Data Narrative II

Name: Saloni Sunil Shinde

Roll No: 22110242

Discipline: CSE

I. OVERVIEW OF THE DATASET

The AAUP dataset, contains information on faculty salaries for 1,161 colleges and universities in the United States during the academic year of 1994-1995. The American Association of University Professors (AAUP) collected the data through a survey of institutions. It has information about faculties of ranks with their average salaries and compensations. The dataset is presented in Comma-Separated Values (CSV) format, which facilitates a wide range of data analysis tools. The AAUP2 dataset, has same the data arranged in fixed columns, with two data lines for each school.

The US News dataset, presents data on different characteristics of colleges and universities in the United States, as reported in the 1995 edition of the U.S. News & World Report annual college rankings. The data was collected from a range of sources, including the institutions themselves and the U.S. Department of Education. USNEWS and USNEWS3 are presented in Comma-Separated Values (CSV) format and column format.

These datasets are valuable resources for researchers interested in various aspects of higher education, as it provides comprehensive information on colleges and universities in the mid-1990s.

II. SCIENTIFIC QUESTIONS / HYPOTHESES

1. Is there any relation between type of institution and average salary of professors?
2. Find if there is any correlation between the number of full professors and the number of associate professors at each institution.
3. Find if there is any correlation between total number of institutions and state of the institution.
4. Is there a significant correlation between the total number of faculty members and the average salary of professors of all ranks at each institution?
5. Is there a correlation between the number of instructors at an institution and the average compensation paid?
6. Is there any relation between the average ACT score and the graduation rate?
7. Find the probability of getting a good average combined sat score being in private and public institution.
8. Find if there is any correlation between the number of applicants received and the number of applicants accepted at each institution.
9. Is there a difference in the number of part-time undergraduates between public and private institutions?
10. How strong is the correlation between the percentage of faculty members with PhDs and the graduation rate at each institution? Can we infer that a higher percentage of faculty members with PhDs leads to a higher graduation rate?

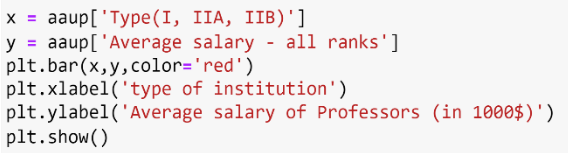
III. DETAILS OF LIBRARIES AND FUNCTIONS

- pandas [Library]: Pandas is a Python library used for working with data sets. It has functions for analysing, cleaning, exploring, and manipulating data.[1]
- read_csv: Read a comma-separated values (csv) file into DataFrame.[1]
- value_counts: Return a Series containing counts of unique values.[1]
- head: Return the first *n* rows.[1]
- replace: Returns a copy of the string with a specified substring replaced specified number of times.[4]
- astype: Cast a pandas object to a specified dtype dtype.[1]
- len: Finds length.[4]
- groupby: Group DataFrame using a mapper or by a Series of columns.[1]
- matplotlib [Library]: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.[3]
- pyplot: mainly intended for interactive plots and simple cases of programmatic plot generation.[3]
- plot: Plot y versus x as lines and/or markers.[3]
- hist: Compute and plot a histogram.[3]
- bar: Make a bar plot.[3]
- xlabel: Labels the x-axis.
- ylabel: Labels the y-axis.
- show: Display all open figures.[3]
- xticks: Get or set the current tick locations and labels of the x-axis.[3]
- scatter: A scatter plot of *y* vs. *x* with varying marker size and/or color.[3]
- numpy [Library]: NumPy is a Python library used for working with arrays.[2]
- corrcoef: Return Pearson product-moment correlation coefficients.[2]
- mean: Compute the arithmetic mean along the specified axis.[2]
- std: Compute the standard deviation along the specified axis.[2]

IV. ANSWERS TO THE QUESTIONS

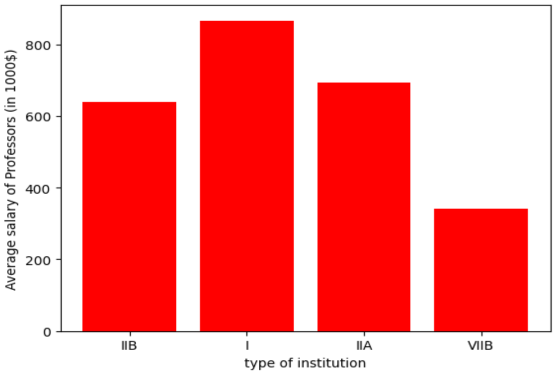
1. To answer this question, we can plot a bar graph to observe distribution of average salaries of faculties and their type of institutions.

   This is the python code for it:



```
x = aaup['Type(I, IIA, IIB)']
y = aaup['Average salary - all ranks']
plt.bar(x,y,color='red')
plt.xlabel('type of institution')
plt.ylabel('Average salary of Professors (in 1000$)')
plt.show()
```

This is the bar graph:

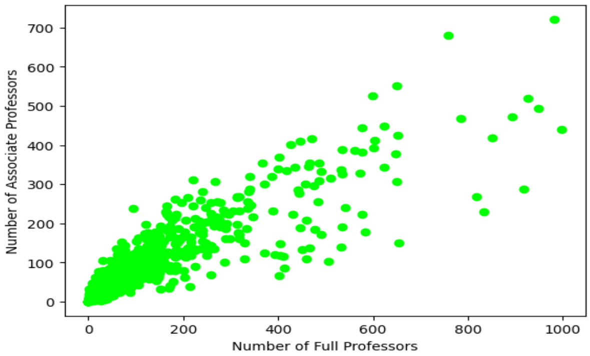


From the above information, we can conclude that there is a relationship between faculty salaries and the type of institution they work for, and that salaries tend to be higher at type I institutions.

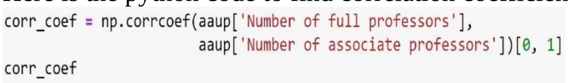
2. To answer this question, we can use a scatter plot to visualize the relationship between the number of full professors and the number of associate professors at each college, and calculate the correlation coefficient to determine if there is a significant correlation.

   First, we can use the "Number of full professors" and "Number of associate professors" columns from the dataset as our random variables. We can then create a scatter plot with "Number of associate professors" on the y-axis and "Number of full professors" on the x-axis to visualize the relationship between these two variables.

   This is the scatter plot:



   Here is the python code to find correlation coefficient:

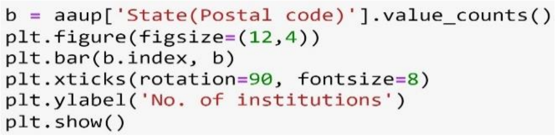


```
corr_coef = np.corrcoef(aaup['Number of full professors'],
                        aaup['Number of associate professors'])[0, 1]
corr_coef
```

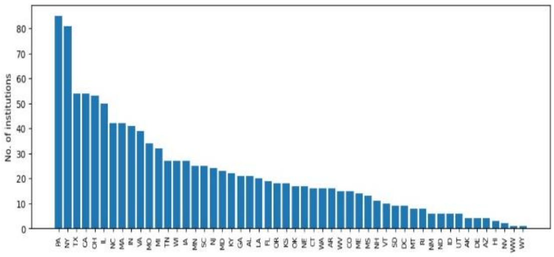
   The scatter plot shows that there is a positive correlation between the number of full professors and the number of associate professors at each college, as we can see from the trend of the data points. The correlation coefficient of 0.90 further confirms this strong correlation.

3. To answer this question, let’s plot the distribution of states and number of institutions in that state. For this we will use one column: ‘State (Postal Code)’ and value_counts function.

   Here is the python code and plot of required distribution:



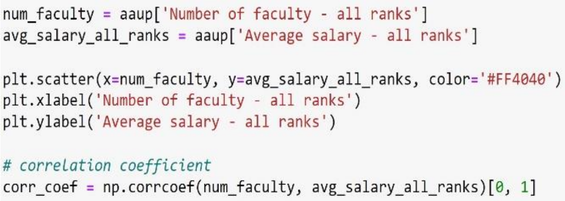
```
b = aaup['State(Postal code)'].value_counts()
plt.figure(figsize=(12,4))
plt.bar(b.index, b)
plt.xticks(rotation=90, fontsize=8)
plt.ylabel('No. of institutions')
plt.show()
```



   By plot we can see, most of the institutes are in Pennsylvania (PA) and New York (NY).

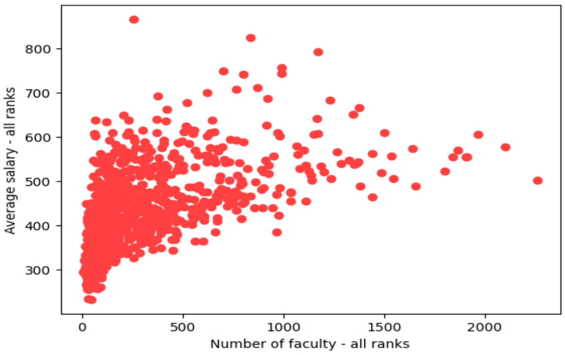
4. To answer this question, we can plot a scatter plot between the total number of faculty members and the average salary of professors of all ranks at each institution. We can then calculate the correlation coefficient to determine if there is a significant correlation between the two variables.

   Here is the Python code to plot the scatter plot and calculate the correlation coefficient:



```
num_faculty = aaup['Number of faculty - all ranks']
avg_salary_all_ranks = aaup['Average salary - all ranks']

plt.scatter(x=num_faculty, y=avg_salary_all_ranks, color='#FF4040')
plt.xlabel('Number of faculty - all ranks')
plt.ylabel('Average salary - all ranks')

# correlation coefficient
corr_coef = np.corrcoef(num_faculty, avg_salary_all_ranks)[0, 1]
```

   This is the scatter plot:



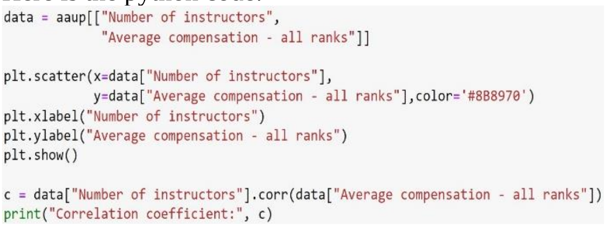
The scatter plot shows the relationship between the total number of faculty members and the average salary of professors of all ranks at each institution. The x-axis shows the number of faculty members, while the y-axis shows the average salary of professors of all ranks. Each point in the plot represents an institution.

The correlation coefficient between the two variables is 0.755, which indicates a strong positive correlation between the total number of faculty members and the average salary of professors of all ranks. Therefore, we can conclude that there is a significant correlation between the two variables.

5. To answer the question of whether there is a correlation between the number of instructors at an institution and the average compensation paid, we can use the "Number of instructors" and "Average compensation - all ranks" columns from the provided dataset.

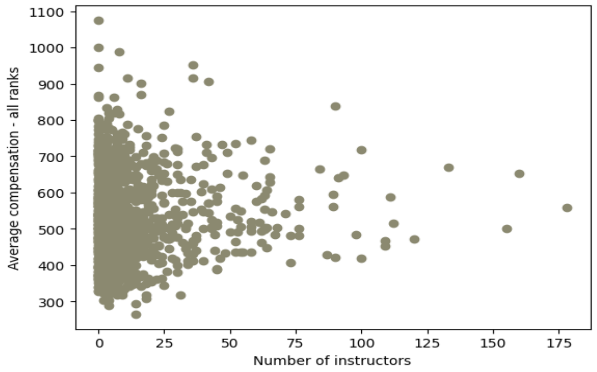
   First, we can extract these columns from the dataset and create a scatter plot with "Number of instructors" on the x-axis and "Average compensation - all ranks" on the y-axis. This will allow us to visualize the relationship between these two variables.

   Here is the python code:



```
data = aaup[["Number of instructors",
             "Average compensation - all ranks"]]

plt.scatter(x=data["Number of instructors"],
            y=data["Average compensation - all ranks"],color='#8B8970')
plt.xlabel("Number of instructors")
plt.ylabel("Average compensation - all ranks")
plt.show()

c = data["Number of instructors"].corr(data["Average compensation - all ranks"])
print("Correlation coefficient:", c)
```

   This is the scatter plot:



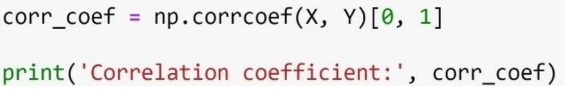
   A correlation coefficient of 0.066 indicates a very weak positive correlation between the number of instructors at an institution and the average compensation paid. This suggests that there is little to no relationship between these two variables in the given dataset.

6. To investigate whether there is a relationship between the average ACT score and the graduation rate, we can use random variables. We can define two random variables:
   X = Average ACT score
   Y = Graduation rate

   We can then compute the sample mean and sample standard deviation for X and Y from the college dataset. The sample mean for X is 12.13 and the sample standard deviation is 11.18. The sample mean for Y is 55.86 and the sample standard deviation is 24.17.

   We can then compute the correlation coefficient between X and Y to determine if there is a relationship between the two random variables and visualize the relationship between X and Y using a scatter plot.

   Here is the python code:

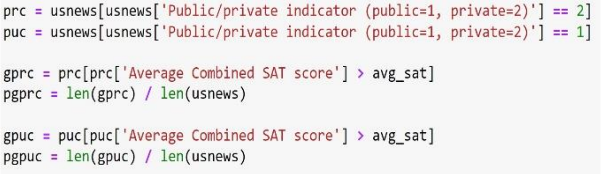


```
corr_coef = np.corrcoef(X, Y)[0, 1]

print('Correlation coefficient:', corr_coef)
```

   The correlation coefficient is -0.02, which indicates that there is a very weak negative relationship between the two variables, X and Y. A correlation coefficient of -0.02 suggests that as the values of X increase, the values of Y tend to decrease very slightly, or vice versa. In this case, we can say that there is no meaningful relationship between the average ACT score and the graduation rate.

7. To answer this question, first we will find the mean of average combined SAT score. Let us assume scores greater than the mean of average combined SAT score as good score. So, we will find probability of getting a good score being in a private college and a public college.

   Here is the python code for it:

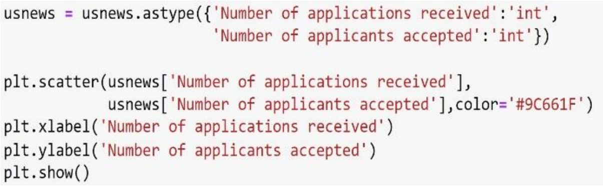


```
prc = usnews[usnews['Public/private indicator (public=1, private=2)'] == 2]
puc = usnews[usnews['Public/private indicator (public=1, private=2)'] == 1]

gprc = prc[prc['Average Combined SAT score'] > avg_sat]
pgprc = len(gprc) / len(usnews)

gpuc = puc[puc['Average Combined SAT score'] > avg_sat]
pgpuc = len(gpuc) / len(usnews)
```

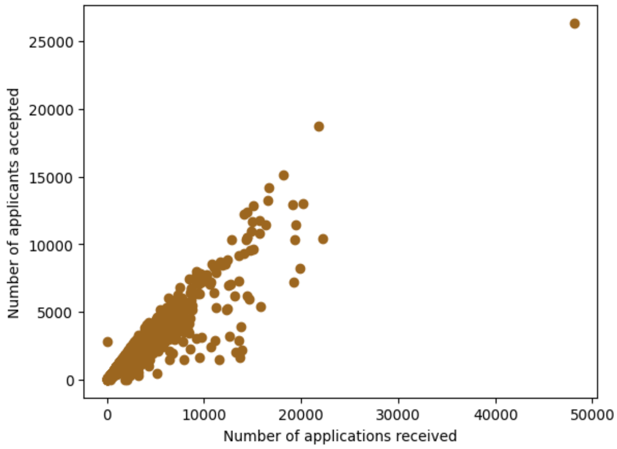
   Probability of getting good SAT score being in a private college is 0.41. Probability of getting good SAT score being in a public college is 0.19.

8. One way to explore the relationship between the number of applicants received and applicants accepted at each college is to create a scatter plot. By using the "Number of applicants received" and "Number of applicants accepted" columns from the dataset, we can visually examine if there is a correlation between these two by using a scatter plot. Additionally, we can calculate the correlation coefficient to determine the strength and direction of the relationship between these variables.

   Here is the scatter plot and its python code:



```
usnews = usnews.astype({'Number of applications received':'int',
                        'Number of applicants accepted':'int'})

plt.scatter(usnews['Number of applications received'],
            usnews['Number of applicants accepted'],color='#9C661F')
plt.xlabel('Number of applications received')
plt.ylabel('Number of applicants accepted')
plt.show()
```

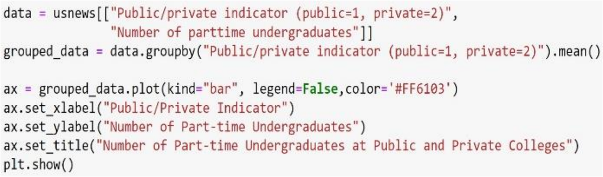


Based on the scatter plot, it can be observed that there is a positive association between the number of applicants received and the number of applicants accepted at each college, which is supported by the upward trend in the data points. The strong correlation is further supported by a high correlation coefficient of 0.93.

9. To answer this question, we can use a bar chart to compare the number of part-time undergraduates at public and private colleges. We can first create a subset of the dataset that includes only the "Public/private indicator" and "Number of parttime undergraduates" columns. We can then group the data by the public/private indicator and calculate the mean of the "Number of parttime undergraduates" column for each group.

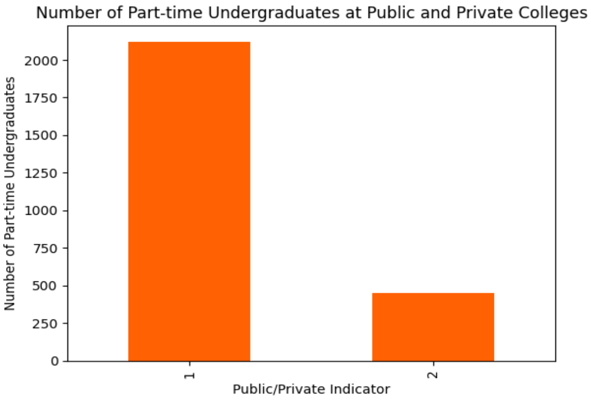
   Next, we can create a bar chart with "Public/Private Indicator" on the x-axis and "Number of Part-time Undergraduates" on the y-axis to visualize the difference between public and private colleges.

   Here is the code:



```
data = usnews[["Public/private indicator (public=1, private=2)",
               "Number of parttime undergraduates"]]
grouped_data = data.groupby("Public/private indicator (public=1, private=2)").mean()

ax = grouped_data.plot(kind="bar", legend=False,color='#FF6103')
ax.set_xlabel("Public/Private Indicator")
ax.set_ylabel("Number of Part-time Undergraduates")
ax.set_title("Number of Part-time Undergraduates at Public and Private Colleges")
plt.show()
```

   This is the bar graph:

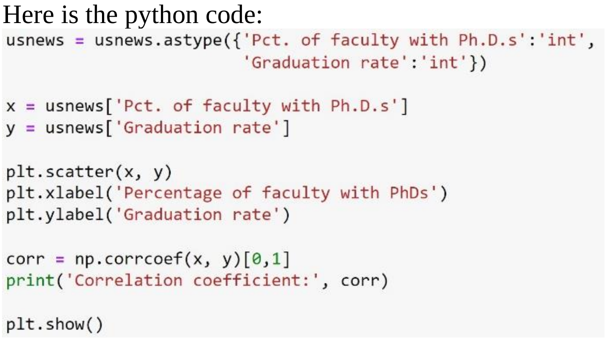


   In the graph:
   1 is for public college and 2 is for private college.

   The resulting bar graph shows that there is a significant difference in the number of part-time undergraduates between public and private colleges.

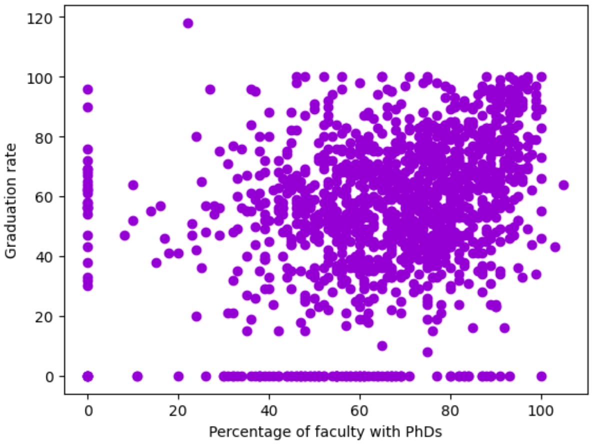
10. To answer this question, we can use a scatter plot to visualize the relationship between the percentage of faculty members with PhDs and the graduation rate at each college, and calculate the correlation coefficient to determine the strength of the correlation. We can use the "Pct. of faculty with Ph. D. s" and "Graduation rate" columns from the dataset as our random variables. We can then create a scatter plot with "Pct. of faculty with Ph. D. s" on the x-axis and "Graduation rate" on the y-axis to visualize the relationship between these two variables.

    Here is the python code:



```
usnews = usnews.astype({'Pct. of faculty with Ph.D.s':'int',
                        'Graduation rate':'int'})

x = usnews['Pct. of faculty with Ph.D.s']
y = usnews['Graduation rate']

plt.scatter(x, y)
plt.xlabel('Percentage of faculty with PhDs')
plt.ylabel('Graduation rate')

corr = np.corrcoef(x, y)[0,1]
print('Correlation coefficient:', corr)

plt.show()
```

    This is the scatter plot:



The correlation coefficient of 0.290 suggests a weak positive correlation between the graduation rate and the percentage of faculty members with PhDs at each college. It is essential to keep in mind that correlation does not necessarily indicate causation, so it is not possible to infer that a higher percentage of faculty members with PhDs leads to a higher graduation rate.

V. Summary of the Observations

In AAUP dataset, average salaries tend to be higher at type I colleges. There is a positive correlation between the number of full professors and the number of associate professors at each college. Most of the colleges are in Pennsylvania (PA) and New York (NY). There is a significant correlation between the total number of faculty members and the average salary of professors of all ranks at each college. There is no relation between the number of instructors at an college and the average compensation paid.

In USNEWS dataset, there is no relation between the average ACT score and the graduation rate. Probability of getting good SAT score being in a private college is 0.41. Probability of getting good SAT score being in a public college is 0.19. There is a positive association between the number of applicants received and the number of applicants accepted at each college. There is a significant difference in the number of part-time undergraduates between public and private colleges. It is not possible to infer that a higher percentage of faculty members with PhDs leads to a higher graduation rate of the university or college.

VI. References

[1] "Pandas Documentation — Pandas 1.0.1 Documentation." n.d. Pandas.pydata.org. https://pandas.pydata.org/docs/

[2] "NumPy Documentation." n.d. Numpy.org. https://numpy.org/doc/.

[3] "Matplotlib: Python Plotting — Matplotlib 3.3.4 Documentation." n.d. Matplotlib.org. https://matplotlib.org/stable/index.html.

[4] "3.7.4 Documentation." 2019. Python.org. 2019. https://docs.python.org/.

VII. Acknowledgements

This analysis of the AAUP and USNEWS dataset was only possible due to the hard work of its creators who collected and organized the data. I would like to thank them and many researchers who have contributed to this field, making it possible to get the information from large datasets like this one. Also, I would like to thank Prof. Shanmuga to provide us this dataset.