DATA NARRATIVE 2

ES 114 PROBABILITY,STATISTICS AND DATA VISUALIZATION

Purva Kaushalbhai Shah
Chemical Engineering
Indian Institute of Technology Ganghinagar
Ganghinagar,India
purva.shah@iitgn.ac.in

***Abstract*—Abstract—This document is the report on Data Narrative. The data given in form of 6 files. Three files contain the data in csv format and two files are word files. All the files provided are analyzed and the report is generated.**

I. INTRODUCTION

In this task of Data Narrative, we are analyzing the data present in csv files by using few visuals and databases. The database is about universities present in the USA and this data is collected by the U.S. News & World Report's Guide to America's Best Colleges (1995) to help people choose the universities. The dataset contains all the basic information about the universities which helps the reader to choose universities properly.

II. OVERVIEW

In this data narrative we are reading two files named aaup and usnews. The file contains different data and can be merged for the better data analysis. The data provided here is collected by the U.S. News & World Report's Guide to America's Best Colleges (1995) to help people choose the universities.
We can group this data by the filter we want like state, type, public/private etc.

The data aaup contains columns 'FICE (Federal ID number)','College name','State (postal code)','Type (I, IIA, or IIB)','Average salary - full professors','Average salary - associate professors','Average salary - assistant professors','Average salary - all ranks','Average compensation - full professors','Average compensation - associate professors','Average compensation - assistant professors','Average compensation - all ranks','Number of full professors','Number of associate professors','Number of assistant professors','Number of instructors','Number of faculty - all ranks' and data us news contains columns 'FICE (Federal ID number)','College name', 'State (postal code)','Public/private indicator (public=1, private=2)','Average Math SAT score', 'Average Verbal SAT score',' Average Combined SAT score', 'Average ACT score', 'First quartile - Math SAT', 'Third quartile - Math SAT', 'First quartile - Verbal SAT', 'Third quartile - Verbal SAT', 'First quartile - ACT', 'Third quartile - ACT', 'Number of applications received',' Number of applicants accepted',' Number of new students enrolled', 'Pct. new students from top 10% of H.S. class', 'Pct. new students from top 25% of H.S. class', 'Number of fulltime undergraduates', 'Number of parttime undergraduates', 'In-state tuition', 'Out-of-state tuition', 'Room and board costs', 'Room costs', 'Board costs', 'Additional fees', 'Estimated book costs', 'Estimated personal spending', "Pct. of faculty with Ph.D.'s", 'Pct. of faculty with terminal degree',' Student/faculty ratio', 'Pct. alumni who donate', 'Instructional expenditure per student',' Graduation rate'.

We can link this data by using common columns like 'FICE (Federal ID number)','College name',
'State (postal code)'.

III. QUESTIONS / HYPOTHESIS

A. How does the number of professors and their pay vary? What is the average salary of professors and how many professors are present in each type? If a professor is chosen by random what is the probability that he is an associate professor?

B. Which states has the higher number of universities (top 5)? Show the distribution of type of universities present in top 3 states having higher number of universities. Which state has highest universities of type IIB?

C. Show the difference between the average compensation and average salary of professors. Who has the highest difference in their mean?

D. What proportion of the staff of university is professors,instructors and faculty?If a staff member is chosen at random what is the probability that it is a full time professor given that he is a professor?

E. What is the difference between the average salary given to full professors, assosiate professors and assistant professors.Which one has the higher frequency of getting paid highly?

F. How does the mean of Average Math SAT score,Average Verbal SAT score,Average Combined SAT score of a univerity vary according to the type(public/private) of university? Whose average combined score is higher?

G. How are Number of applications received,Number of applicants accepted,Number of new students enrolled related? If a student is chosen at random what is the probaiblity that he enrolled given that he accepted the application?

H. Show the distibution of fulltime undergraduates and parttime undergraduates present in the university? What is the probability that the person chosen at random is a parttime undergraduate?

I. Which type(public/private) of collge has more faculties? And what is the the avg number of sudents alloted to a faculty?

J. How does Instate tution and out of state tution vary from state to state? Which state has highest In-state tution?Which state has the max difference between the In-state tution and out-of-state tution?

K. How many parttime alumni donated the money to university?

IV. SOFTWARE / LIBRARIES USED

A. Google colab

A platform provided by google where we can code. In this task google collab was used to read the uploaded file, code and write some texts in between. The code in google collab is written in python language.

Fig 1 – Google colab [1]

B. Pandas [2]

Pandas is a software which provides us with various inbuilt function. In this task, we are using pandas to read the csv file and operate on it. The csv files are viewed are operated like we operate the Data frames.

1. Functions used:
    a. pd.read_csv(): to read the csv file
    b. dtypes: to check the type of columns present in the dataframe
    c. replace(): to replace the old value with new value in dataframe
    d. astype(): to change the datatype of the column present in dataframe
    e. sum(): use to calculate the sum of data present in rows/columns of dataframe
    f. groupby() : to group the identical values and separate out the data
    g. mean(): used to calculate the mean of the data present in rows/columns of dataframe
    h. pd.head(): to return few rows of the dataframe from the top
    i. count(): used to count the occurrence of any data
    j. sort_values(): it is used to sort the data using a specific label
    k. unique(): used to give unique values present in the row/column/series.

Fig 2 – Pandas [2]

C. Matplotlib [3]

Matplotlib is the software used to plot the graphs. In the task of data narrative, it was used to plot the various type of graphs like bar graph, pie chart etc.

1. Functions used:
    a. figure(): to open a panel where graph can be plotted
    b. plot(): to plot the given graph
    c. xlabel(): to label x axis
    d. ylabel(): to label y axis
    e. bar(): to plot the bar graph
    f. pie(): to plot the pie chart
    g. show(): to show the plotted graph
    h. plot.scatter() : to plot the scattered graph
    i. explode : to make one of the wedge present in pie chart stand out of others
    j. legend(): to show the color or symbols represents what in the graph
    k. autopct: to show the percent of data in pie chart

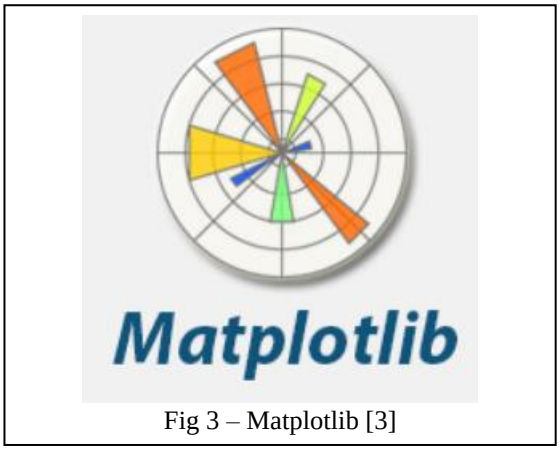


Fig 3 – Matplotlib [3]

Seaborn is a software which help in plotting and visualizing the data in different styles. It can be used to plot the graph both in 2D and 3D form.

1. Functions used:
    a. pairplot(): creates the data provided in the form of graph by using the columns as axis in grid format. In this the graphs present at diagonals are treated differently.
    b. fig.set_size_inches(): to set the size of the plot
    c. subplots_adjust(): to adjust the spacing between different subplots

D. Seaborn [4]

Seaborn is a software which help in plotting and visualizing the data in different styles. It can be used to plot the graph both in 2D and 3D form.

2. Functions used:
    d. pairplot(): creates the data provided in the form of graph by using the columns as axis in grid format. In this the graphs present at diagonals are treated differently.
    e. fig.set_size_inches(): to set the size of the plot
    f. subplots_adjust(): to adjust the spacing between different subplots

Fig 4 – Seaborn [4]

V. ANSWERS

A. The graph below shows how the salary of professors vary. The scatter plot of each color represents different class of professors. The average salary of the assistant professors is less compared to other professors and the frequency of them getting paid is also low. The form a cluster in graph. The more associate professors are there and they are paid good salary. The most scttered payment is given to full professors. They are high in number and their salary payment has high variation.

The average salary of professors is:

a. Full professor: 493.445306
b. Associate professor: 403.484065
c. Assistant professor: 344.652024

Number of professors are:

a. Full professor: 110407
b. Associate professor: 84039
c. Assistant professor: 79685

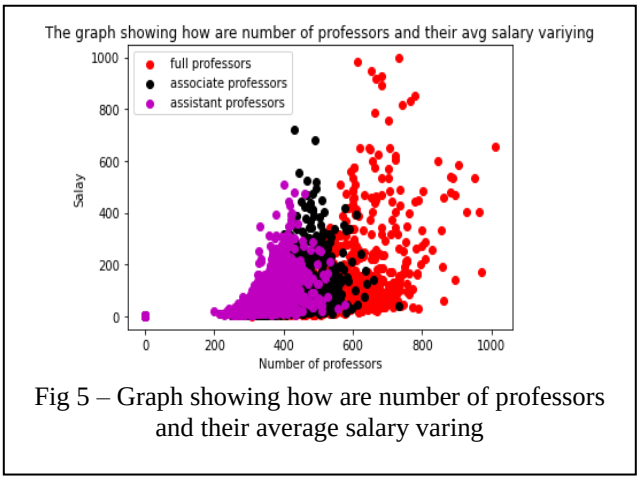


Fig 5 – Graph showing how are number of professors and their average salary varing

When a professor is chosen at random probability that a randomly chosen professor is an associate professor is 0.08477390498229148.

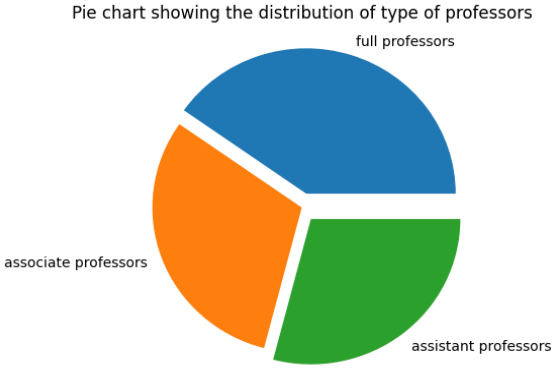


Fig – 6 Pie chart showing the distribution of type of professors

B. The states having highest universities are:

State	Number of Universities
PA	85
NY	81
TX	54
CA	54
OH	53

The top 3 states contains 18.9% of the total universities present in the country and rest 81.1% is present in the other states.

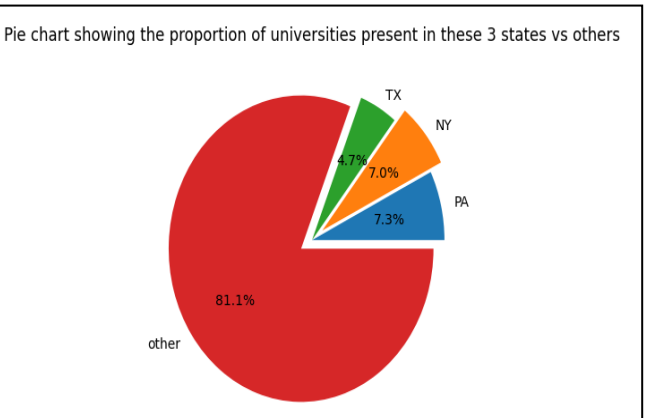


Fig 7 – Pie chart showing the proportion of universities present in these 3 states vs others

The state with highest number of universities IIB is having postal code PA.

C. The professors present in the universities are provided the compensation depending upon their type. The graph below shows the distribution of compensation and salary to different type of professors. Here all data is averaged.

Fig 8 – Graph showing the salary and compensation of professors

The category of professor having highest difference in the mean of salary and compensation are full professors.

D. The university constits of variaous staff members. The staff members here are the professors, the instructors and the faculty. The graph below shows the distribution of the staff in universities (since the mean values is considered, the distribution can be taken into consideration for all universities).

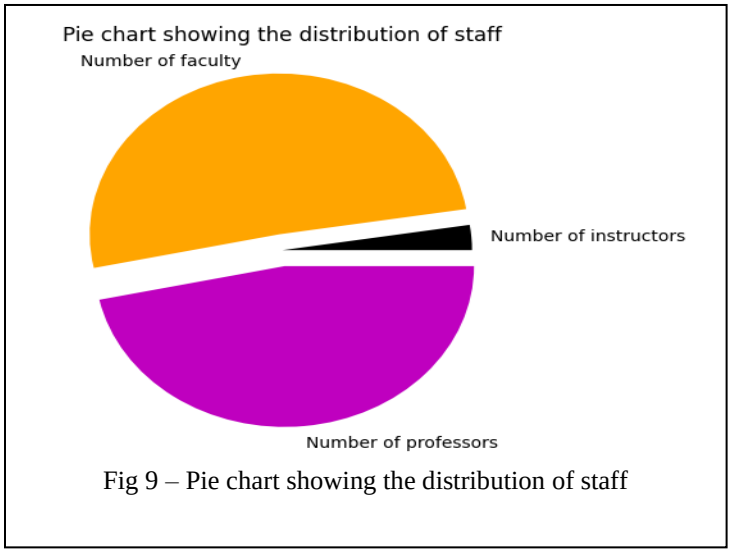


Fig 9 – Pie chart showing the distribution of staff

There is further division in type of professors shown in figure below.

Probability of choosing professor at random is 0.4664447858867489
Probability of choosing full professor at random is 0.4027526985273464.

Probaility of choosing full time professor given that he is a professor is 0.863452032723833

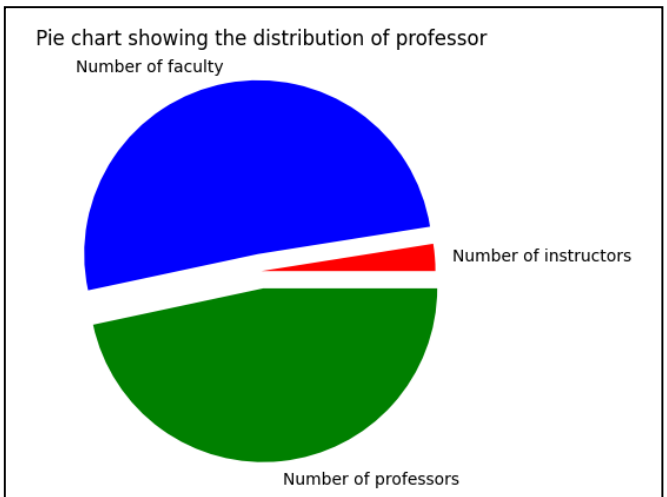


Fig 10 – Pie chart showing the distribution of professor

E. There are 3 types of professors present in the university.

Each has different salaries and are present in different numbers. The graph below consists of 16 plots out of which we get most of the information from 4 graphs.

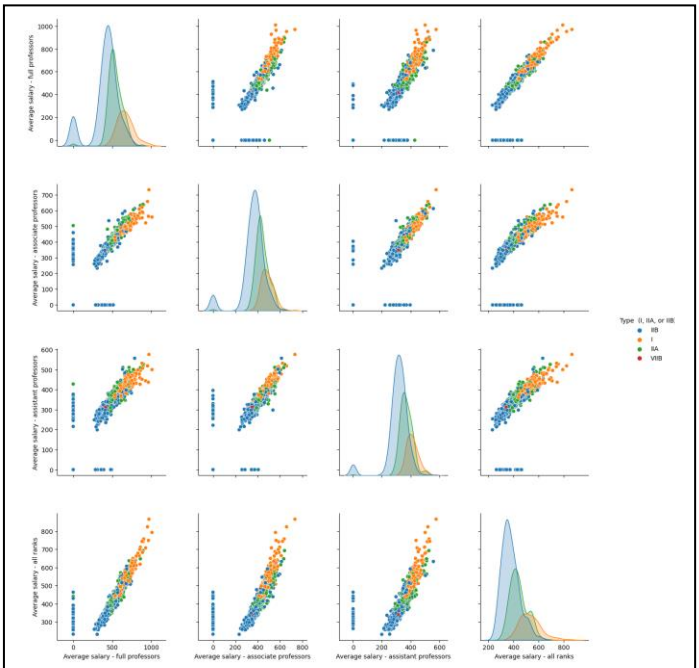


Fig 11 – Graph

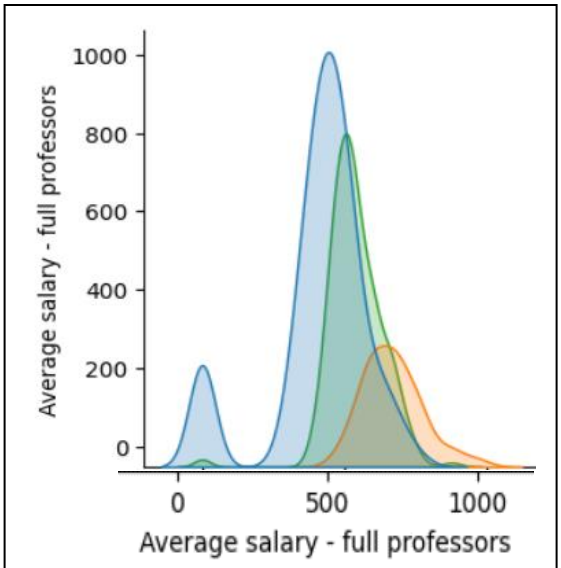


Fig 11.1
X axis – The number of professors
Y axis – the salary of professor

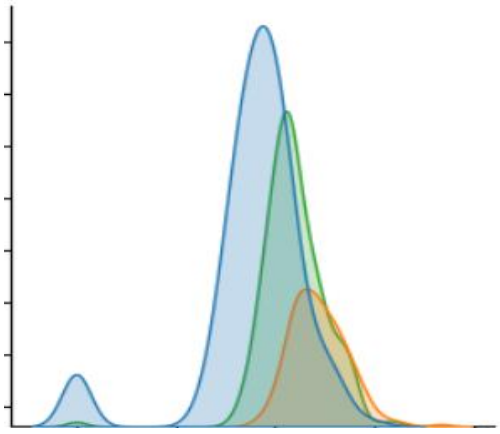


Fig 11.2
x- axis => Average salary- associate professors (Number of professors)
y-axis => Average salary- associate professors (the salary of professor)
scale: same as fig 11.1

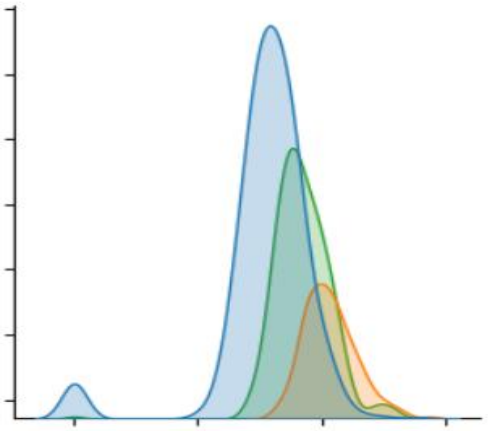
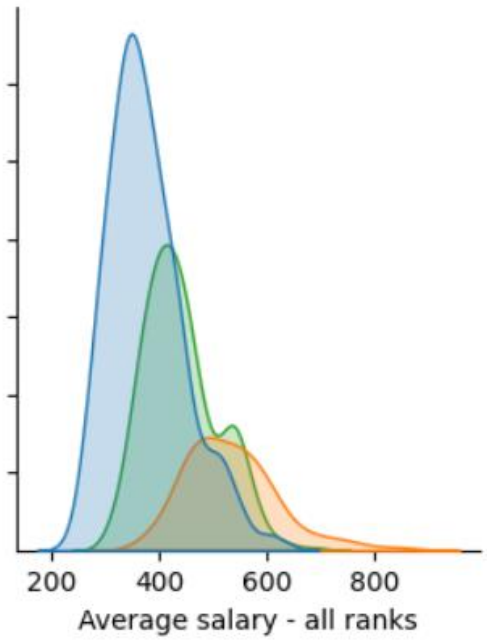


Fig 11.3
x- axis => Average salary- assistant professors
(Number of professors)
y-axis => Average salary- assistant professors
(the salary of professor)

scale: same as fig 11.1



X axis => Average salary – all ranks
(Number of professors)

y- axis => Average salary – all ranks
(the salary of professor)

scale: vertical same as 11.1
Fig 11.4

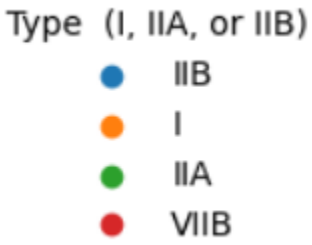


Fig 11.5

The figure 11.5 shows the legend or represents which color represents which type of universities.

Figure 11.1 shows the how full professors are distributed in different type of universities and what is their salary.
We are not marking it by point graph here because we are taking the average, measuring the lowest and highest salary.

Figure 11.2 shows the how associate professors are distributed in different type of universities and what is their salary.
We are not marking it by point graph here because we are taking the average, measuring the lowest and highest salary.

Figure 11.3 shows the how assistant professors are distributed in different type of universities and what is their salary.
We are not marking it by point graph here because we are taking the average, measuring the lowest and highest salary.

Figure 11.4 shows the how all ranks of professor are distributed in different type of universities and what is their salary.
We are not marking it by point graph here because we are taking the average, measuring the lowest and highest salary.

Here the y – axis shows the PMF.

The blue graph shows the type of university to be IIB which is always the highest and if we see all the graphs and their scale, we can easily interpret that the Full professors are paid the highest.

The professors who are paid highly are full professors of type IIB.

F. The Math SAT score, the Verbal SAT score and combined SAT score vary for the universities. If a person in genral wants to take admission in a private college then his scores required will be different and for the private college it would be different.

The graph below shows the avreage SAT score a person should score to get into the type of university he wants.

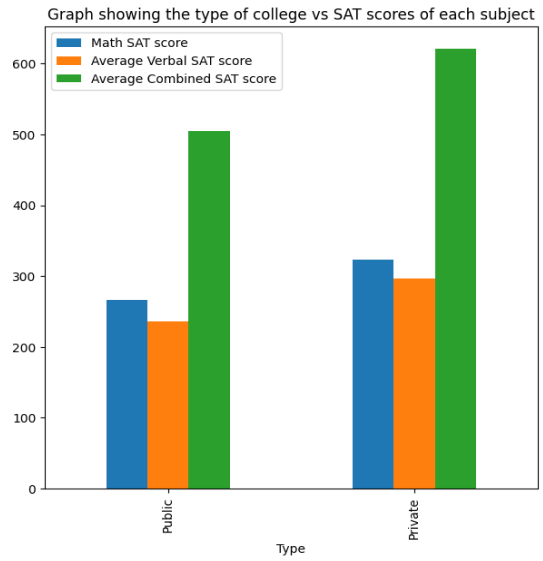


Fig 12 – Graph showing the type of college vs SAT scores of each subject.

If a person wants to go to a private college, he requires more marks. The average combined score of SAT of private university is higher than the public university.

G. The universites provide the application to students to enroll. They provide application to few students out of which some accpet the application and some did not.

The students who accepted the application has two choices, to enroll or not.

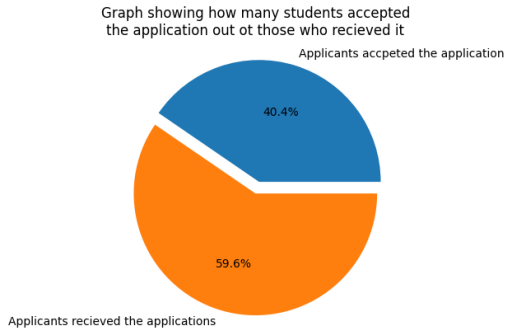


Fig 13.1

The graph 13.1 shows when application is sent, how many students accept it.

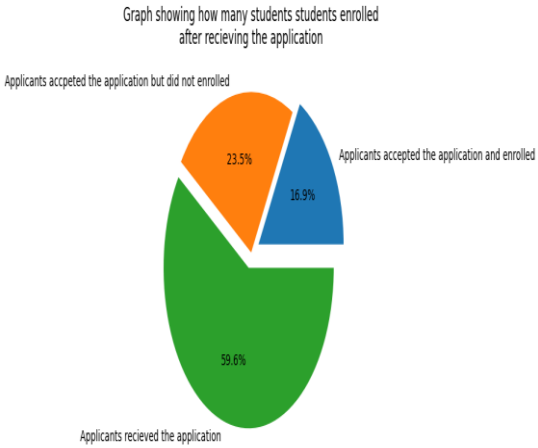


Fig 13.2

Blue – Applicants accepted the application and enrolled
Orange - Applicants accepted the application and did not enrolled
Green – Applicants did not accept the application

The graph 13.2 shows how many students accepted the application and enrolled, how many accepted the application and did not enroll and how many did not accept the application.

Probability of choosing a person who accepted application is: 0.6792038720818064
Probability of choosing a person who enrolled is: 0.2841086590301234
Probability of choosing a person who accepted application and is enrolled is 41829658326197533

So, when a person is chosen at random, the probability of choosing a person who accepted application and is enrolled is 0.41829658326197533

H. The students can join two type of universities public and private. The graph showing how students are distributed amongst the public/private universities is 14.1.

From the graph we can say that more students are present in the public universities than private.

There are two types of undergraduated present in the university, the parttime undergraduates and the full time undergraduates.

The graph 14.2 shows how the undergraduates are divided into parttime and fulltime in public university.

The graph 14.3 shows how the undergraduates are divided into parttime and fulltime in public university.

So from the graphs 14.2 and 14.3 we can say that there are less fulltime undergraduates than parttime undergraduates.

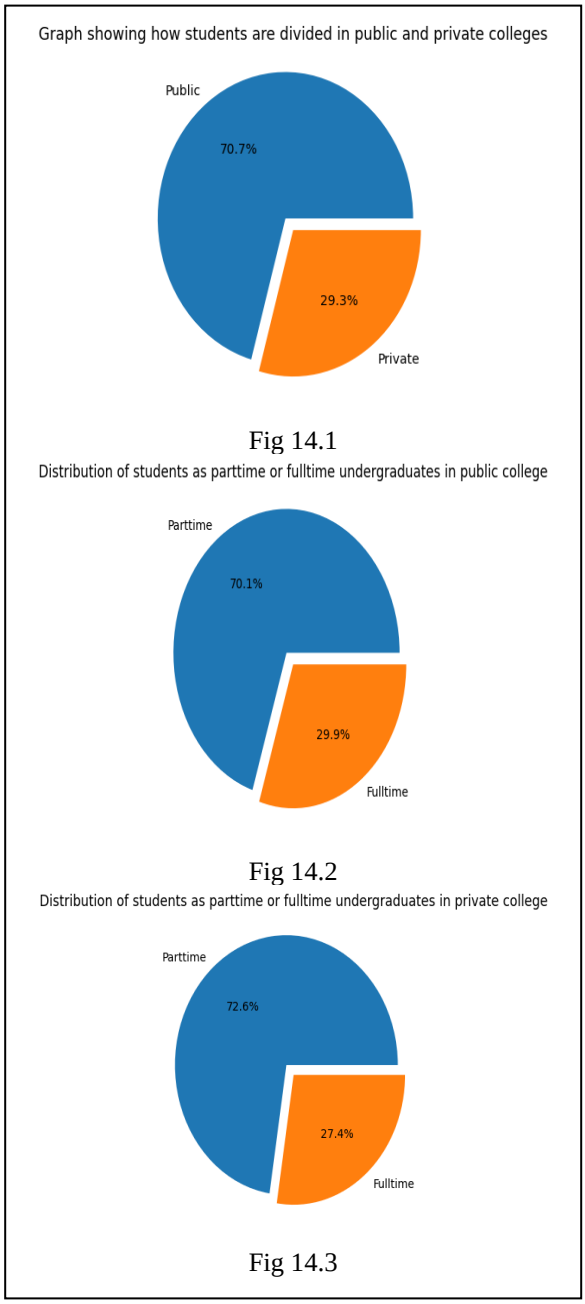


Fig 14.1

Fig 14.2

Fig 14.3

Total number of undergraduates present in all universities are 6170311
Number of parttime undergraduates present in all universities are 1373539
Probability of choosing a parttime undergraduate at random is 0.22260450081041297.

I. The ratio of student/faculty tells us that how many students are alloted to 1 faculty. If this ratio is less means the faculties presnt in university is more.

The graph 15.1 shows the distribution of teacher available for students in public and private universities.

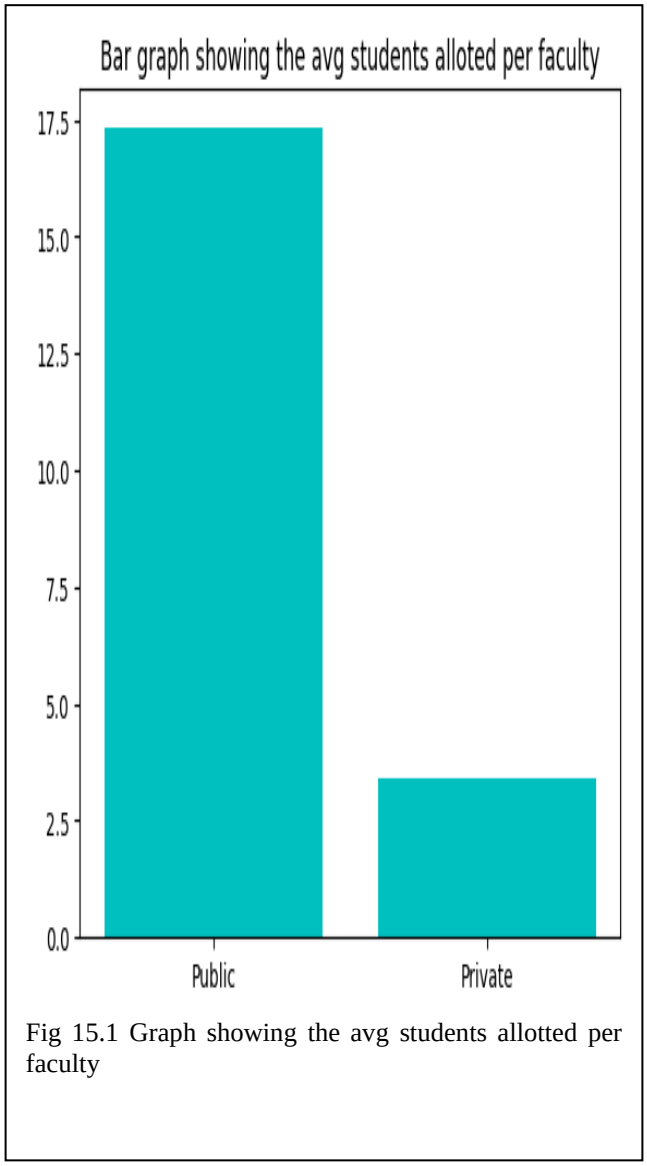


Fig 15.1 Graph showing the avg students allotted per faculty

From the graph we can see that in public universities more students are present for 1 faculty. The number of students alloted to faculty in public college is almosrt 6 times of students alloted to faculty in private college.

From this graph we can also interpret that the public universities have less faculties.

The avreage number of students a faculty teaches in all universities is 15.

J. Depending upon the state in which the university is present, the in state tution fee and out of the state tution fee varies.

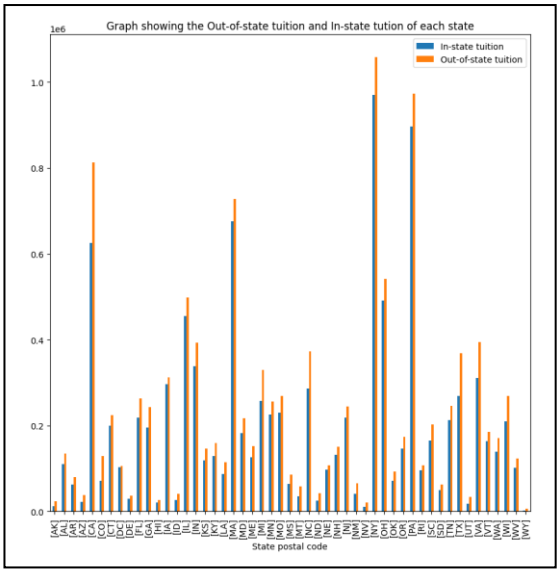


Fig 16 – Graph showing in state and out of the state tuition fee

The highest in state tution fee is of state having postal code NY.

The state having postal code CA has the highest variance in the in state and out of the state tution fee.

VI. OBSERVATION

A. The average salary of professors vary depending upon the type.

The full professors have average salary 493.445306
The associate professors have average salary 403.484065
The assistant professors have average salary 344.652024

The type of professors present in maximum number is the full professor.

When a professor is chosen at random probability that a randomly chosen professor is an associate professor is 0.08477390498229148.

B. The states having maximum number or funiversities are PA,NY,TX,CA,OH.

The top 3 states contain 18.9 percent of the total universities present in the country.
The state having maximum number of IIB type university is PA.

C. The professors are given compensation and salary acc. To their type. The professors who have maximum difference in their compensation and salary arre the full professorss.

D. The unviresity contains professors,faculty members and instructors as their staff. Out of which only professors are further divided.

E. The full professors are present in more numbers and they have the maximum probability to get a high pay.

F. The avg MATH SAT score, Verbal SAT score and combined SAT score vary according to the type of univerity you want to go.

The private requires more score as compared to the public.

G. Only 16.9% percent of the students who gor application enrolled. 59.6% students did not accpet the application.

Probability of choosing a person who accepted application and is enrolled is: 0.41829658326197533

H. 70.7% students are enrolled in public university and 29.3% are enrolled in private university.

In the 70.7% students 70.1% are parttime undregraduates and 29.1% are fulltime undergraduates.

In the 29.3% students 72.6% are parttime undregraduates and 27.4% are fulltime undergraduates.

I. Public colleges have less faculy as compared to private

The faculty in public teaches 17 students while faculty in private teaches 3 students.

The avg number of students the faculty teaches is 15.

J. The highest in state tution fee is of state having postal code NY.

The state having postal code CA has the highest variance in the in state and out of the state tution fee.

VII. UNANSWERED QUESTION

The number of parttime graduates who donated money to the university is unanswered. The question remains unanswered because we just know how many students graduated and how many of them donate money, we don't know the precent in which the parttime undergraduates graduate. So, we can't say who are donating the money and in what percentage.

VIII. ACKNOWLEDGEMENT

I would like to thank to prof. Shanmuga for providing me with the data to do the analysis. The links provided were very helpful.

IX. REFRENCES

[1] Google colab. Google. Accessed March 30, 2023. https://colab.research.google.com/.

[2] Pandas. “Python Data Analysis Library — Pandas: Python Data Analysis Library.” *Pydata.org*, 2018, pandas.pydata.org/.

[3] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007

[4] seaborn. “Seaborn: Statistical Data Visualization — Seaborn 0.9.0 Documentation.” *Pydata.org*, 2012, seaborn.pydata.org/.

[5] Python. “Welcome to Python.org.” *Python.org*, Python.org, 29 May 2019, www.python.org/.

[6] Data: “Index of /Datasets/Colleges.” *Lib.stat.cmu.edu*, lib.stat.cmu.edu/datasets/colleges/.

[7] “US News Education | Best Colleges | Best Graduate Schools | Online Schools.” *@USNews*, 2014, www.usnews.com/education.