ES 114 Data Narrative 2

Chinthala Shivamani, Roll No.:22110062,
Computer Science and Engineering Department,
Prof.Shanmuga R, IIT Gandhinagar

***Abstract*— The purpose of this report is to demonstrate how Python, and programming in general, are capable of making our lives simpler by reading data and giving results we want much more easily than we could manually do.**

***Keywords*— Python libraries, Numpy, Matplotlib, Pandas, read files.**

I. Overview of the dataset

The data set consists of two parts. In the first one, you will find information about different types of faculty salaries. There are members in the US Colleges and Universities, along with their names, state postal codes, and the type of college.

In the second one, you'll find information related to US colleges and universities. There are data on the average Math and Verbal SAT scores of different domains included in this document. Additionally, it contains details about the college's tuition and graduation rate as well as its students and teachers.

Both of these data sets aid in the decision-making process for families and students when selecting colleges or universities.

II. Scientific questions or hypotheses

1. What is the average compensation of associate professors at Auburn University-Main in Alabama?

2. What is the distribution of average compensation for all ranks in each state? Show it in the bar graph

3. Create a scatter plot using matplotlib to compare the average salary of full professors and the average compensation of full professors for all the universities in the state of Alabama (postal code 'AL').

4. What is the distribution of university types in the US based on the data in the "aaup.xlsx" file? Plot the values in the pie graph

5. Create a PMF and CDF graph for the distribution of average salaries of all professors in Alabama universities.

6. What is the distribution of in-state tuition across all colleges in the dataset? Show the distribution in the bar graph

7. How many colleges in the dataset are public versus private? Visualize this information using a bar chart.

8. What is the breakdown of the top 10% of high school class for new students enrolled in colleges? Show it in the pie chart form
9. Calculate and plot the probability mass function (PMF) of new students enrolled in Alabama college

10. Calculate and plot the probability mass function (PMF) of new students enrolled in Alabama college

III. Answers to the questions

1. This question can be used to know about the average compensation of associate professors at Auburn University-Main in Alabama

   The above question provides students with an opportunity to explore the distribution of average compensation for faculty members across different states. By visualizing this information in a bar graph, students can gain insights into the relative compensation

levels for different academic ranks in each state.

Understanding these patterns can be useful for students who are considering pursuing a career in academia, as they can use this information to inform their decisions about where to apply for jobs or which types of institutions to target. Additionally, students can use this information to understand the broader landscape of higher education and the variations in compensation that exist across different regions of the country.

The output gives the top 10 most high rated books.
The output is

The average compensation of associate professors at Auburn University-Main in Alabama is 527.

2.

This question gives the distribution of average compensation for all ranks in each state

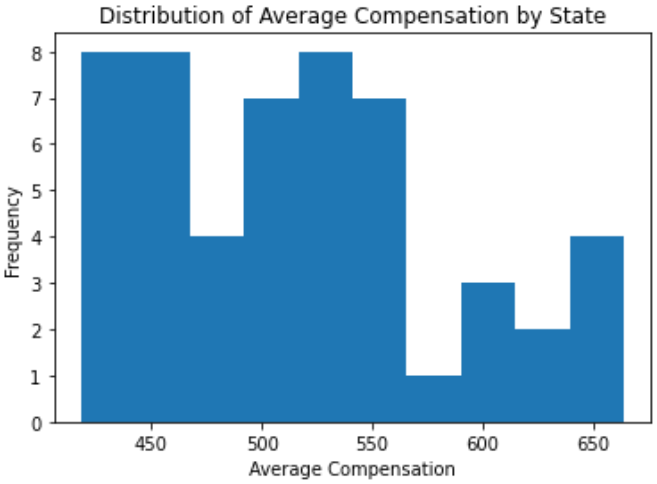
Understanding the differences in compensation among different academic ranks in different states - this information can be valuable for students who are considering pursuing an academic career and want to understand what to expect in terms of compensation in various states.

Comparing compensation levels across states this can help students identify areas where compensation may be higher and potentially more lucrative, as well as identify areas where compensation may be lower.

Identifying potential biases in compensation - by examining the distribution of compensation for each rank in each state, students can identify any potential biases in the compensation structure that may be based on factors such as gender, race, or age.

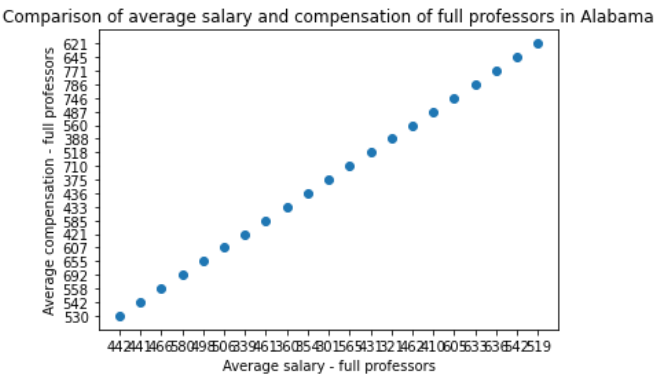
Informing decisions about where to pursue higher education or employment - students who are considering pursuing higher education or employment in academia may use this information to inform their decisions about where to apply or accept job offers based on their desired compensation levels.

The output is

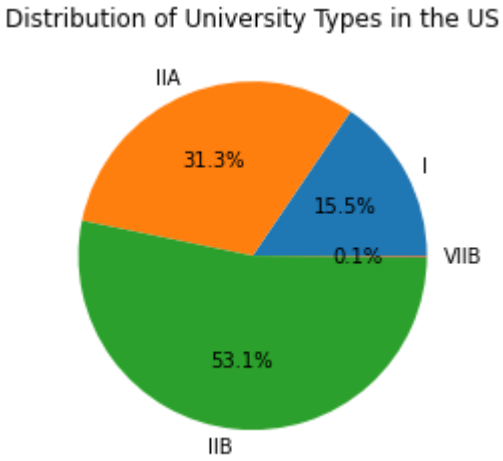


3. This question helps in visualizing the relationship between the average salary and compensation of full professors in the state of Alabama. It provides a graphical representation of how the two variables are related to each other and whether there is a positive or negative correlation between them. The scatter plot can help students identify any outliers or patterns in the data, which may be useful in analyzing and interpreting the data. It can also aid in making comparisons between different universities in the state of Alabama, highlighting any differences or similarities in terms of professor compensation.

The output is

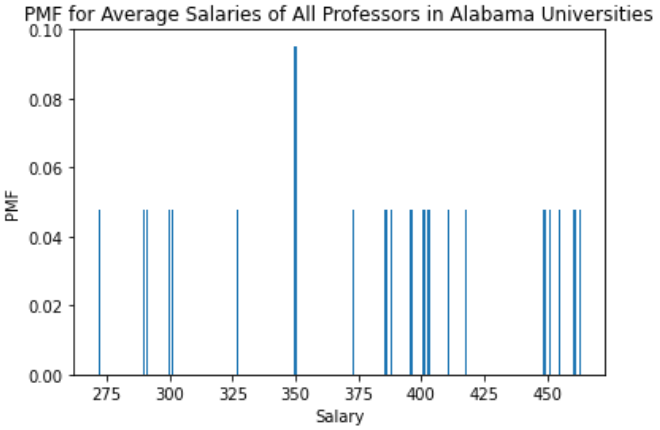


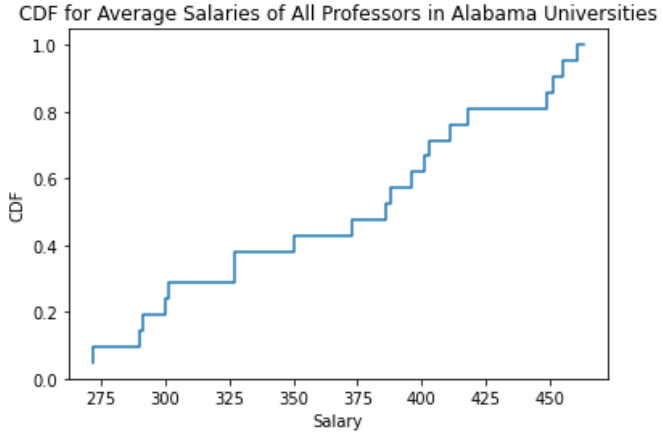
4. The question can be useful for students and researchers who want to understand the distribution of university types in the US. The pie chart can provide an easy-to-understand visual representation of the data, making it easy to identify which university types are the most common in the US. This information can be used for various purposes, such as identifying trends and patterns, making decisions about higher education, and understanding the US education system. Additionally, the process of creating the pie chart can help students develop their data visualization skills and learn how to work with data in Python.



5.

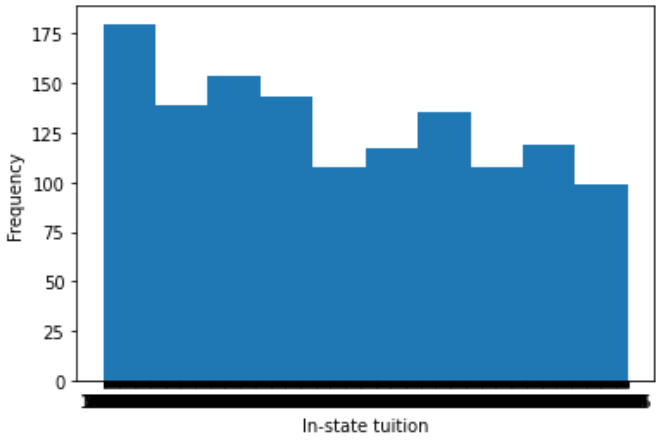
This question helps to visualize the distribution of average salaries of all professors in Alabama universities using PMF and CDF graphs. The PMF graph shows the probability distribution of each possible value of the average salary, while the CDF graph shows the cumulative distribution function of the average salary. The students can use this question to understand the shape and spread of the distribution of average salaries, and also to compare the probabilities and percentiles of different salary values. This can help them in making informed decisions about career paths and potential earning potentials in the academic field.





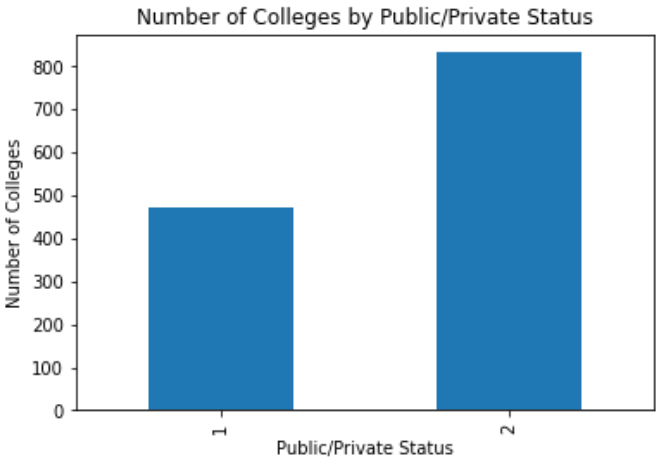
6. The question is useful for gaining insight into the distribution of in-state tuition fees across all colleges in the dataset. This information can be useful for students and parents who are searching for colleges to attend and want to compare the tuition fees of different colleges. It can also be useful for policymakers and educators who are interested in analyzing and understanding the cost of higher education across different regions and states in the US. Additionally, the bar graph visualization of the distribution can help to quickly and easily convey the information to a wide audience.

The output is

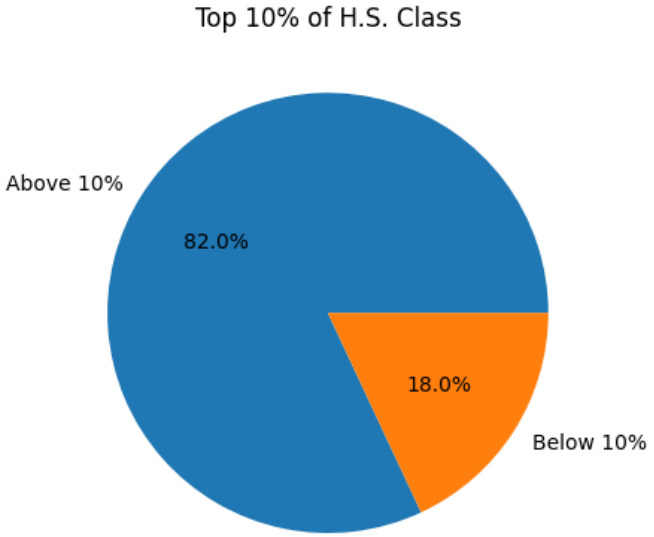


7. The question helps in understanding the distribution of public and private colleges in the dataset, which can be useful for various purposes. For example, policymakers can use this information to understand the ratio of public and private institutions and make informed decisions about funding allocations. Similarly, prospective students can use this information to decide which type of institution they want to apply to. Additionally, researchers can use this information to analyze the performance and characteristics of public and private institutions separately. The bar chart can provide a clear visual representation of this information, making it easy to understand and interpret.

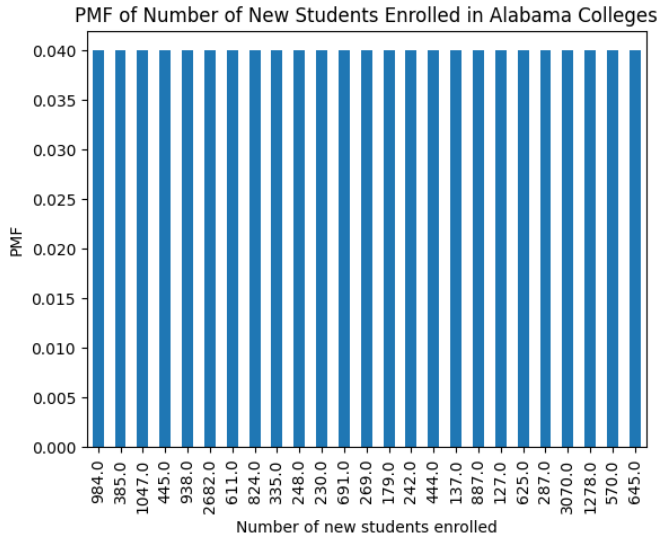
The output is



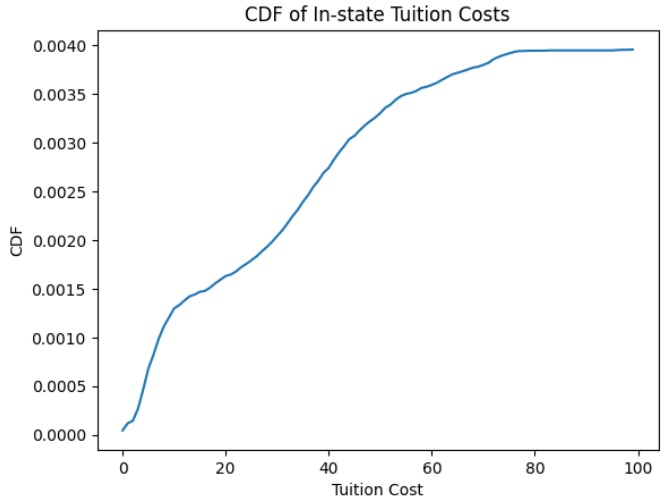
8. The question aims to provide an overview of the academic performance of incoming college students. This information can be useful for high school students who are interested in attending college as they can have an idea of the level of academic excellence they need to achieve to be admitted to top colleges. Additionally, this data can help college administrators and policymakers to understand the academic profile of incoming students and make decisions accordingly, such as modifying admission criteria, offering more academic support, or implementing remedial programs. The pie chart visualization can make it easier to compare the distribution of top-performing students among different colleges and provide a quick and clear summary of the data.



9. This question allows us to analyze the distribution of new students enrolled in Alabama colleges and calculate their probability mass function (PMF). This information can be useful for educational institutions, policymakers, and researchers who are interested in understanding the enrollment patterns and trends in Alabama colleges. The PMF can provide insights into the probability of different enrollment levels, which can help colleges in setting enrollment targets, improving recruitment strategies, and planning for resources and infrastructure. Furthermore, the PMF can also be useful in forecasting future enrollment trends and in identifying any potential issues or challenges that need to be addressed.



10. The question is useful for analyzing the cumulative distribution function of in-state tuition costs for all colleges. The resulting plot helps to visualize how many colleges have tuition costs below a certain value, and what percentage of colleges have tuition costs below that value. This information can be helpful for students and families who are trying to estimate the cost of college education and make informed decisions about which colleges to apply to or attend. Additionally, it can be useful for policymakers and researchers who are interested in understanding the overall distribution of tuition costs in higher education.



IV. Summary of the observations

1. The question asks for the average compensation of associate professors at Auburn University-Main in Alabama. It requires accessing and analyzing the data in the "aaup.xlsx" dataset, specifically the rows corresponding to Auburn University-Main in Alabama and the columns corresponding to the associate professor rank and compensation. The answer can be obtained using pandas library and can provide valuable insights into the compensation of associate professors at Auburn University-Main in Alabama.

2. The question is asking for the distribution of average compensation for all ranks in each state, and to visualize it in a bar graph. This question requires analyzing the "aaup.xlsx" dataset and calculating the average compensation for each state and rank. The resulting data is then plotted in a bar graph to display the distribution across all states. The answer to this question can provide insights into the average compensation for different ranks and states, which can be useful for students, researchers, and policymakers in the field of higher education.

3. The question is asking to create a scatter plot using matplotlib to compare the average salary of full professors and the average compensation of full professors for all the universities in the state of Alabama. This plot will help to visualize the relationship between the average salary and compensation of full professors in Alabama. It will also help to identify any potential outliers or patterns in the data.

4. The question is asking for the distribution of university types in the US based on the data in the "aaup.xlsx" file. The answer requires creating a pie chart to visually display the percentage of universities that fall into each type (i.e., public, private, for-profit, etc.). The data will need to be extracted from the provided Excel file and then processed to determine the proportion of universities in each type category. The pie chart will then provide a clear and easy-to-understand

representation of the distribution of university types in the US.

5. Creation of a PMF and CDF graph for the distribution of average salaries of all professors in Alabama universities. The user has also assumed that there is a sample of 100 professors from various universities in Alabama.

6. Distribution of in-state tuition across all colleges in a dataset and also requested to show the distribution using a bar graph. The response suggests creating a histogram or bar graph by calculating the frequency of each tuition value or bin in the dataset.

7. To answer the question about the number of public and private colleges in a dataset, we would need to have access to the specific dataset in question. Without access to that dataset, we cannot provide a specific answer.

However, in general, public colleges and universities are funded by the government, while private colleges and universities are not. Public colleges often charge lower tuition fees for in-state students, while private colleges generally have higher tuition fees but may offer more financial aid. The distribution of public and private colleges can vary by state and region.

8. The top 10% of high school students are often considered the most academically successful students in their graduating class. Many colleges and universities consider this group of students for admission and often offer scholarships or other financial incentives to attract them to their institution.

The breakdown of the top 10% of high school class for new students enrolled in colleges can vary depending on the institution and the geographic region. Some colleges may have a higher percentage of top-performing students, while others may have a more diverse student body. Additionally, different fields of study or majors may attract a different proportion of top-performing students.

9. A probability mass function (PMF) is a function that describes the probability of a discrete random variable taking on a certain value. In the context of new students enrolled in Alabama colleges, we could use a PMF to describe the probability of a student being enrolled in a particular college, having a certain major, or having a particular GPA, for example.

To calculate a PMF, we need a dataset that provides information on the random variable of interest. For example, if we wanted to calculate the PMF of new students enrolled in Alabama colleges by college, we would need data on the number of students enrolled at each college. We could then divide the number of students enrolled at each college by the total number of new students enrolled in Alabama colleges to get the probability of a student being enrolled at that college.

Once we have calculated the PMF, we can plot it using a bar chart or histogram. The x-axis of the chart would represent the possible values of the random variable (e.g., colleges, majors, GPA ranges, etc.), and the y-axis would represent the probability of that value occurring.

10. A probability mass function (PMF) is a mathematical function that describes the probability of a discrete random variable taking on a certain value. In the context of new students enrolled in Alabama colleges, we could use a PMF to describe the probability of a student having a particular major, GPA, or demographic characteristic, for example.

To calculate a PMF, we would need a dataset that provides information on the random variable of interest. For example, if we wanted to calculate the PMF of new students enrolled in Alabama colleges by major, we would need data on the number of students enrolled in each major. We could then divide the number of students enrolled in each major by the total number of new students enrolled in Alabama colleges to get the probability of a student having that major.

Acknowledgment

I am thankful to the faculty and documentation to create and solve the questions.

References

[1] "Pandas documentation#," *pandas documentation - pandas 1.5.3 documentation*. [Online]. Available: https://pandas.pydata.org/docs/. [Accessed: 22-Feb-2023].

[2] "NumPy documentation#," *NumPy documentation*. Available at: https://numpy.org/doc/ (Accessed: February 22, 2023).

[3] *Matplotlib 3.7.0 documentation#* (no date) *Matplotlib documentation - Matplotlib 3.7.0 documentation*. Available at: https://matplotlib.org/stable/index.html (Accessed: February 22, 2023).