Data Narrative – Assignment 2

Kaveri Visavadiya (22110114)
Electrical Engineering
IIT Gandhinagar

I. Overview of the dataset

The given dataset was the subject of scrutiny of the 1995 Data Analysis Exposition by the American Statistical Association. The datasets are drawn from two sources, U.S. News and the AAUP (American Association of University Professors).

The first dataset, the AAUP dataset, contains information on faculty salaries for 1161 American colleges and universities. The data is obtained in two formats:

aaup.data contains the raw data in comma-delimited fields with a single data line for each college, whereas aaup2.data contains data arranged in fixed columns, with two data lines for each college and a maximum line length of 80 characters. Here, missing values are denoted with * and all salary and compensation figures are yearly in $100's.

The key of the dataset aaup2.data (which has fixed column format with two lines per college) is as follows:

Line #1
- 1 - 5 FICE (Federal ID number)
- 7 - 37 College name
- 38 - 39 State (postal code)
- 40 - 43 Type (I, IIA, or IIB)
- 44 - 48 Average salary - full professors
- 49 - 52 Average salary - associate professors
- 53 - 56 Average salary - assistant professors
- 57 - 60 Average salary - all ranks
- 61 - 65 Average compensation - full professors
- 66 - 69 Average compensation - associate professors
- 70 - 73 Average compensation - assistant professors
- 74 - 78 Average compensation - all ranks

Line #2
- 1 - 4 Number of full professors
- 5 - 8 Number of associate professors
- 9 - 12 Number of assistant professors
- 13 - 16 Number of instructors
- 17 - 21 Number of faculty - all ranks

The second dataset, the USNEWS dataset, contains information on over 1300 American colleges and universities, mostly for the 1993-94 school year. The data is obtained in two formats:

usnews.data contains the raw data in comma-delimited fields with a single data line for each college, whereas usnews3.data contains data arranged in fixed columns, with three data lines for each college and a maximum line length of 80 characters. Here, missing values are denoted with *.

The key of the dataset usnews3.data (which has fixed column format with three lines per college) is as follows:

Line #1
- 1 - 5 FICE (Federal ID number)
- 7 - 51 College name
- 53 - 54 State (postal code)

Line #2

1 - 2 Public/private indicator (public=1, private=2)
3 - 6 Average Math SAT score
7 - 10 Average Verbal SAT score
11 - 15 Average Combined SAT score
16 - 18 Average ACT score
19 - 22 First quartile - Math SAT
23 - 26 Third quartile - Math SAT
27 - 30 First quartile - Verbal SAT
31 - 34 Third quartile - Verbal SAT
35 - 37 First quartile - ACT
38 - 40 Third quartile - ACT
41 - 46 Number of applications received
47 - 52 Number of applicants accepted
53 - 57 Number of new students enrolled
58 - 61 Pct. new students from top 10% of H.S. class
62 - 65 Pct. new students from top 25% of H.S. class

Line #3
1 - 6 Number of fulltime undergraduates
7 - 12 Number of part-time undergraduates
13 - 18 In-state tuition
19 - 24 Out-of-state tuition
25 - 29 Room and board costs
30 - 34 Room costs
35 - 39 Board costs
40 - 44 Additional fees
45 - 49 Estimated book costs
50 - 54 Estimated personal spending
55 - 58 Pct. of faculty with Ph.D.'s
59 - 62 Pct. of faculty with terminal degree
63 - 67 Student/faculty ratio
68 - 70 Pct. alumni who donate
71 - 76 Instructional expenditure per student
77 - 80 Graduation rate

II. Scientific questions/Hypotheses

1. Which colleges have the highest estimated personal spending? Are they mostly public or private? What could be the reason for high personal spending?
2. What is the relation between instructional expenditure per student and graduation rate?
3. Is there a correlation between % of faculty with PhD and average salary of professors?
4. What percent of alumni donate to private colleges compared to public colleges? How is it related to room and board costs and instructional expenditure per student?
5. For the colleges which have a graduation rate equal to 100%, what is the median salary of professors and the number of professors? How does it compare with the median salary of professors and the number of professors of all colleges?
6. Is there a correlation between graduation rate and percent of new students from the top 10% of high school class? Top 25% of high school class?
7. Is there a relation between graduation rate and average ACT score?
8. Among the top 50 colleges which have a high graduation rate, highest number of students from the top 10% of high school class, highest percent alumni who donate and instructional expenditure per student, pick the colleges which are in all of these.
9. What is the relation between number of faculty (< 500 faculty) and number of undergraduate students?
10. How does the number of private colleges in a state relate with the average in-state and out-of-state tuition of that state?

III. Details of the Libraries and Functions

The libraries matplotlib and pandas are required for answering the above posed scientific questions.

A. *Matplotlib*

Matplotlib is a scientific visualization library in Python for 2D plots of arrays. It allows us visual access to huge amounts of data in the form of plots such as line, bar, scatter, histogram, etc. Here we will be using the pyplot module which provides a MATLAB-like interface. Each pyplot function makes some change to a figure, i.e., creates a figure, plots some lines in the plotting area, inputs label, etc. It is imported as follows:

```
import matplotlib.pyplot as plt
```

In the code, we use the functions title(), xlabel() and ylabel().

The title() method in matplotlib module is used to specify the title of the plot. Similarly, xlabel() and ylabel() specify the labels on the x and y axes.

The legend() function is used to label different graphs on the same plot.

B. *Pandas*

pandas works with tabular data such as data stored in spreadsheets and databases. pandas helps in cleaning, exploring and processing data. In pandas, a mapping from keys to values is called a Series and a mapping from a Series to column name, or equivalently a data table, is called a DataFrame.

We use several pandas functions in answering the scientific questions, the most important of which are mentioned below:

- read_csv(url, names = ‘list’) will import a data file into the namespace and store the contents of the file into a variable. This variable can be used to access the columns of the file. The columns of the file are named using names = [list containing names of the columns].
- sort_values(column, ascending = True) will sort values in a DataFrame along either axis.
- DataFrame.head(x) returns the topmost x values of the DataFrame.
- Series.value_counts() returns a Series containing counts of unique values in descending order so that the first element is the most frequently occurring element.
- DataFrame.groupby() groups a DataFrame according to the values in a column to create a GroupBy object. It can then be operated on using aggregate functions like mean, median, max, etc.
- DataFrame.dropna() is used to drop rows that have NaN values.
- Series.astype(‘type’) is used to convert the values in a Series to another type
- A new column can be added to a DataFrame using the syntax:

  ```
  df[‘New Col’] = series
  ```

IV. Answers to the Questions (With Appropriate Illustrations)

1. **Which colleges have the highest estimated personal spending? Are they mostly public or private? What could be the reason for high personal spending?**

The answer to the question can be found in usnews.data. We import the file and use pd.read_csv to create a variable usnews that stores the contents of the books. The variable col2 contains the names of the columns.

```
import pandas as pd
url2 = "http://lib.stat.cmu.edu/datasets/colleges/usnews.data"
usnews = pd.read_csv(url2, names = col2)
```

For this question, we require the column 'Estimated personal spending'. The rows with '*' can be removed using masking and the corresponding series can be converted to int type.

```
spend = usnews[usnews['Estimated personal spending'] != '*']['Estimated personal spending'].astype(int)
```

The corresponding column in usnews is replaced with the updated column.

```
usnews['Estimated personal spending'] = spend
```

Then, we sort the DataFrame in descending order of 'Estimated personal spending' and retrieve the required columns.

```
usnews = usnews.sort_values('Estimated personal spending', ascending = False)['College name', 'Public(1)/Private(2)']].head(10)
```

The output, displayed in the form of a table, is as follows:

TABLE I.

	College name	Public(1)/Private(2)
807	Molloy College	2
607	Saint Louis University	2
385	MidAmerica Nazarene College	2
619	Hannibal-LaGrange College	2
612	University of Missouri at Saint Louis	1
594	Lindenwood College	2
222	Oglethorpe University	2
184	Nova Southeastern University	2
136	Colorado Christian University	2
617	William Woods University	2

As we can see, 9 out of the 10 colleges are private. Another observation is that many of these colleges are Christian and located in Missouri. A possible reason for such a high estimated personal spending is that they are located in the middle of the city/metropolitan area such as Fulton, Connecticut, etc.

2. **What is the relation between instructional expenditure per student and graduation rate?**

   The modules required are pandas and matplotlib.pyplot. The dataset required is usnews.

```
  import matplotlib.pyplot as
plt
  import pandas as pd
  b = pd.read_csv('books.csv')
url2 =
"http://lib.stat.cmu.edu/dataset
s/colleges/usnews.data"
usnews = pd.read_csv(url2, names
= col2)
```

For this question, the columns required are ‘Instructional expenditure per student’ and ‘Grad. rate’. As before, we remove the rows with ‘*’ values and convert the values in the columns to int after retrieving the columns.

```
  exp =
usnews[usnews['Instructional
expenditure per student'] !=
'*']['Instructional expenditure
per student'].astype(int)
  grad = usnews[usnews['Grad.
rate'] != '*']['Grad.
rate'].astype(int)
```

Then, we store the two series in a DataFrame and remove an anomaly in the dataset using masking; this is because a college has a graduation rate above 100%.

```
  df = pd.DataFrame([exp,
grad]).T.dropna()
  df = df[df['Grad. rate'] <
101]
```

Then, we group it according to values in the ‘Grad. rate’ and take the maximum and median instructional expenditure per student. We group it so that colleges that have the same graduation rate do not overcrowd the final graph with their different instructional expenditures.

```
  x = df.groupby('Grad. rate',
as_index = False)['Instructional
expenditure per
student'].median()
  y = df.groupby('Grad. rate',
as_index = False)['Instructional
expenditure per student'].max()
```

We can plot a line graph showing the Graduation rate vs Instructional Expenditure per Student (median and max) as follows.

```
  plt.plot(x['Grad. rate'],
x['Instructional expenditure per
student'], label = 'Median
expenditure')
  plt.plot(y['Grad. rate'],
y['Instructional expenditure per
student'], label = 'Max
expenditure')
  plt.title('Grad. rate vs
Instructional expenditure per
student')
  plt.xlabel('Graduation rate')
  plt.ylabel('Instructional
expenditure per student')
  plt.grid()
  plt.legend()
```

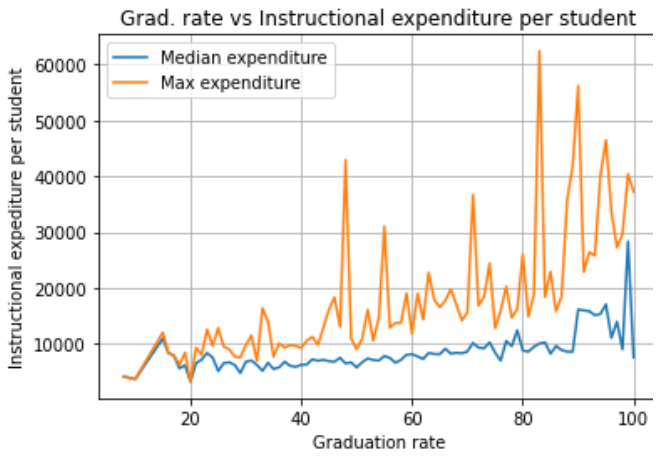

Fig. 1. Grad. Rate vs Instructional expenditure per Student

As we infer, from the median expenditure graph, graduation rate, in general, increases with increase in instructional expenditure per student.

3. Is there a correlation between % of faculty with PhD and average salary of professors?

This question can be answered by using the module pandas and datasets usnews and aaup. As before, we import the modules and the dataset. Since the columns required are 'Average salary - ful prof.' and '% of faculty with PhD', we remove the rows with '*' and convert them to int. Then, we replace/add the columns in usnews with the updated columns without '*' values.

Since we are comparing private to public colleges, we group the dataset usnews according to values in the column 'Public(1)/Private(2)' then take the median of the values in the column 'Avg. math SAT score', 'Avg. verbal SAT score', 'Avg. SAT score', and 'Avg. ACT score'. Then we group usnews by '% of faculty with PhD' and take the mean of the values in the column 'Average salary - full prof.'

```
a = usnews.groupby('% of
faculty with PhD', as_index
= False)['Average salary -
full prof.'].mean().dropna()

print(a)
```

These can then be printed to obtain the following results.

```
% of faculty with PhD
Average salary - full prof.

0                        8.0
497.000000

1                       10.0
512.000000

2                       11.0
489.500000

3                       14.0
446.000000

4                       15.0
745.000000

..                       ...
...

83                      96.0
568.571429

84                      97.0
469.714286
```

```
85                    98.0
517.000000

86                    99.0
467.714286

87                   100.0
574.250000


[86 rows x 2 columns]
```

As we can infer, there is no correlation between % of faculty with PhD and average salary of a full professor.

4. **What percent of alumni donate to private colleges compared to public colleges? How is it related to room and board costs and instructional expenditure per student?**

We require the modules pandas and matplotlib and the dataset usnews.

```
import pandas as pd

usnews = pd.read_csv(url2,
names = col2)
```

The columns required are ‘Public(1)/Private(2)’, ‘% alumni who donate’, ‘Room and board costs’ and ‘Instructional expenditure per student’. We follow the same procedure as before of removing rows with ‘*’ values and updating the columns in usnews.

```
pp =
usnews[usnews['Public(1)/Pri
vate(2)'] !=
'*']['Public(1)/Private(2)']
.astype(int)

usnews['Public(1)/Private(2)
'] = pp

donate = usnews[usnews['%
alumni who donate'] !=
'*']['% alumni who
donate'].astype(int)

usnews['% alumni who
donate'] = donate
```

Since we want to compare the percent of alumni who donate to public and private colleges, we group the dataset usnews based on column ‘Public(1)/Private(2)’ and take the median of values in ‘% alumni who donate’.

```
a =
usnews.groupby('Public(1)/Pr
ivate(2)')['% alumni who
donate'].median()
```

After removing rows with ‘*’ values and columns with int values in usnews, we perform similarly for ‘Room and board costs’ and ‘Instructional expenditure per student.’

```
b =
usnews.groupby('Public(1)/Pr
ivate(2)')['Room and board
costs'].median()
```

```
c =
pd.DataFrame({'Public/Privat
e': ['Public', 'Private'],
'Instructional expenditure
per student': [c[1], c[2]]})
```

We convert a, b and c to DataFrames for easier conversion to bar plot.

```
a =
pd.DataFrame({'Public/Privat
e': ['Public', 'Private'],
'% alumni who donate':
[a[1], a[2]]})
```

Then, finally we plot the subplots.

```
fig, axs =
plt.subplots(nrows=3,
ncols=1, figsize=(6, 12))

a.plot(x='Public/Private',
y='% alumni who donate',
kind='bar', ax=axs[0])

axs[0].set_xlabel('Public/Pr
ivate')

axs[0].set_ylabel('% alumni
who donate')
```

The output is displayed below.

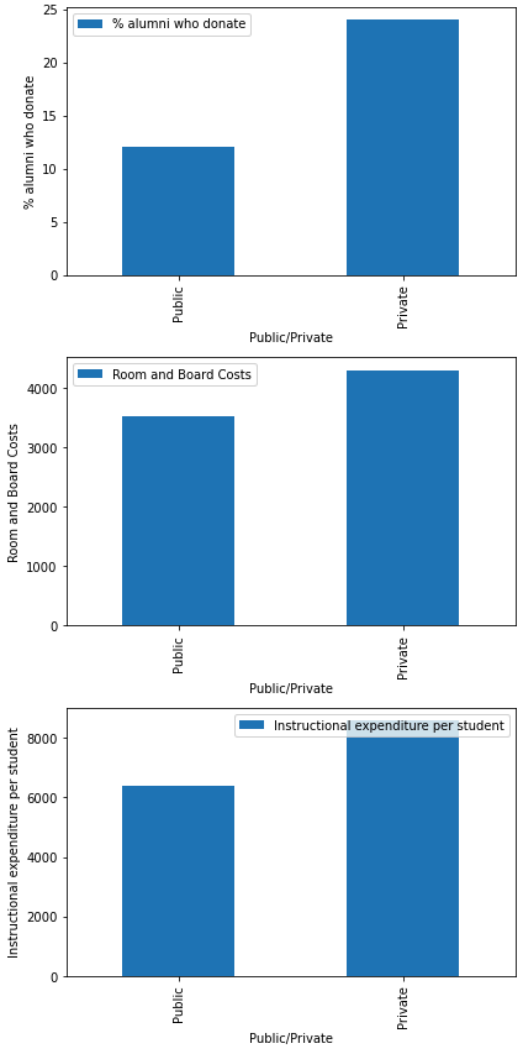

Fig. 2. Comparing the % alumni who donate (income) to public and private colleges and relating the income with other parameters related to investment

As we can infer, the percent of alumni who donate to private colleges is almost double the percent of alumni who donate to public colleges. Now, if room services are costly and hence good, more alumni will donate back to their institute hence keeping up the cycle of high room costs in private colleges. Moreover, higher donations

also mean that more money is spent on the instruction of a student.

5. **For the colleges which have a graduation rate equal to 100%, what is the median salary of professors and the number of professors? How does it compare with the median salary of professors and the number of professors of all colleges?**

We require the module pandas and the datasets usnews and aaup.

```
import pandas as pd
import matplotlib.pyplot as plt
usnews = pd.read_csv(url2, names = col2)
aaup = pd.read_csv(url, names = col)
```

We require the columns 'Grad. rate', 'Average compensation - all ranks', 'Average salary - all ranks' and 'No. of full prof.' After removing '*' values, we sort usnews in descending order of 'Grad. rate' and retrieve the rows with a graduation rate equal to 100%.

```
usnews = usnews.dropna().drop(771).sort_values('Grad. rate', ascending = False)
x = usnews[usnews['Grad. rate'] == 100.0]
```

Then we print the median of all three parameters for x (containing the colleges with 100% graduation rate) and usnews (containing all colleges).

```
print('Average compensation - all ranks for colleges with 100% grad. rate : ', x['Average compensation - all ranks'].median())
print('Average compensation - all ranks for colleges : ', usnews['Average compensation - all ranks'].median())
```

The output is:

```
Average compensation - all ranks for colleges with 100% grad. rate :  537.0
Average compensation - all ranks for colleges :  510.0
```

```
Average salary - all ranks for colleges with 100% grad. rate :  435.0
Average salary - all ranks for colleges :  407.0
```

```
Average no. of full prof. for all colleges with 100% grad. rate :  74.0
Average no. of full prof. for all colleges :  40.0
```

Disappointingly, there is not much difference between average salary/compensation of professors in colleges with 100% graduation rate and all colleges. However, colleges with a 100% graduation rate have a higher number of professors on average.

6. **Is there a correlation between graduation rate and percent of new students from the top 10% of high school class? Top 25% of high school class?**

We require the modules pandas and matplotlib.pyplot and the dataset usnews.

```
import pandas as pd

import matplotlib.pyplot as plt

usnews = pd.read_csv(url2, names
= col2)
```

We require the columns 'Grad. rate', '% new students from top 25% of HS class' and '% new students from top 10% of HS class'. After removing '*' values and updating usnews, we group usnews by 'Grad. rate' and take the mean of values in '% new students from top 25% of HS class' and '% new students from top 10% of HS class' for every integer in the column 'Grad. rate'.

```
a = usnews.groupby('Grad. rate',
as_index = False)['% new
students from top 25% of HS
class'].mean()

a =
a.dropna().drop(88).sort_values(
'Grad. rate', ascending = False)

b = usnews.groupby('Grad. rate',
as_index = False)['% new
students from top 10% of HS
class'].mean()

b =
b.dropna().drop(88).sort_values(
'Grad. rate', ascending = False)
```

We plot using variables a and b as follows, with appropriate labels to distinguish the two graphs on the same plot.

```
plt.plot(a['Grad. rate'], a['%
new students from top 25% of HS
class'], label = 'Top 25%')

plt.plot(b['Grad. rate'], b['%
new students from top 10% of HS
class'], label = 'Top 10%')

plt.title('Grad. rate vs % of
new students from top 10% and
25% of HS class')

plt.xlabel('Graduation rate')

plt.ylabel('% of new students
from top 10% or 25% of HS
class')

plt.legend()
```

The output is obtained as:

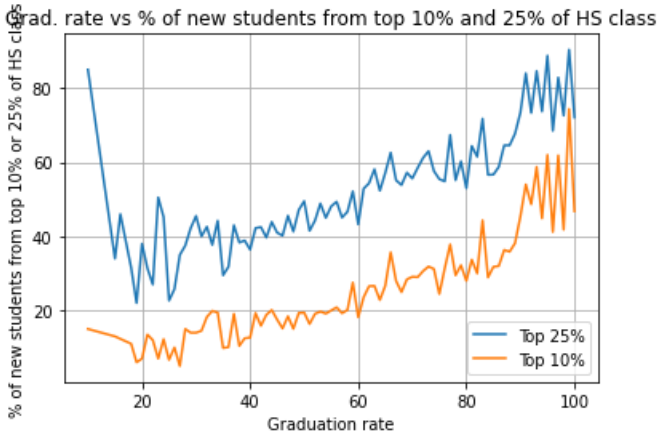

Fig. 3. Grad. Rate vs Percent. of new students from top 10% and top 25% of high school class

Disregarding the few outliers, there is a clear correlation between graduation rate and quality of students. Since students at the

top of their class are more likely to be dedicated to finishing their degree, the graduation rate of colleges with these students is likely to be higher.

7. Is there a relation between graduation rate and average ACT score?

We require the modules pandas and matplotlib.pyplot and the dataset usnews.

We require the columns 'Grad. rate', 'Avg. ACT score', '1st quartile – ACT', and '3rd quartile – ACT'. After removing '*' values and updating usnews, we group usnews by 'Grad. rate' and take the mean of values 'Avg. ACT score', '1st quartile – ACT', and '3rd quartile – ACT' for every integer in the column 'Grad. rate'. Then, we sort the variables where 'Avg. ACT score', '1st quartile – ACT', and '3rd quartile – ACT' are stored in descending order of 'Grad. rate' and drop the NaN values. We also drop the extraneous row in variable a where the graduation rate is greater than 100% (an error in the dataset).

```
a = usnews.groupby('Grad. rate', as_index = False)['Avg. ACT score'].mean()

a = a.dropna().drop(88).sort_values('Grad. rate', ascending = False)

b = usnews.groupby('Grad. rate', as_index = False)['1st quartile - ACT'].mean()

b = b.dropna().sort_values('Grad. rate', ascending = False)

c = usnews.groupby('Grad. rate', as_index = False)['3rd quartile - ACT'].mean()

c = c.dropna().sort_values('Grad. rate', ascending = False)
```

We plot as follows.

```
plt.plot(a['Grad. rate'], a['Avg. ACT score'], label = 'Avg. ACT')

plt.plot(b['Grad. rate'], b['1st quartile - ACT'], label = '1st quartile')

plt.plot(c['Grad. rate'], c['3rd quartile - ACT'], label = '3rd quartile')
```

The graph is obtained as follows:

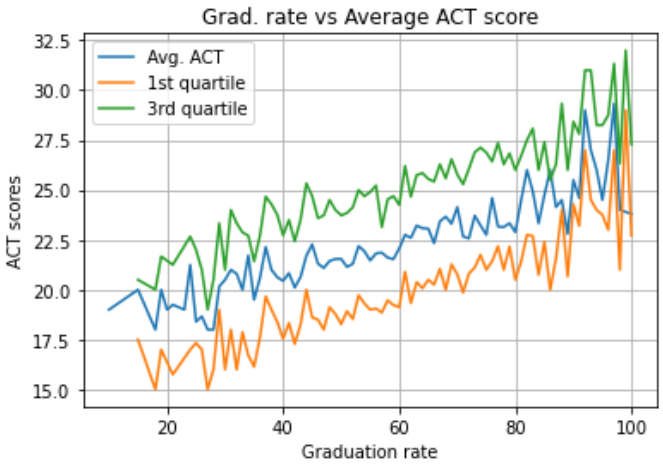

Fig. 4. Grad. Rate vs Average ACT score (including 1st quartile and 3rd quartile)

Similar to the previous hypothesis, graduation rate, in general, increases with ACT score for the reason as before.

8. **Among the top 50 colleges which have a high graduation rate, highest number of students from the top 10% of high school class, highest percent alumni who donate and instructional expenditure per student, pick the colleges which are in all of these.**

We require the module pandas and the dataset usnews. We also require the columns 'Grad. rate' , '% new students from top 10% of HS class', '% alumni who donate', and 'Instructional expenditure per student'. Subsequently, we choose the top 100 college names by sorting in descending order of the values in the 4 columns as shown below.

```
grad = usnews[usnews['Grad. rate'] != '*']['Grad. rate'].astype(int)

usnews['Grad. rate'] = grad

highest_grad = usnews.drop(771).sort_values('Grad. rate', ascending = False)['College name'].head(100)



top10 = usnews[usnews['% new students from top 10% of HS class'] != '*']['% new students from top 10% of HS class'].astype(int)

usnews['% new students from top 10% of HS class'] = top10

highest_top10 = usnews.sort_values('% new students from top 10% of HS class', ascending = False)['College name'].head(100)



donate = usnews[usnews['% alumni who donate'] != '*']['% alumni who donate'].astype(int)

usnews['% alumni who donate'] = donate

highest_donate = usnews.sort_values('% alumni who donate', ascending = False)['College name'].head(100)



exp = usnews[usnews['Instructional expenditure per student'] != '*']['Instructional expenditure per student'].astype(int)

usnews['Instructional expenditure per student'] = exp

highest_exp = usnews.sort_values('Instructional expenditure per student', ascending = False)['College name'].head(100)
```

Then, we select the common college names by converting the pandas Series to a set and taking their intersection. We take 3 intersections for the 4 Series.

```
a = pd.Series(list(set(highest_grad) & set(highest_top10)))

b = pd.Series(list(set(a) & set(highest_donate)))

c = pd.Series(list(set(b) & set(highest_exp)))
```

The output is obtained as follows:

0	Middlebury College
1	Yale University
2	Davidson College
3	Swarthmore College
4	Carleton College
5	Colby College
6	Princeton University
7	Washington and Lee University
8	Harvey Mudd College
9	Wellesley College
10	Harvard University
11	Dartmouth College
12	Bates College
13	Amherst College
14	Duke University
15	Williams College
16	Smith College
17	Bryn Mawr College
18	Bowdoin College
19	Pomona College
20	Haverford College

dtype: object

Many of these colleges are the best in the world even now, such as Harvard, Dartmouth, Amherst, Duke, Yale, Princeton, etc. Thus, the parameters I considered in the question are suitable to measure the quality of a college/university.

9. What is the relation between number of faculty (< 500 faculty) and number of undergraduate students?

We require the modules pandas and matplotlib.pyplot and the datasets aaup and usnews. We need the columns 'Number of faculty – all ranks' from dataset aaup and 'No. of fulltime UG' from dataset usnews (considering only fulltime undergraduate students because part-time undergraduates are few in number). As before, we select only those rows without '*' values and update dataset usnews with the new columns.

```
nof = aaup[aaup['Number of faculty - all ranks'] != '*']['Number of faculty - all ranks'].astype(int)

usnews['Number of faculty - all ranks'] = nof


noug = usnews[usnews['No. of fulltime UG'] != '*']['No. of fulltime UG'].astype(int)

usnews['No. of fulltime UG'] = noug
```

To obtain the correlation between number of UG students and number of faculty, we group usnews according to 'Number of faculty – all ranks', take the average of the number of UG students for a particular number of faculty, and drop the NaN values.

```
a = usnews.groupby('Number of faculty - all ranks', as_index = False)['No. of fulltime UG'].mean().dropna()
```

Upon plotting, however, it is observed that there is overcrowding of values for number of faculty below 500 and sparseness of values above 1000. This implies that most colleges have number of faculty below 500. To obtain a clear graph, only number of faculty below 500 is plotted.

```
a = a[a['Number of faculty - all ranks'] < 500]

plt.plot(a['Number of faculty - all ranks'], a['No. of fulltime UG'])

plt.title('No. of fulltime UG students vs No. of faculty')

plt.xlabel('No. of faculty - all ranks < 500')

plt.ylabel('No. of fulltime undergraduate students')

plt.grid()
```

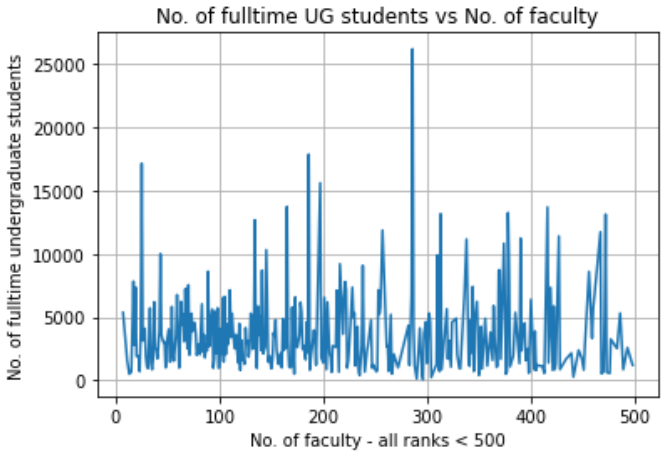

Fig. 5. Number of fulltime undergraduate students vs Number of faculty (< 500)

It is inferred that there is no clear correlation between number of students and faculty. This relates to the fact that each college has a different student/faculty ratio.

10. How does the number of private colleges in a state relate with the average in-state and out-of-state tuition of that state?

We require the modules pandas and matplotlib and the dataset usnews. We require the columns 'Public(1)/Private(2)' and 'In-state tuition'. We remove the rows with '*' (unknown) values and update usnews.

```
instate = usnews[usnews['In-state tuition'] != '*']['In-state tuition'].astype(int)

usnews['In-state tuition'] = instate

outstate = usnews[usnews['Out-of-state tuition'] != '*']['Out-of-state tuition'].astype(int)

usnews['Out-of-state tuition'] = outstate
```

To obtain a series of the average out-of-state tuition of all colleges in the state with 'State' as index, we group usnews by 'State' and take the mean of 'In-state tuition' for every state in the group. Then we count the number of private colleges in that state by grouping it by 'State' again.

```
states = usnews.groupby('State')['In-state tuition'].mean()

private = usnews[usnews['Public(1)/Private(2)'] == 2].groupby('State')['Public(1)/Private(2)'].count()
```

We do the same for 'Out-of-state tuition'. Then, we store these values in a DataFrame and use this DataFrame to plot the graph showing the relation between tuition and no. of private colleges.

```
df = pd.merge(states, private, how = 'outer', left_index = True, right_index = True).dropna().sort_values('In-state tuition', ascending = False)

plt.plot(df['In-state tuition'], df['Public(1)/Private(2)'], label = 'In-state tuition')
```

The result is obtained as:

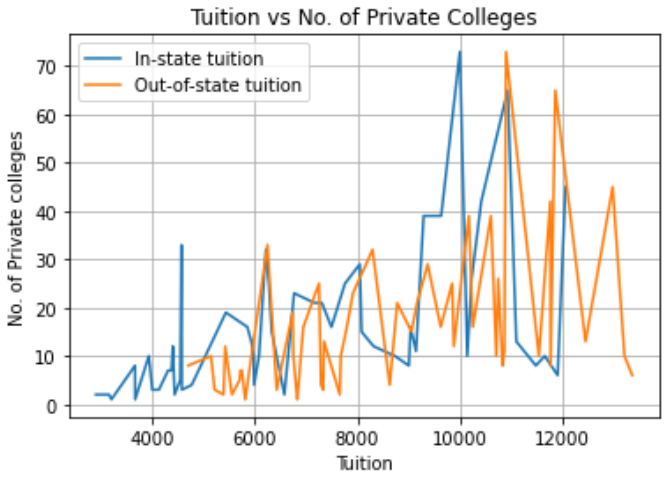

Fig. 6. Tuition vs No. of private colleges for every state

As we can see in the graph, there is a correlation, albeit with several outliers, between the tuition of a state and the number of private colleges it has. The outliers could be due to several reasons such as:

- Lower demand of private colleges in the state due to cultural or economic factors
- Public universities may be well-funded and able to charge higher tuition fees which could be factor in high in-state tuition

V. Summary of the Observations

We can infer from the first answer that private colleges located in the city/metropolitan area are more likely to have higher personal spending.

The second answer says that with an increase in instructional expenditure per student, graduation rate is expected to increase.

In the third answer, we see that contrary to our assumption, there is no relation between the number of faculty with PhD and the average salary of a full professor.

The fourth answer says that the percentage of alumni who donate back to their institutions is higher for private than public colleges. This is likely because of better facilities received at private colleges.

In the fifth question, our hypothesis was that with an increase in graduation rate, the instructors are likely to be more skilled and thus better paid. However, contrary to our expectations, the answer reveals that there is no such correlation.

The sixth answer shows the correlation between graduation rate and the percent of new students graduating at the top of their high school class.

The seventh answer is related to the sixth one and shows the correlation between graduation rate and average ACT scores of UG students.

The eighth answer takes the intersection of the top colleges which satisfy some criteria (which we assume are necessary markers of a good college) and shows them. Indeed, these were the right criteria to take, as many of these universities are at top even now.

The ninth answer shows that there is no relation between the number of students and faculty in a college—a fact that relates to the differing student/faculty ratio of each college.

The tenth answer reveals that if a state has high average in-state or out-of-state tuition, it is likely to have a higher number of private colleges.

This was a fairly simple and straightforward way of gaining meaningful results from a dataset. The data narrative helped gain insight about how to frame relevant scientific questions and how to appropriately answer them using various Python libraries. Moreover, it helped explore various aspects of the dataset like how graduation rate relates with other aspects of the college like quality of students, expenditure per student, etc. and how public and private colleges differ from or are similar to each other.

VI. Unanswerable Questions

Questions such as 'Why are colleges with high personal expenditure mostly Christian private schools?', 'Why do some colleges have a low graduation rate but high test scores or high percent of students from the top of their class?' and 'Why is there no relation between salary of a professor and instructional expenditure per student?' remain unanswerable. This is because the required data to answer them is missing from the dataset, or there were certain discrepancies while collecting the data, or there were some anomalies with the college itself.

References

[1] Chen, Daniel Y. *Pandas for everyone: Python data analysis*. Addison-Wesley Professional, 2017.

[2] VanderPlas, Jake. *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.", 2016.

[3] Pandas. "User Guide." Accessed February 22, 2023. https://pandas.pydata.org/docs/user_guide/index.html

[4] Matplotlib. "Users Guide." Accessed February 22, 2023. https://matplotlib.org/stable/users/index.html#

Acknowledgement

I would like to thank Prof. Shanmuga for giving me the opportunity to gain insight into real-life applications of pandas, numpy and matplotlib. I would also like to thank the providers of the data for the online dataset –Robert Morse (Director of Research for

America's Best Colleges at U.S. News and World Report) and Maryse Eymonerie (Consultant to AAUP). Finally, I thank creators of online resources on pandas and matplotlib such as Jake VanderPlas and John Hunter; this data narrative would have been impossible without them.