Problem 1:

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

Summary:

A specific research lab has conducted trials to ascertain benefits of a proposed medicine. The benefit is measured by the duration of relief caused by a medicine. The medicine is a compound of 2 ingredients A and B.

Now, the trials were conducted over 4 Volunteers only. Each of them were given various combinations of ingredients A and B and the result was recorded.

This exercise is for the research lab's attempts to find the ingredient combination which is most beneficial i.e. with the longest duration of relief.

1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually. [both statement and statistical form like Ho=mu, Ha>mu]

Answer:

Statement Hypothesis for A:

A Ingredient	Relief (Mean hours)
1	3.88
2	7.83
3	9.83

- H0 => The mean relief hours with respect to each trial including ingredient A are equal
- Ha => At least one of the pairs of the mean of relief hours with respect to each trial of ingredient A are unequal

Statistical Hypothesis for A:

- H0 => mu(A1) = mu(A2) = mu (A3)
- Ha => mu(A1) != mu(A2) != mu (A3)

Statement Hypothesis for B:

B Ingredient	Relief (Mean hours)
1	4.63
2	7.93
3	8.98

- H0 => The mean relief hours with respect to each trial including ingredient B are equal
- Ha => At least one of the pairs of mean of relief hours with respect to each trial of ingredient B are unequal

Statistical Hypothesis for B:

- H0 => mu(A1) = mu(A2) = mu (A3)
- Ha => mu(A1) != mu(A2) != mu (A3)

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.

- H0 => The mean relief hours with respect to trials including ingredient A are equal
- Ha => At least one of the pairs of the mean of relief hours with respect to each trial of ingredient A are unequal

Formula = Relief ~ A					
	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
A	2	220.02	110.01	23.465387	4.58E-07
Residual	33	154.71	4.688182	NaN	NaN

The P-value < 0.05, is statistically significant, therefore, we reject the null hypothesis, and accept the alternative hypothesis.

Conclusion: At least one of the means of relief hours with respect to each trial of ingredient A are unequal.

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.

- H0 => The mean relief hours with respect to trials including ingredient B are equal
- Ha => At least one of the pairs of mean of relief hours with respect to each trial of ingredient B are unequal

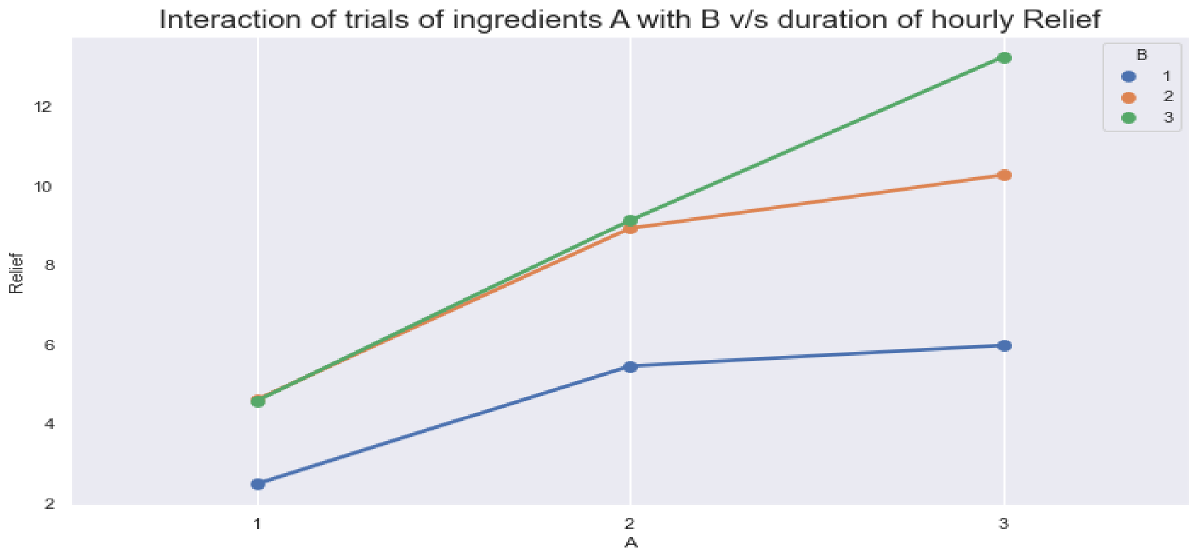
Formula = Relief ~ B					
	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
B	2	123.66	61.83	8.126777	0.00135
Residual	33	251.07	7.608182	NaN	NaN

The P-value < 0.05, is statistically significant, therefore, we reject the null hypothesis, and accept the alternative hypothesis.

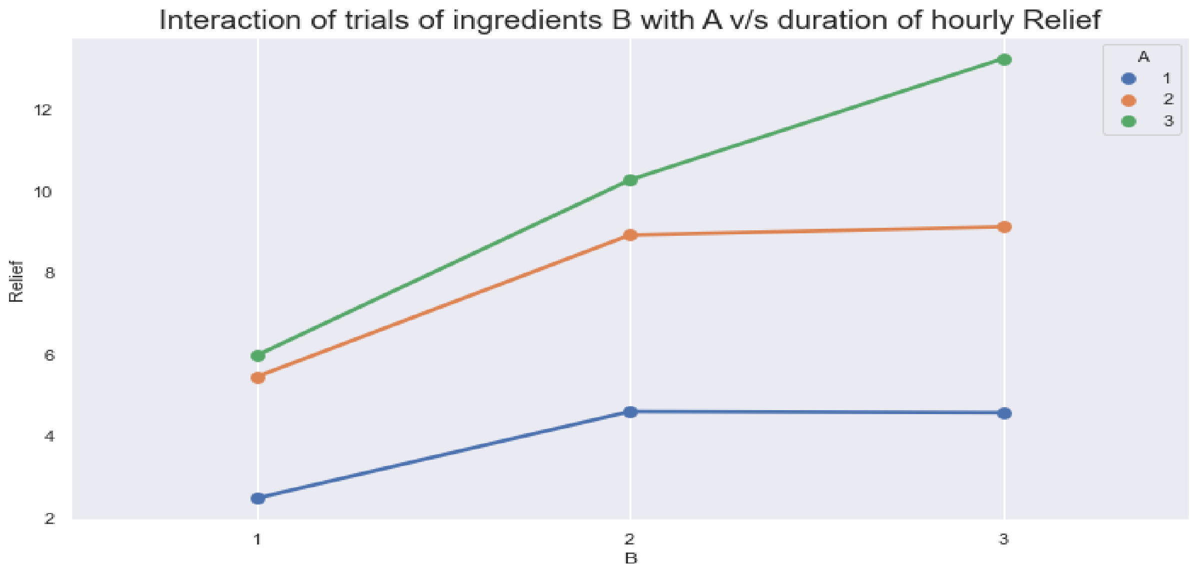
Conclusion: At least one of the means of relief hours with respect to each trial of ingredient B are unequal.

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?

The interaction of presence of both the ingredient A & B is given in the point-plot below:



Alternatively,



Conclusion:

- The trials with combinations of A with B ingredients are interacting between ingredient level 1 and 2 of A and B. Ingredient B with A is not interacting
- More than 10 hours of Relief achieved at ingredient A (Level 3) and ingredient B (Level 2&3). Average relief hours for the trials is 7.18 hours.

  Hence, highest levels of dosage of ingredient A & B gives maximum relief/ benefit.
- Treatment with ingredient A and B at Level 1 gives poor results (below average relief) for the trials.

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.

Here is the ANOVA output table for the 2-way ANOVA:

H0 => The mean relief hours with respect to each trial including both ingredient A & B are equal

Ha => At least one of the means of relief hours with respect including both ingredient A & B are unequal

Formula = Relief ~ A & B					
	df	sum_sq	mean_sq	F	PR(>F)
A	2	220.02	110.01	1827.85846	1.51E-29
B	2	123.66	61.83	1027.32923	3.35E-26
A:B	4	29.425	7.35625	122.226923	6.97E-17
Residual	27	1.625	0.060185	NaN	NaN

- The significance levels of A and B individually is **not statistically significant**.
- The interaction of A & B together is also statistically **not statistically significant** explaining the variation in the mean of ‘Relief’.

Conclusion

It is safe to conclude that the presence of A&B ingredients **presents strong evidence against the null Hypothesis.**

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

After investigating the variables, A and B individually with Volunteers through ANOVA table, following observations are found:

- Both ingredient A & B **provides significant evidence against the H0 (the mean of Relief durations are not equal).** The p-value is statistically significant.
- The interaction of Independent variable A and B is **statistically significant** and presents significant evidence that the mean duration of Relief with various combinations are **not equal**.

Conclusion:

- Vaccine developed by a mixture of ingredient B and ingredient A at higher levels has displayed longer durations of relief in the trials.
- There is higher chance of increased benefit to the end customer with A & B ingredients at Level 3, which can likely translate in to better sales.
- Volunteer is a concomitant Variable. Although doesn’t provide enough evidence to explain the variation in the mean Relief durations. It is not statistically significant.

A summary of ANOVA analysis.

H0 => The mean relief hours of test done at all the three levels including both ingredient A or B are equal.

Ha => At least one of the means of relief hours are unequal.

Formula = Relief ~ Volunteer & A					
	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
Volunteer	3	0.072222	0.024074	0.003741	0.999675
A	2	220.02	110.01	17.094833	0.000024
Volunteer:A	6	0.191111	0.031852	0.00495	0.999999
Residual	24	154.446667	6.435278	NaN	NaN

Formula = Relief ~ Volunteer & B					
	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
Volunteer	3	0.072222	0.024074	0.002305	0.999843
B	2	123.66	61.83	5.919894	0.008131
Volunteer:B	6	0.331111	0.055185	0.005284	0.999999
Residual	24	250.666667	10.444444	NaN	NaN

Formula = Relief ~ Volunteer + (A & B)					
	df	**sum_sq**	**mean_sq**	**F**	**PR(>F)**
Volunteer	**3**	**0.07222**	**0.02407**	**0.372093**	**0.77385**
A	2	220.02	110.01	1700.33345	1.40E-26
B	2	123.66	61.83	955.655098	1.32E-23
A:B	4	29.425	7.35625	113.699463	3.12E-15
Residual	24	1.55278	0.0647	NaN	NaN

p-value of A & B are statistically significant and presents evidence against the H0.

Problem 2:

Summary:
The data contains various parameters of students from various institutions.

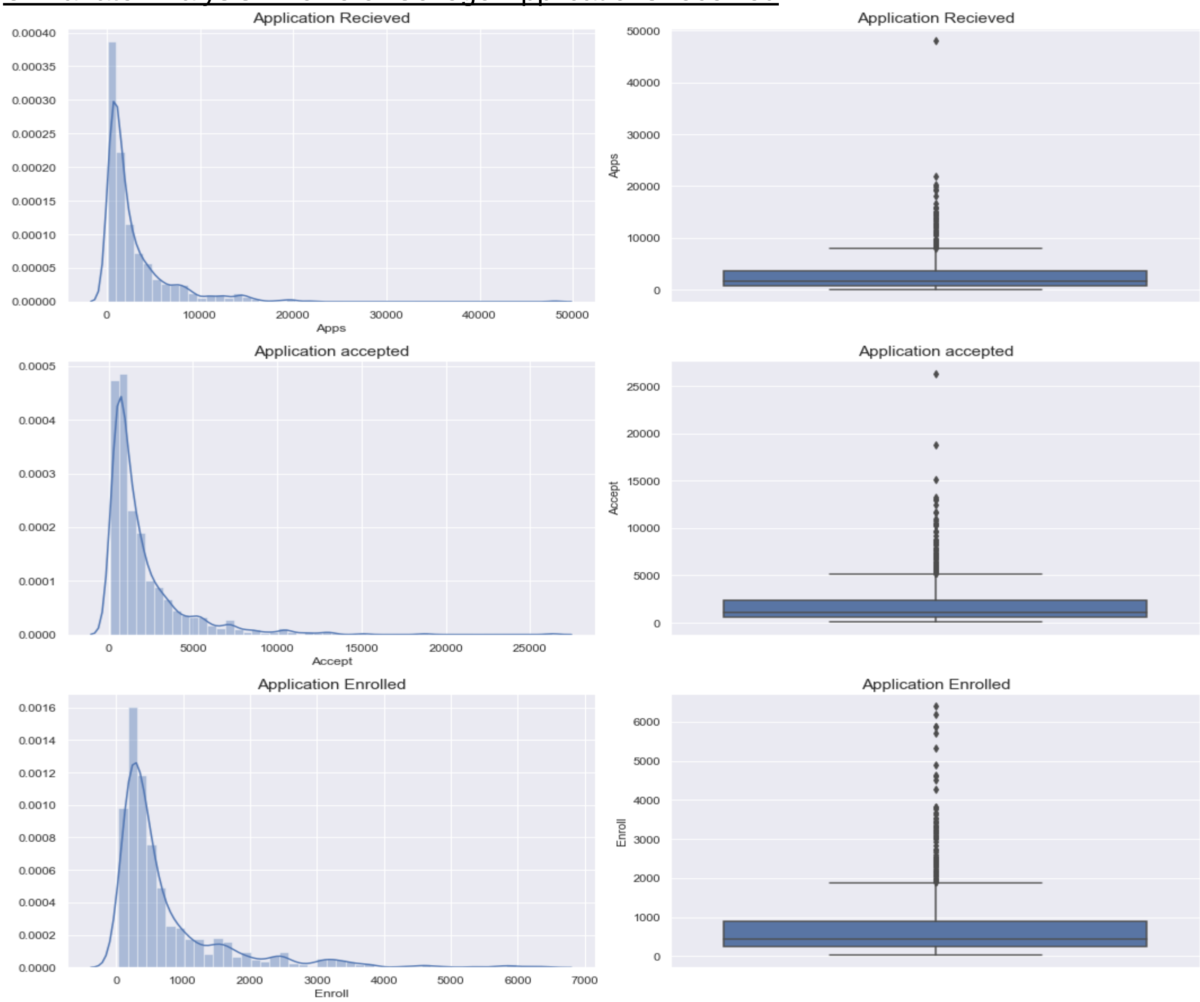
Assumption: The PCA exercise is done with the intention of finding the most relevant parameter which is found in the current students in various universities.

This may help the universities to formulate strategies to target students which may result both into better conversions/admissions for the Universities and better education outcome for the students.

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

Univariate Analysis: We have clubbed all the variables into 4 different categories to look at univariate aspect of analysis.

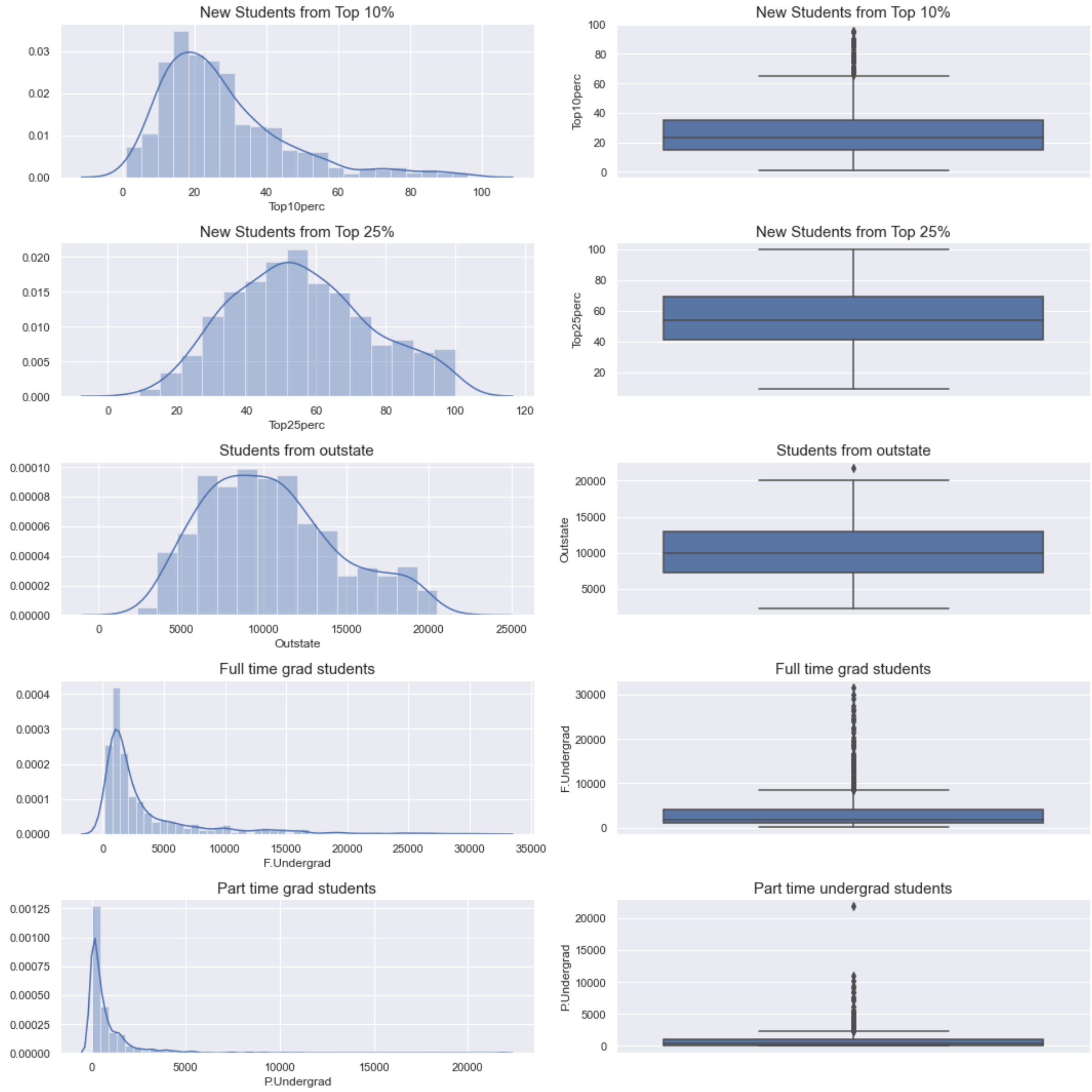
- Univariate Analysis: Profile of College Applications received



Inferences:

- No of applications received / Accepted and Enrolled are heavily right skewed with huge positive outliers.
- There is a gradual decrease of outliers from the stage of applications received -> Accepted -> Enrolled. Evident by the reduced coff of variance.

Univariate Analysis: Profile of Aspiring college students:



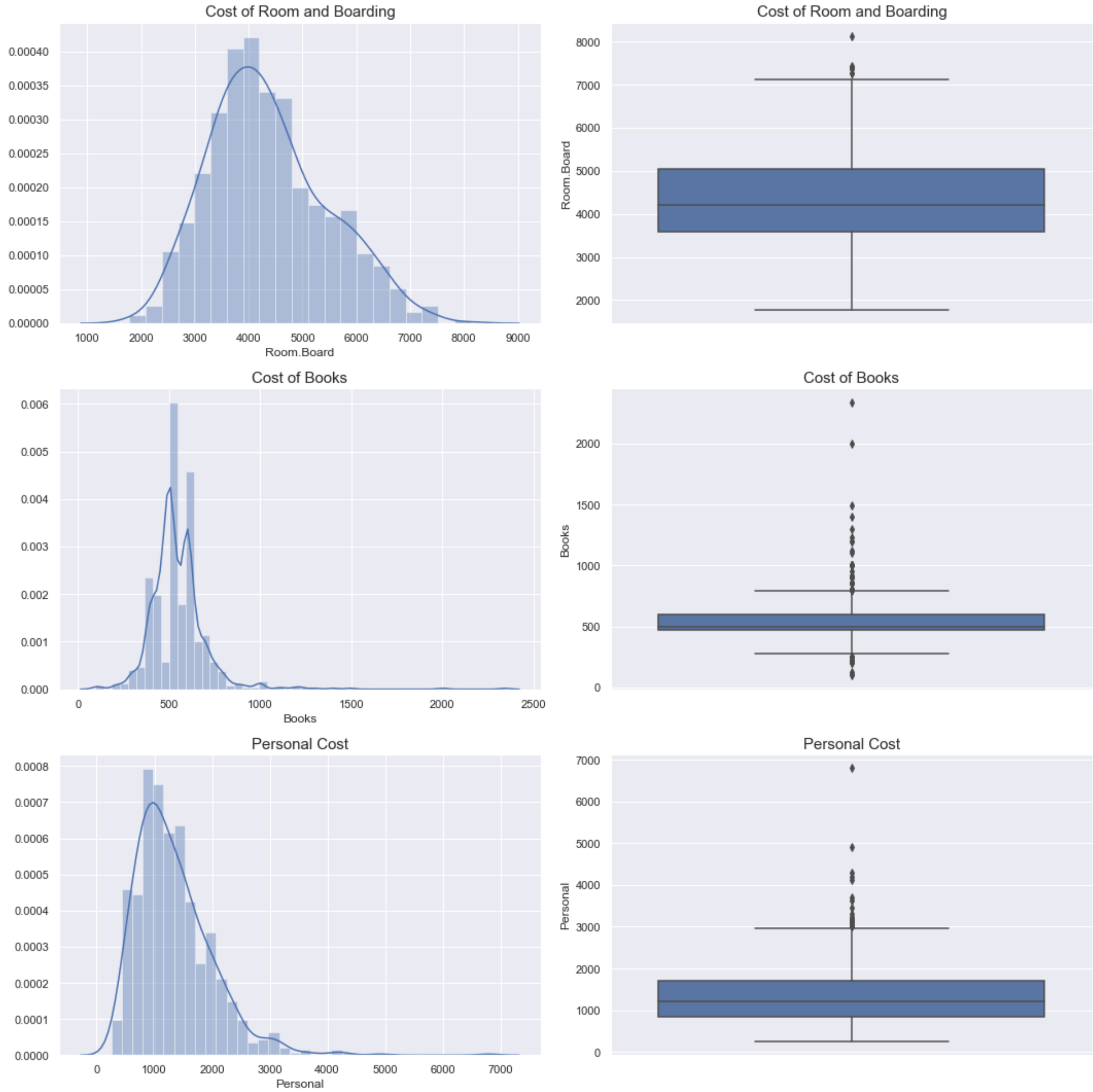
<u>Inferences:</u>

- Students from Top 10% high schools are right skewed with heavy outliers. Less probable event, but high number of students, if there.

- Students from **Top 25% high schools** and from **Outside the state** are slightly right skewed with almost no outliers.

- Full time and Part time grad students are again right skewed and with heavy number of outliers.

- Univariate Analysis: Cost to the students:



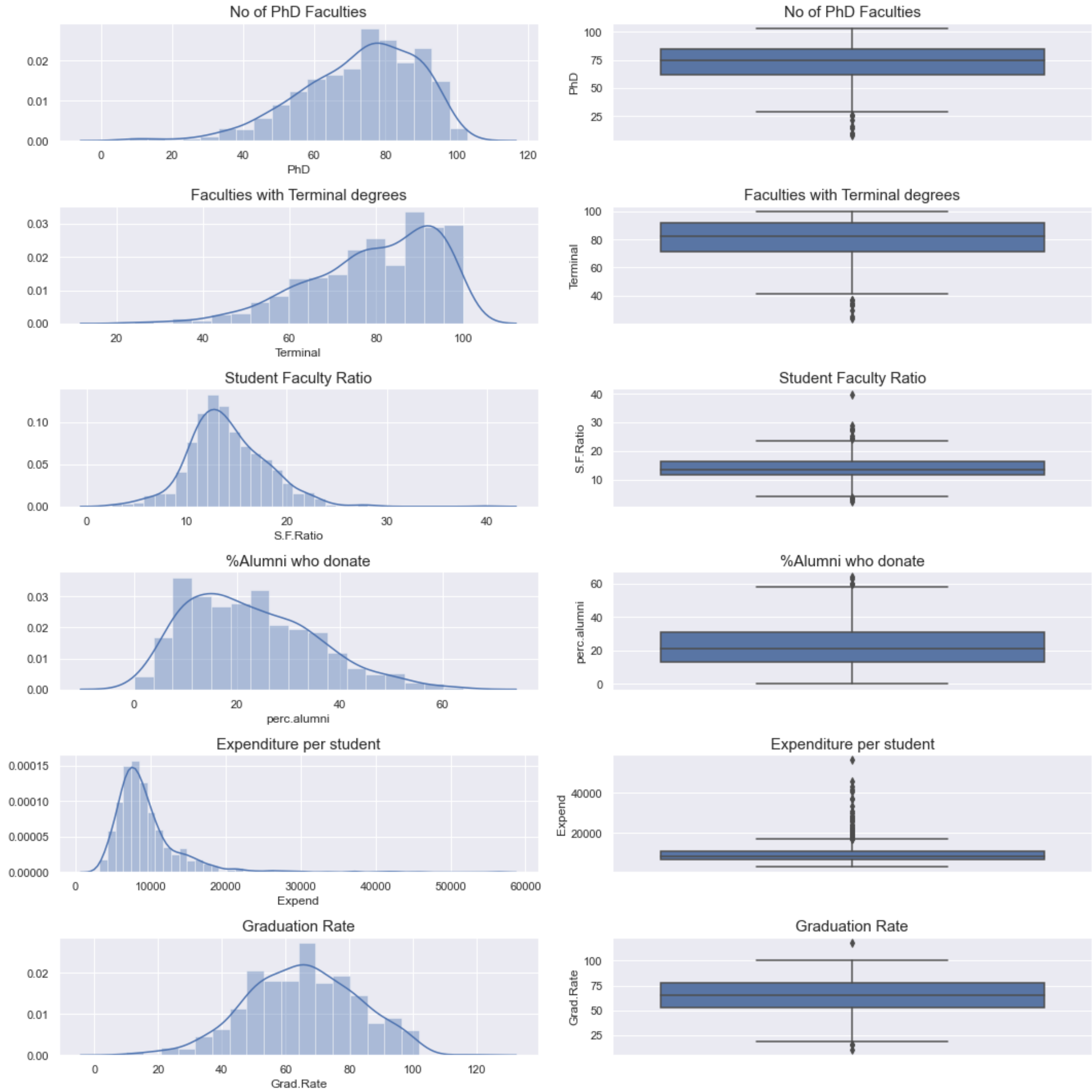
Inference:

- Costs are mostly normally distributed to the students. We can safely assume the education/tuition cost is normally distributed.

- Cost of books displays heavy outliers with slightly right skewed distribution.

- Personal costs have positive outliers which is due to the cost of living of the city the University is situated in.

- <u>Univariate Analysis: Profile of College Faculty and Alumni and Graduation rate:</u>



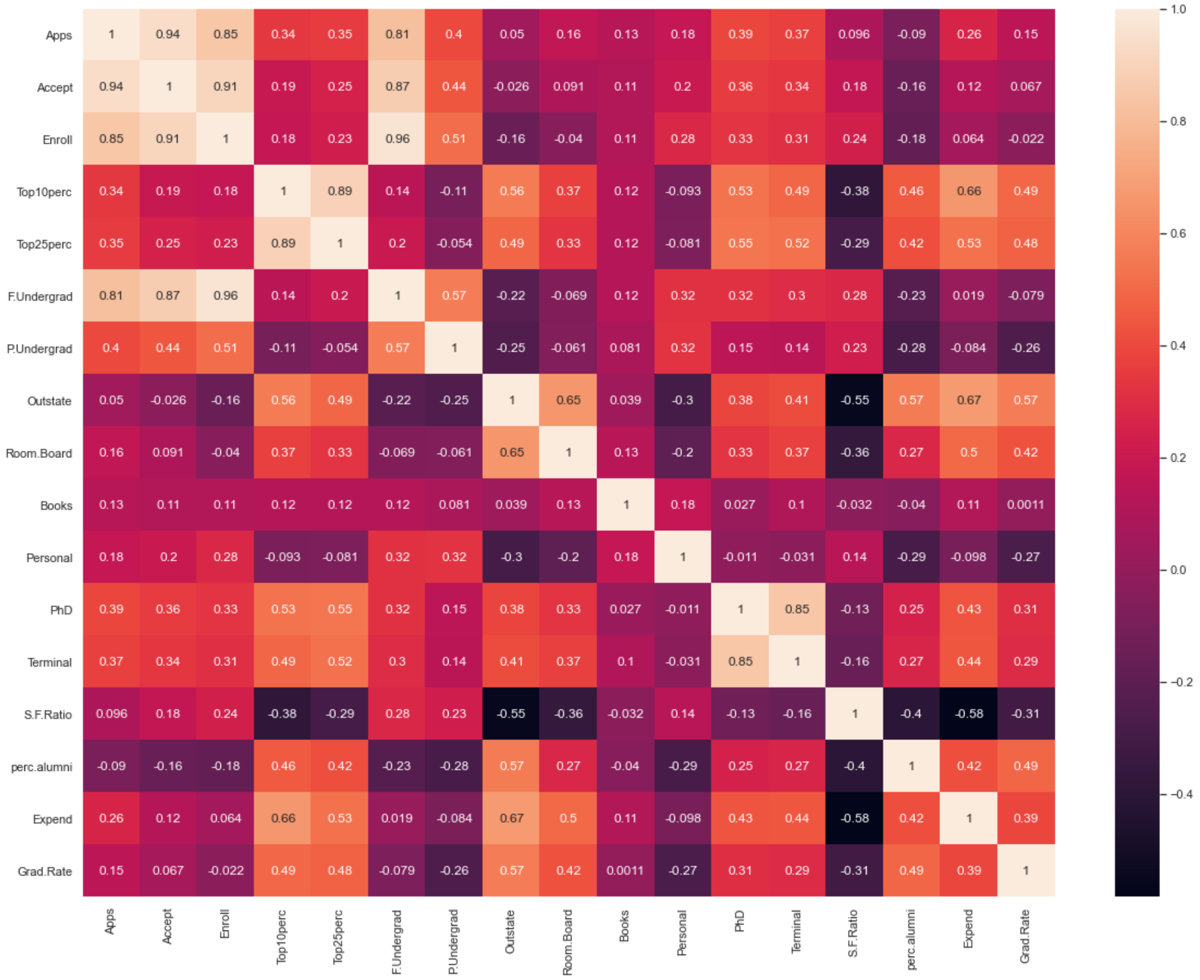
<u>Inferences:</u>

- **No of PhD faculties** and **faculty with Terminal degrees** have negatively skewed distribution, which is evidence of high-level qualification of faculties across universities.

- Student-Faculty ratio: Outliers on both the side, with close to normal distribution.
- Donations from alumni is slightly right skewed.
- **Expenditure per student** has positive outliers and have right skewed distribution
- **Graduation Rate:** Normally distributed; outliers on both the sides.

MultiVariate Analysis

- There is a strong correlation between (1) number of applications, (2) Acceptations and (3) Enrollments (as discussed previously in 2.1.1)
- Major positive correlations include-
    - (Applications accepted & Full time Students),
    - (Students enrolled & Full time Students),
- Major negative correlations include-
    - (Student:Facutly ratio & Outstate students)
    - (Student:Facutly ratio & Instructional spend per student)



Note: Applying a pairplot image is not possible due to it size. Hence to support my inferences please find the heatmap of the correlation matrix.

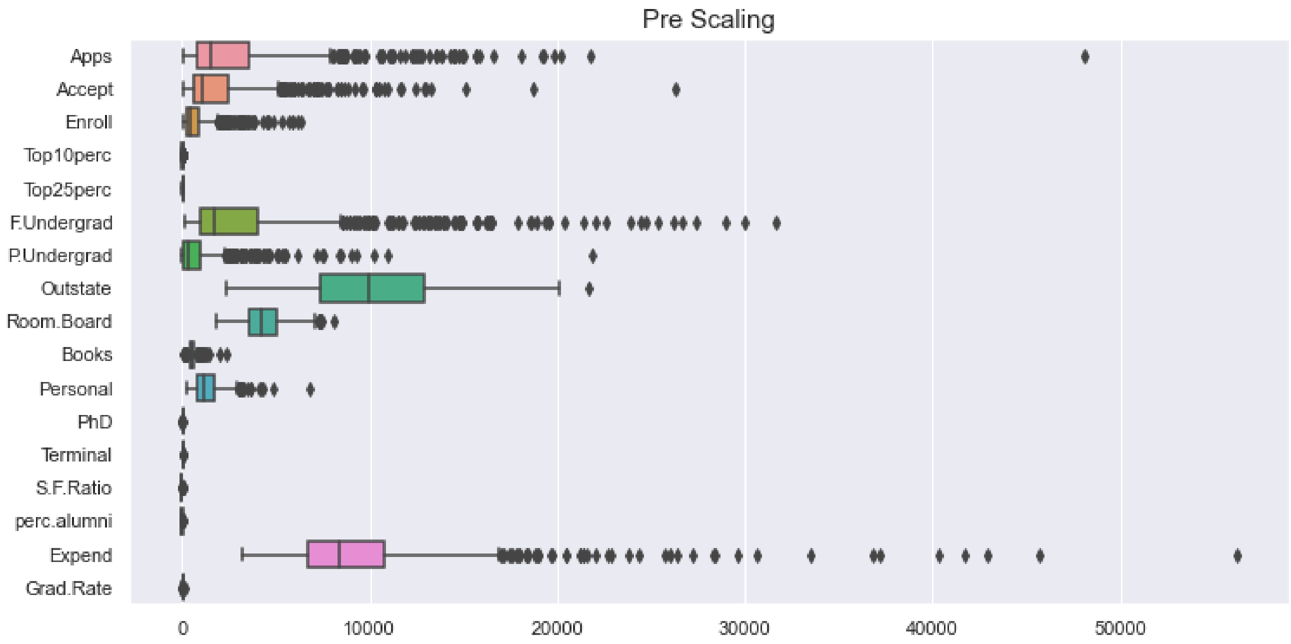
2.2) Scale the variables and write the inference for using the type of scaling function for this casestudy

- All variables are on same scale (continuous numeric), hence we can use Z-scaler to standardise the data.
- After analysing the summary, there are some variables with extreme outliers, which are likely to influence the overall analysis.
- The variables with extreme outliers, hence require adjustments, are highlighted as under.

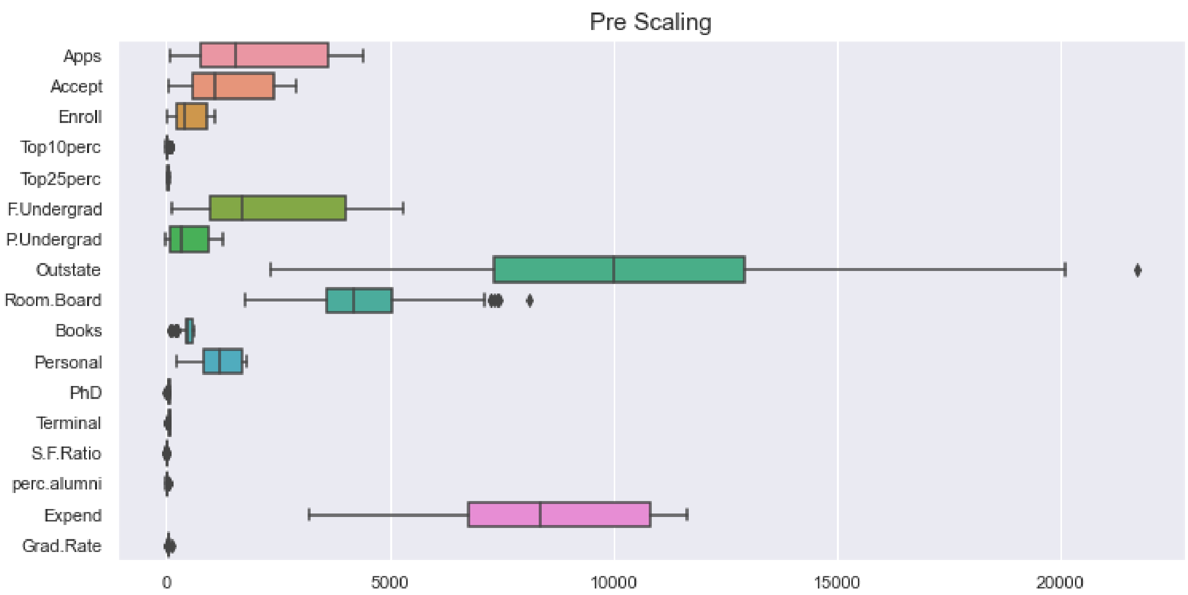
	count	mean	std	min	0.25	0.50	0.75	max
Apps	777.00	3,001.64	3,870.20	81.00	776.00	1,558.00	3,624.00	48,094.00
Accept	777.00	2,018.80	2,451.11	72.00	604.00	1,110.00	2,424.00	26,330.00
Enroll	777.00	779.97	929.18	35.00	242.00	434.00	902.00	6,392.00
Top10perc	777.00	27.56	17.64	1.00	15.00	23.00	35.00	96.00
Top25perc	777.00	55.80	19.80	9.00	41.00	54.00	69.00	100.00
F.Undergrad	777.00	3,699.91	4,850.42	139.00	992.00	1,707.00	4,005.00	31,643.00
P.Undergrad	777.00	855.30	1,522.43	1.00	95.00	353.00	967.00	21,836.00
Outstate	777.00	10,440.67	4,023.02	2,340.00	7,320.00	9,990.00	12,925.00	21,700.00
Room.Board	777.00	4,357.53	1,096.70	1,780.00	3,597.00	4,200.00	5,050.00	8,124.00
Books	777.00	549.38	165.11	96.00	470.00	500.00	600.00	2,340.00
Personal	777.00	1,340.64	677.07	250.00	850.00	1,200.00	1,700.00	6,800.00
PhD	777.00	72.66	16.33	8.00	62.00	75.00	85.00	103.00
Terminal	777.00	79.70	14.72	24.00	71.00	82.00	92.00	100.00
S.F.Ratio	777.00	14.09	3.96	2.50	11.50	13.60	16.50	39.80
perc.alumni	777.00	22.74	12.39	-	13.00	21.00	31.00	64.00
Expend	777.00	9,660.17	5,221.77	3,186.00	6,751.00	8,377.00	10,830.00	56,233.00
Grad.Rate	777.00	65.46	17.18	10.00	53.00	65.00	78.00	118.00

To normalise the impact of extreme outliers, we have:

- Replaced the positive outliers with a relatively lower value i.e. 80^{th} %tile.
- Replaced the negative outliers with a relatively higher value i.e. 10^{th} %tile.
- This is an assumption to mitigate the extreme impact of outliers over the Principal Component Analysis.



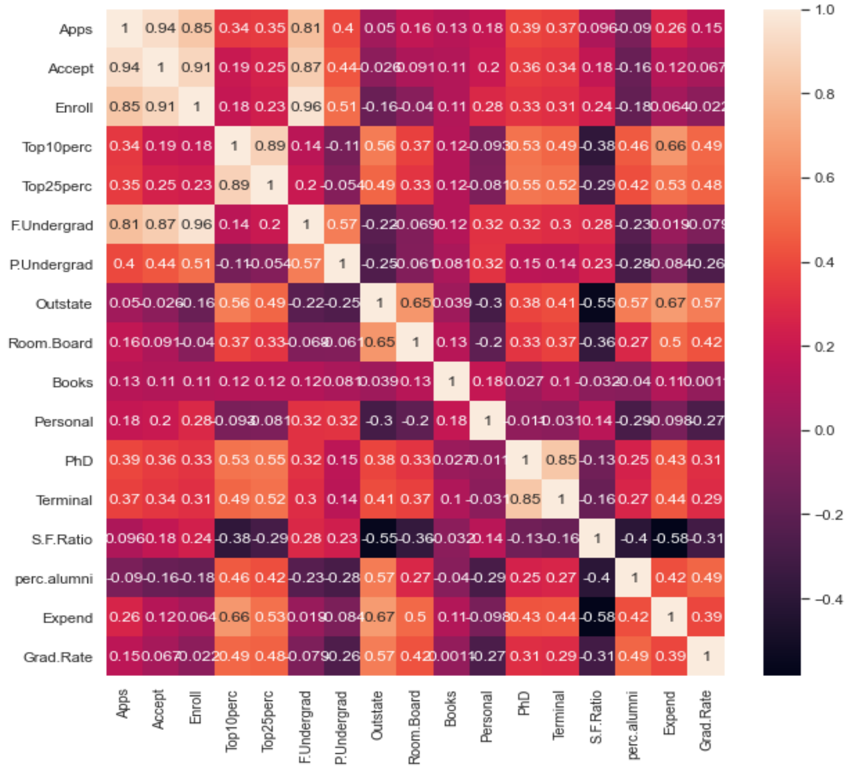
Note: This box-plot is before the scaling and adjustments to the outliers.

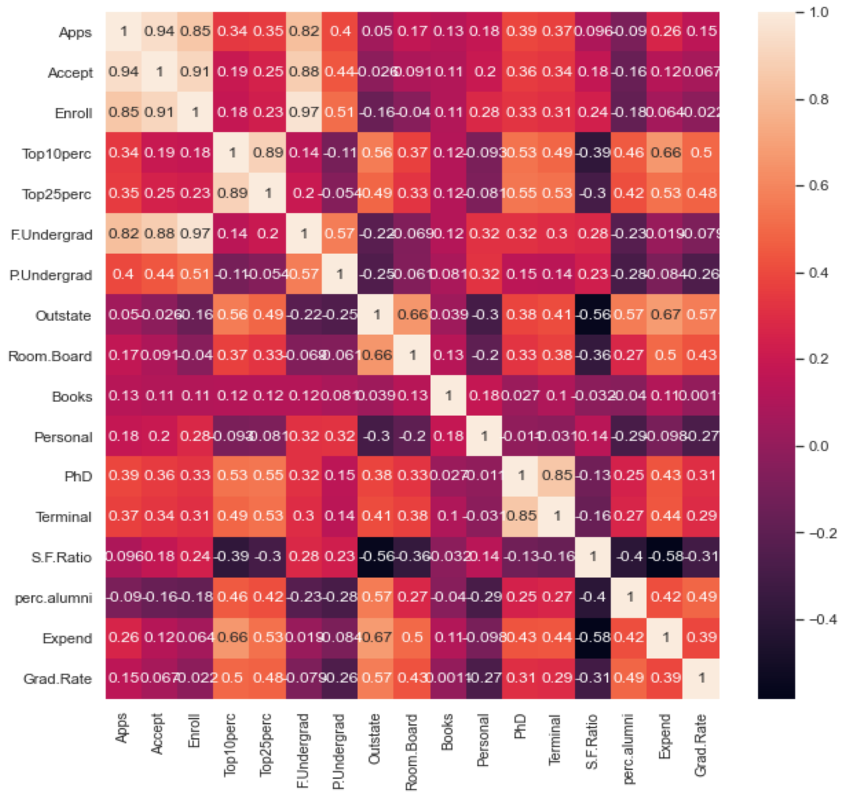


Note: This box-plot is after adjustments to the outliers, but before z-scaling.

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

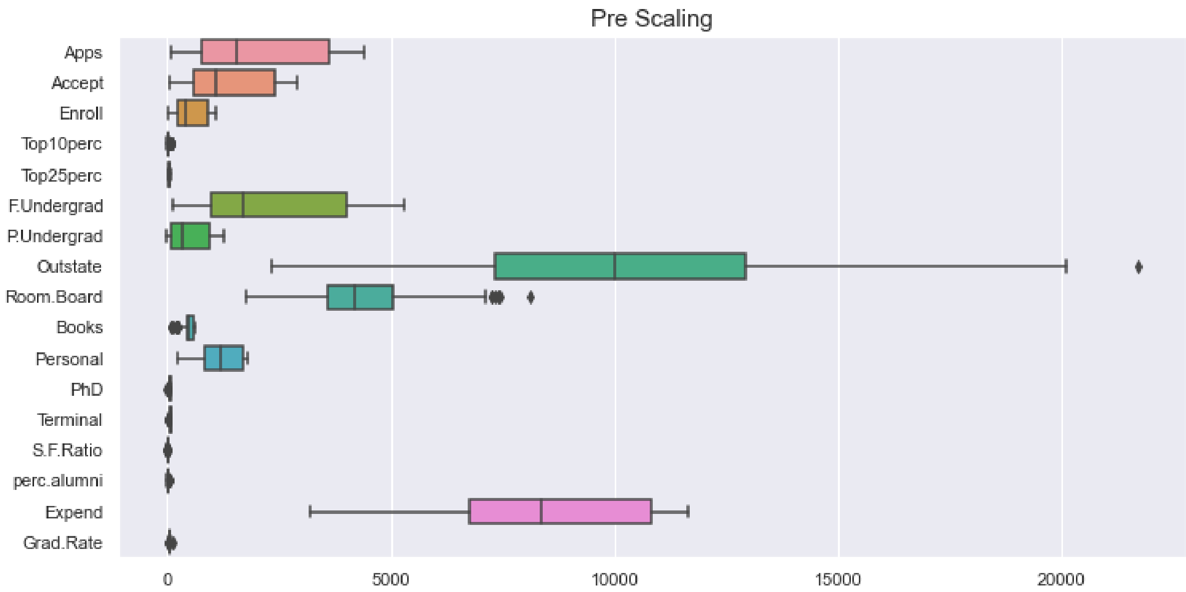
Covariance matrix and **Correlation matrix** of the Scaled database is **exactly similar**.

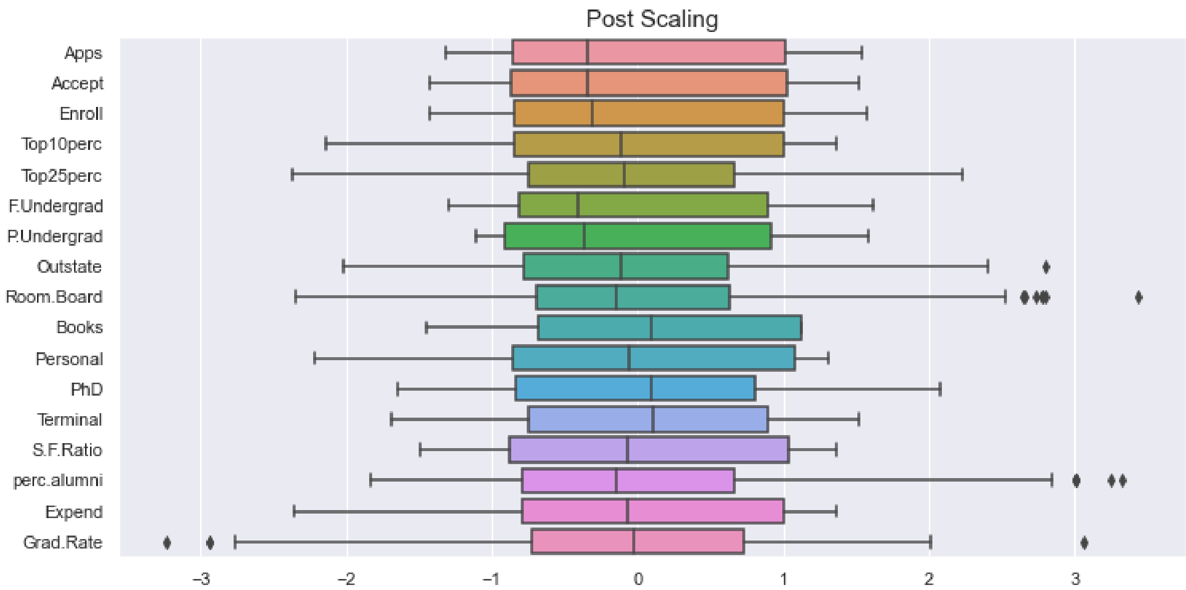




2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

We scaled the data using Z-value Scaler. The pre and post scaling data is as follows:





Conclusions:

1. After adjustment to outliers, the boxplots of each variables correctly represent the skewness.
2. Skewness: Only 2 variables (Faculty with Phd and with Terminal degrees) have negative skew, as inferred by the median and outliers of scaled data.

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

Covariance Matrix:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1.001	0.967	0.907	0.247	0.351	0.878	0.492	0.075	0.189	0.211	0.219	0.482	0.455	0.221	-0.092	0.157	0.159
2	0.967	1.001	0.937	0.202	0.282	0.899	0.521	0.035	0.138	0.185	0.231	0.457	0.431	0.262	-0.135	0.122	0.116
3	0.907	0.937	1.001	0.158	0.249	0.963	0.575	-0.114	0.009	0.203	0.294	0.399	0.370	0.332	-0.202	0.000	0.031
4	0.247	0.202	0.158	1.001	0.893	0.102	-0.206	0.521	0.306	0.070	-0.128	0.529	0.486	-0.344	0.434	0.569	0.464
5	0.351	0.282	0.249	0.893	1.001	0.207	-0.141	0.490	0.332	0.085	-0.089	0.547	0.525	-0.283	0.418	0.513	0.478
6	0.878	0.899	0.963	0.102	0.207	1.001	0.643	-0.187	-0.008	0.232	0.318	0.377	0.347	0.373	-0.276	-0.045	-0.033
7	0.492	0.521	0.575	-0.206	-0.141	0.643	1.001	-0.368	-0.057	0.132	0.314	0.078	0.074	0.401	-0.436	-0.250	-0.238
8	0.075	0.035	-0.114	0.521	0.490	-0.187	-0.368	1.001	0.655	-0.070	-0.341	0.383	0.409	-0.584	0.567	0.741	0.572
9	0.189	0.138	0.009	0.306	0.332	-0.008	-0.057	0.655	1.001	0.034	-0.221	0.330	0.375	-0.360	0.273	0.560	0.425
10	0.211	0.185	0.203	0.070	0.085	0.232	0.132	-0.070	0.034	1.001	0.208	0.084	0.098	0.072	-0.098	0.081	-0.037
11	0.219	0.231	0.294	-0.128	-0.089	0.318	0.314	-0.341	-0.221	0.208	1.001	-0.028	-0.052	0.216	-0.305	-0.189	-0.278
12	0.482	0.457	0.399	0.529	0.547	0.377	0.078	0.383	0.330	0.084	-0.028	1.001	0.851	-0.100	0.249	0.473	0.305
13	0.455	0.431	0.370	0.486	0.525	0.347	0.074	0.409	0.375	0.098	-0.052	0.851	1.001	-0.136	0.267	0.503	0.290
14	0.221	0.262	0.332	-0.344	-0.283	0.373	0.401	-0.584	-0.360	0.072	0.216	-0.100	-0.136	1.001	-0.428	-0.642	-0.292
15	-0.092	-0.135	-0.202	0.434	0.418	-0.276	-0.436	0.567	0.273	-0.098	-0.305	0.249	0.267	-0.428	1.001	0.432	0.492
16	0.157	0.122	0.000	0.569	0.513	-0.045	-0.250	0.741	0.560	0.081	-0.189	0.473	0.503	-0.642	0.432	1.001	0.393
17	0.159	0.116	0.031	0.464	0.478	-0.033	-0.238	0.572	0.425	-0.037	-0.278	0.305	0.290	-0.292	0.492	0.393	1.001

Eigen Vectors:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.269	0.317	0.116	-0.047	-0.151	-0.071	0.156	0.006	-0.145	-0.075	-0.436	-0.451	-0.056	-0.569	-0.099	-0.041	-0.024
2	0.250	0.336	0.144	-0.045	-0.127	-0.084	0.196	0.009	-0.160	-0.052	0.632	0.410	0.030	-0.349	0.139	-0.034	-0.032
3	0.209	0.371	0.086	0.035	-0.125	-0.094	0.188	0.044	-0.068	-0.011	-0.531	0.516	-0.088	0.420	0.106	0.022	0.023
4	0.321	-0.111	-0.252	0.350	-0.083	-0.121	-0.257	0.328	0.050	-0.086	-0.025	-0.128	0.089	-0.022	0.671	0.003	-0.131
5	0.336	-0.068	-0.218	0.354	-0.119	-0.108	-0.291	0.265	0.058	-0.138	0.063	0.160	-0.143	-0.034	-0.649	-0.046	0.182
6	0.184	0.390	0.061	0.003	-0.084	-0.058	0.119	0.077	0.005	0.006	0.349	-0.558	-0.109	0.569	-0.071	0.073	0.023
7	0.009	0.345	0.130	-0.215	0.072	-0.049	-0.275	0.109	0.791	0.115	-0.026	0.046	0.247	-0.092	-0.019	-0.105	-0.007
8	0.284	-0.253	0.129	-0.230	-0.068	-0.070	0.023	-0.068	-0.142	-0.138	-0.013	-0.048	0.427	0.185	-0.039	-0.711	0.071
9	0.244	-0.129	0.214	-0.531	-0.036	0.083	-0.469	-0.076	-0.061	-0.391	0.002	0.034	-0.387	0.052	0.083	0.196	-0.082
10	0.059	0.112	-0.646	-0.300	-0.338	0.578	0.114	0.016	0.065	0.031	0.006	0.028	-0.013	-0.007	0.005	-0.092	-0.018
11	-0.046	0.224	-0.504	-0.080	0.070	-0.532	-0.230	-0.579	-0.084	-0.038	0.007	-0.005	0.040	0.004	0.017	-0.020	0.014
12	0.333	0.050	-0.037	0.130	0.492	0.233	-0.014	-0.170	-0.028	0.169	-0.006	0.033	-0.047	0.032	-0.150	-0.097	-0.691
13	0.331	0.036	-0.029	0.052	0.519	0.264	0.022	-0.181	0.024	0.118	-0.003	-0.020	-0.116	-0.050	0.172	-0.011	0.673
14	-0.138	0.295	0.142	0.300	-0.037	0.404	-0.325	-0.204	-0.200	-0.383	-0.023	-0.006	0.501	0.039	-0.039	0.173	0.037
15	0.187	-0.260	0.043	0.231	-0.170	0.015	0.419	-0.435	0.487	-0.436	0.020	-0.015	-0.052	0.011	0.020	0.123	-0.051
16	0.303	-0.188	-0.130	-0.308	0.129	-0.153	0.184	0.159	-0.061	0.056	-0.031	0.011	0.531	0.027	-0.140	0.593	0.008
17	0.245	-0.157	0.226	0.110	-0.479	0.085	-0.243	-0.373	-0.010	0.629	0.006	-0.008	0.043	0.016	0.010	0.137	0.030

Eigen Values:

	Eigen Values
1	5.518
2	4.869
3	1.097
4	1.058
5	0.866
6	0.772
7	0.572
8	0.606
9	0.437
10	0.395
11	0.022
12	0.034
13	0.246
14	0.106
15	0.097
16	0.181
17	0.148

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

Explicit form of the PC 1:

PC 1= (0.269 * Apps) + (0.317 * Accept) + (0.116 * Enroll) - (0.047 * Top10Perc) - (0.151 * Top25Perc) - (0.071 * F.Undergrad) + (0.156 * P.Undergrad) + (0.006 * Outstate) - (0.145 * Room.Board) - (0.075* Books) - (0.436 * Personal) + (0.451 * PhD) - (0.056 * Terminal) - (0.569 * S.F.Ratio) - (0.099 * percAlumni) - (0.041 * Expand) - (0.024 * GradRate)

Eigen Vectors of 4PCs is as follows:

	Apps	Accept	Enroll	Top10 perc	Top25 perc	F.Under	P.Under	Outstate	Room Board	Books	Personal	PhD	Terminal	S.F. Ratio	perc alumni	Expend	Grad Rate
PC 1	0.269	0.317	0.116	-0.047	-0.151	-0.071	0.156	0.006	-0.145	-0.075	-0.436	-0.451	-0.056	-0.569	-0.099	-0.041	-0.024
PC 2	0.250	0.336	0.144	-0.045	-0.127	-0.084	0.196	0.009	-0.160	-0.052	0.632	0.410	0.030	-0.349	0.139	-0.034	-0.032
PC 3	0.209	0.371	0.086	0.035	-0.125	-0.094	0.188	0.044	-0.068	-0.011	-0.531	0.516	-0.088	0.420	0.106	0.022	0.023
PC 4	0.321	-0.111	-0.252	0.350	-0.083	-0.121	-0.257	0.328	0.050	-0.086	-0.025	-0.128	0.089	-0.022	0.671	0.003	-0.131

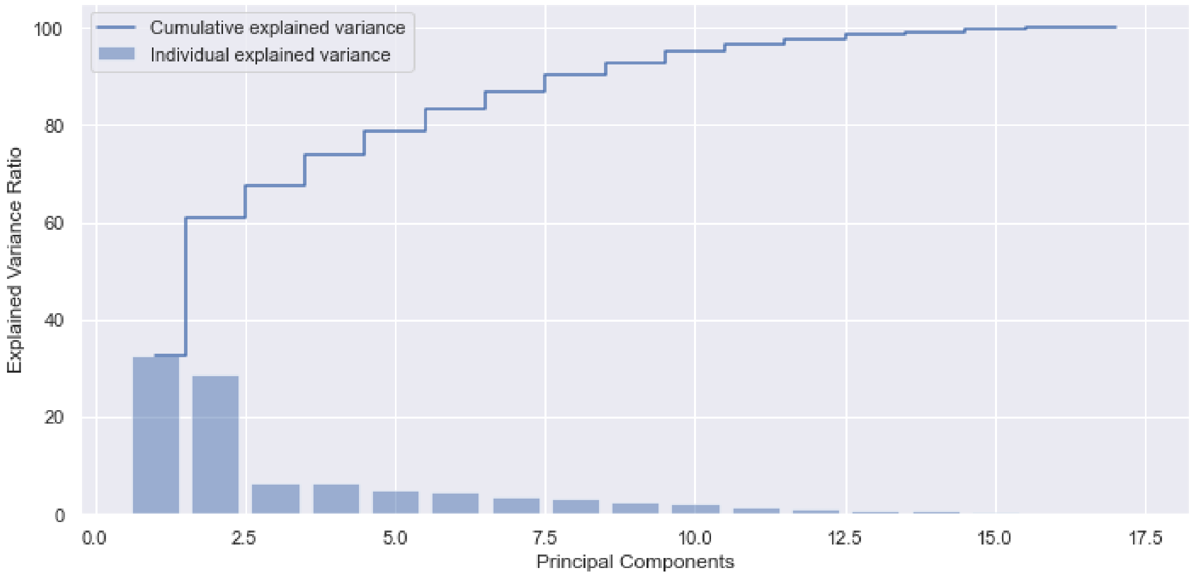
2.7) Discuss the cumulative values of the eigenvalues.

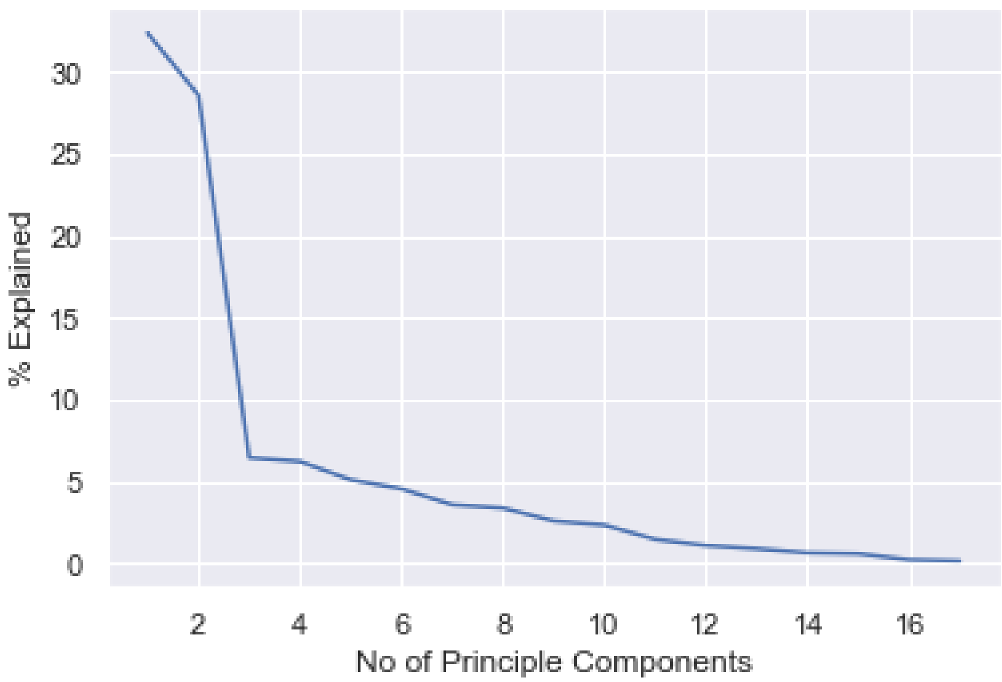
a. How does it help you to decide on the optimum number of principal components?

As per the Scree plot (below), Eigen values helps understand the proportion of the variance explained by Eigen Values.

1. We can see an elbow formation at 3rd Principle component. Which means that beyond 3rd component, the variance explained by successive Principle components is far lower than previous components to explain overall variance.

2. Upto 75% of the variance is explained by 4 Principle components.





b. What do the eigenvectors indicate?

Eigen Vectors indicate the direction of the influence of various Variables towards the Principle components. Principle components further help in reducing the number of variables to capture only those variables which capture the variance to a significant level.

Eigen Vectors are orthogonal to the principles components which incorporate the direction (positive or negative) of various variables, i.e. the impact of variable in explaining the quantum of Principle component.

For example, 3-4 Principle components help explain almost 75% variance in the database.

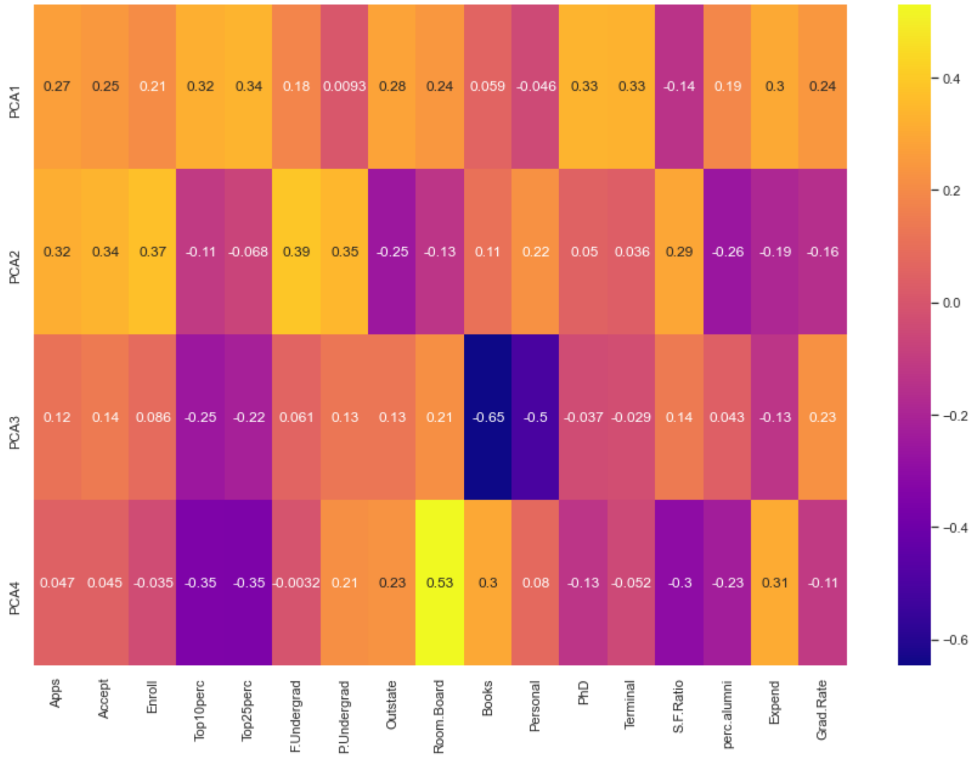
c. Perform PCA and export the data of the Principal Component scores into a data frame.

PCA Component scores

	Apps	Accept	Enroll	Top10 perc	Top25 perc	F.Under grad	P.Under grad	Outstate	Room .Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc. alumni	Expend	Grad. Rate
PCA1	0.269	0.250	0.209	0.321	0.336	0.184	0.009	0.284	0.244	0.059	-0.046	0.333	0.331	-0.138	0.187	0.303	0.245
PCA2	0.317	0.336	0.371	-0.111	-0.068	0.390	0.345	-0.253	-0.129	0.112	0.224	0.050	0.036	0.295	-0.260	-0.188	-0.157
PCA3	0.116	0.144	0.086	-0.252	-0.218	0.061	0.130	0.129	0.214	-0.646	-0.504	-0.037	-0.029	0.142	0.043	-0.130	0.226
PCA4	0.047	0.045	-0.035	-0.350	-0.354	-0.003	0.215	0.230	0.531	0.300	0.080	-0.130	-0.052	-0.300	-0.231	0.308	-0.110

2.8) Mention the business implication of using the Principal Component Analysis for this case study. [**Hint:** Write Interpretations of the Principal Components Obtained]

The interpretation of the Principal Component Analysis can be interpreted through a Heatmap derived from the 4 Principal components and the eigen vectors contributing to the respective components.



Interpretation:

1. Since PC1 explains ~33% of the variation, it is positively influenced by almost all the variables of unsupervised PCA.
2. PC3 is the component with major negative influence of ‘Books’ and ‘Personal’ variables.
3. Overall, all the vectors influence the PCs in a positive manner. A few exceptions are still there.

Thank you!