**EE 325: Probability and Random Processes Programming Assignment 1**

**NOTE:**  GitHub repo link: [github.EE325\_Assignment](https://github.com/rookie-apoorv/Assignment_1.git) . Refer to the README.md for easy navigation

**Part-1**

1. **Is the opinion from the 100 people representative of the opinion of India? Explain.**

* The following is the given opinion

|  |  |
| --- | --- |
| State | NO. of students |
| Andhra and Telangana | 41 |
| Maharashtra | 28 |
| Rajasthan | 10 |
| UP and Bihar | 10 |
| West Bengal | 9 |
| Don't Know | 2 |
| Total | 100 |

* + The above opinion is not representative of the opinion of India.
  + This is so because this set of people does not include people from many parts of the country which are known as prominent education centers, like Delhi NCR, and it also does not have any representatives from far eastern states or the south eastern part.

1. **What is your belief about the home state most IITB students? Write the belief from each member of your batch; it is reiterated that there is no shame in being wrong.**

**View of Member2 (Apoorv)**

* My belief is similar to what opinion we have been given in the first question, only that I disagree with the numbers.
* This is my belief:

|  |  |
| --- | --- |
| **State** | **Percentage of students** |
| The Southern States  (Andhra, Telangana, Tamil Nadu, Kerala) | 30% |
| Rajasthan and Delhi NCR | 25% |
| Central and Northern States  (MP, Punjab, Haryana, UP, Bihar) | 20% |
| Maharashtra and Gujarat | 15 % |
| Others | 10% |

* The above belief is based on my personal experience from the people I have met in the institute.

**View of Member1(Hari)**

|  |  |
| --- | --- |
| State | Percentage Students |
| The Southern states | 30% |
| Maharashtra | 30% |
| Rajasthan & Delhi | 20% |
| Others | 20% |

**3.Do the experiment stated in the assignment for K = 10,20,50,100, 200.There is one scatter plot for every combination of K and the method of selecting the K. Now answer the following questions.**

The following is the code used, also added on the GitHub repo.

1. import pandas as pd

2. import matplotlib.pyplot as plt

3. import numpy as np                                 #importing all important libraries

4. import random

5. import statistics

6.

7. data = pd.read\_csv('JEEDemographics.csv')         #reading the demographic file

8.

9. #print(data.head())

10.

11. # define 3 functions each of which returns a list of 50 top 3 fractions for each way of selecting the k people

12. def firstKfromStart(k,rep) :

13.     frac = []

14.     top33 = pd.DataFrame()

15.     for i in range(rep) :

16.         firstkrows = data.head(k)

17.         state\_count = firstkrows['Origin'].value\_counts()

18.         #print(state\_count)

19.         state\_count\_df = state\_count.reset\_index()

20.         state\_count\_df.columns = ['origin','count']

21.         top3 = state\_count\_df.head(3)

22.         top33 = state\_count\_df

23.         fractop3 = (top3['count'].sum())/k

24.         frac.append(fractop3)

25.     top33.to\_excel('Top3.xlsx')

26.     return frac

27. def firstKfromRand(k,rep) :

28.     frac = []

29.     data\_to\_append  = pd.read\_csv('JEEDemographics.csv')

30.     data\_combined = pd.concat([data,data\_to\_append],ignore\_index = True)

31.     for i in range(rep) :

32.         start = random.randint(1,1500)

33.         krowfromrand = data\_combined[start : start+k]

34.         state\_count = krowfromrand['Origin'].value\_counts()

35.         state\_count\_df = state\_count.reset\_index()

36.         state\_count\_df.columns = ['Origin','Count']

37.         top3 = state\_count\_df.head(3)

38.         fractop3 = (top3['Count'].sum())/k

39.         frac.append(fractop3)

40.     return frac

41. def randomk(k,rep) :

42.     frac = []

43.     for i in range(rep) :

44.         new\_data = data.sample(n=k)

45.         state\_count = new\_data['Origin'].value\_counts()

46.         state\_count\_df = state\_count.reset\_index()

47.         state\_count\_df.columns = ['Origin','Count']

48.

49.         top3 = state\_count\_df.head(3)

50.

51.         fractop3 = (top3['Count'].sum())/k

52.         frac.append(fractop3)

53.     return frac

54.

55. #define lists over which we will iterate the for loop

56. k\_list = [10,20,50,100,200,1200]

57. way\_list = [randomk]

58.

59. rep = 50  # no. of repetitions for each way of choosing k

60.

61. xaxis = list(range(1,rep+1))         # defining the xaxis for scatter plot

62. i= 1

63.

64. for way in way\_list :

65.     for k in k\_list :

66.         fig = way(k,rep)

67.

68.         #print(way1)

69.         plt.scatter(xaxis,fig)

70.         plt.title(f'{i} with k={k}')

71.         plt.xlabel('Repition')

72.         plt.ylabel('Fraction of students from top3 states')

73.         plt.savefig(f'{i}\_k={k}.jpeg',format = 'jpeg')

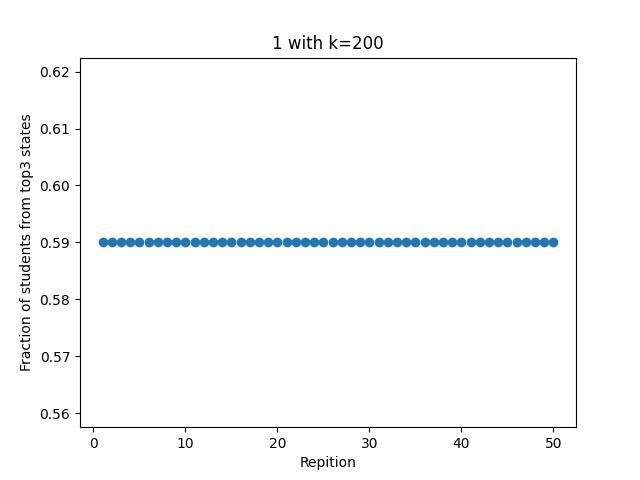
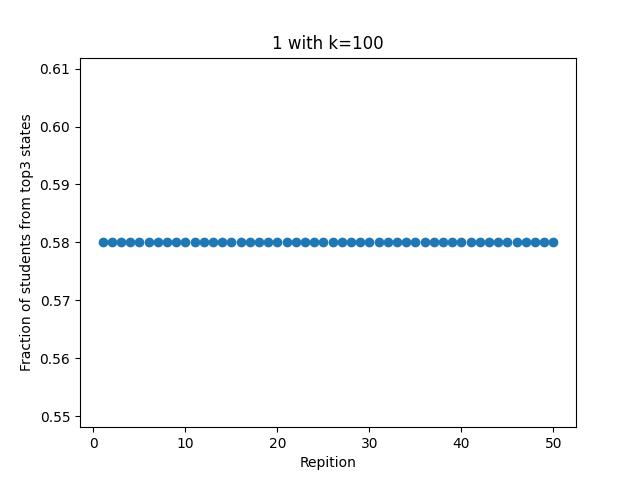
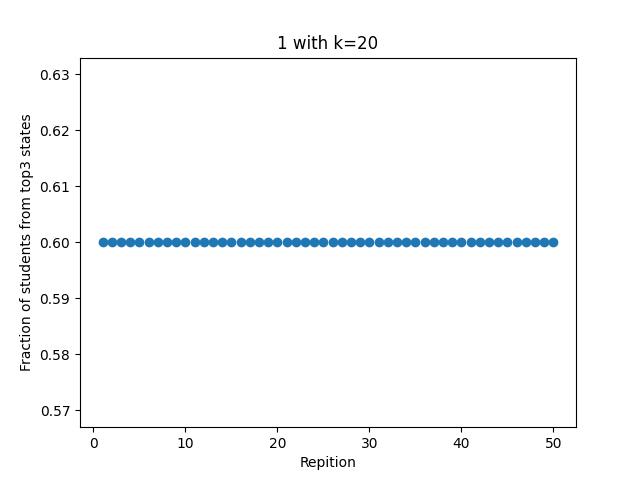
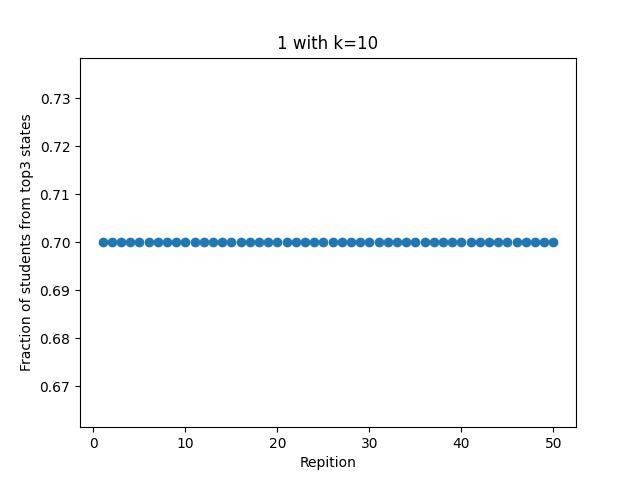
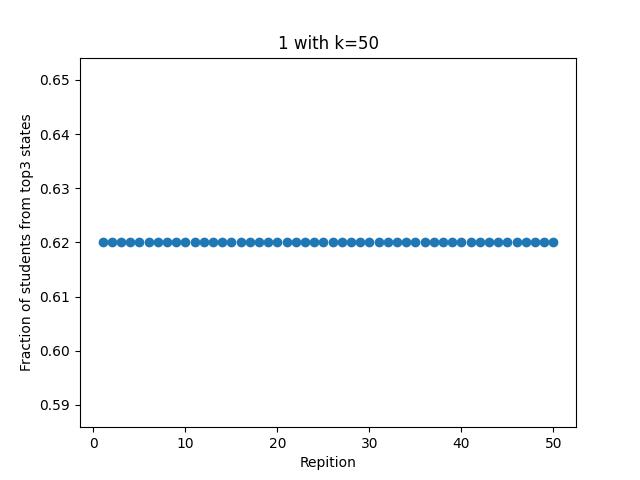
74.         plt.clf()

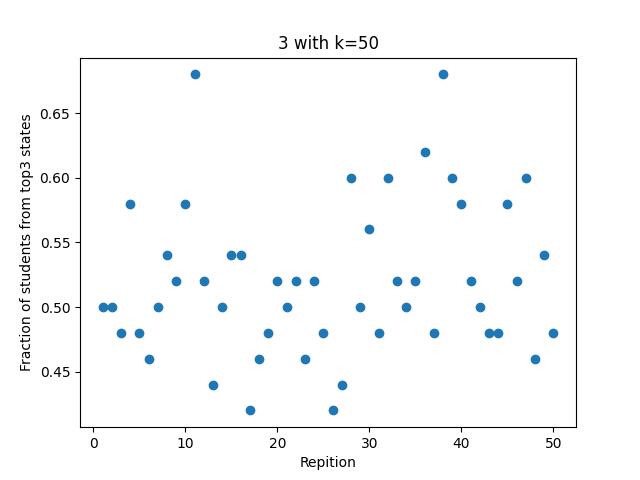
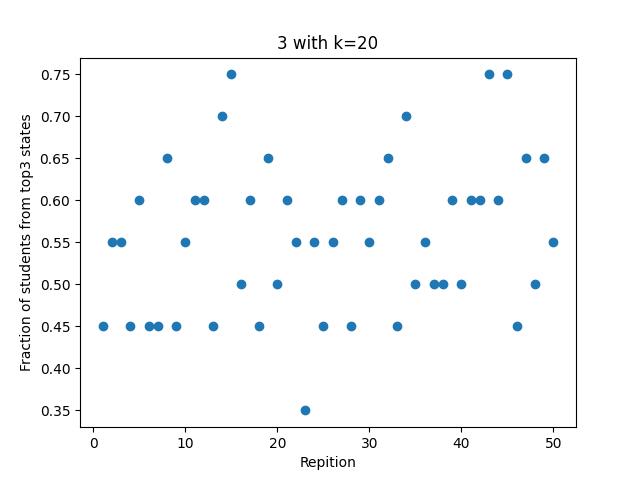
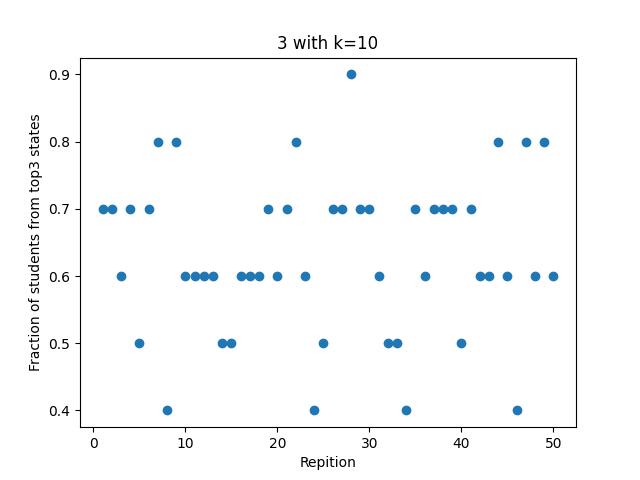
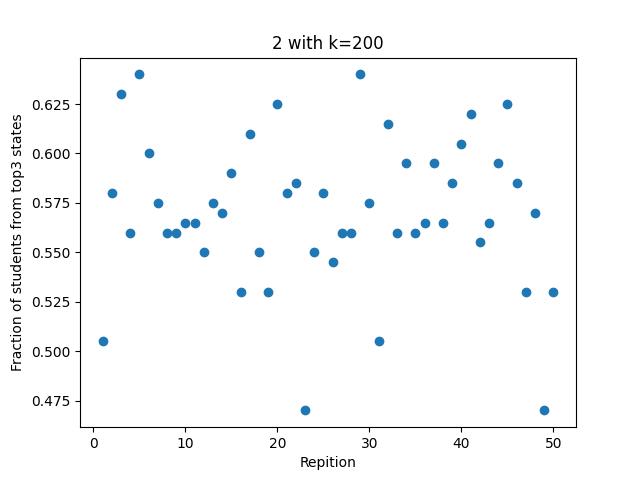
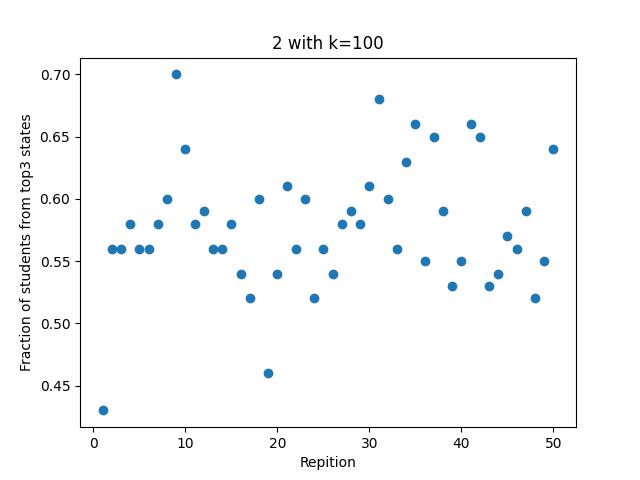
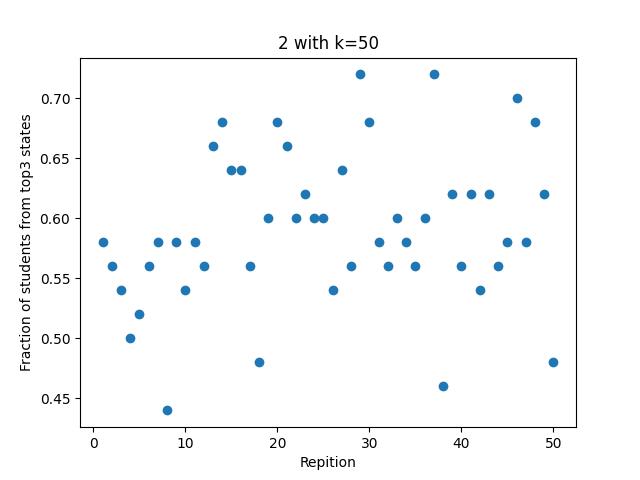
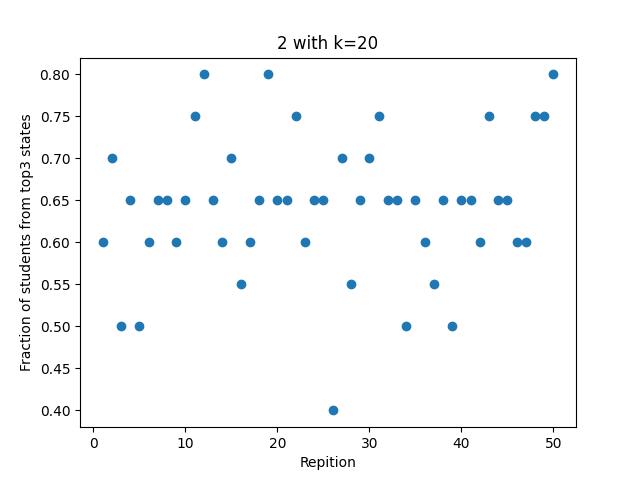
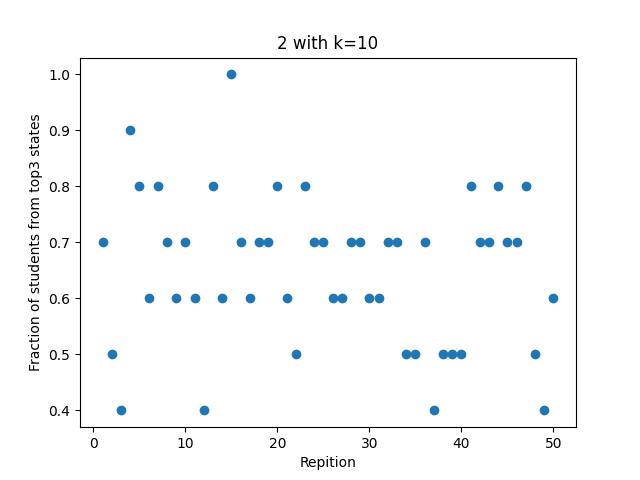
75.         print(np.sum(fig)/50)

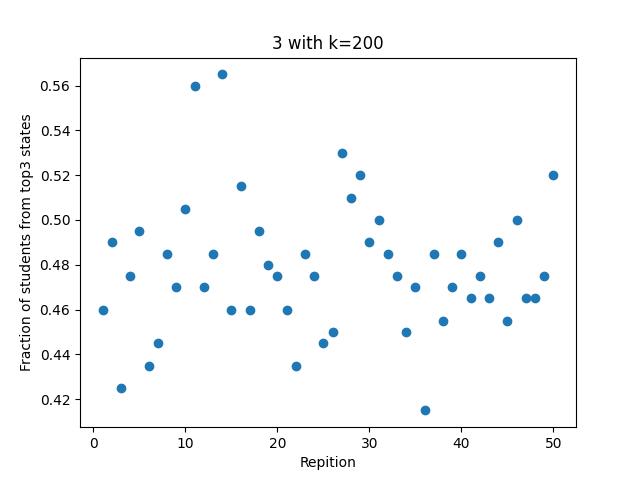
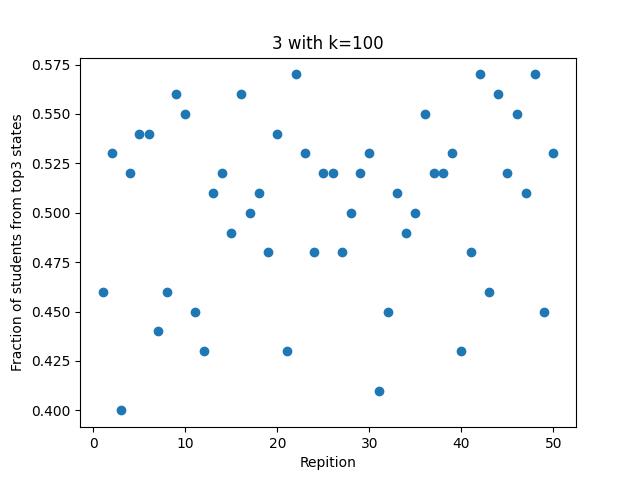
76.         print(statistics.stdev(fig))

77.     i+=1

78.

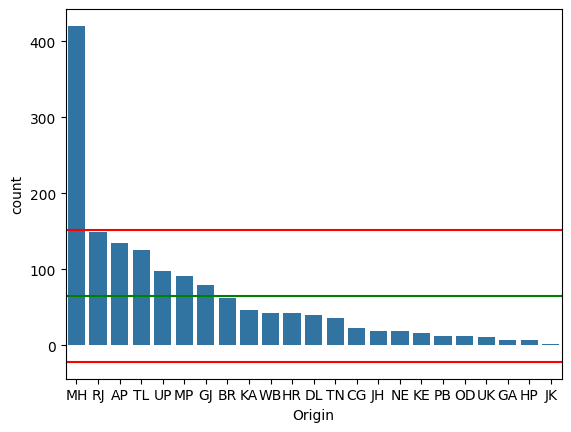
The following are the scatter plots for each way of selecting k people for each k





**NOTE:**  For better clarity these plots along with the code are also uploaded on the GitHub repository whose link is shared at the top of this document.

**3(c)By looking at data of the entire list of 1500 students, what are the top three states and what is the fraction of students from these students?**



The above shows a histogram of no. of students from each state.

* The top 3 states are:

|  |  |  |
| --- | --- | --- |
| State | No. of students | Fraction |
| Maharashtra | 420 | 0.28 |
| Rajasthan | 149 | 0.1 |
| Andhra Pradesh | 135 | 0.09 |
|  |  |  |
| Total | 704 | 0.47 |

**3(d) Each of the fifty repetitions can be seen to be a separate survey. If you could do the survey only once for a given K**

1. **Which of the above three schemes would you use in practice to determine the best guess?**

An important point to note is that our result should be independent of in what order the data is collected.

* First k people: This is not an accurate way of measurement as this might a bias data collection. As we increase k for this model, the fraction obtained approaches the actual value of 0.47, but for this we would have make the value of k very high (maybe even 700 or more). The value greatly depends on the order of data. Hence this is not a good way.
* First K people from a random point: Although this model introduces some degree of randomness to data collection, still taking k subsequent values assigns some weight to the order in which the data was collected. Hence this is also not a good way.
* Random k data points: This has the most randomness, and the values obtained from this can be considered to be very close to the actual value. As you can see from the scatter plots, the range of values for the 5 repetitions contains the actual value and this range keeps centering towards the actual value as we increase k.

Hence, I would choose the 3rd way of selecting random k people, as this may give the best results.

1. **If you were allowed to choose the value of K, what value would you choose? And how sure would you be of the actual values of the average? What kind of quantitative measure would you use to describe your sureness of the estimate from the single survey of K samples?**

* We are choosing the third way of sample selection so we will focus on the results from that only.
* What is the situation?
  + We don’t know the actual value (i.e. the value 0.47).
  + Now we do the survey once (i.e. we get one of the values from the 50 repetitions)
  + We need to determine for what k would we be very sure of the correctness of the obtained value.
  + The following is the table consisting of the average of the 50 repetitions for each value of k in model 3.

|  |  |
| --- | --- |
| Value of K | Average of 50 repetitions |
| 10 | 0.65 |
| 20 | 0.58 |
| 50 | 0.52 |
| 100 | 0.49 |
| 200 | 0.48 |
| 1200 | 0.47 |

* What we are trying to do here is called sampling distribution in statistics.
* In this we don’t know the actual average of the population, so we take various samples, take the average of each sample and then take the expected value of these averages.
* We now hypothesize that the actual mean should be very close to this expected value.
* So, what we have on the right in this table is that expected value, as you can see increasing k has a very nice effect on it and it approaches the actual value.
* But is increasing k after a certain point beneficial?? Not really as you can see from the table.
* Hence 200 is a good enough k, where we are very close to the actual value with not much cost for data collection.

**How sure are we?**

* Our sureness depends on the fact of how many values of those 50 repetitions for k=200 is actually close to the expected value of 0.48 of k=200.
* For this our quantitative measurement would be standard deviation of this data.
* We calculate the standard deviation and it comes out to be
  + Sigma = 0.03
* In statistics we say that with 95% confidence the value lies in the interval expected value +- 2\*sigma.
* So, we say with 95% confidence that our value is in 2\*sigma proximity of the actual value.

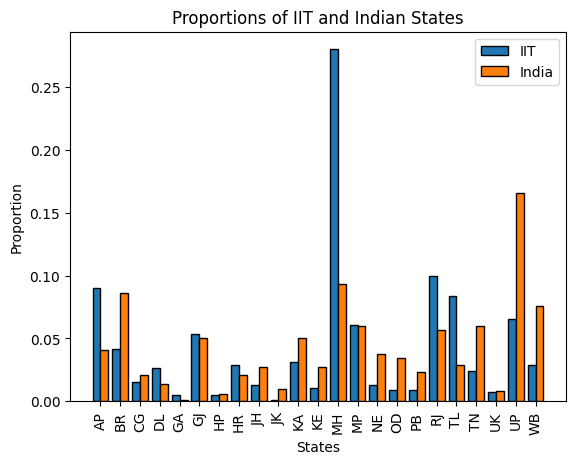
**PART 2:**

**Code:** <https://colab.research.google.com/drive/1BG4FRowG5hxPH3Uxfw-rYMaFYRd_IeDs?usp=sharing>

1. **Is the IITB UG population a good sample of India at the granularity of states? If not, how badly is it skewed in favor of some states? Define your own measure for skew and apply to the data that you have. Justify the measure.**

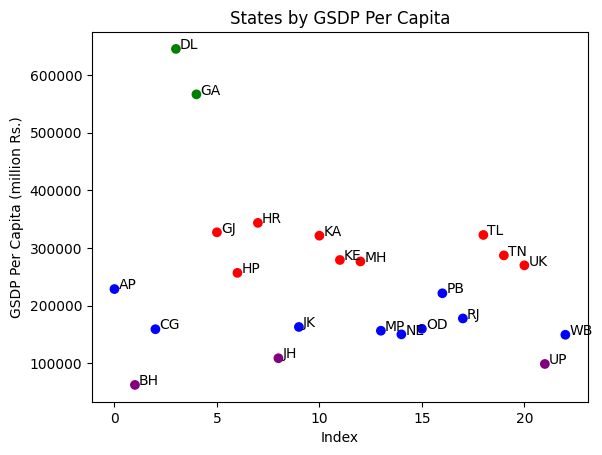
* The Below bar graph shows the population of each state and the number of students in IITB from that state. The data is scaled appropriately.
* Inferences:
  + The third column of below table gives the factor by which we have more/ a smaller number of students in IITB as compared to the actual population of the state.
  + Observe a high factor for Maharashtra, Delhi, Rajasthan, Andhra Pradesh, and several negative factors for other states.
  + This clearly means that IITB population is not a good sample of India at the granularity of states.
  + It is badly skewed towards certain states.

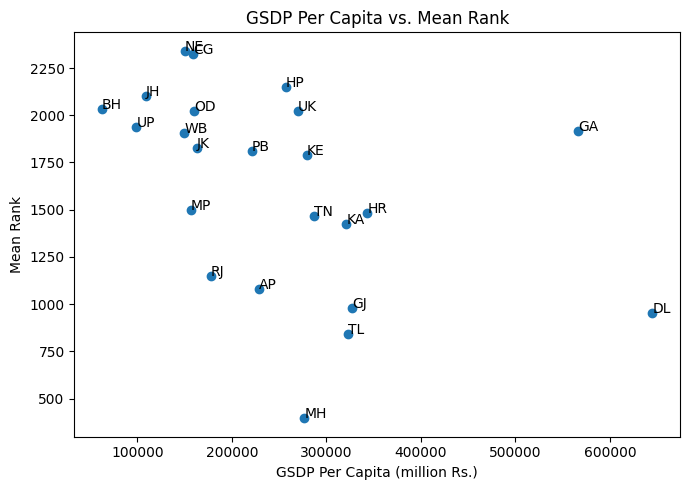
|  |  |  |  |
| --- | --- | --- | --- |
| **State** | **Total Population (scaled to 1500)** | **IITB population** | **(Total – IITB)/IITB** |
| Uttar Pradesh | 248 | 98 | -0.61 |
| Maharashtra | 140 | 420 | 2.01 |
| Bihar | 129 | 62 | -0.52 |
| West Bengal | 113 | 43 | -0.62 |
| Madhya Pradesh | 90 | 91 | 0.01 |
| Tamil Nadu | 90 | 36 | -0.60 |
| Rajasthan | 85 | 149 | 0.75 |
| Karnataka | 76 | 47 | -0.38 |
| Gujarat | 75 | 80 | -0.07 |
| Andhra Pradesh | 62 | 135 | 1.19 |
| North-East States | 57 | 19 | -0.66 |
| Odisha | 52 | 13 | -0.75 |
| Telangana | 44 | 126 | 1.90 |
| Kerala | 42 | 16 | -0.61 |
| Jharkhand | 41 | 19 | -0.54 |
| Punjab | 34 | 13 | -0.62 |
| Chhattisgarh | 32 | 23 | -0.28 |
| Haryana | 32 | 43 | 0.36 |
| Delhi | 21 | 40 | 0.92 |
| Jammu and Kashmir | 15 | 2 | -0.87 |
| Uttarakhand | 13 | 11 | -0.12 |
| Himachal Pradesh | 9 | 7 | -0.18 |
| Goa | 2 | 7 | 2.86 |
|  |  |  |  |



1. **Considering the population and the per capita income of the states, is the distribution of the student body among the states/regions fair? Once again, define your own measure for skew fairness and apply to the data that you have. Justify the measure.**

The following are two graphs which will aide us in answering the above question



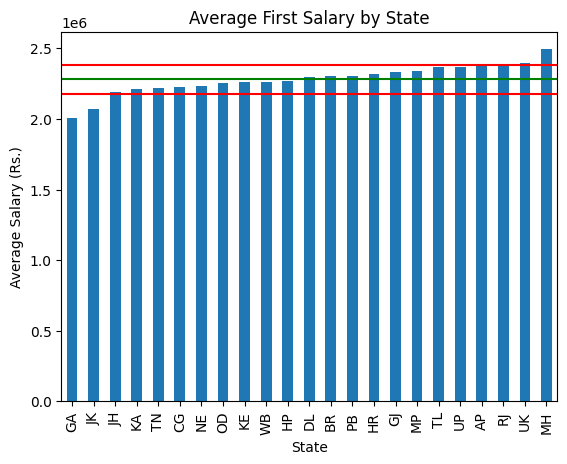


* + The first graph from above is a mapping of states vs Gross State Domestic Product per capita. In this graph, we can see that there is great economic disparity. I have also tried to divide them into four groups,
    - **Group 1(high GDP):** Delhi, Goa, Karnataka, Haryana, Gujarat, Telangana.
    - **Group 2(average GDP):** Tamil Nadu, Maharashtra, Kerala, Uttar Khand, Himachal Pradesh, Andhra Pradesh, Punjab.
    - **Group 3(low GDP):** North-East, Chandigarh, West Bengal, Odisha, Jammu and Kashmir, Madya Pradesh, Rajasthan.
    - **Group 4(lowest GDP):** Jharkhand, Uttar Pradesh, Bihar.
  + After this, it can be clearly seen that students from states with High GSDP are more likely to get into IIT Bombay than students from other states. The exception to this rule is Maharashtra.
  + The data is highly skewed towards Maharashtra, this can be credited to IIT Bombay being in Maharashtra. Thus, students from Maharashtra are far more likely to get into IIT Bombay.
  + The second graph is Gross State Domestic Product Per capita vs JEE Rank. This Graph further solidifies the claim made during the inference of the first graph. The trend is such that (with a few exceptions) The students from states with lesser GSDP per capita are more likely to get a larger rank, while students from states with higher GSDP per capita are more likely to get a better rank.
  + Again, as pointed in Question1, the data is skewed when seen with respect to population or with respect to GSDP per capita.

1. **How strongly does the state affect the JEE rank, the graduating CPI, and the 1st salary?**

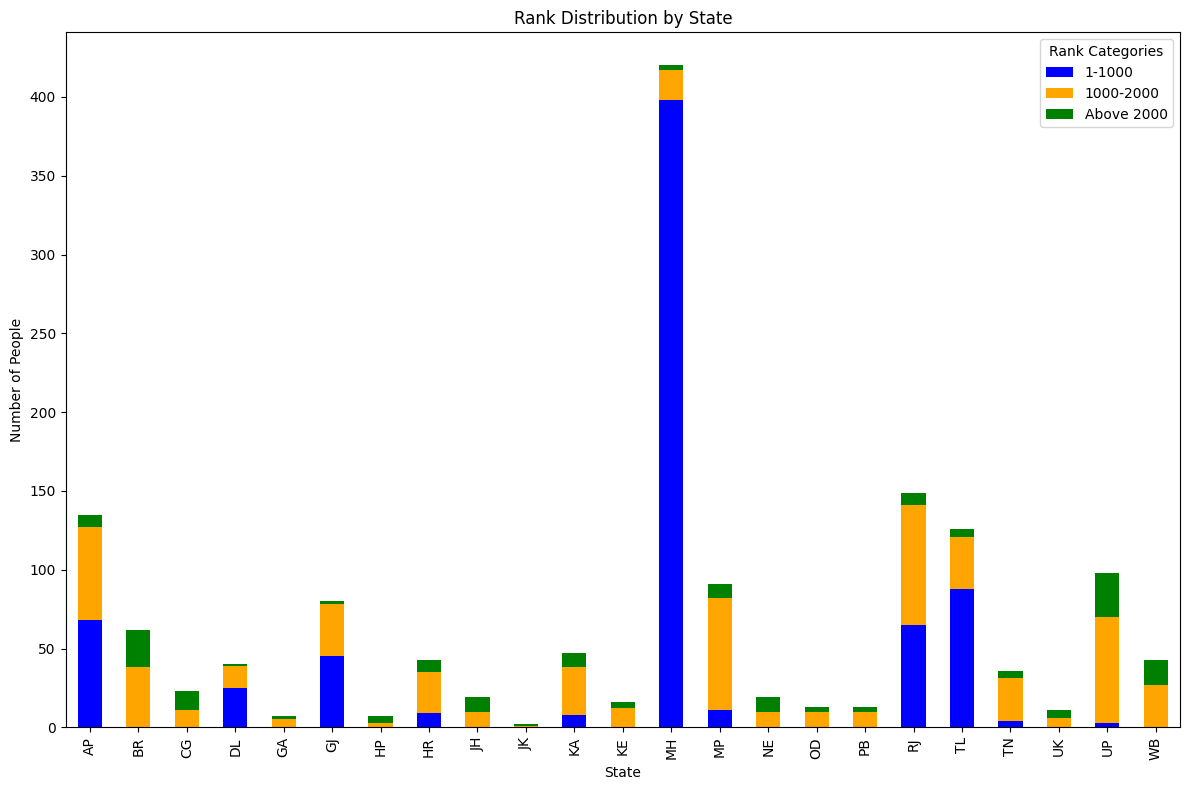
**Code for this in:** [**https://colab.research.google.com/drive/1kW9gLzk2fI-l4qDLglAIfawlxuWuJqZj?usp=sharing**](https://colab.research.google.com/drive/1kW9gLzk2fI-l4qDLglAIfawlxuWuJqZj?usp=sharing)

**Effect of State on 1st salary:**

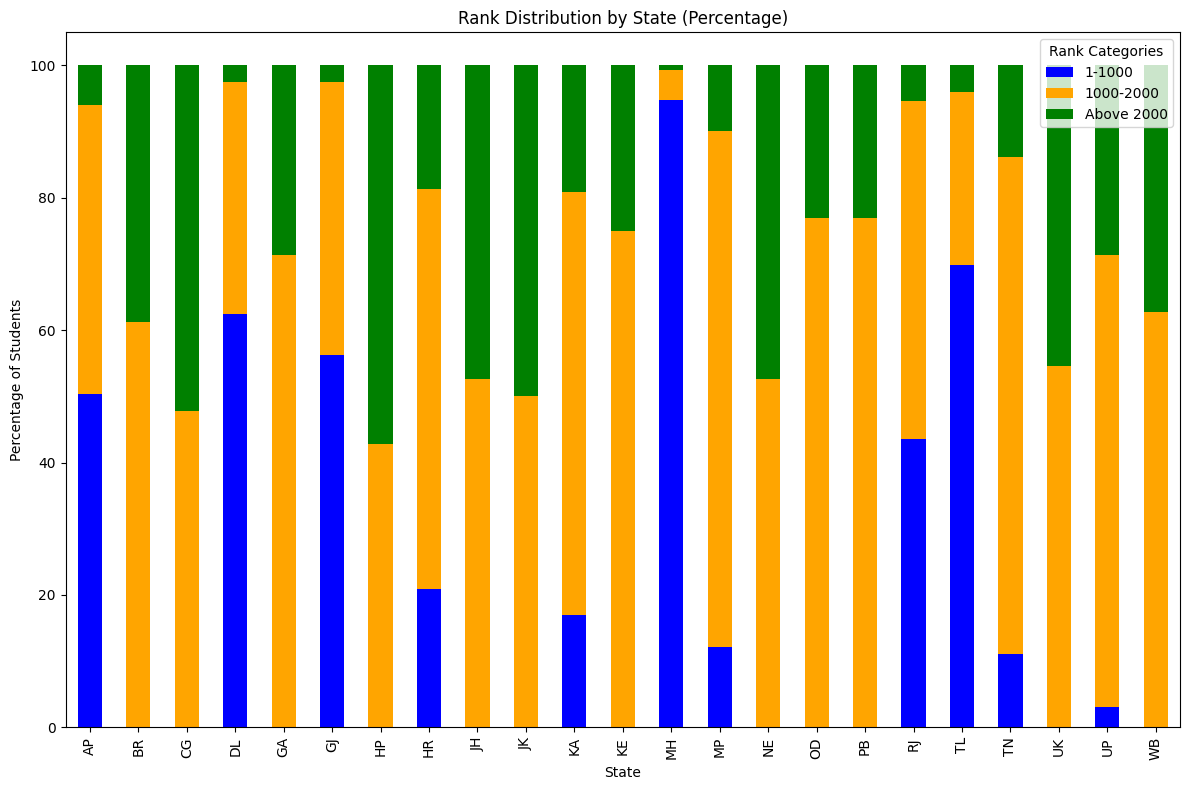


* The above graph was made by taking the average salary (the sum of salaries earned by students by students from a state divided by the number of students from said state in IIT Bombay) and plotting them with respect to the states.
* It can be seen that except a few cases there is not much skew in the first salary with respect to say. So, it is safe to say that state doesn’t affect first salary a lot.

**Effect of State on JEE Rank:**

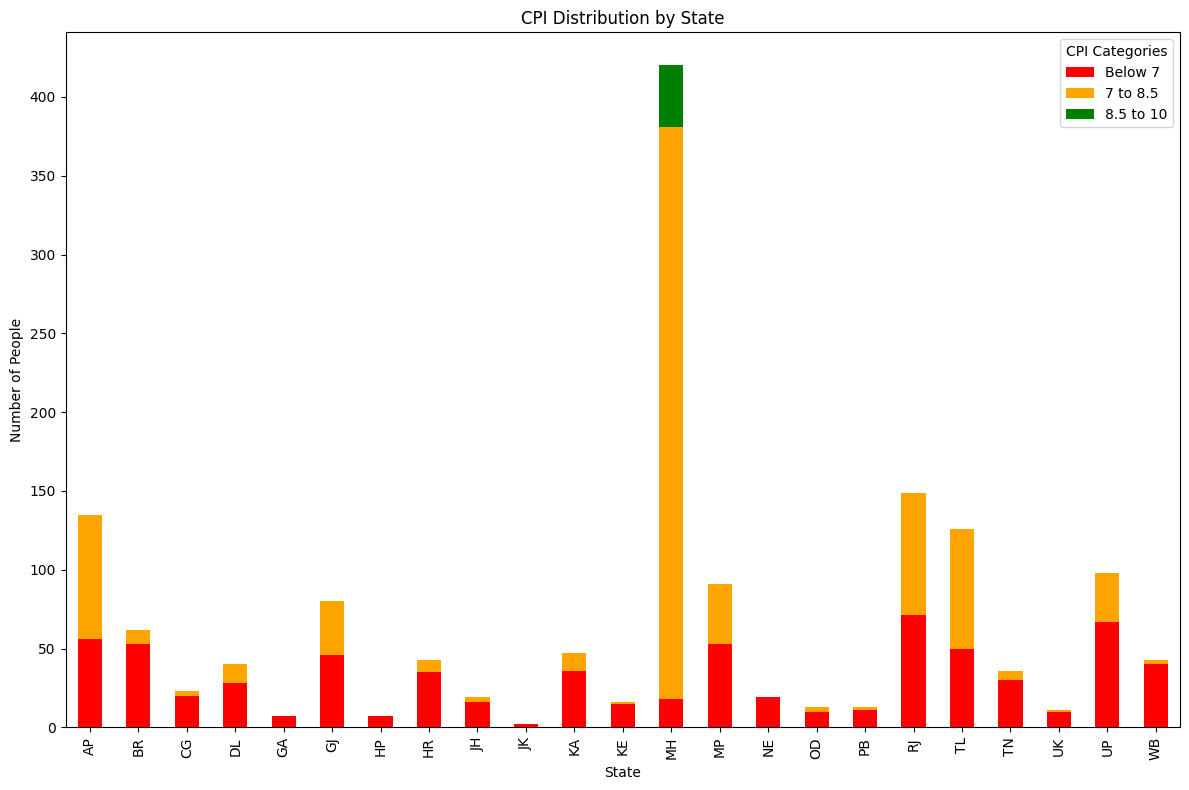


* The above graph was made by plotting the numbers of students from each state from a certain rank category, the rank categories being:
  + I – (1-1000) (Blue color)
  + II – (1000-2000) (Yellow color)
  + III – (1000-2000) (Green color)
* Note: This graph is for all students from the state and is not normalized
* Although the graph is not normalized, we can clearly see that a few states (specifically states that were grouped in to the category of states with “good”- GSDP per capita) have more people with better ranks.
* Again, here too due to Maharashtra’s home state advantage it stays an anomaly in the data.

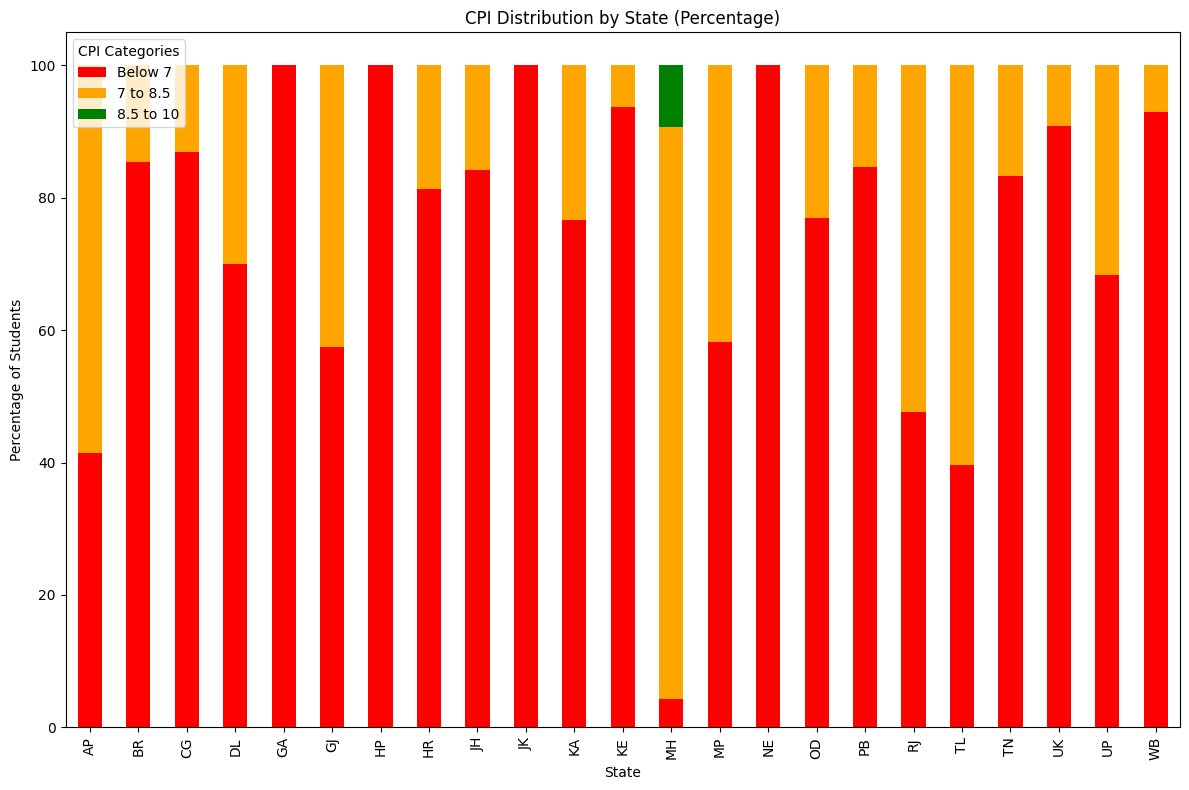


* The above graph was made by plotting the numbers of students from each state from a certain rank category out of the total number of students from that said state in IIT Bombay, the rank categories being:
  + I – (1-1000) (Blue color)
  + II – (1000-2000) (Yellow color)
  + III – (1000-2000) (Green color)
* Note: This graph is normalized so it gives a better view of how rank is distributed among states
* Since the graph is normalized, we can clearly see better that a few states (again states that were grouped into the category of states with “good”- GSDP per capita) have more people with better ranks.
* Here too Maharashtra stays an anomaly.

**Effect of State on CPI**



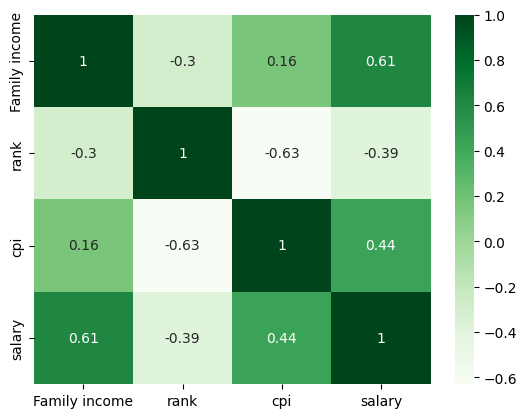
* The above graph was made by plotting the numbers of students from each state from a certain cpi category, the cpi categories being:
  + I – (10 – 8.5) (Green color)
  + II – (8.5 – 7) (Yellow color)
  + III – (below 7) (Red color)
* Note: This graph is for all students from the state and is not normalized
* It can see that a few states (specifically states that were grouped into the category of states with “good”- GSDP per capita) have more students with better cpi.
* Here more students from up have average cpi as opposed to the trend seen in other data points.



* The above graph was made by plotting the numbers of students from each state from a certain cpi category out of the total number of students from that said state in IIT Bombay, the CPI categories being:
  + I – (10 – 8.5) (Green color)
  + II – (8.5 – 7) (Yellow color)
  + III – (below 7) (Red color)
* Note: This graph is normalized so it gives a better view of how CPI is distributed among states
* It can see that a few states (but no real trend that can be associated) have more students with better CPI. Also, only students from Maharashtra have more than 8.5 CPI.

1. **How strongly does the family income affect the JEE rank, the graduating CPI, and the 1st salary?**

Link to code: <https://colab.research.google.com/drive/1wAC8641Iq7UYoB03t9QVdeCPzn0ssqPM?usp=sharing>



The correlation coefficient in statistics in statistics is a method to measure how related two variables are. It is calculated as the covariance of said two variables divided by the square root of the product of their standard deviation. The correlation matrix is a nxn matrix made of correlation coefficients of n variables. The correlation coefficient can range from -1 to 1 (negative of the number to being the number itself)

* The correlation between Family income and first salary is 0.61
* The correlation between Family income and Graduating CPI is 0.16
* The correlation between Family income and Rank is 0.3(since a lesser rank is better, we take the negative sign out).