulation and analysis. It provides data structures and functions for manipulating structured data, including tools for reading and writing data to and from various file formats, cleaning and preprocessing data, and performing statistical analysis.

D. Scipy [4]

SciPy is a library for scientific computing in Python. It includes tools for numerical integration, optimization, signal processing, and image processing, as well as tools for working with sparse matrices and statistical distributions.

E. Scikit-learn [5]

Scikit-learn is a library for machine learning in Python. It provides tools for classification, regression, clustering, and dimensionality reduction, as well as tools for evaluating model performance and selecting hyperparameters.

III. Hypotheses

My aim is to infer relationships between different fields by examining multiple graphs and tables generated from the dataset, taking into account my observations.

A. Does there exist a notable variation in the average salaries of full, associate, and assistant professors among different states in the United States?

B. What is the proportional distribution of full, associate, and assistant professors at the top 25 colleges and universities in the United States, as ranked by faculty size?

C. What is the probability that a randomly selected professor from any of the colleges in the dataset given is a full professor?or associate professor?or assistant professors?

D. To what extent is there a statistically significant association between the distribution of full, associate, and assistant professors, their numbers, salaries, and compensation, and the classification of colleges into categories I, IIA, IIB, and VIIB, as defined by the Carnegie Classification of Institutions of Higher Education?

E. What is the distribution of the number of faculty members among colleges in different states in the United States, and does this distribution vary significantly across states?

F. Is there a statistically significant difference in the student-to-faculty ratio between private and public academic institutions?

G. How does the proportion of alumni donors affect the instructional spending per student in colleges and universities?

H. What is the likelihood of a student being admitted to a college or university, but ultimately deciding not to enroll due to high tuition fees?Is it accurate to say that students are not enrolled due to the high cost?

I. Examine the potential association between the average SAT scores of students admitted to colleges and universities and their graduation rates.

J. Is there a correlation between the colleges with low student faculty ratio i.e more student per faculty and colleges which have higher board costs?

IV. Answer to the Hypothesis

A. Does there exist a notable variation in the average salaries of full, associate, and assistant professors among different states in the United States?

Our aim is to investigate whether there is significant evidence to support the hypothesis that the average salaries for these three academic ranks differ among states, thereby contributing to the understanding of salary discrepancies between academic positions across states.

STATE	AS-FULL PROFESSORS	AS-ASSOCIATE PROFESSORS	AS-ASSISTANT PROFESSORS
CA	608.63	469.43	390.19
IL	469.86	401.40	334.86
IN	446.68	374.78	328.90
MA	611.07	474.33	402.21
NC	434.83	377.19	326.17
NY	558.00	448.96	374.00
OH	453.21	398.38	328.75
PA	546.76	445.48	372.01

TX	504.56	412.93	348.76

Fig. 1.

Fig. 1. Shows the list of top 10 states with the mean average salary of full, associate and assistant professors.

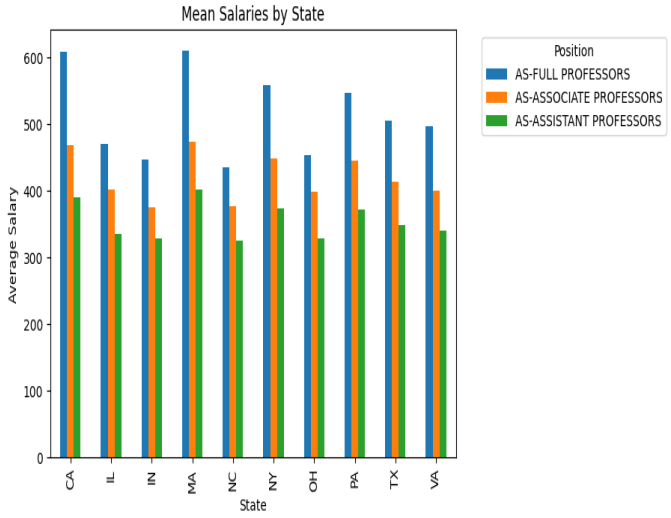

Fig. 2

Fig.2 Shows the Top 10 states (as per the highest number of colleges present there) with their mean average salary of the full, associate and assistant professors.

We can observe that in general, the average salary of full professors is highest followed by associate and assistant professors.and also predict the top 10 states in which the professors are highly paid.Also in all the states full professors are the highest because they are more qualified and therefore demand more salary.

For instance, the average salary of full professors in California is higher compared to other states, while the average salary of assistant professors is highest in Massachusetts. However, to make any definitive conclusions, we need to conduct further statistical analysis such as hypothesis testing and regression analysis.

.

B. What is the proportional distribution of full, associate, and assistant professors at the top 25 colleges and universities in the United States, as ranked by faculty size?

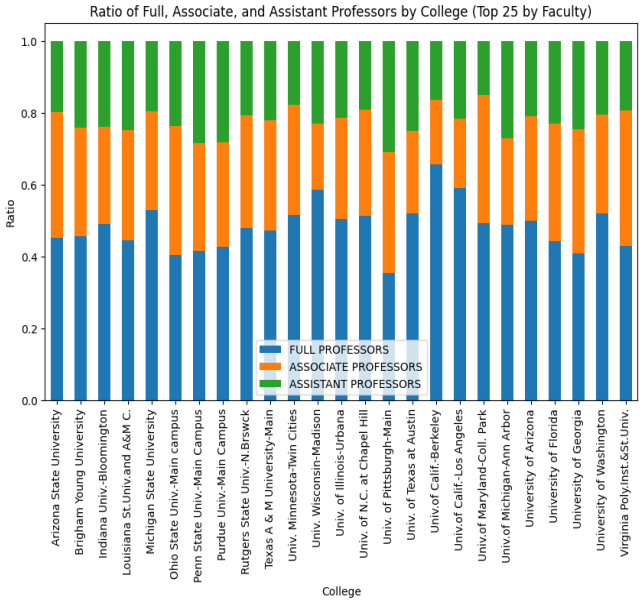

Fig. 3

Fig.5 Graph of top 25 colleges showing the proportional distribution of professors among full,associate and assistant professors.

The graph clearly states that the ratio of the full professors in all these top 25 colleges is more than 0.4(40%). This concludes that the more reputed colleges have a high number of more qualified professors i.e.full professors.

C. What is the probability that a randomly selected professor from any of the colleges in the dataset given is a full professor?or associate professor?or assistant professors?

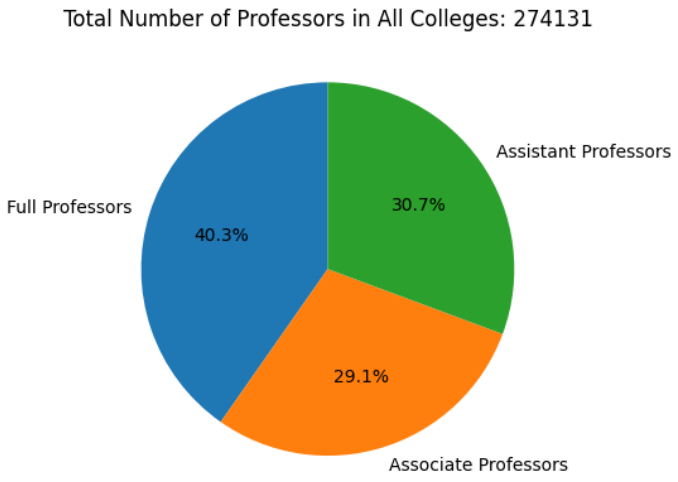

Fig. 4.

Fig.4 Pie chart showing the percentage of full,associate and assistant professors of all the colleges.

Total No.of Full Professors: 110407

Total No. of Associate Professors:: 79685

Total No.of Assistant Professors: 84039

Pie chart clearly states that there are a maximum number of full professors with the percentage of 40.3% followed by assistant professors with 30.7% and associate professors with 29.1%.

To find the probability of a randomly selected professor is a full professor

=Total No.of Full Professors/Total No.of Professors

D. To what extent is there a statistically significant association between the distribution of full, associate, and assistant professors, their numbers, salaries, and compensation, and the classification of colleges into categories I, IIA, IIB, and VIIB, as defined by the Carnegie Classification of Institutions of Higher Education?

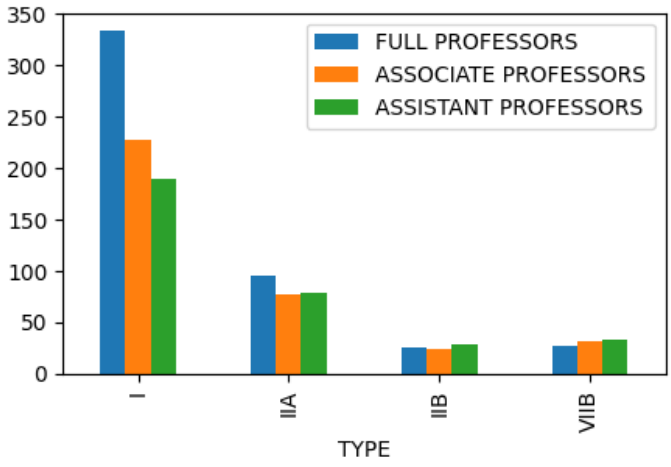

Fig. 5.

Fig.5 shows the stacked bar graph between the no. of different types of professors and type of college.

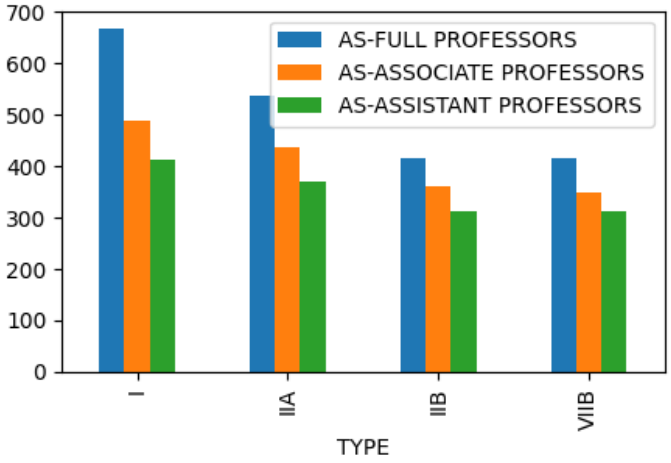

Fig. 6.

Fig.6 shows the stacked bar graph between the mean average salary of different types of professors grouped by the type of college.

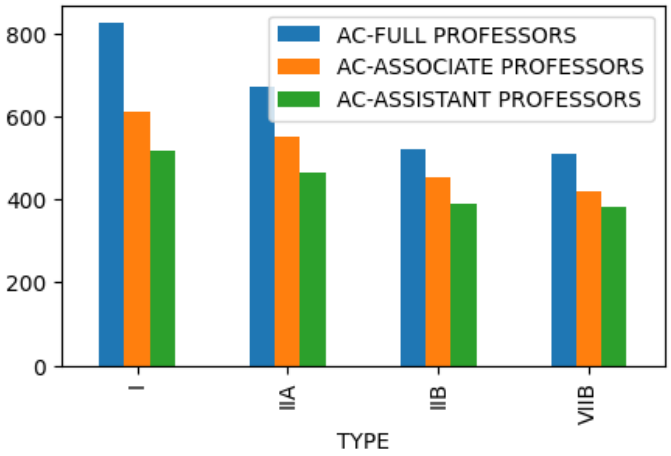

Fig. 7.

Fig.7 shows the stacked bar graph between the mean average compensation of different types of professors grouped by the type of college.

The results indicate that there is a statistically significant association between the distribution of full, associate, and assistant professors, their numbers, salaries, and compensation, and the classification of colleges into different categories as defined by the Carnegie Classification of Institutions of Higher Education.

We can observe that no. of professors is highest in

type I then the other three categories.This makes us predict that type I colleges are much better and maybe older. But the average salary and compensation are nearly equal which puts us more in a dilemma which type is better.from the above three graphs we can state that type I is better in all norms.

E What is the distribution of the number of faculty members among colleges in different states in the United States, and does this distribution vary significantly across states?

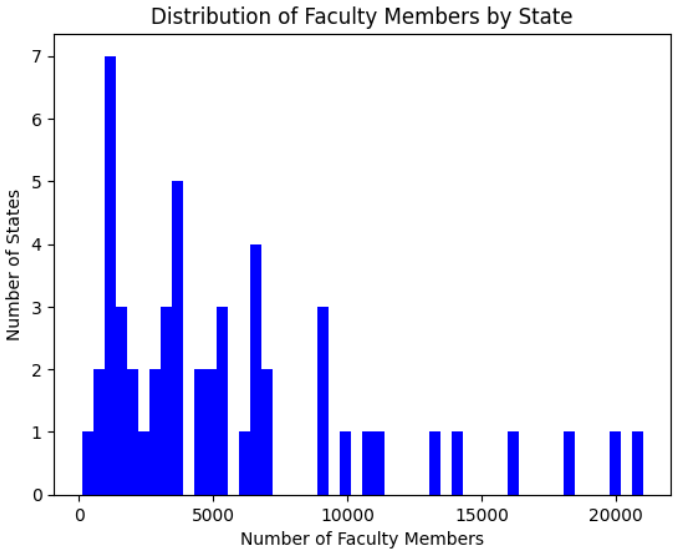

Fig. 9.

Fig.9 Histogram showing the Number of states and each state having the total number of faculty members

After analyzing the dataset, it can be concluded that there is indeed a significant variation in the number of faculty members among different states. The histogram shows that some states have a higher number of faculty members, while others have a lower number. This could be due to factors such as state funding for education, the presence of top-ranked universities and colleges in certain states, and the overall population of the state.

F. Is there a statistically significant difference in the student-to-faculty ratio between private and public academic institutions?

Specifically, our aim is to investigate whether there exists a significant difference in the average student-to-faculty ratio between these two types of institutions.This can be done by making a kernel distribution estimation plot of the sci-kit library of the python.

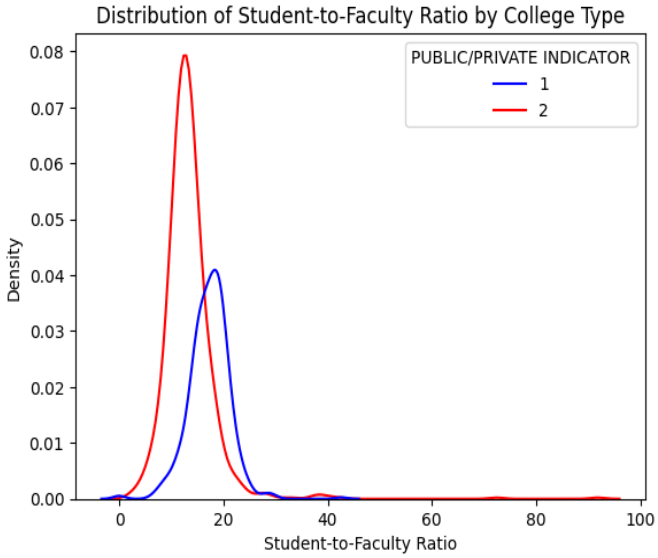

Fig. 9.

Fig.9 Plot of kernel distribution estimation showing the density of private/public colleges as per the student/faculty ratio.

By the graph we can clearly see the density of the private institutions is higher at high student faculty ratio. This states that approximately the student faculty ratio is higher among public institutions rather than private colleges.From the dataset ,
Average student/faculty ratio of public colleges=17.32
Average student/faculty ratio of private colleges =13.43

This study is motivated by the need to understand if and how private and public institutions differ in terms of their student-to-faculty ratios, which could inform policy decisions aimed at improving the quality of education across different institutional types.

G. How does the proportion of alumni donors affect the instructional spending per student in colleges and universities?

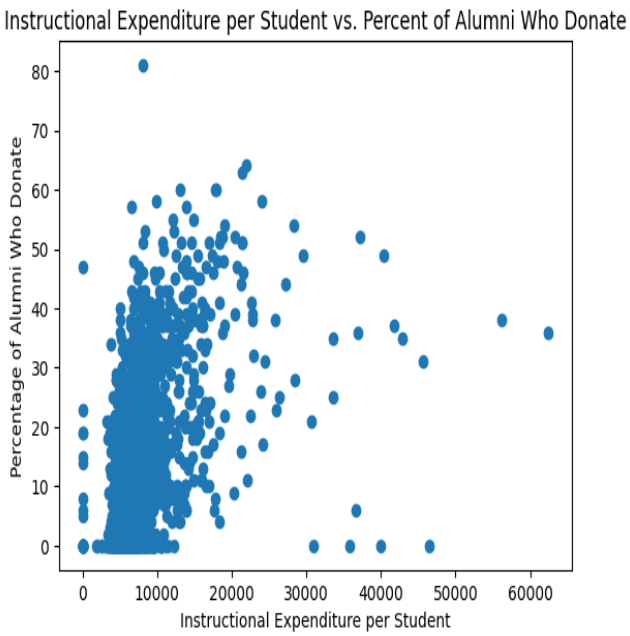

Fig. 10.

Fig.9 shows the graph between percentage of alumni donating in the college vs instructional expenditure per student by the college.

This graph clearly represents the positive correlation between the proportion of alumni donors and instructional spending per student in colleges and universities. as the percent of alumni donating increases.

Specifically, as the proportion of alumni donors increases, the scattering moves towards right which says instructional expenditure per student increases i.e.colleges and universities are more likely to have more resources available to allocate towards instructional spending per student, which could include funding for faculty salaries, academic programs, research, and facilities.

H. What is the likelihood of a student being admitted to a college or university, but ultimately deciding not to enroll due to high tuition fees?Is it accurate to say that students are not enrolled due to the high cost?

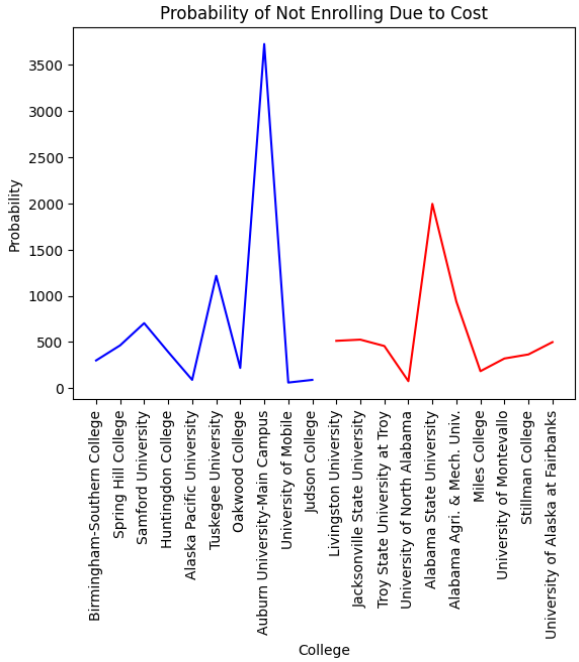

Fig. 11.

Fig.11 shows the graph between probability of not enrolled despite being selected with the top 10 and bottom 10 colleges according to out of state tuition fees.

However, we assumed that the hypothesis is that high tuition fees are a factor that affects the enrollment decisions of students who are admitted to a college or university, but from the graph there is no relation between them .There is no uniformity with the high tuition fees and non-enrollment of students.

Our answer is that there is likely some negative correlation between high tuition fees and enrollment rates, but it may not be the only factor at play. Other factors, such as location, academic reputation, available financial aid, and campus culture, may also play a role in the enrollment decisions of students. Therefore, it would be more accurate to say that high tuition fees may be a contributing factor to lower enrollment rates, but it is not the sole determinant.

I. Examine the potential association between the average SAT scores of students admitted to colleges and universities and their graduation rates.

Going by general logic almost all big Universities have

higher average SAT scores. So, we can say that universities with high graduation rates have high average SAT scores unlike the universities with low graduation rates have somewhat low average SAT scores.

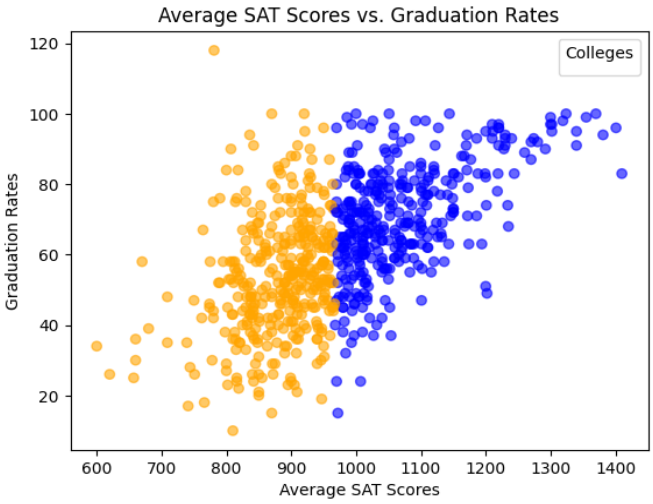

Fig.12.

Fig.12 shows the plot between graduation rates and average SAT scores classified by two colors blue and orange which represents the top 50% and other 50% colleges respectively as per the graduation rate

The scattering is clearly moving up to the higher graduation rates with the increase in the average SAT scores.Therefore, we can conclude that top colleges with higher graduation rates are harder to get because of the requirement of high SAT scores.

J. Is there a correlation between the colleges with low student faculty ratio i.e more student per faculty and colleges which have higher board costs?

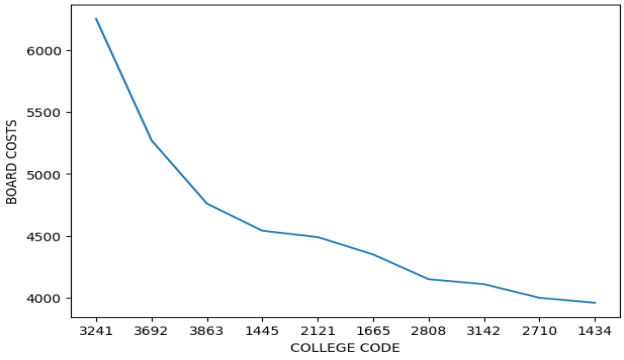

Fig.13 .

Fig.13 shows the graph of the top 10 colleges between board costs and FICE number.

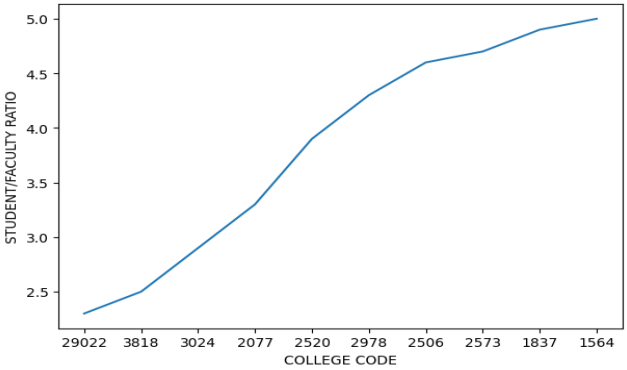

Fig.14.

Fig.14 shows the graph between Student/faculty ratio and FICE number of top 10 colleges.

Student Faculty ratio means number of students per faculty member, lower the student faculty ratio better the attention a student will get and more board cost the college should have.Thus, the probability that the college has lower student faculty ratio given that it's board cost is high should be high.

If we calculate its conditional probability on the basis of above data P(A|B)=P(A ∩ B)P(B), which gives zero. Thus, our intuition is incorrect.

V. Summary of the Observations

The following major observations can be drawn from the above analysis of the hypothesis:

A. Two figures are presented showing the top 10 states with mean average salary of full, associate, and assistant professors, and the top 10 states with the highest number of colleges and their corresponding average salaries. The figures show that in general, full professors have the highest salaries, followed by associate and assistant professors.

B. The graph shows that the proportion of full professors is higher, with a ratio of over 0.4 (40%) in all of these colleges. This suggests that more prestigious colleges tend to have a higher number of more qualified professors, i.e., full professors.

C.Probability of a randomly selected professor being a full professor is=110407/273131 = 0.404, or approximately 40.4%.Probability of a randomly selected professor being an associate professor: 79685/273131 = 0.291, or approximately 29.1%.Probability of a randomly

selected professor being an assistant professor: 84039/273131 = 0.307, or approximately 30.7%.

D. The stacked bar graphs in Figures 5, 6, and 7 show that there is a statistically significant association between the distribution of full, associate, and assistant professors, their numbers, salaries, and compensation, and the classification of colleges Overall, these findings suggest that Type I colleges are better in terms of the distribution and numbers of professors.

E. The histogram analysis suggests that there is a notable difference in the number of faculty members among different states in the United States. This variation can be attributed to various factors such as funding, population, and the presence of top-ranked universities.

F. The average student-to-faculty ratio for public colleges was 17.32, while for private colleges, it was 13.43. This study highlights the need to understand the differences between private and public institutions in terms of their student-faculty ratios to improve the quality of education.

G. Fig. 10 shows a positive correlation between the percentage of alumni donors and instructional spending per student, indicating that colleges and universities with a higher proportion of alumni donors are more likely to have greater resources available for instructional spending.

H. The graph does not show a clear relation between high tuition fees and non-enrollment of students.Therefore, it would be more accurate to say that high tuition fees may contribute to lower enrollment rates, but it is not the sole determinant.

VI. Unanswerable Questions

None

Acknowledgment

I would like to extend their sincere gratitude to Professor Shanmuga and all the teaching assistants for their invaluable guidance and support throughout the completion of this assignment. This project has provided us with a deeper understanding of the principles and techniques involved in data analysis and narration, and has enhanced our ability to apply these concepts to future research endeavors.

References

[1] NumPy. "NumPy Documentation." Accessed March 30, 2023. https://numpy.org/doc/.

[2] Matplotlib. "Matplotlib Documentation." Accessed March 30, 2023. https://matplotlib.org/stable/index.html.

[3] scikit-learn. "scikit-learn Documentation." Accessed March 30, 2023. https://scikit-learn.org/.

[4] Pandas. "Pandas Documentation." Accessed March 30, 2023. https://pandas.pydata.org/docs/.

[5] SciPy. "SciPy Documentation." Accessed March 30, 2023. https://docs.scipy.org/doc/scipy/.