Project Report

Name: Akshata More

Topic: College Score Cards

Data Source: https://data.world/xprizeai-edu/college-scorecard/workspace/file?filename=merged_2013_PP.csv (data.world)

Data Set Link:https://query.data.world/s/55ejgjtervq7xijexg3r74rvdc3xac?dws=00000

Analysis Technique Used: Principal Component Analysis (PCA) + Cluster Analysis

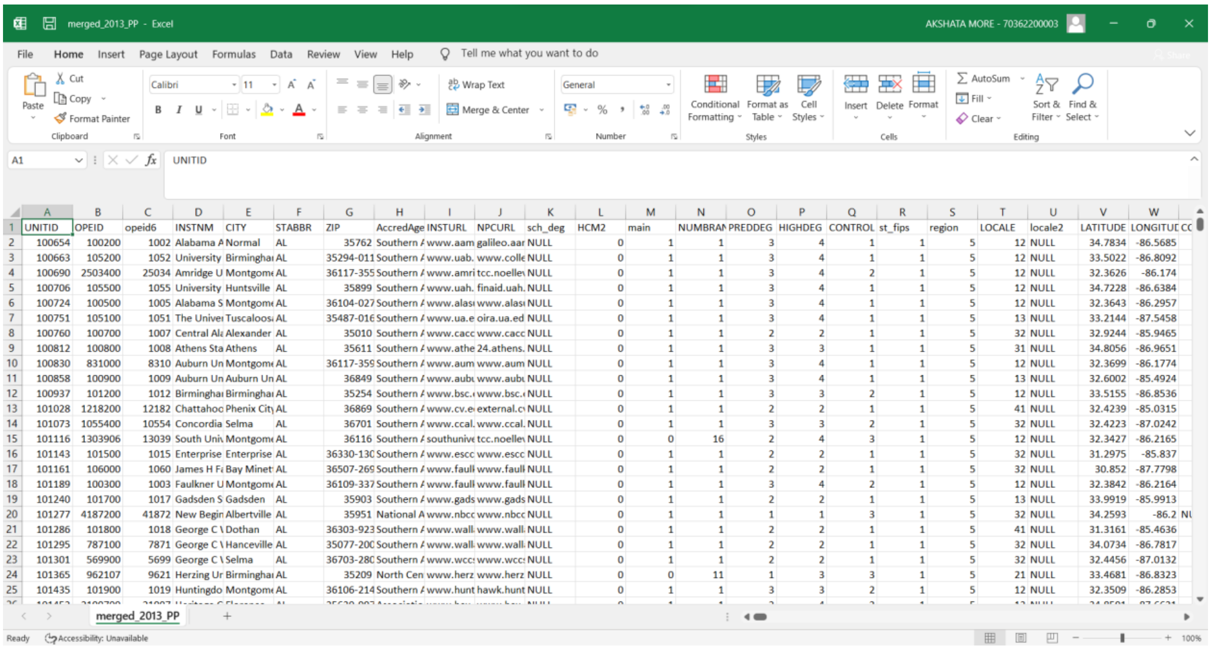
Problem Statement: Using cluster analysis to identify the groups of characteristically similar schools in the College Scorecard dataset.

1. INTRODUCTION:

- This data set of “College Scorecard” was primarily released by U.S. Department of Education which is now available on https://data.world/.
- **Motto:** The Department of Education is focused on ensuring that parents, students, and policymakers are able to use its publicly available data to take to take better college decisions.
- Principal Component Analysis or PCA is a statistical technique that is used to reduce dimensionality of data and to obtain suitable principal components, which explains the maximum variance.
- To select the appropriate data from the vast data set to perform analyses like Cluster Analysis, PCA is used as a prerequisite step.
- Cluster Analysis is a statistical technique used to group similar objects into categories or clusters.

2. Steps Involved:

a. I chose the College Scorecard data set of the year 2013 in csv format



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
1	UNITID	OPEID	opeid6	INSTNM	CITY	STABBR	ZIP	AccredAge	INSTURL	NPCURL	sch_deg	HCM2	main	NUMBRAN	PREDDEG	HIGHDEG	CONTROL	st_fips	region	LOCALE	locale2	LATITUDE	LONGITUD	CC
2	100654	100200	1002	Alabama A	Normal	AL	35762	Southern A	www.aam	galileo.aar	NULL	0	1	1	3	4	1	1	5	12	NULL	34.7834	-86.5685	
3	100663	105200	1052	University	Birminghar	AL	35294-011	Southern A	www.uab.	www.colle	NULL	0	1	1	3	4	1	1	5	12	NULL	33.5022	-86.8092	
4	100690	2503400	25034	Amridge U	Montgome	AL	36117-355	Southern A	www.amri	tcc.noelle	NULL	0	1	1	3	4	2	1	5	12	NULL	32.3626	-86.174	
5	100706	105500	1055	University	Huntsville	AL	35899	Southern A	www.uah.	finaid.uah.	NULL	0	1	1	3	4	1	1	5	12	NULL	34.7228	-86.6384	
6	100724	100500	1005	Alabama S	Montgome	AL	36104-027	Southern A	www.alas	www.alas	NULL	0	1	1	3	4	1	1	5	12	NULL	32.3643	-86.2957	
7	100751	105100	1051	The Univer	Tuscaloosa	AL	35487-016	Southern A	www.ua.e	oira.ua.ed	NULL	0	1	1	3	4	1	1	5	13	NULL	33.2144	-87.5458	
8	100760	100700	1007	Central Ala	Alexander	AL	35010	Southern A	www.cacc	www.cacc	NULL	0	1	1	2	2	1	1	5	32	NULL	32.9244	-85.9465	
9	100812	100800	1008	Athens Sta	Athens	AL	35611	Southern A	www.athe	24.athens.	NULL	0	1	1	3	3	1	1	5	31	NULL	34.8056	-86.9651	
10	100830	831000	8310	Auburn Un	Montgome	AL	36117-359	Southern A	www.aum	www.aum	NULL	0	1	1	3	4	1	1	5	12	NULL	32.3699	-86.1774	
11	100858	100900	1009	Auburn Un	Auburn Un	AL	36849	Southern A	www.aubu	www.aubu	NULL	0	1	1	3	4	1	1	5	13	NULL	32.6002	-85.4924	
12	100937	101200	1012	Birminghar	Birminghar	AL	35254	Southern A	www.bsc.	www.bsc.	NULL	0	1	1	3	3	2	1	5	12	NULL	33.5155	-86.8536	
13	101028	1218200	12182	Chattahoc	Phenix City	AL	36869	Southern A	www.cv.e	external.c	NULL	0	1	1	2	2	1	1	5	41	NULL	32.4239	-85.0315	
14	101073	1055400	10554	Concordia	Selma	AL	36701	Southern A	www.ccal.	www.ccal.	NULL	0	1	1	3	3	2	1	5	32	NULL	32.4223	-87.0242	
15	101116	1303906	13039	South Univ	Montgome	AL	36116	Southern A	southunive	tcc.noelle	NULL	0	0	16	2	4	3	1	5	12	NULL	32.3427	-86.2165	
16	101143	101500	1015	Enterprise	Enterprise	AL	36330-130	Southern A	www.escc	www.escc	NULL	0	1	1	2	2	1	1	5	32	NULL	31.2975	-85.837	
17	101161	106000	1060	James H Fa	Bay Minett	AL	36507-269	Southern A	www.faulk	www.faulk	NULL	0	1	1	2	2	1	1	5	32	NULL	30.852	-87.7798	
18	101189	100300	1003	Faulkner U	Montgome	AL	36109-337	Southern A	www.faulk	www.faulk	NULL	0	1	1	3	4	2	1	5	12	NULL	32.3842	-86.2164	
19	101240	101700	1017	Gadsden S	Gadsden	AL	35903	Southern A	www.gads	www.gads	NULL	0	1	1	2	2	1	1	5	13	NULL	33.9919	-85.9913	
20	101277	4187200	41872	New Begin	Albertville	AL	35951	National A	www.nbcc	www.nbcc	NULL	0	1	1	1	1	3	1	5	32	NULL	34.2593	-86.2	NU
21	101286	101800	1018	George C \	Dothan	AL	36303-923	Southern A	www.wall	www.wall	NULL	0	1	1	2	2	1	1	5	41	NULL	31.3161	-85.4636	
22	101295	787100	7871	George C \	Hanceville	AL	35077-200	Southern A	www.wall	www.wall	NULL	0	1	1	2	2	1	1	5	32	NULL	34.0734	-86.7817	
23	101301	569900	5699	George C \	Selma	AL	36703-280	Southern A	www.wcc	www.wcc	NULL	0	1	1	2	2	1	1	5	32	NULL	32.4456	-87.0132	
24	101365	962107	9621	Herzing Ur	Birminghar	AL	35209	North Cen	www.herz	www.herz	NULL	0	0	11	1	3	3	1	5	21	NULL	33.4681	-86.8323	
25	101435	101900	1019	Huntingdo	Montgome	AL	36106-214	Southern A	www.hunt	hawk.hunt	NULL	0	1	1	3	3	2	1	5	12	NULL	32.3509	-86.2853	

b. Loaded the csv into Google Colab and Imported Libraries.

c. Computed description of the data set.

d. Data Preprocessing: Data Cleaning and Assessing.

i. I have selected the following variables, and created a new dataframe , “df_clean”:

UNITID:

UNITID which is the unique identification number

INSTNM:

institution’s name

CITY, STABBR:

NUMBRANCH:

the number of branch campuses

HIGHDEG:

Highest award

PREDDEG:

Predominant undergraduate award

CONTROL:

institution’s governance structure is public, private nonprofit, or private for-profit.

DISTANCEONLY:

distance education-

TUITFTE:

The net tuition revenue per full-time equivalent (FTE)

AVGFACSAL:

average faculty salary

ADM_RATE_ALL:

Fall admissions rate, represents the admissions rate across all campuses

SATVR25, SATVR75, SATMT25, SATMT75, ACTCM25, ACTCM75:

The files include the 25th and 75th percentiles of SAT

UGDS:

the number of degree/certificate-seeking undergraduates

PCTFLOAN:

the share of undergraduate students who received federal loans in a given year.

CDR3:

The three-year cohort default rate (CDR3) represents a snapshot in time.

e. Checked for null values and cleaned the na values.

f. Many Categorical variables have been encoded as Numeric, as evident from the data set. I converted those back for following variables:

i. STABBR: converted column to categorical type

ii. HIGHDEG and PREDDEG: mapped the numerical value to categorical value

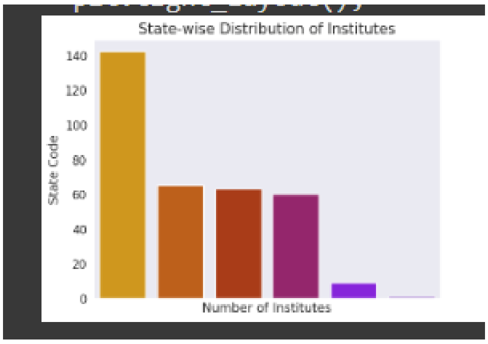
iii. CONTROLD

iv. DISTANCEONLY

v. UGDS

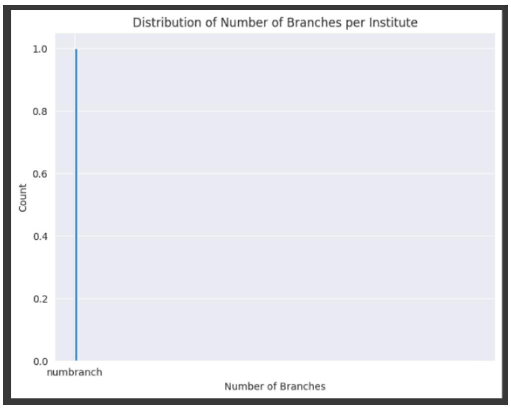
g. Created some graphs/charts for data visualization to answer questions like:

i. How are institutes spread out across the US?

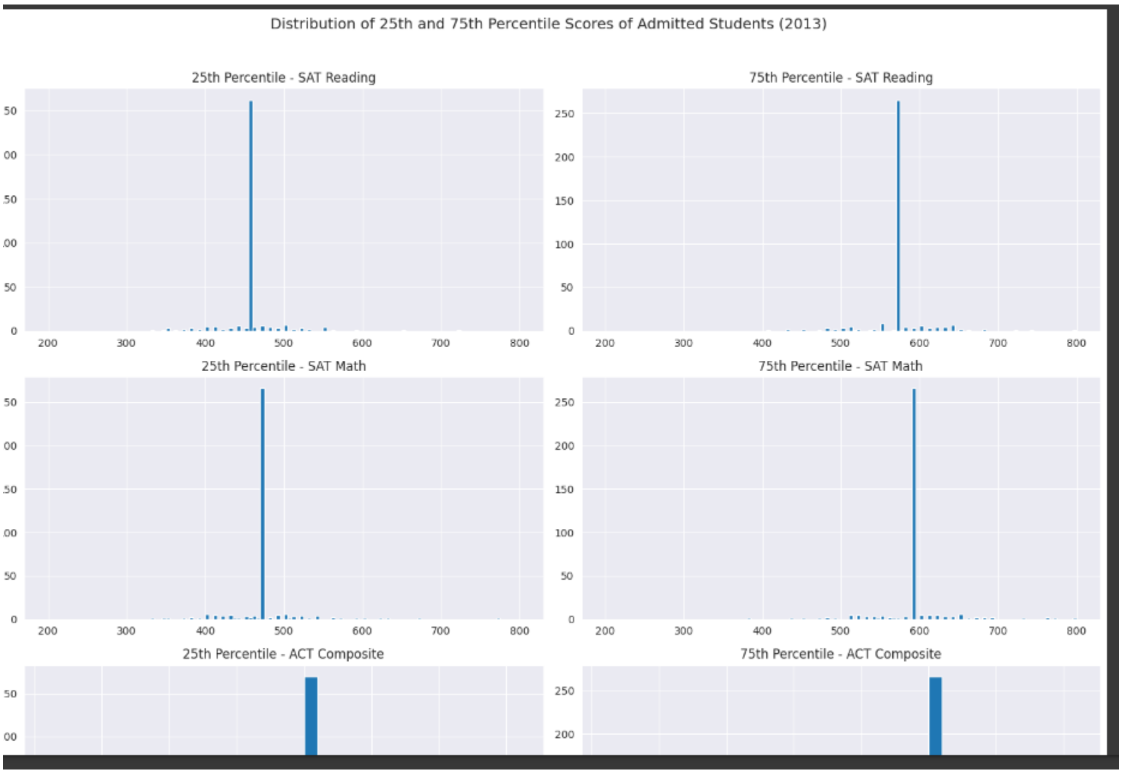


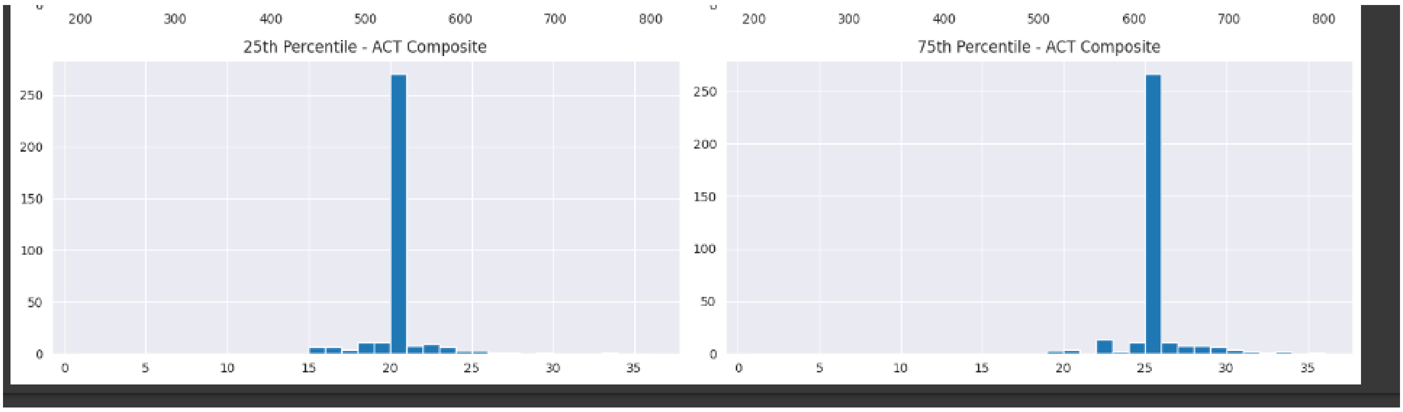
1. The state with code 140 has by far the greatest number of institutes.
2. The states with code 55 to 60 have closely competing numbers.

ii. What is the median number of branches per Institute?

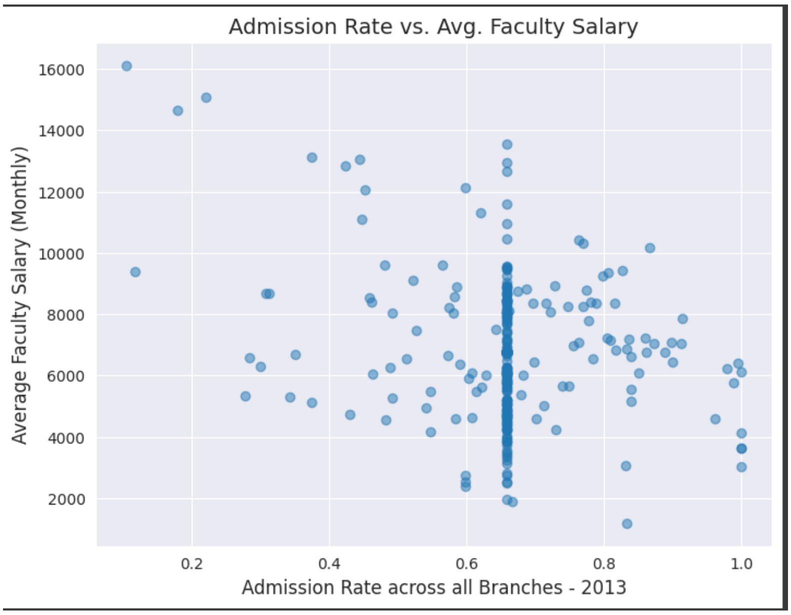


iii. What makes for a good or bad SAT or ACT score? : Distribution of 25th and 75th percentile Scores of Admitted Students

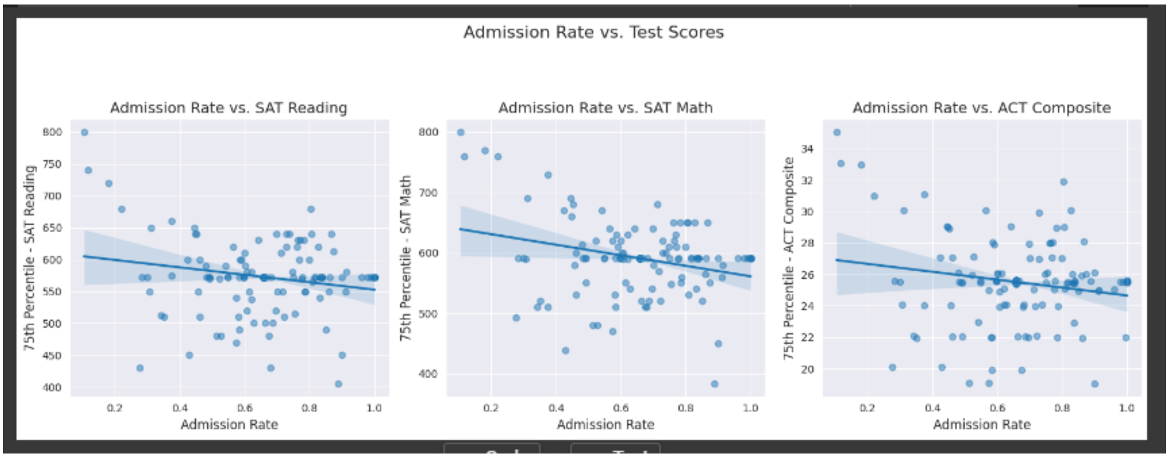




iv. How does admission rate of an institute affect the average faculty salary?



v. One might expect that institutes having a high 75th percentile of test scores for admitted students, would have low admission rates. Does the data support this assumption?

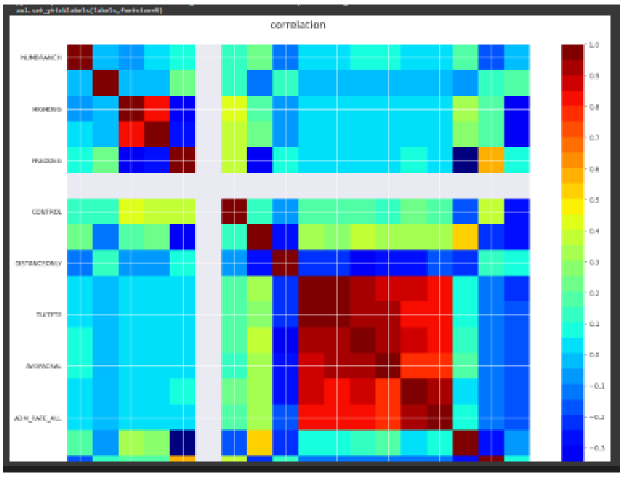


h. Plotted covariance matrix – to check independency

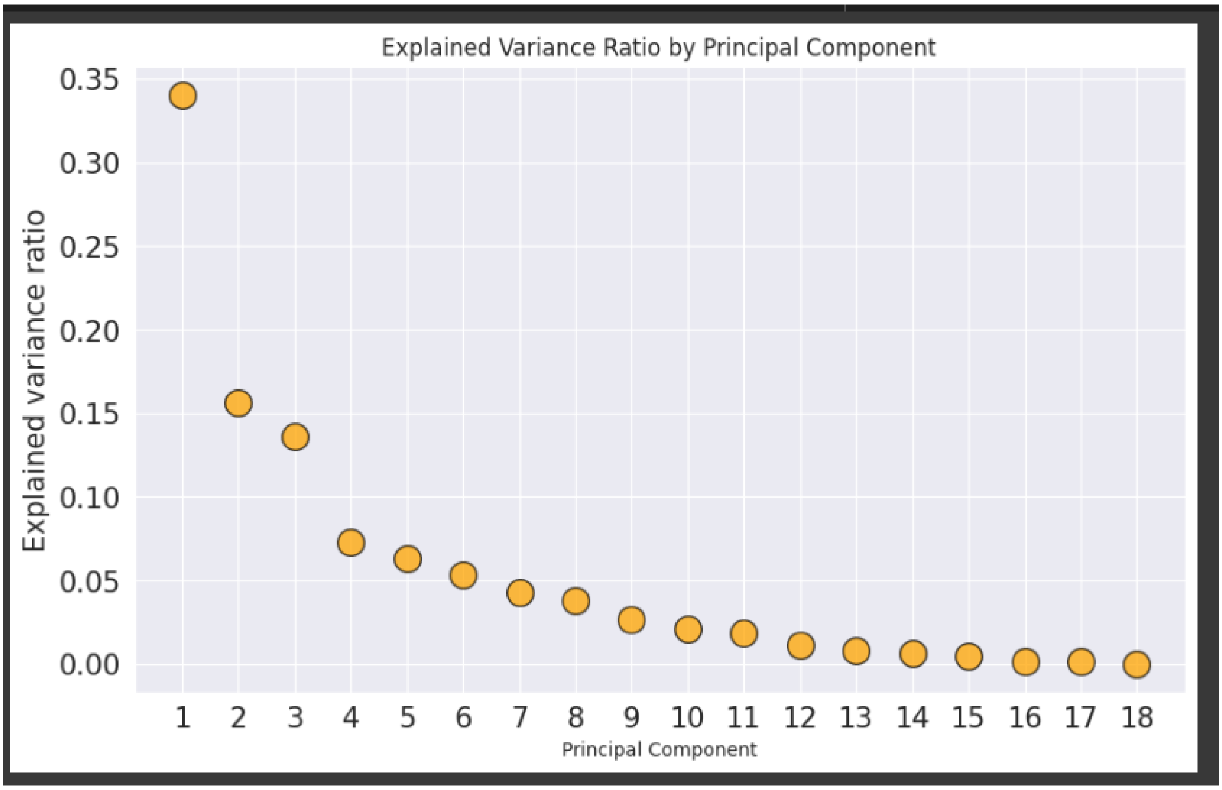
i. Performed PCA

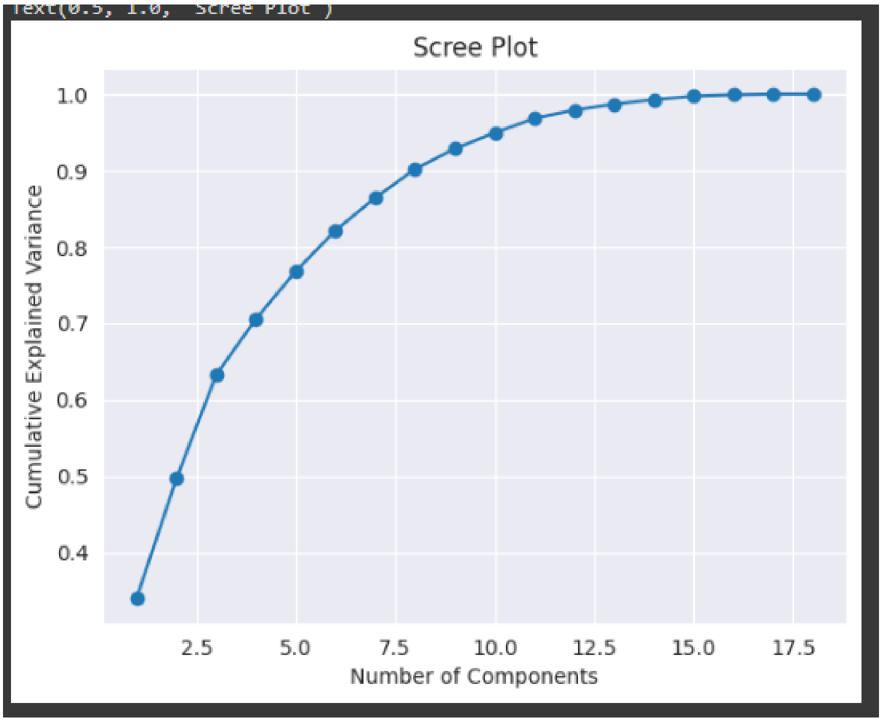
j. Performed Cluster Analysis

3. Results :

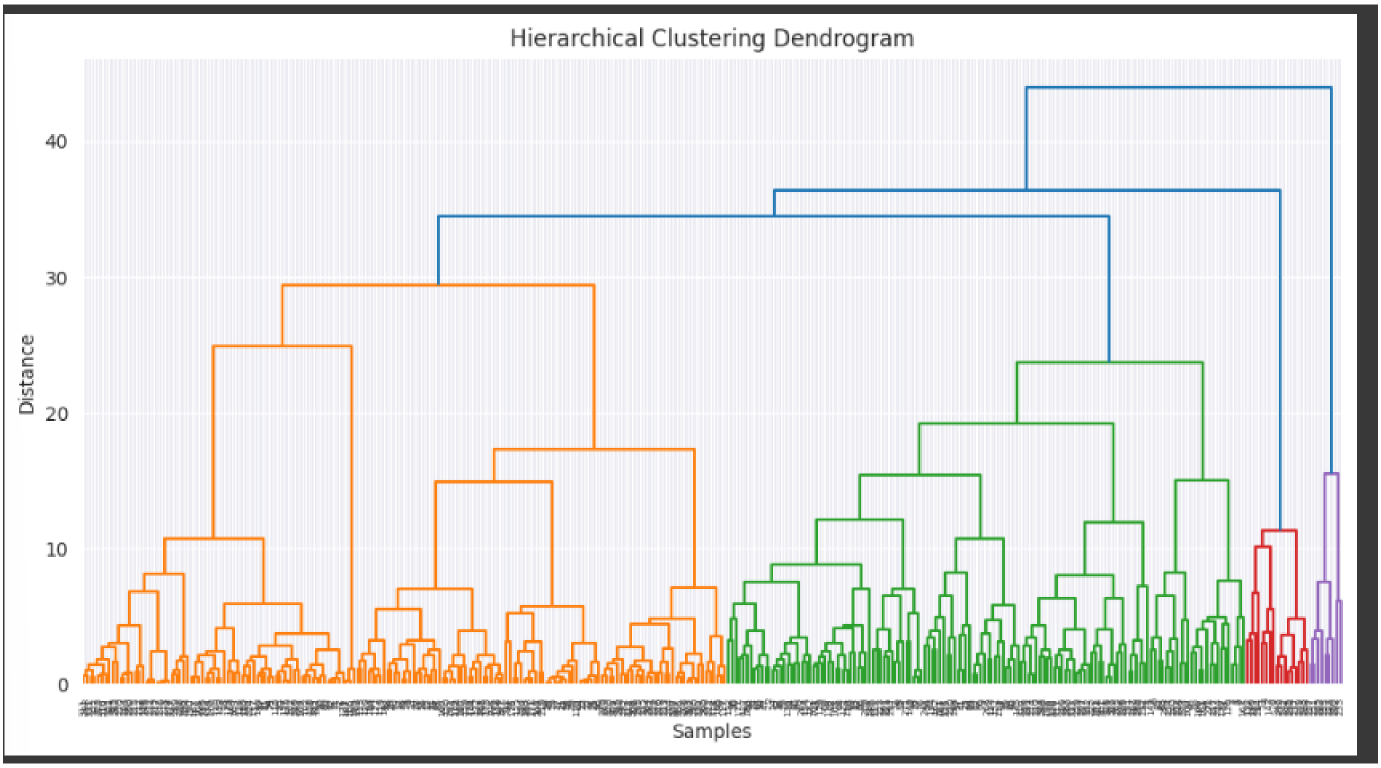


```
ax1.set_yticklabels(labels,fontsize=8)
```





```
Text(0.5, 1.0, 'Scree Plot')
```



4. Conclusion:

Hence, here using PCA and Clustering you can draw conclusions as to which college is better depending on many varibales chosen before like act scores, act scores, distance only, etc. According to the cumulative explained variance we retain 18 components and reject other for clustering.