Project - Advance Statistics

1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually.

H0 the means of relief with regards to ingredient A is equal.
muA=muA1=muA2=muA3
H1 the mean of relief with regards to ingredient A is unequal
muA ≠muA1≠muA2≠muA3
H0 the mean of relief with regards to ingredient B is equal
muB=muB1=muB2=muB3
H1 the mean of relief with regards to ingredient B is unequal
muB ≠muB1≠muB2≠muB3

1.2) Perform one-way ANOVA for variable 'A' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

```
formula  = 'Relief ~ A'
model = ols(formula,df).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
              df  sum_sq     mean_sq          F        PR(>F)
A            2.0  220.02  110.010000  23.465387  4.578242e-07
Residual    33.0  154.71    4.688182        NaN           NaN
```

From the above one-way anova test for variable A with respect to variable relief we see that the P value is less than our assumed significance level of 0.05 (95%). Therefore, we reject the null. Ingredient A is significant ingredient for relief of hay fever.

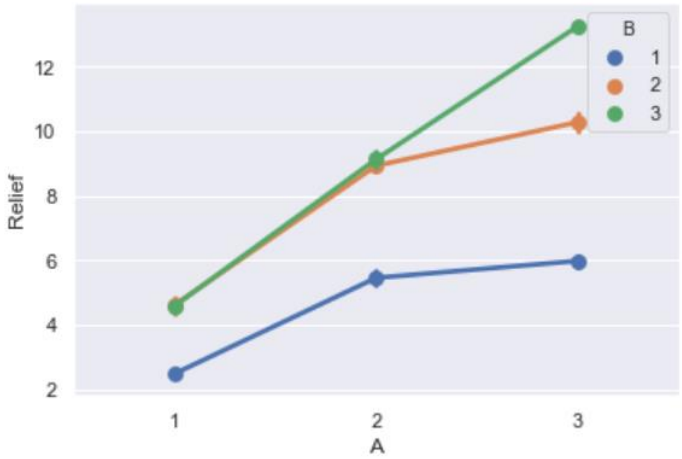
1.3) Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

```
formula  = 'Relief ~ B'
model = ols(formula,df).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
              df  sum_sq    mean_sq         F   PR(>F)
B            2.0  123.66  61.830000  8.126777  0.00135
Residual    33.0  251.07   7.608182       NaN      NaN
```

Since the P value is less than the significance level of 0.05, we can reject the null hypothesis and state that there is a difference in mean between relief and ingredient B

1.4) Analyse the effects of one variable on another with the help of an interaction plot. What is the interaction between the two treatments? [hint: use the 'pointplot' function from the 'seaborn' function]



We can see there is an interaction between ingredients at 2nd and 3rd level. There is some significance since there is interaction. The means of 2 and 3 level are interacting at relief of 9 hours.Level 3 is showing a upward trend and proves to provide more hours of relief compared to 1 and 2 levels.

1.5)Perform a two-way ANOVA based on the different ingredients (variable 'A' & 'B' along with their interaction 'A*B') with the variable 'Relief' and state your results.

```
formula = 'Relief ~ A + B + A:B'
model = ols(formula, df).fit()
aov_table = anova_lm(model)
print(aov_table)
```

```
              df    sum_sq     mean_sq            F        PR(>F)
A            2.0   220.020  110.010000  1827.858462  1.514043e-29
B            2.0   123.660   61.830000  1027.329231  3.348751e-26
A:B          4.0    29.425    7.356250   122.226923  6.972083e-17
Residual    27.0     1.625    0.060185          NaN           NaN
```

We can see that still both A and B ingredients are significant and the interaction of A: B is also significant from the above result. We can also see that the variance (F score) of A is increased to 1827 from 23 earlier. So, variance between A is 1827 times within A. similarly for B ingredient also there is a huge increase for F score.

1.6) Mention the business implications of performing ANOVA for this particular case study.

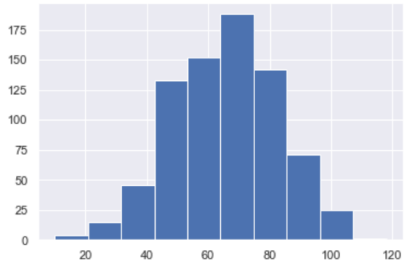
From the various steps performed we can conclude that:

- Both A and B treatment alone and together are significant cause for mean hours of relief.
- In this problem anova is performed because we have been comparing with two or more variables and this can be done by anova and also, we are having categorical data which using anova we can perform.
- A and B ingredient can be approved for the manufacturing of new component that can help in relief of hay fever because the null were rejected.
- The level 3 of the ingredient has shown better relief as per the point plot. It is showing an upward in the graph therefore relief hours are more with this level of ingredient.
- We can see an interaction in A and B at level 2 and 3 in the point plot above.
- A and B interaction is also significant for the relief of hay fever.
- The relief level at 2 and 3 is better than 1. 3rd level of component is the best in term of relief hours we can see that in the point plot.
- A and B are an important ingredient for the development of new component for hay fever.
- Seeing the F value of A, the variance between ingredient A is 23 times within each segment of ingredient A.
- Seeing the F value of B, the variance between ingredient A is 8 times within each segment of ingredient B.

2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

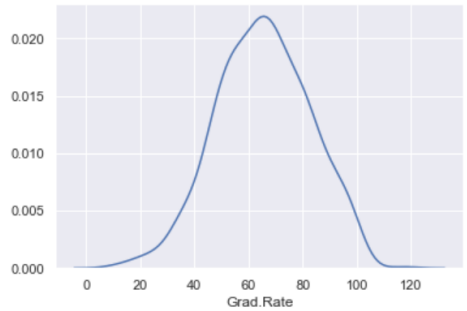
```
plt.hist(edu['Grad.Rate'],bins=10)
```

```
(array([  4.,  15.,  46., 133., 152., 188., 142.,  71.,  25.,   1.]),
 array([ 10. ,  20.8,  31.6,  42.4,  53.2,  64. ,  74.8,  85.6,  96.4,
        107.2, 118. ]),
 <BarContainer object of 10 artists>)
```




```
sns.distplot(edu['Grad.Rate'],bins=10,hist=False)
```

```
<AxesSubplot:xlabel='Grad.Rate'>
```

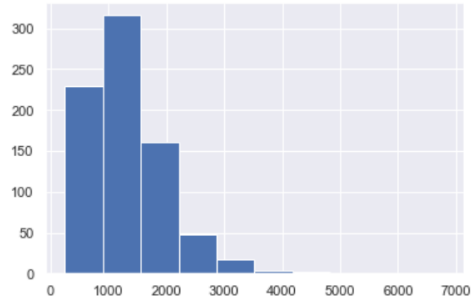


We can see from the above univariate analysis of student's graduate rate. More than half of the data lies in bins 3,4,5, and 6 we can see the class limit 64,74.8 has most of the students who had graduated, around 188 of them. From the distribution plot we can see that the data is almost

normally distributed. The mean of grad rate is 65. We can also see than out of 777 students almost 350 have scored less than 60 grad rate.

```
plt.hist(edu['Personal'],bins=10)
```

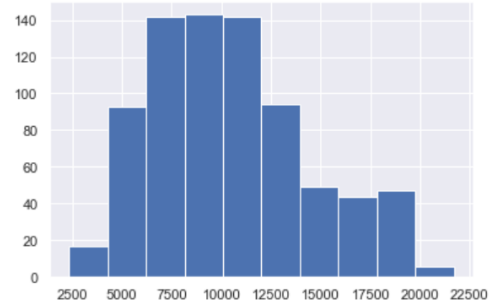
```
(array([229., 316., 160.,  48.,  17.,   3.,   2.,   1.,   0.,   1.]),
 array([ 250.,  905., 1560., 2215., 2870., 3525., 4180., 4835., 5490.,
        6145., 6800.]),
 <BarContainer object of 10 artists>)
```



We can see here that 2 bin has the maximum number of students who spend on personal needs. There are around 300 students who spend between 905 and 1560. From above we can also see the mean of personal is around the same range of 1340. After 1560 spending the students spending reduces drastically.

```
plt.hist(edu['Outstate'],bins=10)
```

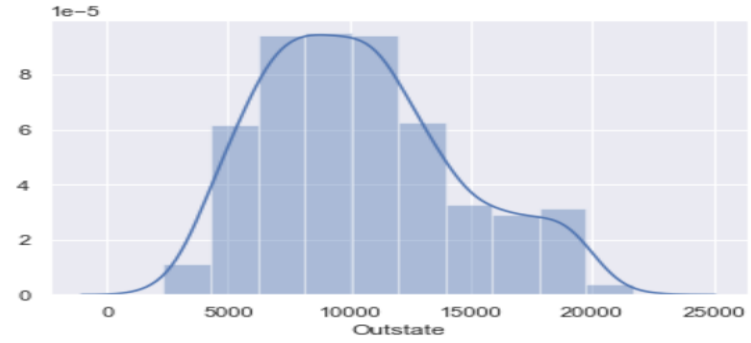
```
(array([ 17.,  93., 142., 143., 142.,  94.,  49.,  44.,  47.,   6.]),
 array([ 2340.,  4276.,  6212.,  8148., 10084., 12020., 13956., 15892.,
        17828., 19764., 21700.]),
 <BarContainer object of 10 artists>)
```



We can see a huge number of students are from coming to study from outstate also. we can see bins 3 4 and 5 have almost same amount of data.The class limit of bins 3 4 and 5 are 6212,8148,10084 and have data of 142,143, and 142 respectively. We can also see from the dist plot that the data is not normally distributed.

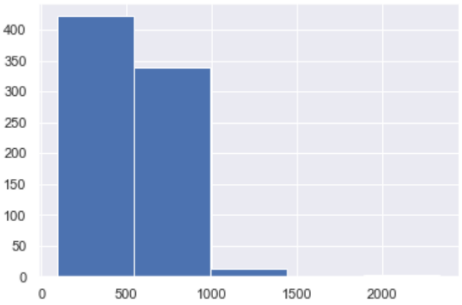
```
sns.distplot(edu.Outstate,bins=10)
```

```
<AxesSubplot:xlabel='Outstate'>
```




```
plt.hist(edu['Books'],bins=5)
```

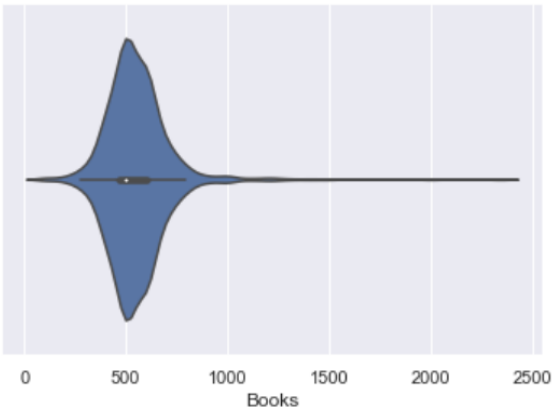
```
(array([422., 339.,  13.,   1.,   2.]),
 array([  96. ,  544.8,  993.6, 1442.4, 1891.2, 2340. ]),
 <BarContainer object of 5 artists>)
```



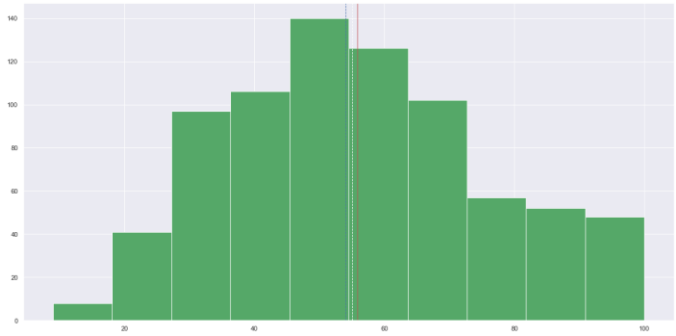
we can see that students spending on books are around the first bin which with class limit of 96 to 544.8.we can see that the two bars represent the maximum data.And the 2nd bin from class limit 544.8 to 993.6 has the next largest data.

```
sns.violinplot(edu.Books)
```

```
<AxesSubplot:xlabel='Books'>
```



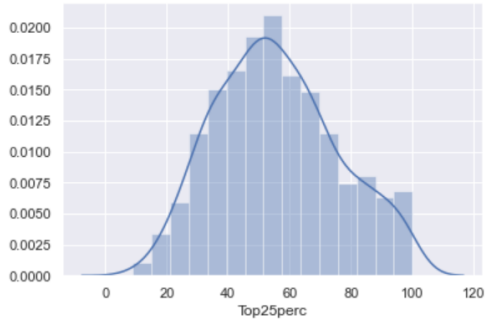
we can see the violen plot of the books variable above and as per the image we can see maximum data are in 500 range.



we can see here students with “Top25 perc” are more in bin 5 with class limit 45.4, 54.5. we can also see that the data looks normally distributed. we can also see the mean median and mode are close to each one but not at same point.

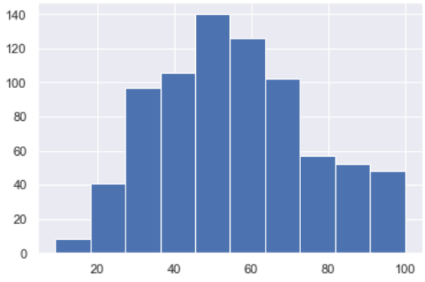
```
sns.distplot(edu['Top25perc'])
```

```
<AxesSubplot:xlabel='Top25perc'>
```



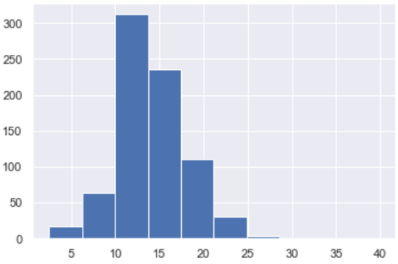

```
plt.hist(edu['Top25perc'])
```

```
(array([  8.,  41.,  97., 106., 140., 126., 102.,  57.,  52.,  48.]),
 array([  9. ,  18.1,  27.2,  36.3,  45.4,  54.5,  63.6,  72.7,  81.8,
         90.9, 100. ]),
 <BarContainer object of 10 artists>)
```

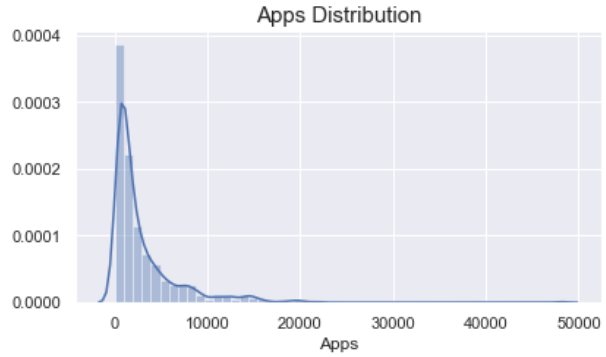


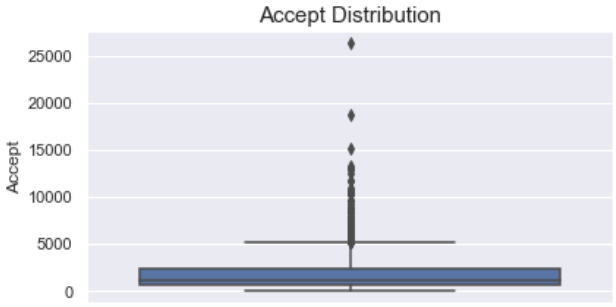

```
plt.hist(edu['S.F.Ratio'],bins=10)
```

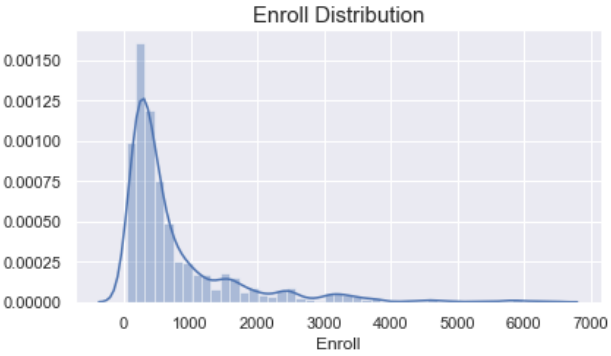
```
(array([ 17.,  64., 312., 236., 111.,  31.,   4.,   1.,   0.,   1.]),
 array([ 2.5 ,  6.23,  9.96, 13.69, 17.42, 21.15, 24.88, 28.61, 32.34,
        36.07, 39.8 ]),
 <BarContainer object of 10 artists>)
```

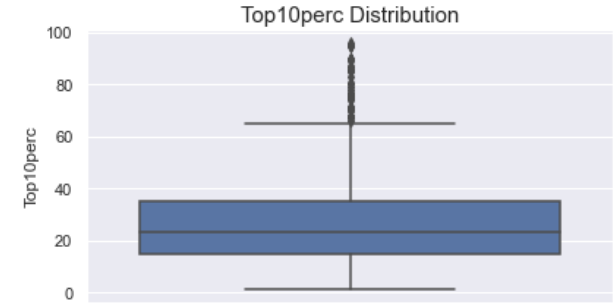


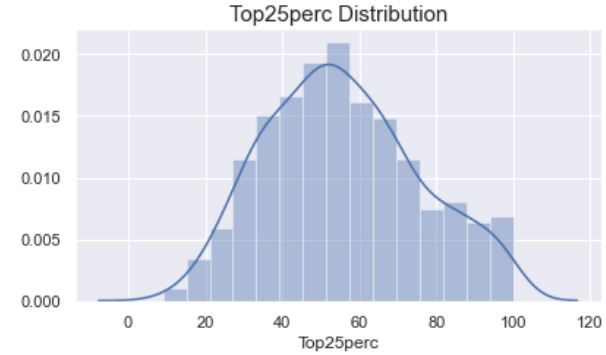
Here we look at the students to faculty ratio and from the histogram we can say that the 3 bin has the most ratio of class limit 9.96 – 13.69 and number of students around 312.

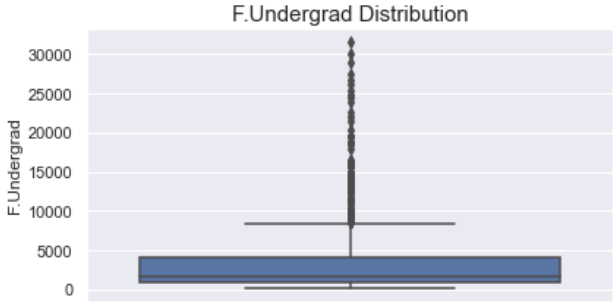

Apps Distribution
0.0004
0.0003
0.0002
0.0001
0.0000
0
10000
20000
30000
40000
50000
Apps

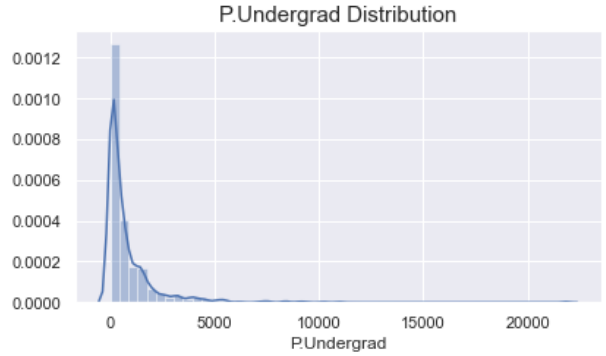

Accept Distribution
25000
20000
15000
10000
5000
0
Accept

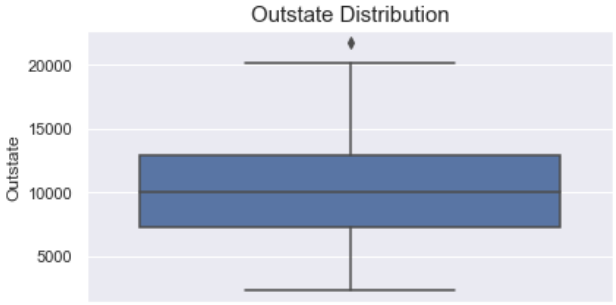

Enroll Distribution
0.00150
0.00125
0.00100
0.00075
0.00050
0.00025
0.00000
0
1000
2000
3000
4000
5000
6000
7000
Enroll

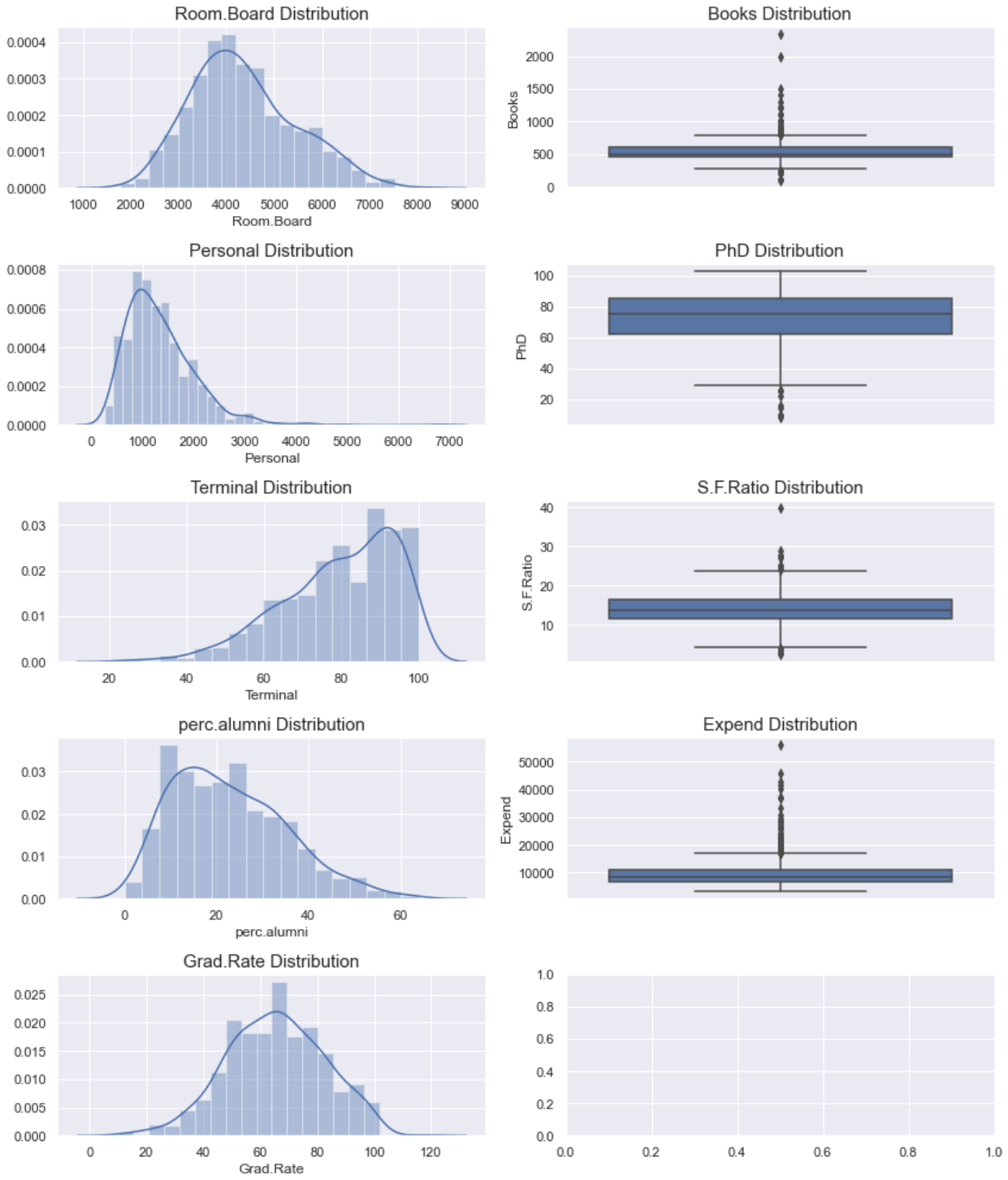

Top10perc Distribution
100
80
60
40
20
0
Top10perc


Top25perc Distribution
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
120
Top25perc


F.Undergrad Distribution
30000
25000
20000
15000
10000
5000
0
F.Undergrad

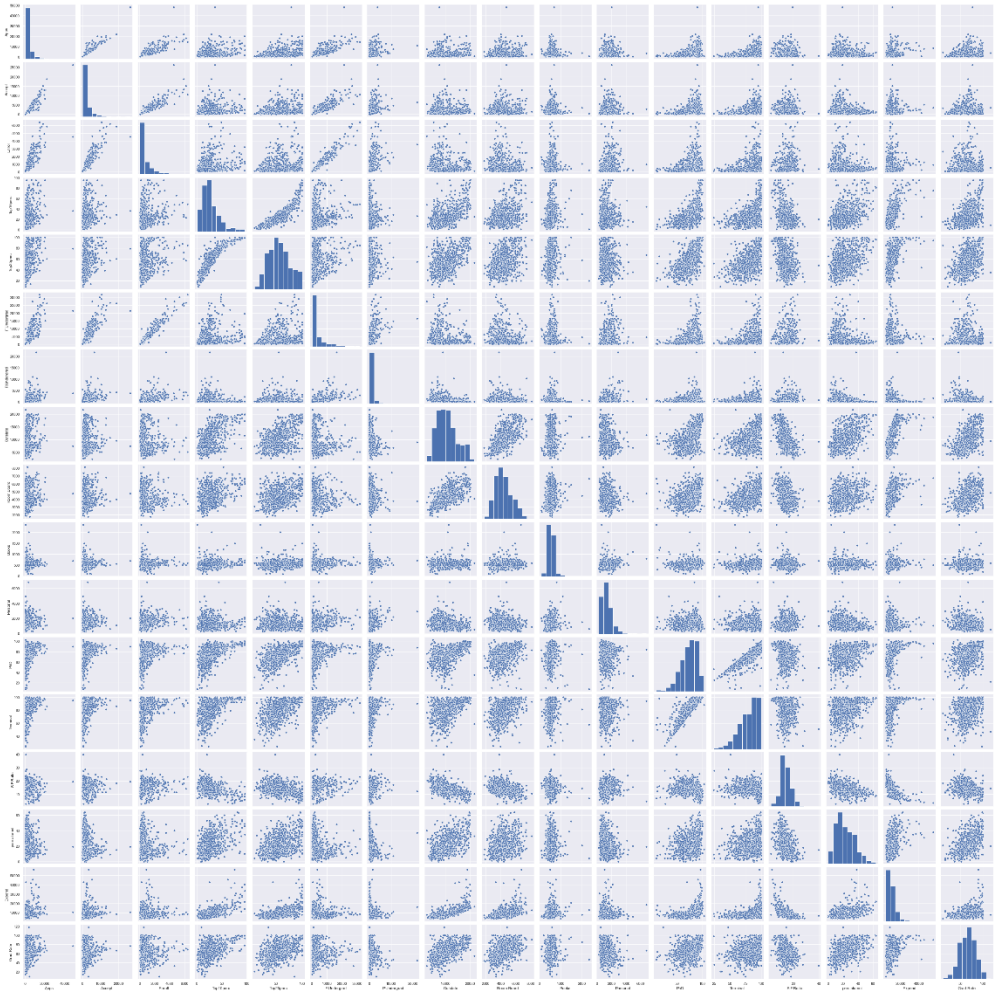

P.Undergrad Distribution
0.0012
0.0010
0.0008
0.0006
0.0004
0.0002
0.0000
0
5000
10000
15000
20000
P.Undergrad


Outstate Distribution
20000
15000
10000
5000
Outstate

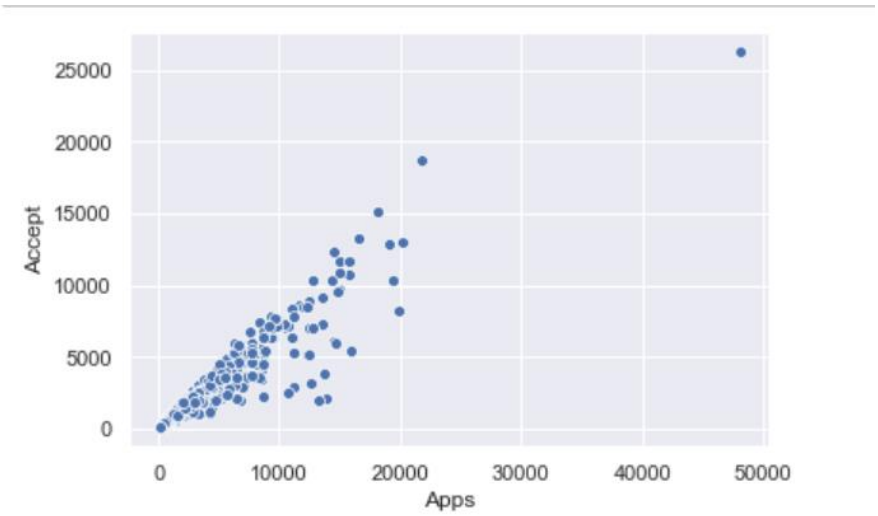


From the above distributions we can see that only few of the variables looks like having a normal distribution through we cannot confirm it. Grad rate, room board, Top 25perc, are the ones that looks like normally distributed data. We can also see there are numerous outliers in our data set.

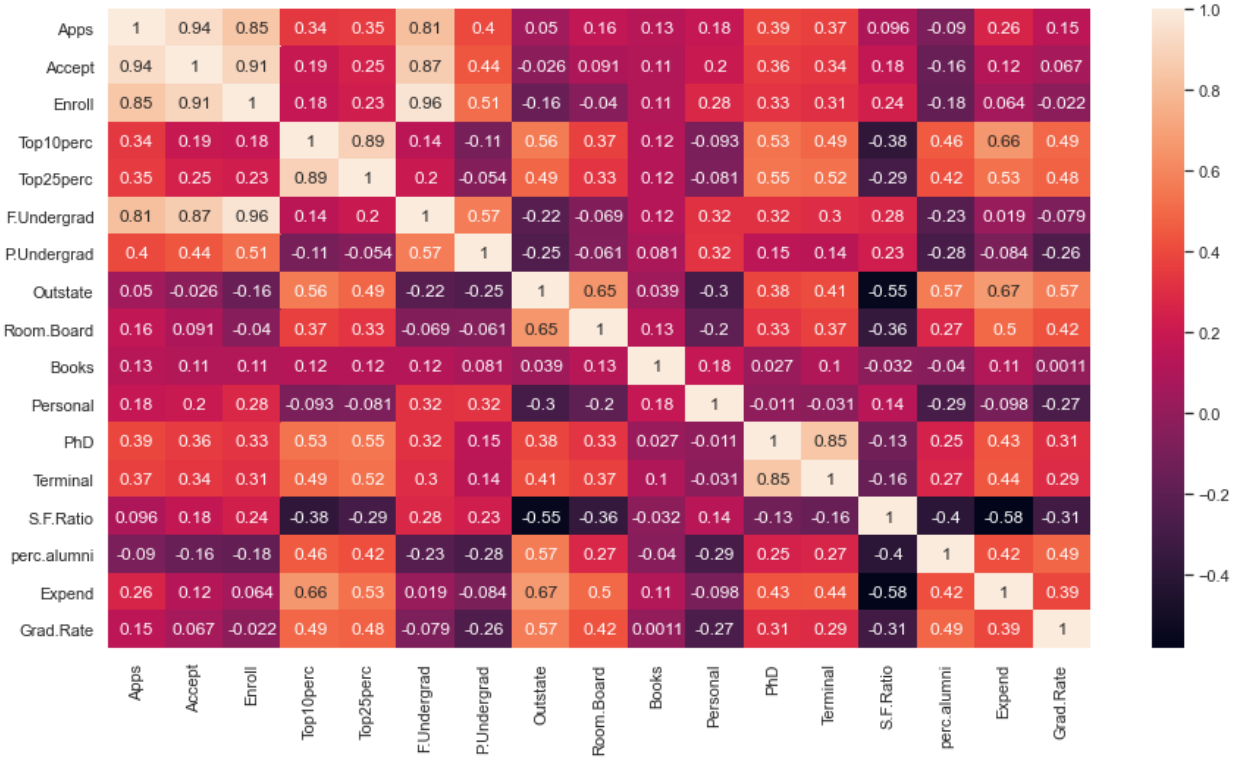
MULTIVARIATE:



Above is the pair plot where we can see the joint scatter plot and the histogram for our dataset edu.



From the above scatterplot we can see that most of the universities have application of 10000 range. Except few which got application above 20000 and one university got 50000 applications.



We can see above the correlation heatmap. From the above heatmap we can see few columns that are correlated. The highest correlation is between enrollment of students who wants to study the course full time. They have a correlation of 0.96%. Next is followed by number of applications accepted and number of applications received at 0.94%.

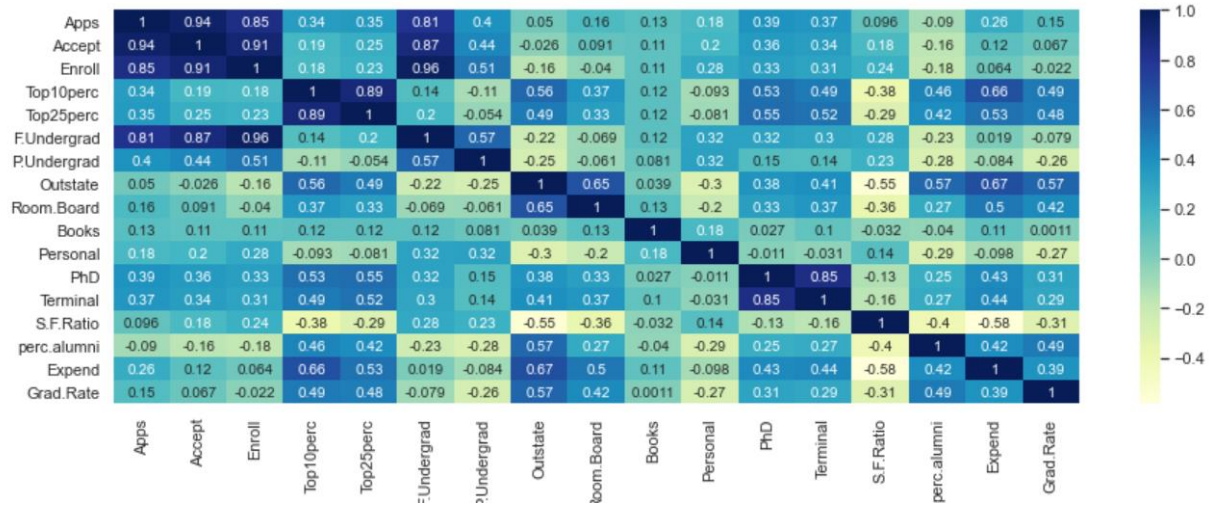
2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ra
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729	1.013
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176	-0.477
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341	-0.300
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657	-1.615
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535	-0.553

We have scaled the variables here using the standard scaler function where the values are transformed to Zscore. We can see that when we do this all the values in each variable are on same scale and the standard deviation is 1 and mean is 0 for all the variables which can be seen below. We can also use the normalization scaling function where the min and max will be 0 and 1, respectively.

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	1.000644	-1.506526	-0.712380	-0.258583	0.422113	3.882319
Top25perc	777.0	-1.546739e-16	1.000644	-2.364419	-0.747607	-0.090777	0.667104	2.233391
F.Undergrad	777.0	-1.661405e-16	1.000644	-0.734617	-0.558643	-0.411138	0.062941	5.764674
P.Undergrad	777.0	-3.029180e-17	1.000644	-0.561502	-0.499719	-0.330144	0.073418	13.789921
Outstate	777.0	6.515595e-17	1.000644	-2.014878	-0.776203	-0.112095	0.617927	2.800531
Room.Board	777.0	3.570717e-16	1.000644	-2.351778	-0.693917	-0.143730	0.631824	3.436593
Books	777.0	-2.192583e-16	1.000644	-2.747779	-0.481099	-0.299280	0.306784	10.852297
Personal	777.0	4.765243e-17	1.000644	-1.611860	-0.725120	-0.207855	0.531095	8.068387
PhD	777.0	5.954768e-17	1.000644	-3.962596	-0.653295	0.143389	0.756222	1.859323
Terminal	777.0	-4.481615e-16	1.000644	-3.785982	-0.591502	0.156142	0.835818	1.379560
S.F.Ratio	777.0	-2.057556e-17	1.000644	-2.929799	-0.654660	-0.123794	0.609307	6.499390
perc.alumni	777.0	-6.022638e-17	1.000644	-1.836580	-0.786824	-0.140820	0.666685	3.331452
Expend	777.0	1.213101e-16	1.000644	-1.240641	-0.557483	-0.245893	0.224174	8.924721
Grad.Rate	777.0	3.886495e-16	1.000644	-3.230876	-0.726019	-0.026990	0.730293	3.060392

2.3) Comment on the comparison between covariance and the correlation matrix after scaling.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	1
0	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.09034
1	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.176456	-0.16019
2	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.237577	-0.18102
3	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.385370	0.45607
4	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.295009	0.41840
5	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300406	0.280064	-0.22975
6	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142086	0.232830	-0.28115
7	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408509	-0.555536	0.56699
8	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375022	-0.363095	0.27271
9	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100084	-0.031970	-0.04026

We can see that after doing standardized scaling (Z score) both the correlation and covariance are the same. Since we standardized the variances it is correlation now. We can see above in correlation Apps variable the values are 1, 0.94, 0.85 and in the correlation matrix 0 variable column also represent the same values 1.00, 0.944, 0.847……. we can also see the diagonals are 1 in the covariance matrix after standardization.

Covariance shows us the direction of the linear relationship between the variables whereas correlation measures the strength and direction of the linear relationship between two variables.

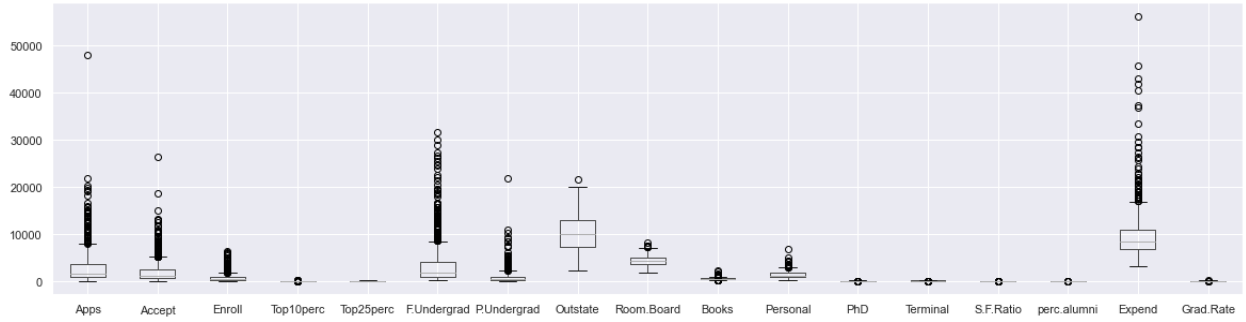
Co-relation = co-variance of standardized variables.

Below is the co-variance matrix before the standardization. We can see the difference before the standardization and after the standardization.

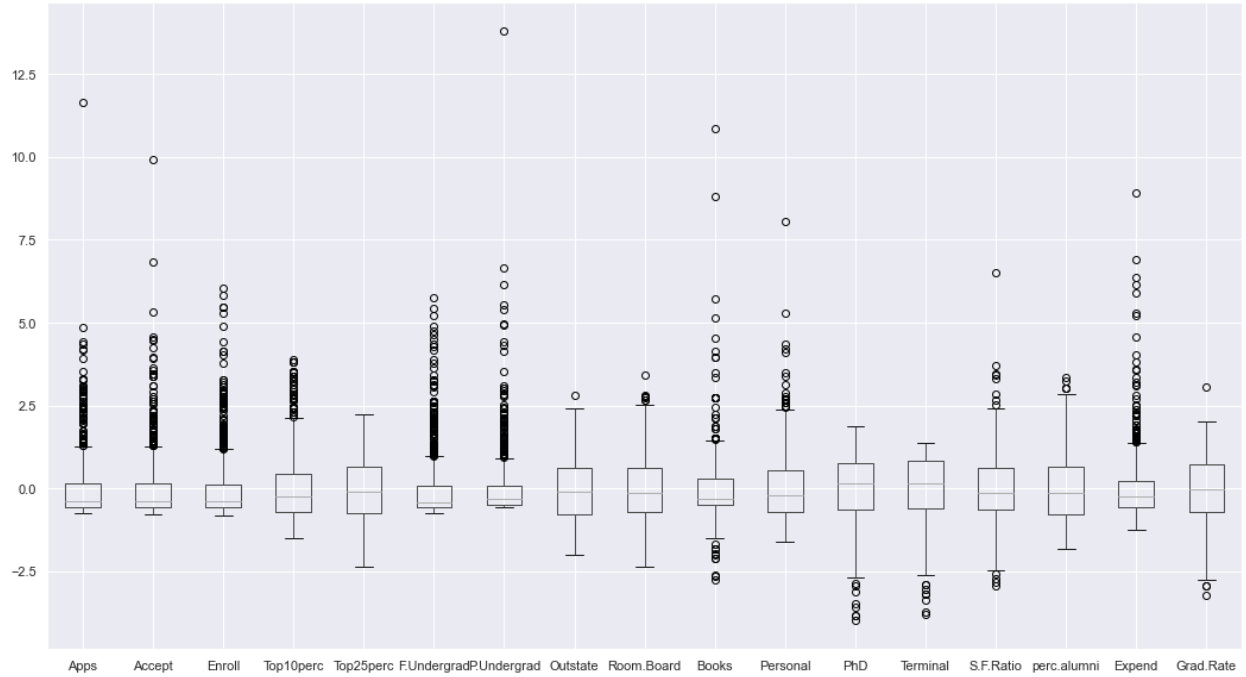
0	1.497846e+07	8.949860e+06	3.045256e+06	23132.773138	26952.663479	1.528970e+07	2.346620e+06	7.809704e+05	7.000729e+05	84703.752639	4.68346
1	8.949860e+06	6.007960e+06	2.076268e+06	8321.124872	12013.404757	1.039358e+07	1.646670e+06	-2.539623e+05	2.443471e+05	45942.807867	3.33556
2	3.045256e+06	2.076268e+06	8.633684e+05	2971.583415	4172.592435	4.347530e+06	7.257907e+05	-5.811885e+05	-4.099706e+04	17291.199742	1.76738
3	2.313277e+04	8.321125e+03	2.971583e+03	311.182456	311.630480	1.208911e+04	-2.829475e+03	3.990718e+04	7.186706e+03	346.177405	-1.11455
4	2.695266e+04	1.201340e+04	4.172592e+03	311.630480	392.229216	1.915895e+04	-1.615412e+03	3.899243e+04	7.199904e+03	377.759266	-1.08360
5	1.528970e+07	1.039358e+07	4.347530e+06	12089.113681	19158.952782	2.352658e+07	4.212910e+06	-4.209843e+06	-3.664582e+05	92535.764728	1.04170
6	2.346620e+06	1.646670e+06	7.257907e+05	-2829.474981	-1615.412144	4.212910e+06	2.317799e+06	-1.552704e+06	-1.023919e+05	20410.446674	3.29732
7	7.809704e+05	-2.539623e+05	-5.811885e+05	39907.179832	38992.427500	-4.209843e+06	-1.552704e+06	1.618466e+07	2.886597e+06	25808.242145	-8.14673
8	7.000729e+05	2.443471e+05	-4.099706e+04	7186.705605	7199.903568	-3.664582e+05	-1.023919e+05	2.886597e+06	1.202743e+06	23170.313390	-1.48083
9	8.470375e+04	4.594281e+04	1.729120e+04	346.177405	377.759266	9.253576e+04	2.041045e+04	2.580824e+04	2.317031e+04	27259.779946	2.00430

2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

Before Scaling:



After Scaling



Since all the variable have a mean of zero and standard deviation of 1, we can see the box plot of variables are more visible than before scaling. Also, some looks normally distributed. We can see that the data after scaling are having same magnitude of variables. We can see the difference in the Top 25 per earlier IQR range before scaling and after scaling is different. Every value in the variables are now in same magnitude. It is also noticed the Y axis data is same for all variables so the distribution of data in each variable is more visible since they all are scaled equally. We can see the data after scaling is more distributed.

2.5) Build the covariance matrix, eigenvalues, and eigenvector.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	1
0	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.09034
1	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.176456	-0.16019
2	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.237577	-0.18102
3	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.385370	0.45607
4	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.295009	0.41840
5	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300406	0.280064	-0.22975
6	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142086	0.232830	-0.28115
7	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408509	-0.555536	0.56699
8	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375022	-0.363095	0.27271
9	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100084	-0.031970	-0.04026

```
Eigen Values
%s [5.45052162 4.48360686 1.17466761 1.00820573 0.93423123 0.84849117
0.6057878  0.58787222 0.53061262 0.4043029  0.02302787 0.03672545
0.31344588 0.08802464 0.1439785  0.16779415 0.22061096]
```

Eigen Values:

	0
0	5.450522
1	4.483607
2	1.174668
3	1.008206
4	0.934231
5	0.848491
6	0.605788
7	0.587872
8	0.530613
9	0.404303
10	0.023028
11	0.036725
12	0.313446
13	0.088025
14	0.143978
15	0.167794
16	0.220611

We can see above that there are 4 eigen values that are greater than 1 which can be considered for our analysis and how many Principal components that we can take for our business solution based on our required percentage of achievement. This is explained below in detail. We should also ensure all the variables are continuos and scaled when doing PCA

Eigen Vectors:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
0	-0.248766	0.331598	0.063092	-0.281311	0.005741	0.016237	0.042486	0.103090	0.090227	-0.052510	0.358970	-0.459139	0.043046	-0.133406	0.080
1	-0.207602	0.372117	0.101249	-0.267817	0.055786	-0.007535	0.012950	0.056271	0.177865	-0.041140	-0.543427	0.518569	-0.058406	0.145498	0.033
2	-0.176304	0.403724	0.082986	-0.161827	-0.055694	0.042558	0.027693	-0.058662	0.128561	-0.034488	0.609651	0.404318	-0.069399	-0.029590	-0.085
3	-0.354274	-0.082412	-0.035056	0.051547	-0.395434	0.052693	0.161332	0.122678	-0.341100	-0.064026	-0.144986	0.148739	-0.008105	-0.697723	-0.107
4	-0.344001	-0.044779	0.024148	0.109767	-0.426534	-0.033092	0.118486	0.102492	-0.403712	-0.014549	0.080348	-0.051868	-0.273128	0.617275	0.151

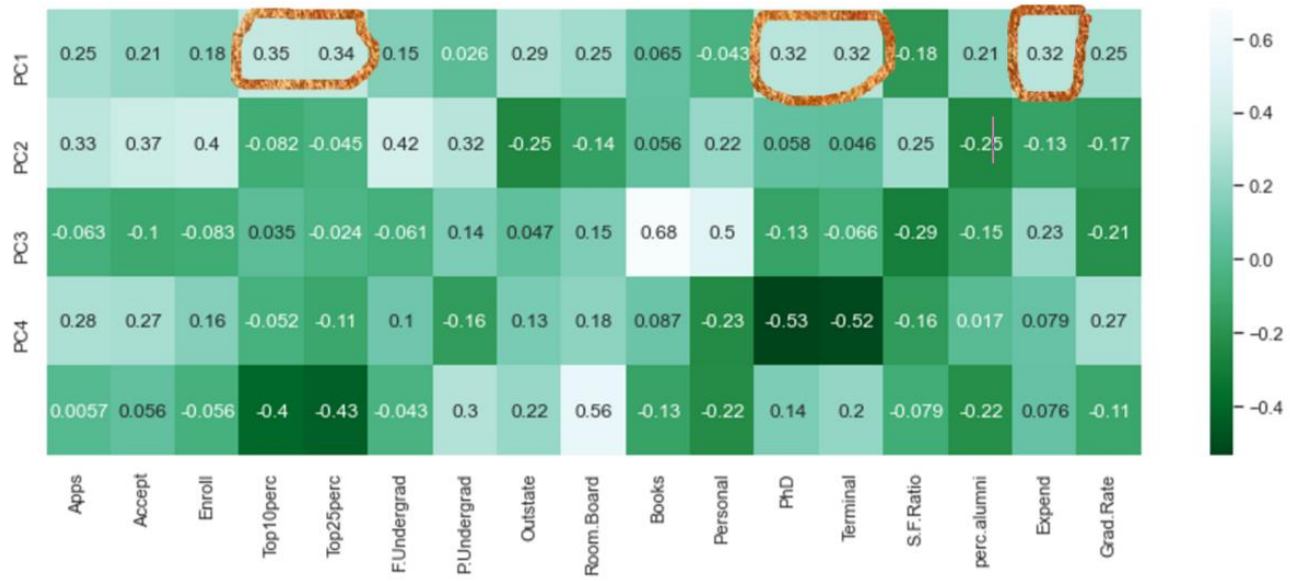
2.6) Write the explicit form of the first PC (in terms of Eigen Vectors).

Explicit form:

we can see from the explicit data of the first PC that top 10 perc has the highest value of eigen vector which contributing to PC1 which is at 0.35 followed by top 25 perc 0.34 and expend at 0.318

PC1=(Apps*0.248766)+(Accept*0.207602)+(Enroll*0.176304)+(Top10perc*0.354274)+(Top25perc*0.344001)+(F.Undergrad*0.154641)+(P.Undergrad*0.026443)+(Outstate*0.294736)+(Room.Board*0.249030)+(Books*0.064758)+(Personal*0.042529)+(PhD*0.318313)+(Terminal*0.317056) (S.F.Ratio*0.176958)+ (perc.alumni*0.205082) +(Expend*0.318909)+(Grad.Rate*0.252316)

Below we can see the contribution of each eigen vectors to their respective variables. And the first PC explicit form is the values of PC1 in below image.



The main contributors for the first principal component are marked in the above image.

2.7) Discuss the cumulative values of the eigenvalues. How does it help you to decide on the optimum number of principal components? What do the eigenvectors indicate? Perform PCA and export the data of the Principal Component scores into a data frame.

Eigen Values

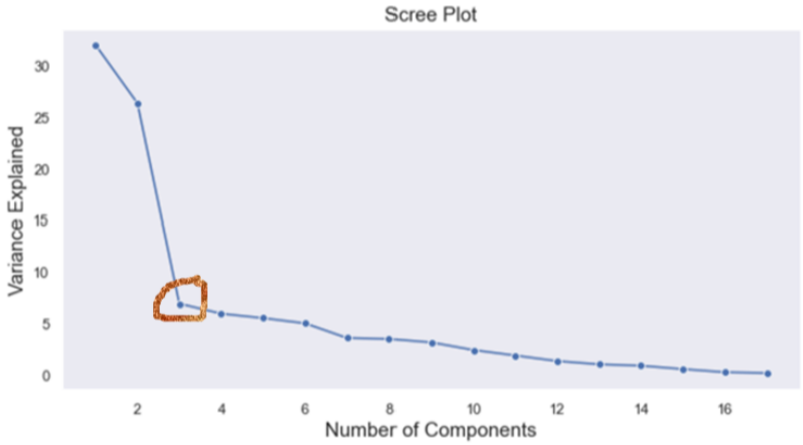
	0
0	5.450522
1	4.483607
2	1.174668
3	1.008206
4	0.934231
5	0.848491
6	0.605788
7	0.587872
8	0.530613
9	0.404303
10	0.023028
11	0.036725
12	0.313446
13	0.088025
14	0.143978
15	0.167794
16	0.220611

These are the different principal components(PCA) that we have got from our dataset. These represent the total number of columns we had in dataset which is 17 and we have 17PCA. The first PCA of 5.45 means that the amount of data that this contains. Which means there are almost 5.45

variables present in this particular PCA. So, 5 columns of data are reduced to 1 eigen value. Furthermore, the 2nd PCA contains 4.48 variable data in this particular PCA.

Cumulative distribution of eigen values(PCA)

```
Cumulative Variance Explained [ 32.0206282   58.36084263  65.26175919  71.18474841  76.67315352
  81.65785448  85.21672597  88.67034731  91.78758099  94.16277251
  96.00419883  97.30024023  98.28599436  99.13183669  99.64896227
  99.86471628 100.        ]
```



Above we can see the cumulative values of the PCA and the scree plot. The cumulative values of PCA are got by adding each PCA and from this we can decide how much percentage of data do we need for our business. For e.g. from the values if we decide we need around 71% of data for our business we go with 4 PCA which gives us 71% of information of our dataset. So, from 18 variables in our original dataset we can now say that we need only 4 PCA which will give us 71% of the information of the same dataset.

In the scree plot the point where the steep slope start is called the elbow, and this gives us an indication how many PCA are good to take for the business. Here in the above scree plot the elbow(circled in the image) start after the 3rd PCA and we have taken one more extra PCA for this particular data.

Eigen Vectors:

Each row corresponds to the principal component. Row 0 below corresponds to PC1(5.45) which we saw above. Eigen vector is the contribution of each variable making towards each Principal component. Each PCA will have the data of all the variables. For e.g. 1st eigen vector will have contribution of each 17 variables towards the first principal component. That is why we see 17 values in each vector and there are total 17 vectors for each principal component. Each column corresponds to the original variables.

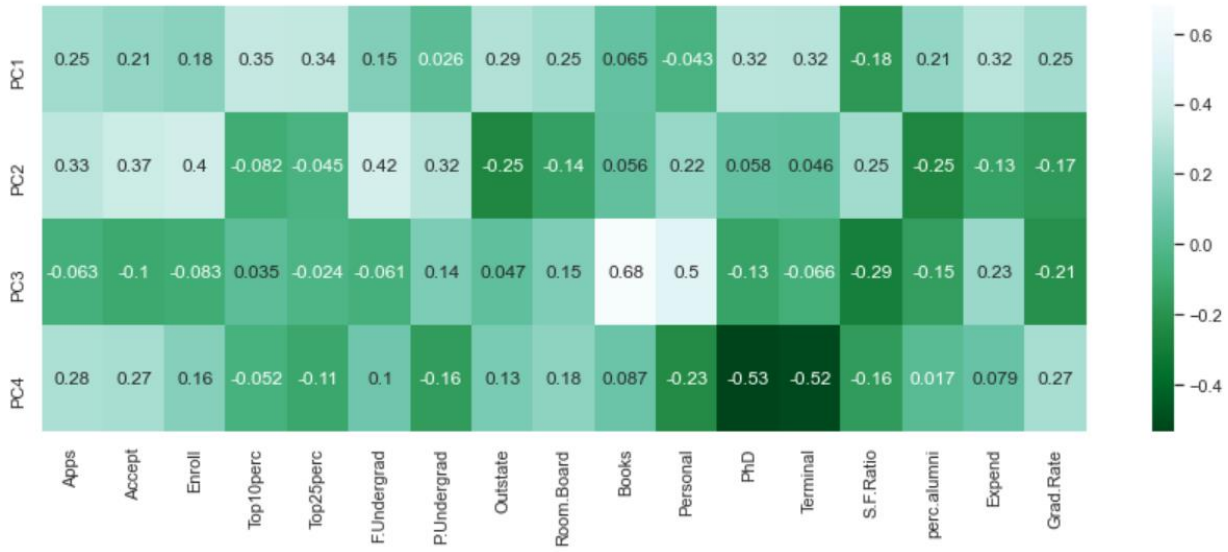
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
0	-0.248766	0.331598	0.063092	-0.281311	0.005741	0.016237	0.042486	0.103090	0.090227	-0.052510	0.358970	-0.459139	0.043046	-0.133406	0.0
1	-0.207602	0.372117	0.101249	-0.267817	0.055786	-0.007535	0.012950	0.056271	0.177865	-0.041140	-0.543427	0.518569	-0.058406	0.145498	0.0
2	-0.176304	0.403724	0.082986	-0.161827	-0.055694	0.042558	0.027693	-0.058662	0.128561	-0.034488	0.609651	0.404318	-0.069399	-0.029590	-0.0
3	-0.354274	-0.082412	-0.035056	0.051547	-0.395434	0.052693	0.161332	0.122678	-0.341100	-0.064026	-0.144986	0.148739	-0.008105	-0.697723	-0.1
4	-0.344001	-0.044779	0.024148	0.109767	-0.426534	-0.033092	0.118486	0.102492	-0.403712	-0.014549	0.080348	-0.051868	-0.273128	0.617275	0.1
5	-0.154641	0.417674	0.061393	-0.100412	-0.043454	0.043454	0.025076	-0.078890	0.059442	-0.020847	-0.414705	-0.560363	-0.081158	-0.009916	-0.0
6	-0.026443	0.315088	-0.139682	0.158558	0.302385	0.191199	-0.061042	-0.570784	-0.560673	0.223106	0.009018	0.052731	0.100693	-0.020952	0.0
7	-0.294736	-0.249644	-0.046599	-0.131291	0.222532	0.030000	-0.108529	-0.009846	0.004573	-0.186675	0.050900	-0.101595	0.143221	-0.038354	-0.0
8	-0.249030	-0.137809	-0.148967	-0.184996	0.560919	-0.162755	-0.209744	0.221453	-0.275023	-0.298324	0.001146	0.025929	-0.359322	-0.003402	-0.0
9	-0.064758	0.056342	-0.677412	-0.087089	-0.127289	-0.641055	0.149692	-0.213293	0.133663	0.082029	0.000773	-0.002883	0.031940	0.009439	-0.0
10	0.042529	0.219929	-0.499721	0.230711	-0.222311	0.331398	-0.633790	0.232661	0.094469	-0.136028	-0.001114	0.012890	-0.018578	0.003090	0.0
11	-0.318313	0.058311	0.127028	0.534725	0.140166	-0.091256	0.001096	0.077040	0.185182	0.123452	0.013813	-0.029808	0.040372	0.112056	-0.6

PCA Decomposition:

```
array([[-1.59285540e+00, -2.19240182e+00, -1.43096371e+00, ...,
        -7.32560607e-01,  7.91932731e+00, -4.69508058e-01],
       [ 7.67333500e-01, -5.78829915e-01, -1.09281891e+00, ...,
        -7.72352084e-02, -2.06832871e+00,  3.66660917e-01],
       [-1.01072915e-01,  2.27879364e+00, -4.38091508e-01, ...,
        -4.07947944e-04,  2.07355479e+00, -1.32891372e+00],
       [-9.21755807e-01,  3.58895679e+00,  6.77229073e-01, ...,
         5.43340182e-02,  8.52135015e-01, -1.08036052e-01]])
```

PCA components:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ra
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056	-0.1769
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429	0.2466
2	-0.063092	-0.101249	-0.082986	0.035054	-0.024147	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038	-0.2898
3	0.281314	0.267815	0.161827	-0.051534	-0.109778	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534724	-0.519443	-0.1611



Above we can see the data frame of the 4 PCA component that we have selected. The above heatmap shows the values of each PC and the amount of contribution by each variable to that particular principal component.

Shape after reduced data:

```
data_reduced.shape
```

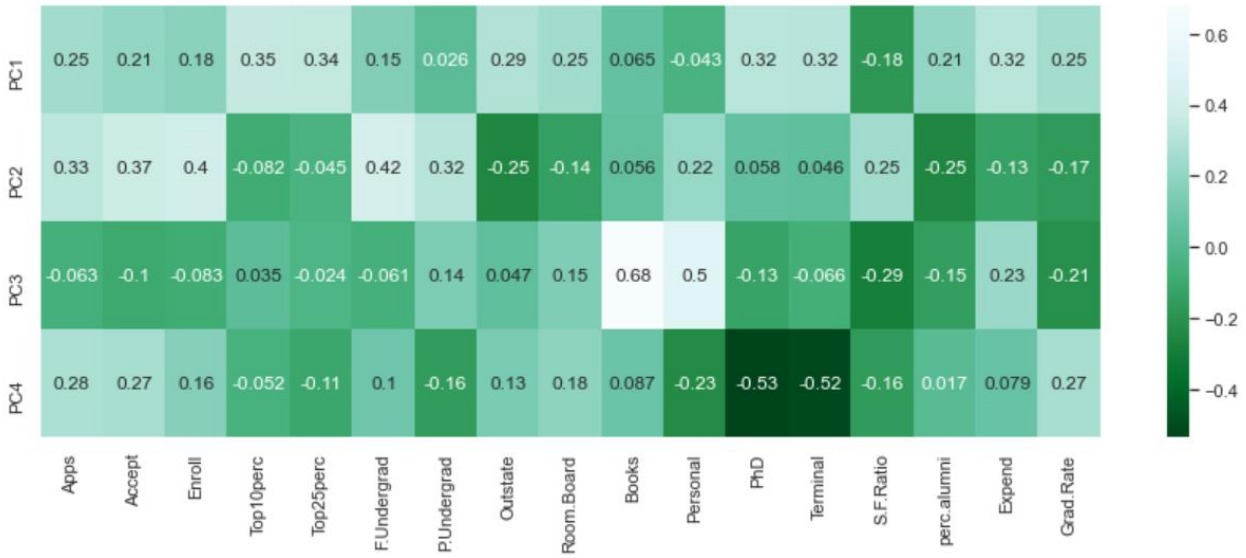
```
(777, 4)
```

2.8) Mention the business implication of using the Principal Component Analysis for this case study.

From the principal component analysis, we can say that PC1 has captured 32% of the data from all the variables. From the loading values of the PCA we can see PC1 is having a constant loading there are no variables which are highly correlated or devoted to this particular principal component. For the 2nd PC most of the information is contributed by F. undergrad and perc alumni and outside. PC 2 is contributing 26% of the information. PC3 contains major data or books which is at 0.68 and PC3 is contributing around 7% of the data out of which books is the most contributing to this PCA.

The major implication of PCA is that business can achieve the required amount of information by reducing the number of variables they have to go through. In this study our original dataset had 17 variables and we have reduced it to 4 PCA without losing much of the information that are required for the business. With the 4 PCA we are achieving 71% of the information which was earlier distributed in 17 columns. Even though the 17 columns contained 100% data we extracted 71% of that information with 4PCA.

Based on the need for different business scenario we can select what % of information do we require to achieve our business goals.


PC1
PC2
PC3
PC4
0.25 0.21 0.18 0.35 0.34 0.15 0.026 0.29 0.25 0.065 -0.043 0.32 0.32 -0.18 0.21 0.32 0.25
0.33 0.37 0.4 -0.082 -0.045 0.42 0.32 -0.25 -0.14 0.056 0.22 0.058 0.046 0.25 -0.25 -0.13 -0.17
-0.063 -0.1 -0.083 0.035 -0.024 -0.061 0.14 0.047 0.15 0.68 0.5 -0.13 -0.066 -0.29 -0.15 0.23 -0.21
0.28 0.27 0.16 -0.052 -0.11 0.1 -0.16 0.13 0.18 0.087 -0.23 -0.53 -0.52 -0.16 0.017 0.079 0.27
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
Personal
PhD
Terminal
S.F.Ratio
perc.alumni
Expend
Grad.Rate
0.6
0.4
0.2
0.0
−0.2
−0.4