DATA NARRATIVE

#1 Vipul Sunil Patil (22110189)
Btech First Year Student (CSE),
IIT Gandhinagar,
Gandhinagar, Gujrat, India.
vipul.patil@iitgn.ac.in

I. Overview of the Dataset

The dataset consists of 2 file. we get an big data that we had to analyze and make questions of The dataset contains in total 35 variables, including 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 and Graduation rate in on data set.

On the other hand, second data set contains 17 variables, 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 and Number of faculty - all ranks

Strings, integers, and floats are all present in this data. However, although being large, this data is insufficient in certain ways. Certain things are not adequately described. Several missing information had needed to be filled in. Data was usable after all, being easily transformed, calculated according to our need; this data was ready to create questions.

In my situation, I make use of data visualization and statistics to make sense of the data. It also helps us understand the plot and what data tried to convey us. On the other hand, exploring this data and seeing patterns and trends explains to me how the admission process is carried out.

Statistical analysis and charts came to provide numerous conclusions in front of me, but this data did not support some of them. With this information, we were able to provide several answers. Below questions and their answers will explain the data much closer and will make it easy to understand the data.

II. Scientific Questions/Hypotheses

1. Which region or state has the highest number of top grade colleges?

2. What is the average cost of living in each state for a student? Subsequently, for each state selected by the user, plot the Average Cost of Living for each college for both types of students (out-of-state and in-state).

3. Could you figure out the likelihood that you submitted an application and were chosen, and What should be the pattern of this probability from top to bottom collages?

4. What are the grade requirements for each college? Observe the trend of marks required for different colleges (arranged from top to bottom)?

5. How are Public Colleges different from Private Colleges?

6. Does the state influence teachers' salaries?

7. Are there colleges which do not follow minimum wage law?

8. Show in pie plot total number of different types of collage (I, IIA, IIB) and also in that same calculate the percentage of number of each professor ('Number of faculty - all ranks', 'Number of associate professors', 'Number of assistant professors', 'Number of instructors') in that particular type of collage.

9. Display the states of America together with their postal codes and the number of each type of institute present in that state.

10. What is the PDF and CDF of the data set's average salary for full professors?

III. Details of Libraries and Functions

Libraries Used in the Code are:

1. Pandas can be defined as a Python package that offers user-friendly data structures and tools for data analysis. It is constructed on top of NumPy, a scientific computing library for Python, and presents an effective and simple environment for data manipulation and analysis to Python developers. Pandas includes two primary data structures, namely the Series and DataFrame objects. The Series is a one-dimensional array-like object that can contain any data type, whereas the DataFrame is a two-dimensional table-like data structure with rows and columns capable of holding various data types.[1]

2. Matplotlib is a versatile and easy-to-use visualization library in Python that offers powerful tools for creating different types of plots such as line, bar, scatter, histogram, and many more. It is based on NumPy arrays and has seamless integration with the broader SciPy ecosystem.[2]

3. plotly.graph objs is a Python module that allows you to create interactive and dynamic data visualizations. It's part of the Plotly ecosystem, which also contains libraries like Plotly Express and Dash.

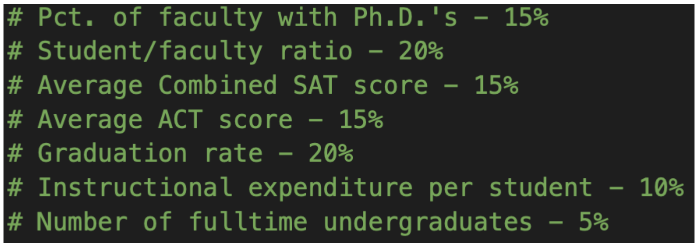
4. Numpy is a key Python module for scientific computing. It supports massive, multi-dimensional arrays and matrices, as well as a wide library of mathematical functions for working with these arrays.

Functions used in code.

1. For making dataframe we use Dataframe().
2. To sum the values we use sum().
3. To group data as we need to access we use groupby(). 4. reset_index() method to reset the index of the dataframe.
5. using the plot() we can create a plot of graphs and specify kind using kind.
6. We show the chart using the show().
7. sort_values() is used to sort values ascending or descending accordingly.
8. scatter() helps us in plotting scatter plots.
9. For the bar plot we use barh() where h means horizontal.
10. We set the title for the chart using the title() and x and y labels using xlabel() and ylabel() respectively.

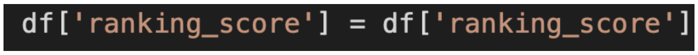
IV. Answers to the Questions (with Appropriate Illustrations)

1. We had to first create a criteria for top colleges,



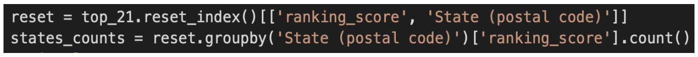
```
# Pct. of faculty with Ph.D.'s - 15%
# Student/faculty ratio - 20%
# Average Combined SAT score - 15%
# Average ACT score - 15%
# Graduation rate - 20%
# Instructional expenditure per student - 10%
# Number of fulltime undergraduates - 5%
```

then sort by that criteria (new column) through this we would get all the colleges in ascending order.



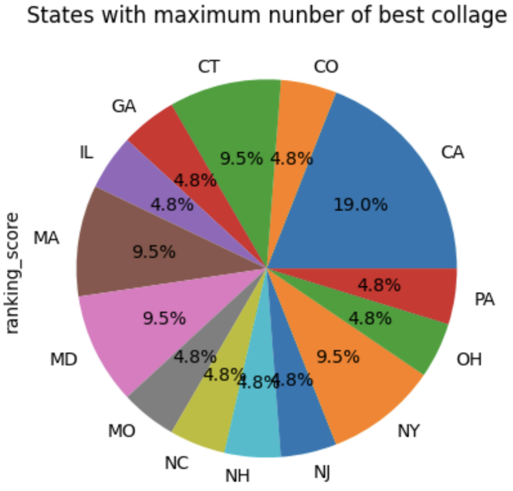
```
df['ranking_score'] = df['ranking_score']
```

To shorten the list and get the top colleges we will use head function and to order colleges according to states we will use groupby function.

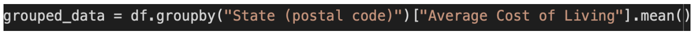


```
reset = top_21.reset_index()[['ranking_score', 'State (postal code)']]
states_counts = reset.groupby('State (postal code)')['ranking_score'].count()
```

To create a graph, we will use pie charts as it will be easy to identify the states with the most number of best colleges.

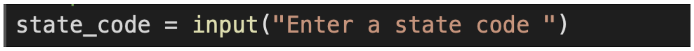


2. To do this, we first add all of the columns of data that provide Cost of Living and average them. The data must then be grouped by state and plotted for each.

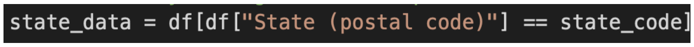


```
grouped_data = df.groupby("State (postal code)")["Average Cost of Living"].mean()
```

For the second question, we need to collect input from the user

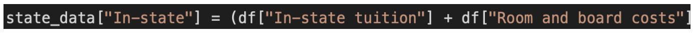
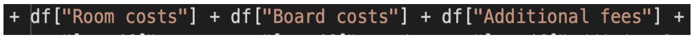
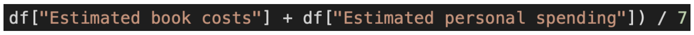


```
state_code = input("Enter a state code ")
```

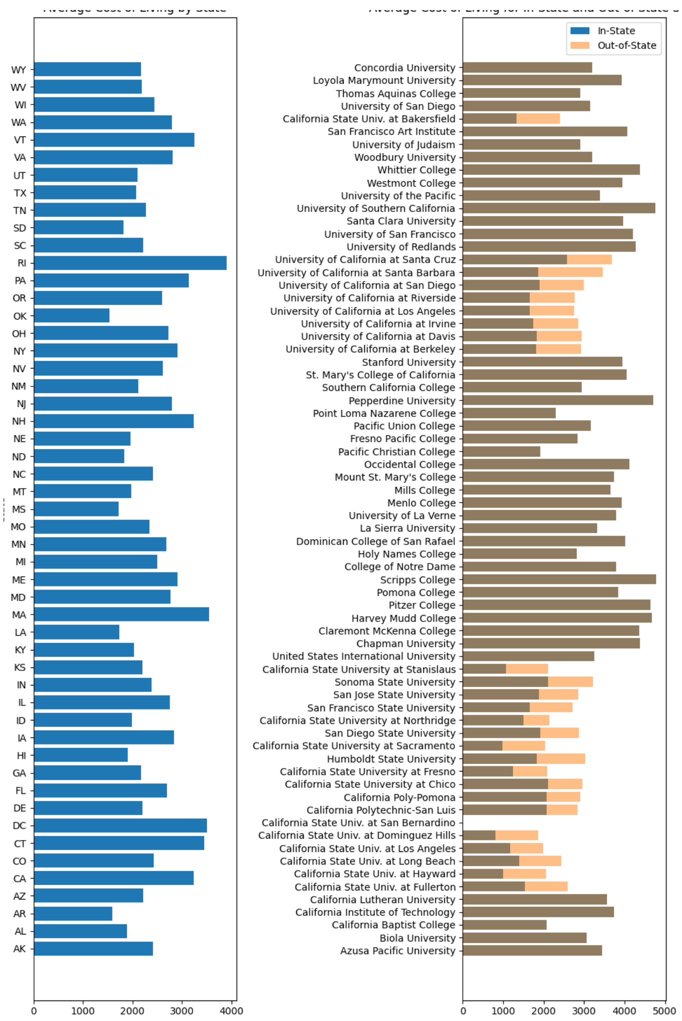


```
state_data = df[df["State (postal code)"] == state_code]
```

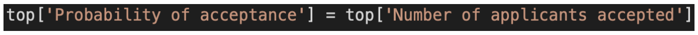
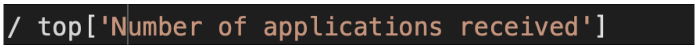
(as the name of the state) and then create a bar graph for that state that shows two typical prices of living for each college in that state, one for persons leaving that state and the other for those living in the other state.

```
state_data["In-state"] = (df["In-state tuition"] + df["Room and board costs"]
+ df["Room costs"] + df["Board costs"] + df["Additional fees"] +
df["Estimated book costs"] + df["Estimated personal spending"]) / 7
```

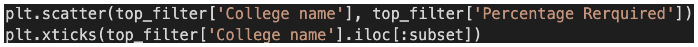


3. To calculate the possibility that you gave an application and got selected, we need to divide the elements of the column Number of applicants accepted and Number of applications received.

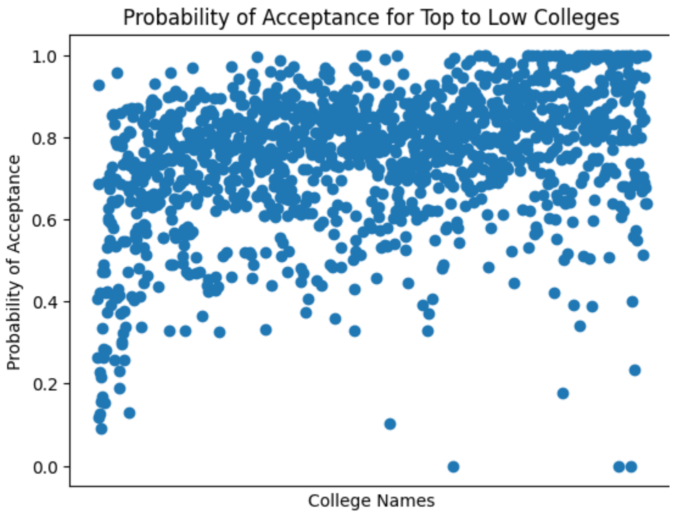



```
top['Probability of acceptance'] = top['Number of applicants accepted']
/ top['Number of applications received']
```

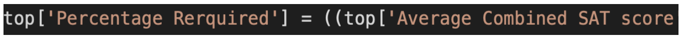
Now, for the second step, we need to construct a scatter plot between the probability we estimated previously and for all the collages (arranged in top to bottom order)



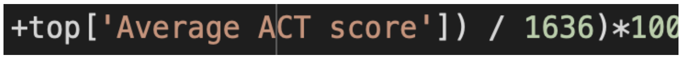
```
plt.scatter(top_filter['College name'], top_filter['Percentage Rerquired'])
plt.xticks(top_filter['College name'].iloc[:subset])
```



4. Secoundy, To compute the percentage of total marks required for admission to a top college, first calculate the percentage of total marks (ACT and SAT combined);

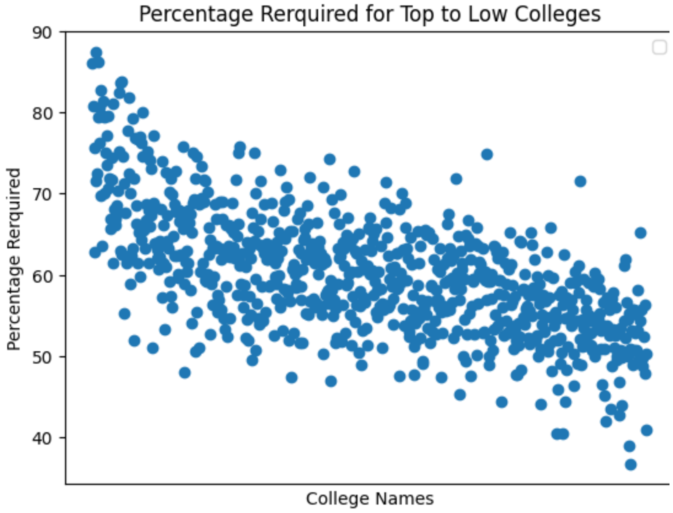


```
top['Percentage Rerquired'] = ((top['Average Combined SAT score
```

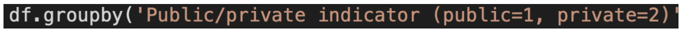


```
+top['Average ACT score']) / 1636)*100
```

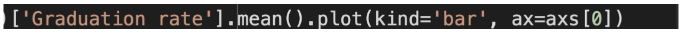
then draw the graph (scatter) between colleges and the percentage required for the college we calculated before. With the graph, we can easily observe that to get into top colleges, we need a top percentage; we can also see that as your percentage improves, your chances of getting into a top college increase as well.



5. For this code we need to determine the trend of many columns like Graduation rate

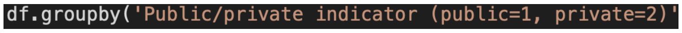


```
df.groupby('Public/private indicator (public=1, private=2)'
```

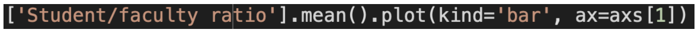


```
['Graduation rate'].mean().plot(kind='bar', ax=axs[0])
```

Student per professor

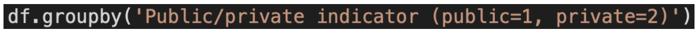


```
df.groupby('Public/private indicator (public=1, private=2)'
```

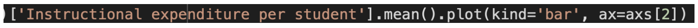


```
['Student/faculty ratio'].mean().plot(kind='bar', ax=axs[1])
```

spending per student

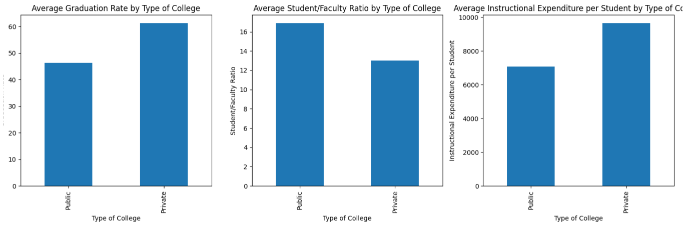


```
df.groupby('Public/private indicator (public=1, private=2)')
```

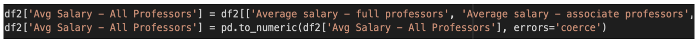


```
)['Instructional expenditure per student'].mean().plot(kind='bar', ax=axs[2])
```

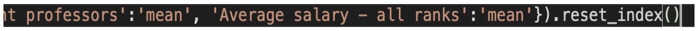
for both types of colleges (Private and Public).



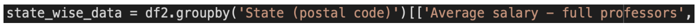
6. Definitely, the state has an impact on teacher compensation. For this question, we must first compute the average income of all professors in that state, and then, by producing a line graph, we can clearly observe that the salary of all types of professors varies by state.



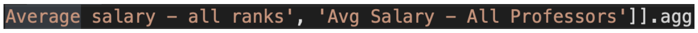
```
df2['Avg Salary - All Professors'] = df2[['Average salary - full professors', 'Average salary - associate professors',
df2['Avg Salary - All Professors'] = pd.to_numeric(df2['Avg Salary - All Professors'], errors='coerce')
```



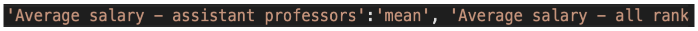
```
t professors':'mean', 'Average salary - all ranks':'mean'}).reset_index()
```



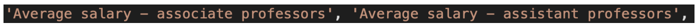
```
state_wise_data = df2.groupby('State (postal code)')[['Average salary - full professors',
```



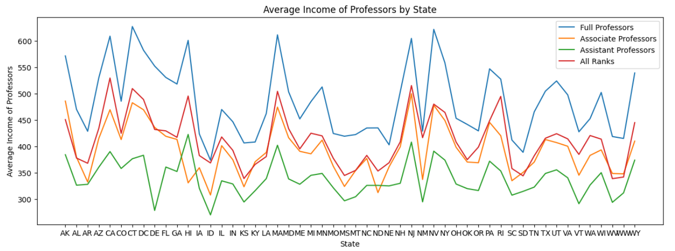
```
Average salary - all ranks', 'Avg Salary - All Professors']].agg
```



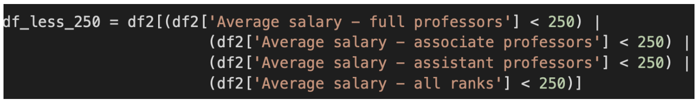
```
'Average salary - assistant professors':'mean', 'Average salary - all rank
```



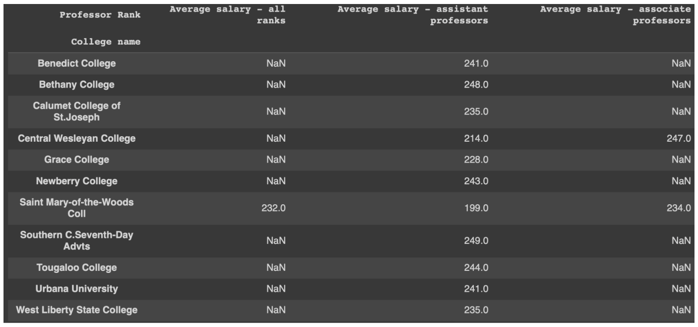
```
'Average salary - associate professors', 'Average salary - assistant professors',
```



7. For the second part, we must determine whether there is a college that pays less than $100 to its instructors; if so, the college has violated the US minimum wage statute.



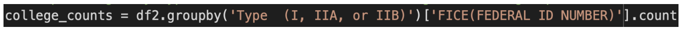
```
df_less_250 = df2[(df2['Average salary - full professors'] < 250) |
                  (df2['Average salary - associate professors'] < 250) |
                  (df2['Average salary - assistant professors'] < 250) |
                  (df2['Average salary - all ranks'] < 250)]
```



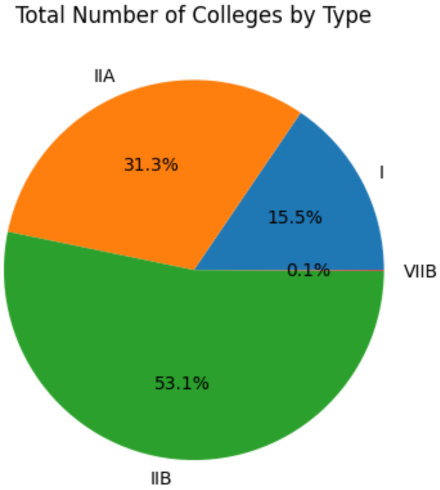
Professor Rank	Average salary - all ranks	Average salary - assistant professors	Average salary - associate professors
College name			
Benedict College	NaN	241.0	NaN
Bethany College	NaN	248.0	NaN
Calumet College of St.Joseph	NaN	235.0	NaN
Central Wesleyan College	NaN	214.0	247.0
Grace College	NaN	228.0	NaN
Newberry College	NaN	243.0	NaN
Saint Mary-of-the-Woods Coll	232.0	199.0	234.0
Southern C.Seventh-Day Advts	NaN	249.0	NaN
Tougaloo College	NaN	244.0	NaN
Urbana University	NaN	241.0	NaN
West Liberty State College	NaN	235.0	NaN

8. For this code, I was unable to make progress until the second step; I just estimated the percent or total number of

colleges in that category.



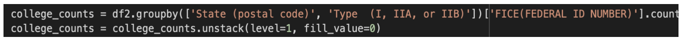
```
college_counts = df2.groupby('Type  (I, IIA, or IIB)')['FICE(FEDERAL ID NUMBER)'].count
```



9. For this code I need help of new function ploty.graph_objs

```
import plotly.graph_objs as go
```

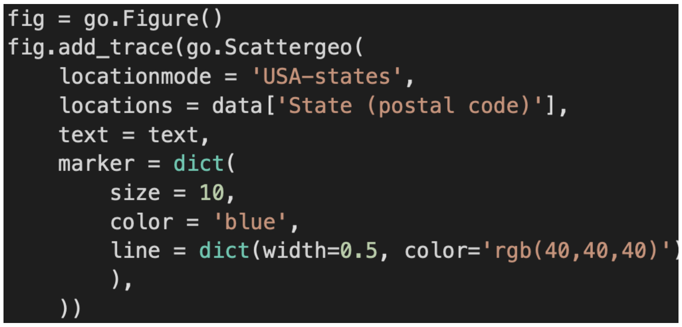
taken from resource [6] from this I imported the entire graph of the United States and then added the names of the states to it. For the following step, we need to use groupby to group the sets by states, count the number of each type of collage, and then print it by producing a short list.



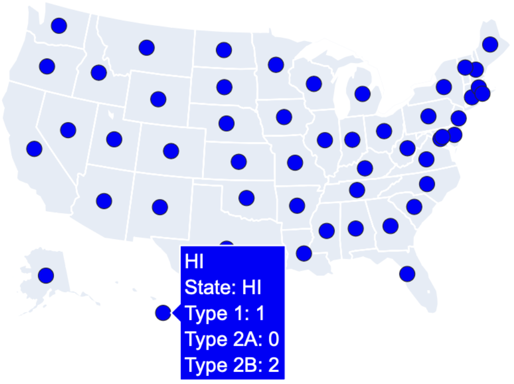
```
college_counts = df2.groupby(['State (postal code)', 'Type  (I, IIA, or IIB)'])['FICE(FEDERAL ID NUMBER)'].count
college_counts = college_counts.unstack(level=1, fill_value=0)
```

```
text = []
for state in data['State (postal code)']:
```

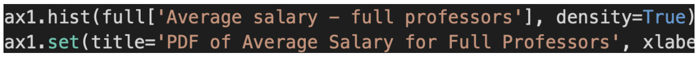
At last we need to plot the dig



```
fig = go.Figure()
fig.add_trace(go.Scattergeo(
    locationmode = 'USA-states',
    locations = data['State (postal code)'],
    text = text,
    marker = dict(
        size = 10,
        color = 'blue',
        line = dict(width=0.5, color='rgb(40,40,40)')
        ),
    ))
```



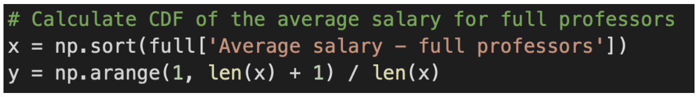
10. For this code we need to calculate PDF value



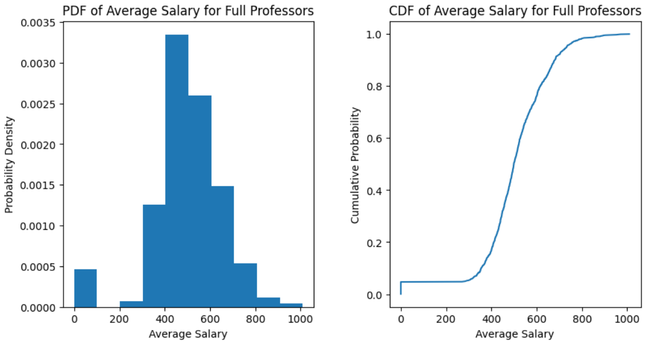
```
ax1.hist(full['Average salary - full professors'], density=True)
ax1.set(title='PDF of Average Salary for Full Professors', xlabe
```

```
, xlabel='Average Salary', ylabel='Probability Density')
```

and CDF value for Average Salary for full Professors



```
# Calculate CDF of the average salary for full professors
x = np.sort(full['Average salary - full professors'])
y = np.arange(1, len(x) + 1) / len(x)
```



VI. Unanswerable Questions, if any

Are there any strategies that colleges follow so as to boost their admission rate? Whether these actions have a negative impact on the students.

We need to sort to acquire trends of other columns when we increase Acceptance rate = ['Number of applicants accepted'] / ['Number of applications received']. Now we can see trends of fee structure and ACT and SAT scores by plotting a two line graph where y axes will have Acceptance rate. By analyzing we could conclude that generally colleges with high acceptance rates have higher fees and need lower ACT and SAT scores.

Can we argue that "as the number of colleges grows, so does the popularity of a city/state?"

For this we need to groupby states and make a sum of the total number of colleges in each state. We also need to take into account the fees(both tuition, and room and board) and take the mean. Now we can easily view the histogram of these three data points.

As we can see, the city with the most colleges has the highest room and board prices, indicating that the city has grown in popularity eventually. On the other hand the tuition fee has also significantly increased.

Can the Graduation rate of colleges be predicted on the basis of change in the student-faculty ratio and faculty with PhD?

Summary of the Observation that could be created by Unanswerable Questions

VIII. References

[1]The pandas development team. 2023. *Pandas-Dev/Pandas: Pandas*. Zenodo.

[2]“Matplotlib Tutorial.” 2021. GeeksforGeeks. February 8, 2021. https://www.geeksforgeeks.org/matplotlib-tutorial/.

This are some web sources that I used in my code:

1. “Python � Sort Grouped Pandas Dataframe by Group Size.” n.d. Tutorialspoint.com. Accessed March 30, 2023.

2. “How to Sort Data by Column in a CSV File in Python ?” 2021. GeeksforGeeks. March 30, 2021. https://www.geeksforgeeks.org/how-to-sort-data-by-column-in-a-csv-file-in-python/.

3. “Difference between ‘&’ and ‘and’ in Pandas.” n.d. Stack Overflow. Accessed March 30, 2023. https://stackoverflow.com/questions/54315627/difference between-and-and-in-pandas.

4. “Range Filtering with BETWEEN.” n.d. Peachpit.com. Accessed March 30, 2023. https://www.peachpit.com/articles/article.aspx?p=1276352&seqNum=8.

5. “Pandas DataFrame Count() Method.” n.d. W3schools.com. Accessed March 30, 2023. https://www.w3schools.com/python/pandas/ref_df_count. asp.

6. “Map Configuration and Styling in Python.” n.d. Plotly.com. Accessed March 30, 2023. https://plotly.com/python/map-configuration/.

IX. Acknowledgements

My sincere gratitude goes out to Sir Shanmuga R, our course lecturer for ES 114 Probability, Statistics, and Data Visualization, as well as to all of the TAs who helped me out whenever I needed assistance with this assignment.

First and foremost, I want to express my gratitude to all the online resources that provided me with the necessary resources to finish this task. Also, I want to thank all of the organisations that have helped me by lending me their knowledge and insight in a variety of sectors through a variety of additional channels. Finally, I want to express my gratitude for the chance you offered me to ask questions, learn how to interpret facts, and acquire this incredible skill.