ADVANCE STATISTICS PROJECT REPORT

By Karthik Sreeram R

SEPTEMBER 12, 2020

UNIVERSITY OF TEXAS AT AUSTIN AND GREAT LAKES

Purpose

This document is the business report for my final project in the subject “Advanced Statistics “.

This document gives us a detailed explanation of various approaches used, their insight and inferences.

Tools used analysis: Python and Jupiter notebook.
Packages used: NumPy, pandas, seaborn, os, matplotlib, SciPy, stats model , sklearn and sweetviz

Contents

Problem 1

Business scenario

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. Perform Anova.

a.) Dataset Head

	A	B	Volunteer	Relief
0	1	1	1	2.4
1	1	1	2	2.7
2	1	1	3	2.3
3	1	1	4	2.5
4	1	2	1	4.6

b.) Type of the variables in dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   A          36 non-null     int64
 1   B          36 non-null     int64
 2   Volunteer  36 non-null     int64
 3   Relief     36 non-null     float64
dtypes: float64(1), int64(3)
memory usage: 1.2 KB
```

c.) Summary of the dataset:

	A	B	Volunteer	Relief
count	36.000000	36.000000	36.000000	36.000000
mean	2.000000	2.000000	2.500000	7.183333
std	0.828079	0.828079	1.133893	3.272090
min	1.000000	1.000000	1.000000	2.300000
25%	1.000000	1.000000	1.750000	4.675000
50%	2.000000	2.000000	2.500000	6.000000
75%	3.000000	3.000000	3.250000	9.325000
max	3.000000	3.000000	4.000000	13.500000

d) Data preprocessing

Converting the column “A” and column “B” into categorical and printing the value counts.

```
3    12
2    12
1    12
Name: A, dtype: int64
```

```
3    12
2    12
1    12
Name: B, dtype: int64
```

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

For Variable A

NULL hypothesis: The mean of all the three levels in the ingredient "A" are equal

H0: μA1 = μA2 = μA3

where A = levels of ingredient A

Alternative hypotheses: At least, one level mean is different from other levels for the ingredient "A"
H1: Means are not all equal.

For Variable B

NULL hypothesis: The mean of all the three levels in the ingredient "B" are equal

H0: μB1 = μB2 = μB3

where B = levels of ingredient B

Alternative hypotheses: At least, one level mean is different from other levels for the ingredient "B"

H1: Means are not all equal.

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

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

Insights

```
alpha =  0.05
 p < alpha  . so  Null hypothesis is rejected. At least, one level mean is different from other level means for the  ingredient A
```

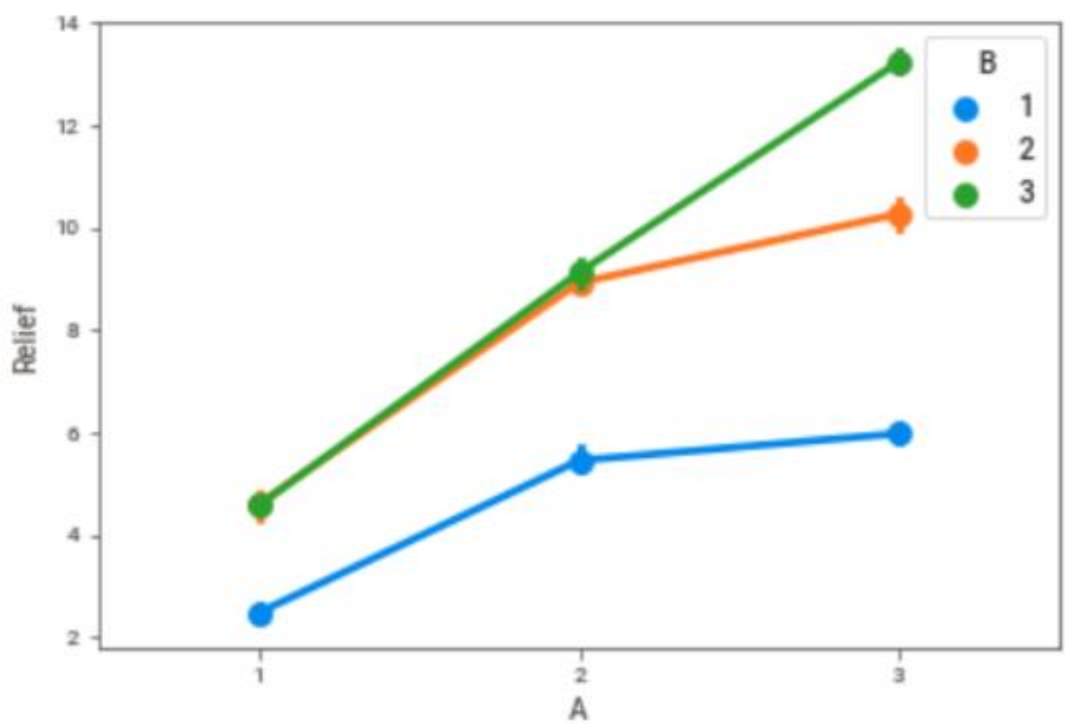
1.3) Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. Stating whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

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

Insights

alpha = 0.05
p < alpha . so Null hypothesis is rejected. At least, one level mean is different from other level means for the ingredient B

1.4) Analyze the effects of one variable on another with the help of an interaction plot.



Insights :

From the above plot we can see that by using level three in both the active ingredicant (A and B) the Relief increases . when we reduce the level, Relief decreases .

1.5) Performing a two-way ANOVA based on the different ingredients (variable ‘A’ & ‘B’ along with their interaction 'A*B') with the variable 'Relief' and stating the results.

From two-way ANOVA, we can test three hypotheses

1) effect of ingredient "A” on Relief.
2) effect of ingredient "B" on Relief.
3) effect of ingredients "A" and "B" interactions on Relief

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

Insights

```
alpha =  0.05
 is PValue is less than Alpha
C(A)          True
C(B)          True
C(A):C(B)     True
```

The P-value obtained from ANOVA analysis for ingredient "A",ingredient "B" and interaction (A & B) are statistically significant (P < alpha)

1.6) The business implications of performing ANOVA for this particular case study

We conclude that Levels of ingredient " A " significantly affects the Relief outcome, levels of ingredient " B " significantly affects the Relief outcome, and interaction of both ingredients " A and B " significantly affects the Relief outcome.

Problem 2

Business scenario

The dataset Education - Post 12th Standard.csv is a dataset that contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. Perform PCA

2.1) Performing Exploratory Data Analysis

a.) Dataset Head

6]:

	Names	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	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

b.) Dataset has any null values.

]:

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
dtype: int64
```

c.) Type of the variables in dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Names        777 non-null    object 
 1   Apps         777 non-null    int64  
 2   Accept       777 non-null    int64  
 3   Enroll       777 non-null    int64  
 4   Top10perc    777 non-null    int64  
 5   Top25perc    777 non-null    int64  
 6   F.Undergrad  777 non-null    int64  
 7   P.Undergrad  777 non-null    int64  
 8   Outstate     777 non-null    int64  
 9   Room.Board   777 non-null    int64  
 10  Books        777 non-null    int64  
 11  Personal     777 non-null    int64  
 12  PhD          777 non-null    int64  
 13  Terminal     777 non-null    int64  
 14  S.F.Ratio    777 non-null    float64
 15  perc.alumni  777 non-null    int64  
 16  Expend       777 non-null    int64  
 17  Grad.Rate    777 non-null    int64  
dtypes: float64(1), int64(16), object(1)
memory usage: 109.4+ KB
```

d.) Summary of the dataset:

73]:

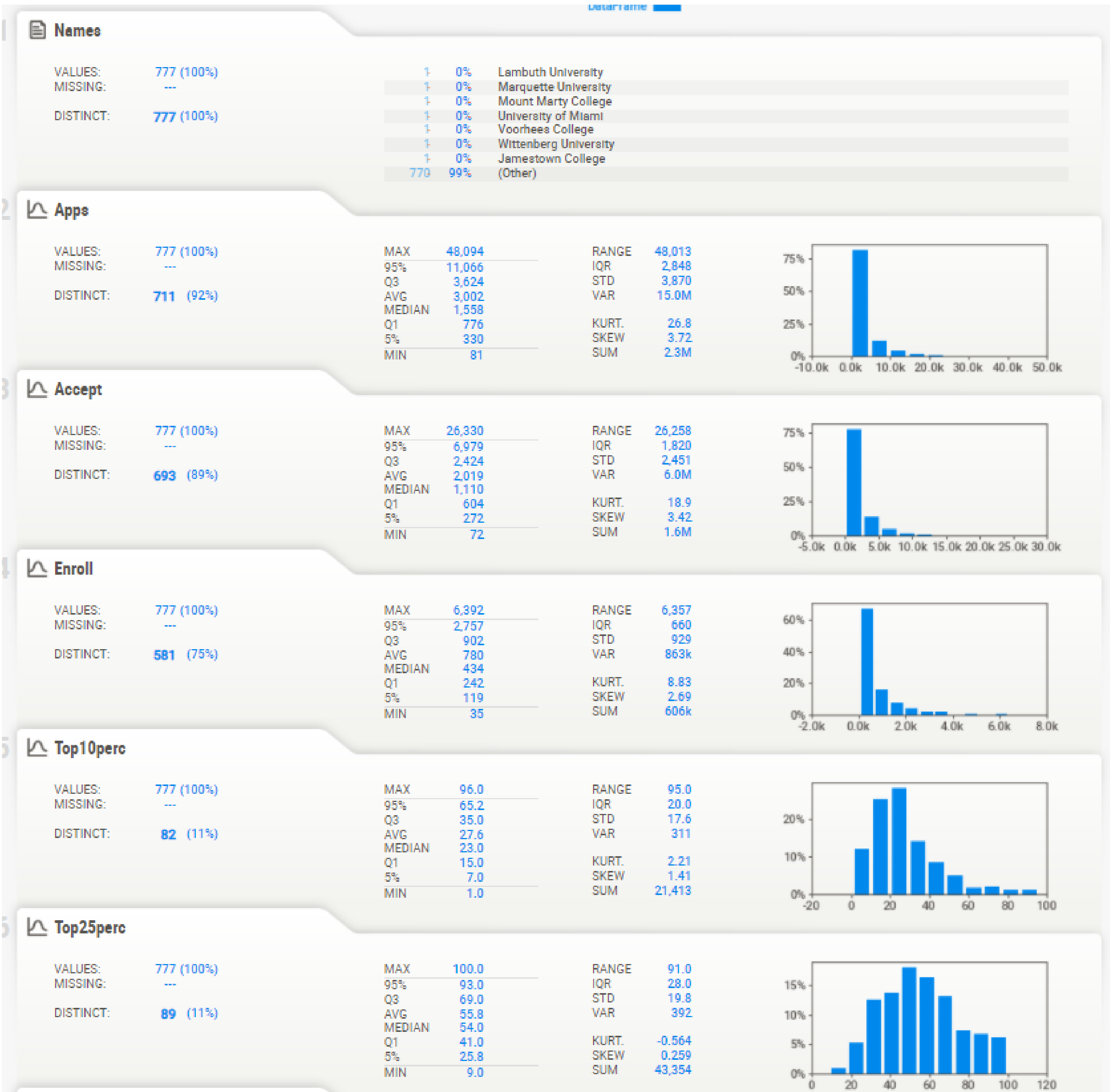
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384	549.380952	1340.642214	72.660232	79.702703	14.089704	22.743887	9660.171171	65.46332
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416	165.105360	677.071454	16.328155	14.722359	3.958349	12.391801	5221.768440	17.17771
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000	96.000000	250.000000	8.000000	24.000000	2.500000	0.000000	3186.000000	10.00000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000	470.000000	850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.00000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000	500.000000	1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.00000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000	600.000000	1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.00000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000	2340.000000	6800.000000	103.000000	100.000000	39.800000	64.000000	56233.000000	118.00000

e.) Check for duplicates.

```
Number of duplicate rows = 0
```

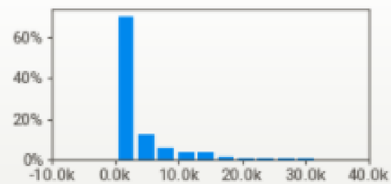
f.) EDA using sweet viz to visualize the summary for each variable as well to underrated the data

Univariate Analysis



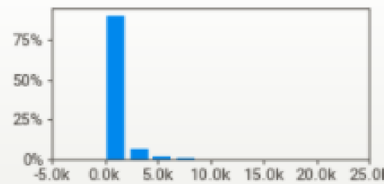
F.Undergrad

VALUES:	777 (100%)	MAX	31,643	RANGE	31,504
MISSING:	---	95%	14,478	IQR	3,013
DISTINCT:	714 (92%)	Q3	4,005	STD	4,850
		AVG	3,700	VAR	23.5M
		MEDIAN	1,707	KURT.	7.70
		Q1	992	SKEW	2.61
		5%	510	SUM	2.9M
		MIN	139		



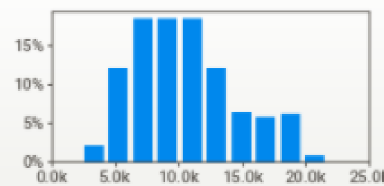
P.Undergrad

VALUES:	777 (100%)	MAX	21,836	RANGE	21,835
MISSING:	---	95%	3,304	IQR	872
DISTINCT:	566 (73%)	Q3	967	STD	1,522
		AVG	855	VAR	2.3M
		MEDIAN	353	KURT.	55.0
		Q1	95	SKEW	5.69
		5%	20	SUM	665k
		MIN	1		



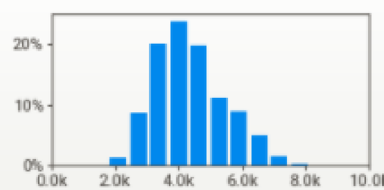
Outstate

VALUES:	777 (100%)	MAX	21,700	RANGE	19,360
MISSING:	---	95%	18,498	IQR	5,605
DISTINCT:	640 (82%)	Q3	12,925	STD	4,023
		AVG	10,441	VAR	16.2M
		MEDIAN	9,990	KURT.	-0.414
		Q1	7,320	SKEW	0.509
		5%	4,602	SUM	8.1M
		MIN	2,340		



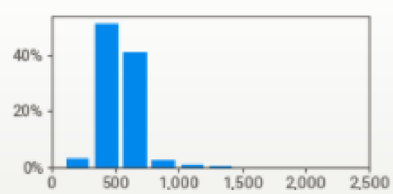
Room.Board

VALUES:	777 (100%)	MAX	8,124	RANGE	6,344
MISSING:	---	95%	6,382	IQR	1,453
DISTINCT:	553 (71%)	Q3	5,050	STD	1,097
		AVG	4,358	VAR	1.2M
		MEDIAN	4,200	KURT.	-0.188
		Q1	3,597	SKEW	0.477
		5%	2,736	SUM	3.4M
		MIN	1,780		



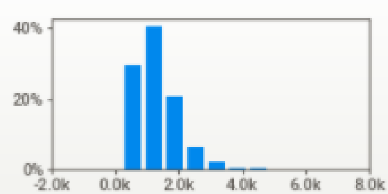
Books

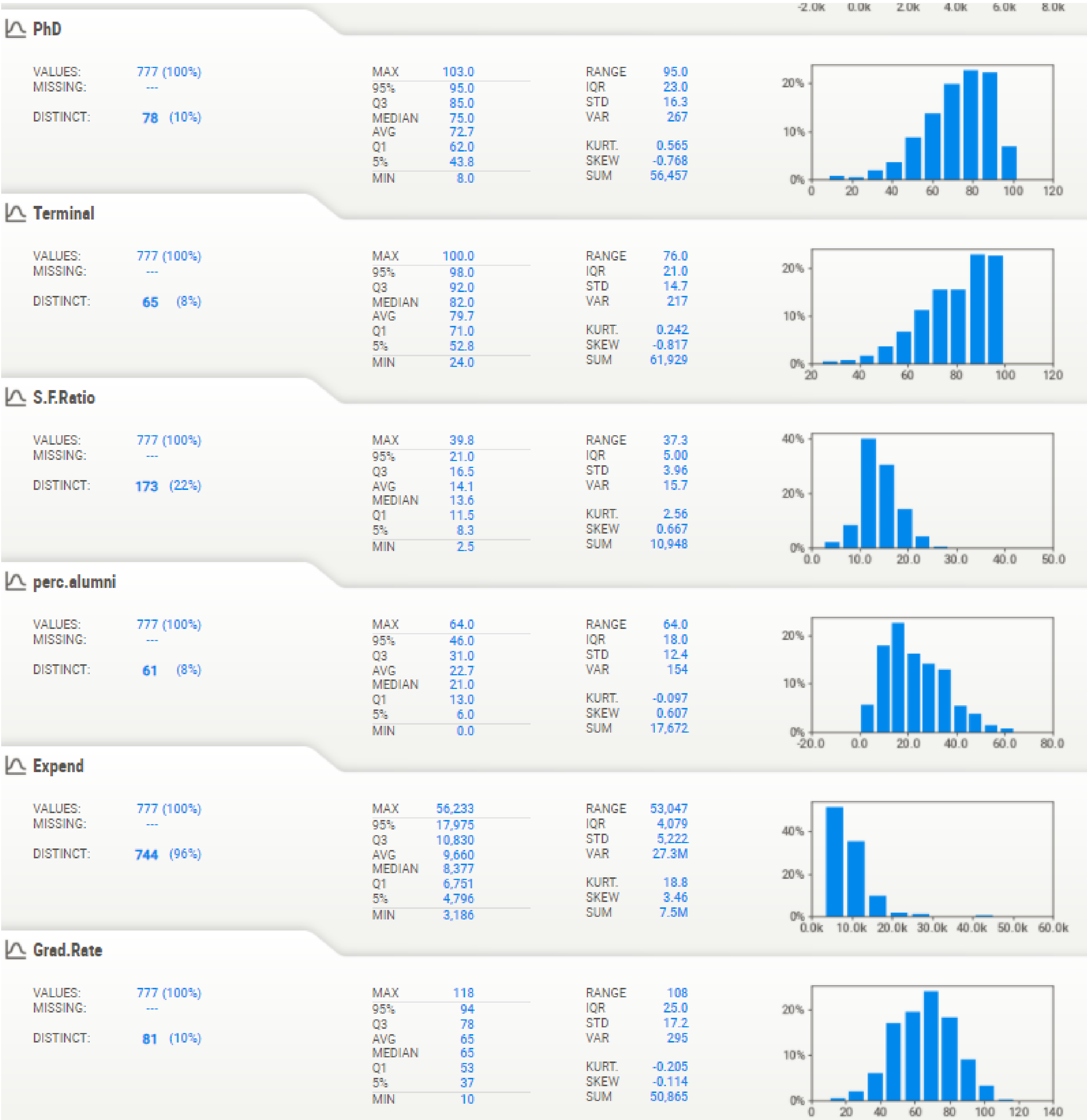
VALUES:	777 (100%)	MAX	2,340	RANGE	2,244
MISSING:	---	95%	766	IQR	130
DISTINCT:	122 (16%)	Q3	600	STD	165
		AVG	549	VAR	27,260
		MEDIAN	500	KURT.	28.3
		Q1	470	SKEW	3.49
		5%	350	SUM	427k
		MIN	96		



Personal

VALUES:	777 (100%)	MAX	6,800	RANGE	6,550
MISSING:	---	95%	2,489	IQR	850
DISTINCT:	294 (38%)	Q3	1,700	STD	677
		AVG	1,341	VAR	458k
		MEDIAN	1,200	KURT.	7.12
		Q1	850	SKEW	1.74
		5%	500	SUM	1.0M
		MIN	250		

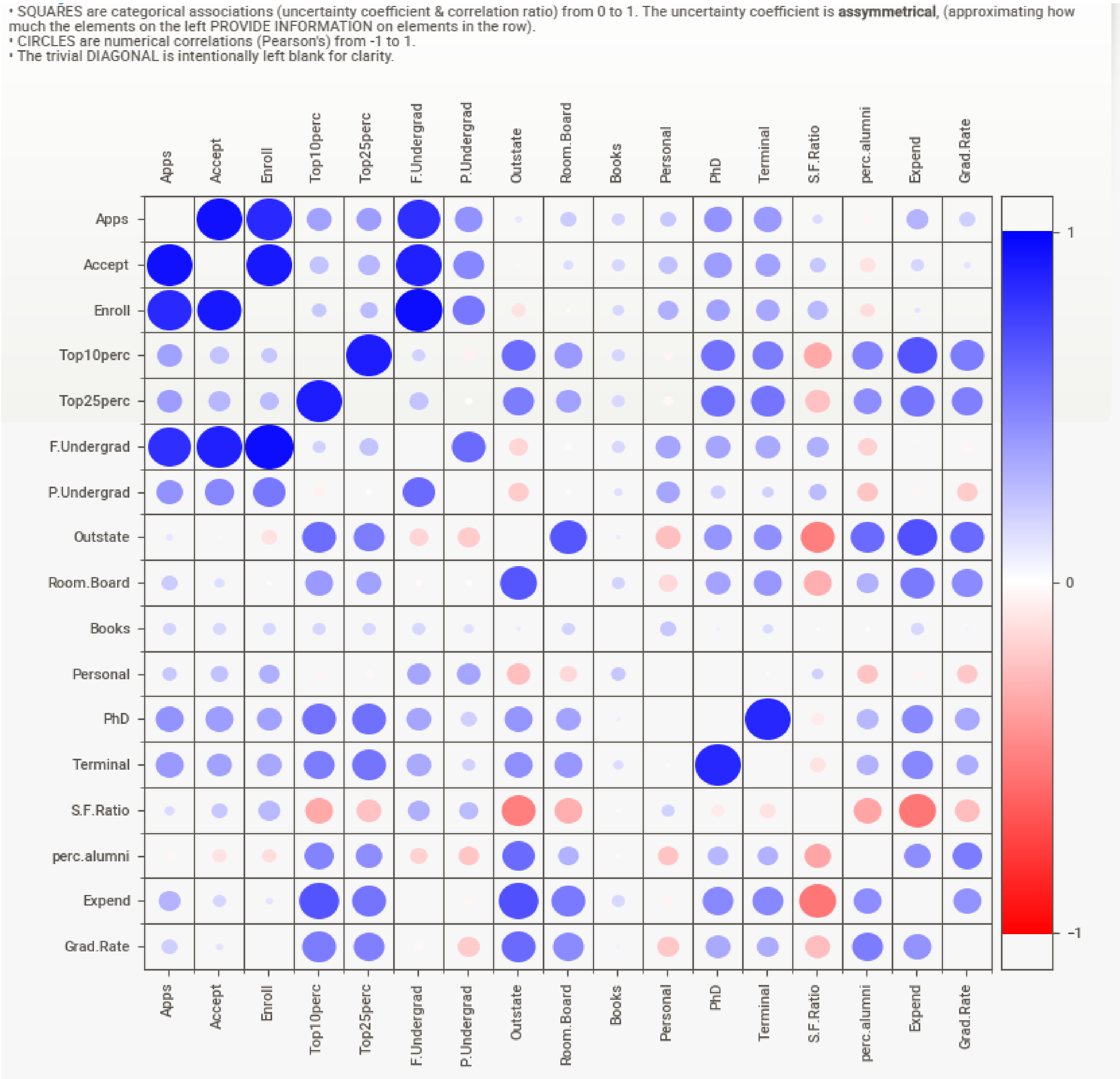


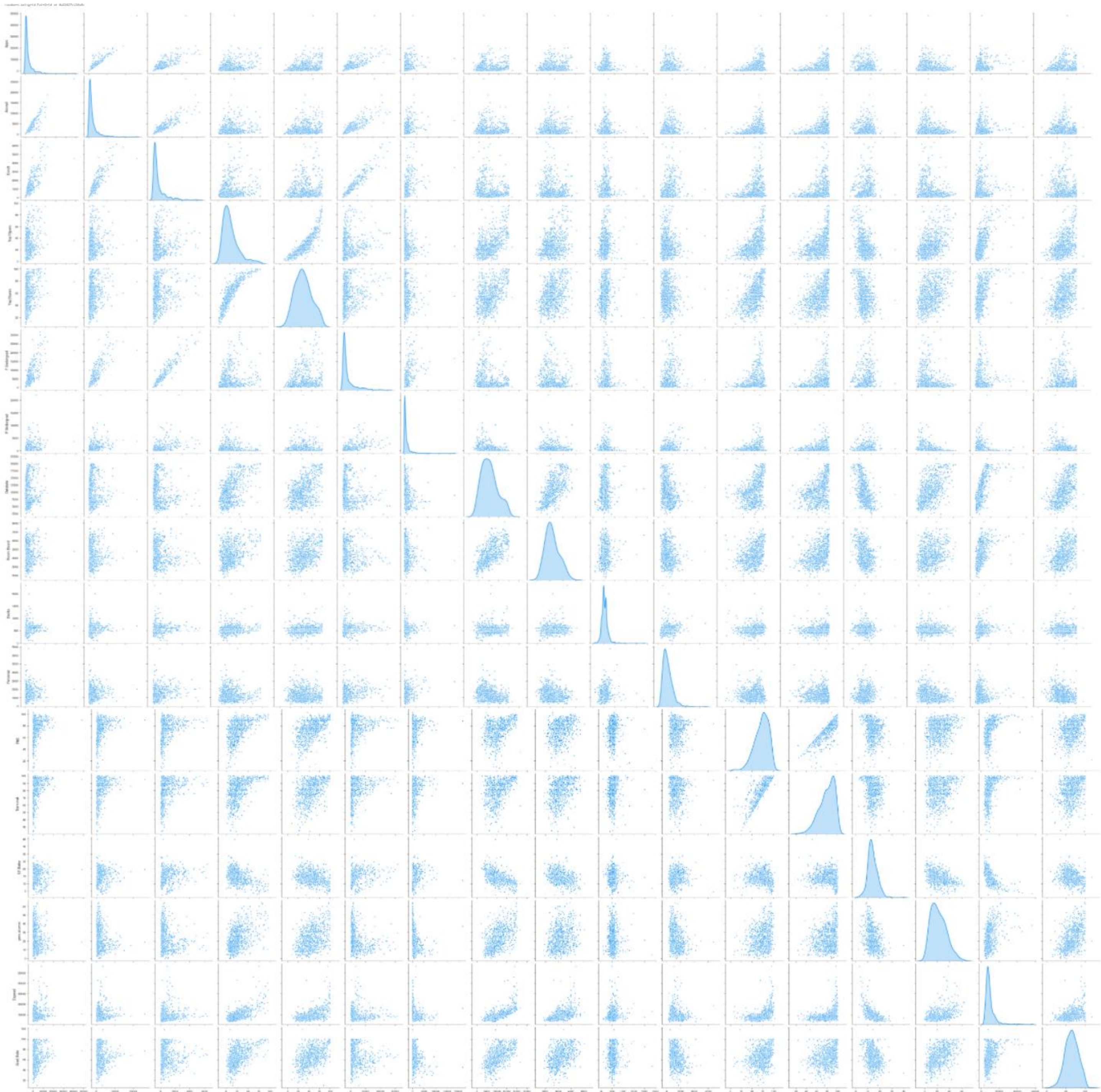


Insights:

Variables Apps, Accept, Enroll, F. Undergrad, P. Undergrad, Top10perc, Books, Personal and Expend looks highly right skewed
Variables Top25pero, Outstate, Room board, Grad. Rate looks like normally distributed
Variables perc. Alumni, S.F ratio are near normal distribution
Variables PhD, Terminal are left skewed
Only variable Name is categorical variable.

Multivariate Analysis:





Insights

Distribution of variables shows that most of the values are concentrated on the lower side. Possible chance of outliers.
F. undergrad, Apps, Accept and Enroll are highly corelated with each other. PhD and terminal are highly corelated with each other. Top10 Perc and Top25 Perc are highly corelated with each other. Outstate and expend are highly corelated. Outstate and S.F ratio are Highly negatively corelated.

g.) Preprocessing:

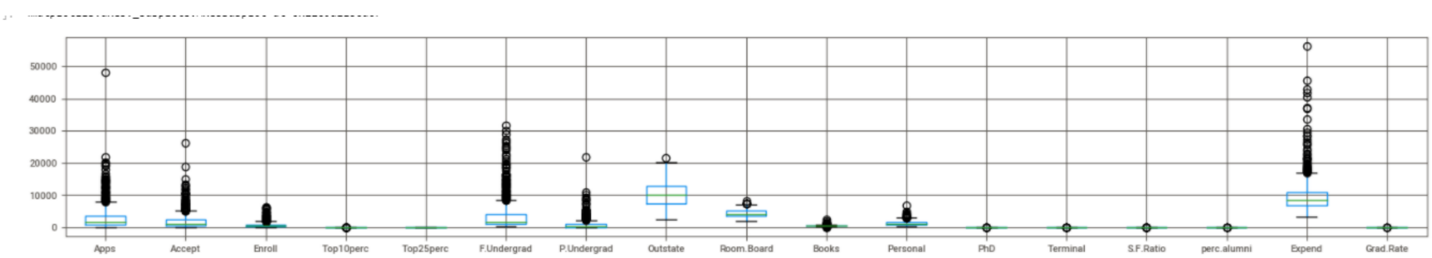
Removing Unwanted columns. In this case-study column "Names" Is removed.

Head of dataset after removing the Names column

180]:

	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	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
1	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
2	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
3	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
4	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15

h.) Box plot to identify outliers



Insights: Outliers are present in the dataset.

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

Note: Here variables outlier is not removed. It is done in heading 2.4 and again scaling of variables is redone.

We have used zscore to scale the variables

Head of dataset after scaling:

To standardize the process we will use zscaler

]:

	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	-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	1.013776	-0.867574	-0.501910	-0.318252
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	-0.477704	-0.544572	0.166110	-0.551262
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	-0.300749	0.585935	-0.177290	-0.667767
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	-1.615274	1.151188	1.792851	-0.376504
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	-0.553542	-1.675079	0.241803	-2.939613

Summary of data after scaling:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16	-1.661405e-16	-3.029180e-17	6.515595e-17	3.570717e-16	-2.192583e-16	4.765243e-17	5.954768e-17	-4.481615e-16	-2.057556e-17	-6.022638e-17	1.213101e-16	3.886495e-16
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00	-7.346169e-01	-5.615022e-01	-2.014878e+00	-2.351778e+00	-2.747779e+00	-1.611860e+00	-3.962596e+00	-3.785982e+00	-2.929799e+00	-1.836580e+00	-1.240641e+00	-3.230876e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01	-5.586426e-01	-4.997191e-01	-7.762035e-01	-6.939170e-01	-4.810994e-01	-7.251203e-01	-6.532948e-01	-5.915023e-01	-6.546598e-01	-7.868237e-01	-5.574826e-01	-7.260193e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02	-4.111378e-01	-3.301442e-01	-1.120949e-01	-1.437297e-01	-2.992802e-01	-2.078552e-01	1.433889e-01	1.561419e-01	-1.237939e-01	-1.408197e-01	-2.458933e-01	-2.698956e-02
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01	6.294077e-02	7.341765e-02	6.179271e-01	6.318245e-01	3.067838e-01	5.310950e-01	7.562224e-01	8.358184e-01	6.093067e-01	6.666852e-01	2.241735e-01	7.302926e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00	5.764674e+00	1.378992e+01	2.800531e+00	3.436593e+00	1.085230e+01	8.068387e+00	1.859323e+00	1.379560e+00	6.499390e+00	3.331452e+00	8.924721e+00	3.060392e+00

Inference

The variables in given dataset has different scale of values and different standard deviation. So, if we don't scale the data, PCA will give high weight for the variable with Higher standard deviation. So, if we scale the data all data will have same standard deviation so will be given same weight. To scale the data, we used zscore. Now all variables follow same scale between – 2.5 to 12.5 and with same standard deviation of 1.0006.

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

Correlation of the dataset (before scaling):

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.090226	0.259592	0.146755
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.176229	-0.159990	0.124717	0.067313
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.237271	-0.180794	0.064169	-0.022341
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135	-0.384875	0.455485	0.660913	0.494989
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749	-0.294629	0.417864	0.527447	0.477281
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.300019	0.279703	-0.229462	0.018652	-0.078773
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.141904	0.232531	-0.280792	-0.083568	-0.257001
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.407983	-0.554821	0.566262	0.672779	0.571290
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.374540	-0.362628	0.272363	0.501739	0.424942
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.099955	-0.031929	-0.040208	0.112409	0.001061
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.030613	0.136345	-0.285968	-0.097892	-0.269344
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.849587	-0.130530	0.249009	0.432762	0.305038
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.000000	-0.160104	0.267130	0.438799	0.289527
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.160104	1.000000	-0.402929	-0.583832	-0.306710
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130	-0.402929	1.000000	0.417712	0.490898
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799	-0.583832	0.417712	1.000000	0.390343
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.289527	-0.306710	0.490898	0.390343	1.000000

Covariance of the dataset (before scaling):

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.R
Apps	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	4.683468e+05	24689.433666	21053.067602	1465.060576	-4327.122381	5.246171e+06	9756.42
Accept	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	3.335566e+05	14238.201489	12182.093828	1709.838189	-4859.487022	1.596272e+06	2834.16
Enroll	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	1.767380e+05	5028.961166	4217.086027	872.684773	-2081.693787	3.113454e+05	-356.58
Top10perc	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	-1.114551e+03	153.184870	127.551581	-26.874525	99.567208	6.087931e+04	149.99
Top25perc	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	-1.083605e+03	176.518449	153.002612	-23.097199	102.550946	5.454648e+04	162.37
F.Undergrad	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	1.041709e+06	25211.784197	21424.241746	5370.208581	-13791.929691	4.724040e+05	-6563.30
P.Undergrad	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	3.297324e+05	3706.756219	3180.596615	1401.302563	-5297.337090	-6.643512e+05	-6721.06
Outstate	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	-8.146737e+05	25157.515051	24164.147673	-8835.253539	28229.553066	1.413324e+07	39479.68
Room.Board	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	-1.480838e+05	5895.034749	6047.299735	-1574.205914	3701.431379	2.873308e+06	8005.36
Books	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	2.004303e+04	72.534242	242.963918	-20.867207	-82.263132	9.691258e+04	3.00
Personal	4.683468e+05	3.335566e+05	1.767380e+05	-1114.551186	-1083.605065	1.041709e+06	3.297324e+05	-8.146737e+05	-1.480838e+05	20043.025650	4.584258e+05	-120.898783	-305.154186	365.415770	-2399.310824	-3.460978e+05	-3132.61
PhD	2.468943e+04	1.423820e+04	5.028961e+03	153.184870	176.518449	2.521178e+04	3.706756e+03	2.515752e+04	5.895035e+03	72.534242	-1.208988e+02	266.608636	204.231332	-8.436492	50.383230	3.689806e+04	85.55
Terminal	2.105307e+04	1.218209e+04	4.217086e+03	127.551581	153.002612	2.142424e+04	3.180597e+03	2.416415e+04	6.047300e+03	242.963918	-3.051542e+02	204.231332	216.747841	-9.330256	48.734327	3.373346e+04	73.22
S.F.Ratio	1.465061e+03	1.709838e+03	8.726848e+02	-26.874525	-23.097199	5.370209e+03	1.401303e+03	-8.835254e+03	-1.574206e+03	-20.867207	3.654158e+02	-8.436492	-9.330256	15.668528	-19.764109	-1.206756e+04	-20.85
perc.alumni	-4.327122e+03	-4.859487e+03	-2.081694e+03	99.567208	102.550946	-1.379193e+04	-5.297337e+03	2.822955e+04	3.701431e+03	-82.263132	-2.399311e+03	50.383230	48.734327	-19.764109	153.556744	2.702892e+04	104.49
Expend	5.246171e+06	1.596272e+06	3.113454e+05	60879.310196	54546.483305	4.724040e+05	-6.643512e+05	1.413324e+07	2.873308e+06	96912.580326	-3.460978e+05	36898.058233	33733.456882	-12067.564601	27028.921473	2.726687e+07	35012.96
Grad.Rate	9.756422e+03	2.834163e+03	-3.565880e+02	149.992164	162.371398	-6.563308e+03	-6.721062e+03	3.947968e+04	8.005360e+03	3.008837	-3.132615e+03	85.557109	73.220396	-20.854888	104.493815	3.501297e+04	295.07

Comment

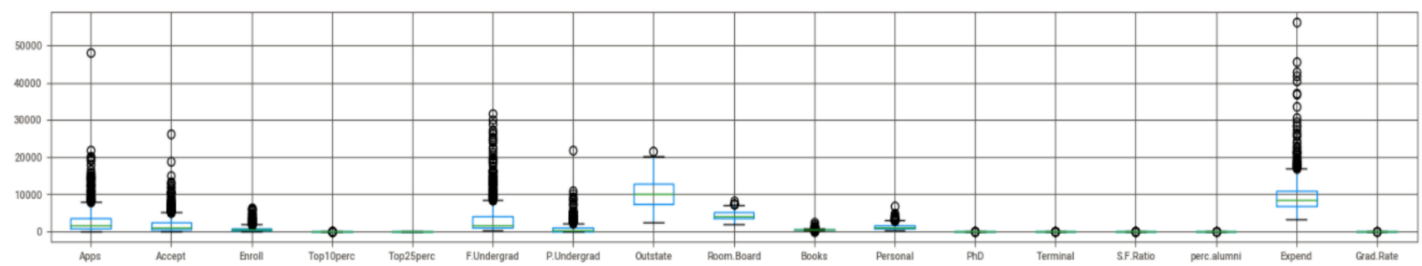
Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance. What sets them apart is the fact that correlation values are standardized whereas, covariance values are not.

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

a.) Comparing the data with outliers

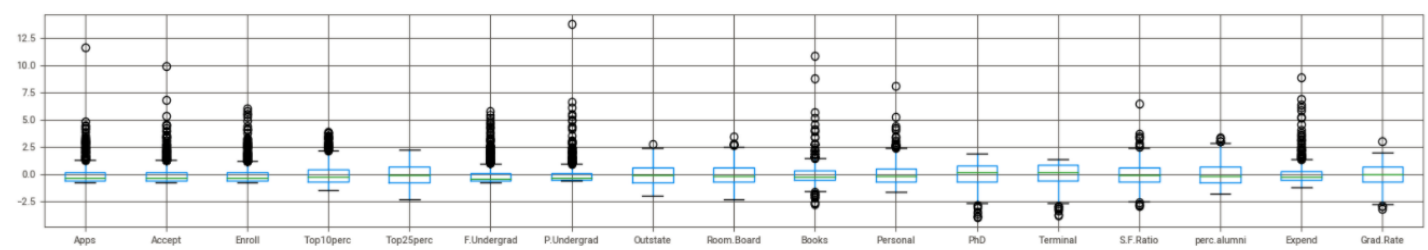
a.) Box plot of data before scaling with outliers

```
Box plot for data before scaling with outliers
<matplotlib.axes._subplots.AxesSubplot at 0x2269edb8a00>
```



b.) Box plot of data after scaling with outliers

```
Box plot for data after scaling with outliers
<matplotlib.axes._subplots.AxesSubplot at 0x2269ec0fd60>
```



Summary of data before scaling with outliers:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000
mean	3001.638353	2018.804376	779.972973	27.558559	55.796654	3699.907336	855.298584	10440.669241	4357.526384
std	3870.201484	2451.113971	929.176190	17.640364	19.804778	4850.420531	1522.431887	4023.016484	1096.696416
min	81.000000	72.000000	35.000000	1.000000	9.000000	139.000000	1.000000	2340.000000	1780.000000
25%	776.000000	604.000000	242.000000	15.000000	41.000000	992.000000	95.000000	7320.000000	3597.000000
50%	1558.000000	1110.000000	434.000000	23.000000	54.000000	1707.000000	353.000000	9990.000000	4200.000000
75%	3624.000000	2424.000000	902.000000	35.000000	69.000000	4005.000000	967.000000	12925.000000	5050.000000
max	48094.000000	26330.000000	6392.000000	96.000000	100.000000	31643.000000	21836.000000	21700.000000	8124.000000

	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
count	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.000000	777.00000
mean	549.380952	1340.642214	72.660232	79.702703	14.089704	22.743887	9660.171171	65.46332
std	165.105360	677.071454	16.328155	14.722359	3.958349	12.391801	5221.768440	17.17771
min	96.000000	250.000000	8.000000	24.000000	2.500000	0.000000	3186.000000	10.00000
25%	470.000000	850.000000	62.000000	71.000000	11.500000	13.000000	6751.000000	53.00000
50%	500.000000	1200.000000	75.000000	82.000000	13.600000	21.000000	8377.000000	65.00000
75%	600.000000	1700.000000	85.000000	92.000000	16.500000	31.000000	10830.000000	78.00000
max	2340.000000	6800.000000	103.000000	100.000000	39.800000	64.000000	56233.000000	118.00000

Summary of data after scaling with outliers

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16	-1.661405e-16	-3.029180e-17	6.515595e-17	3.570717e-16	-2.192583e-16	4.765243e-17	5.954768e-17	-4.481615e-16	-2.057556e-17	-6.022638e-17	1.213101e-16	3.886495e-16
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00	-7.346169e-01	-5.615022e-01	-2.014878e+00	-2.351778e+00	-2.747779e+00	-1.611860e+00	-3.962596e+00	-3.785982e+00	-2.929799e+00	-1.836580e+00	-1.240641e+00	-3.230876e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01	-5.586426e-01	-4.997191e-01	-7.762035e-01	-6.939170e-01	-4.810994e-01	-7.251203e-01	-6.532948e-01	-5.915023e-01	-6.546598e-01	-7.868237e-01	-5.574826e-01	-7.260193e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02	-4.111378e-01	-3.301442e-01	-1.120949e-01	-1.437297e-01	-2.992802e-01	-2.078552e-01	1.433889e-01	1.561419e-01	-1.237939e-01	-1.408197e-01	-2.458933e-01	-2.698956e-02
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01	6.294077e-02	7.341765e-02	6.179271e-01	6.318245e-01	3.067838e-01	5.310950e-01	7.562224e-01	8.358184e-01	6.093067e-01	6.666852e-01	2.241735e-01	7.302926e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00	5.764674e+00	1.378992e+01	2.800531e+00	3.436593e+00	1.085230e+01	8.068387e+00	1.859323e+00	1.379560e+00	6.499390e+00	3.331452e+00	8.924721e+00	3.060392e+00

Inference 1 : comparison between scaled and non-scaled Data which has outliers :

Without removing the outliers if we scale the data using z score it will affect the mean and the standard deviation of the data . From the above analysis we can see that the standard deviation for scaled data with outliers is 1.00644.

b.) Comparing the data without outliers:

Identified the outlier using interquartile range and replaced it with 95 percentile (upper range) and 5 percentile (lower range) values to remove the outliers.

Head of the data without outliers :

	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	1660.0	1232.0	721.0	23.0	52	2885.0	537.0	7440.0	3300.0	450.0	2200.0	70.0	78.0	18.1	12.0	7041.0	60.0
1	2186.0	1924.0	512.0	16.0	29	2683.0	1227.0	12280.0	6450.0	750.0	1500.0	29.0	52.8	12.2	16.0	10527.0	56.0
2	1428.0	1097.0	336.0	22.0	50	1036.0	99.0	11250.0	3750.0	400.0	1165.0	53.0	66.0	12.9	30.0	8735.0	54.0
3	417.0	349.0	137.0	60.0	89	510.0	63.0	12960.0	5450.0	450.0	875.0	92.0	97.0	7.7	37.0	17974.8	59.0
4	193.0	146.0	55.0	16.0	44	249.0	869.0	7560.0	4120.0	765.6	1500.0	76.0	72.0	11.9	2.0	10922.0	37.0

Scaling the data without outliers.

Scale the above data (without outliers) using zscore.

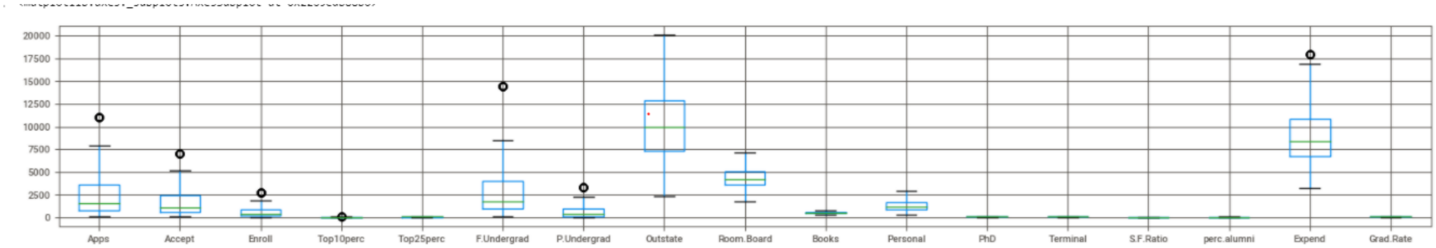
Head of the scaled data without outliers.

scaling data after removing outliers

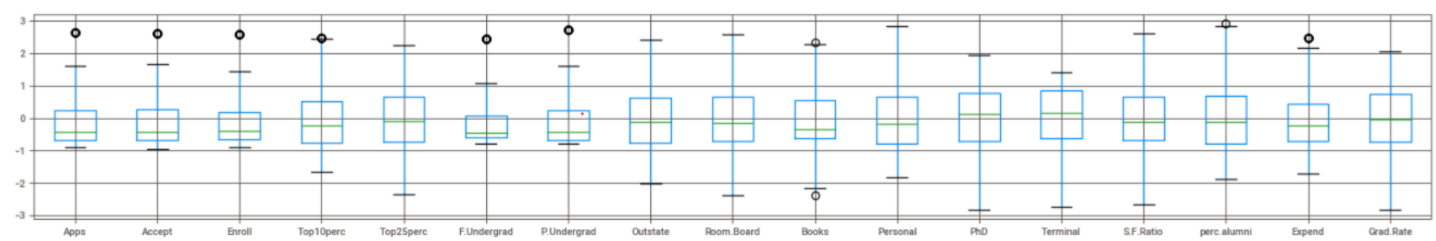
	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.383829	-0.353198	-0.035012	-0.247034	-0.191827	-0.179951	-0.220908	-0.747173	-0.976849	-0.807146	1.534067	-0.189281	-0.135735	1.107193	-0.877388	-0.621979	-0.327324
1	-0.215156	0.003214	-0.302696	-0.695834	-1.353911	-0.225740	0.513397	0.459655	1.959843	1.912681	0.325766	-2.826575	-1.917777	-0.502762	-0.547691	0.361374	-0.564340
2	-0.458225	-0.422730	-0.528115	-0.311148	-0.292878	-0.599082	-0.687032	0.202830	-0.557322	-1.260450	-0.252493	-1.282793	-0.984326	-0.311750	0.606247	-0.144125	-0.682848
3	-0.782423	-0.807984	-0.782992	2.125195	1.677612	-0.718316	-0.725344	0.629209	1.027560	-0.807146	-0.753075	1.225853	1.207869	-1.730694	1.183216	2.462296	-0.386578
4	-0.854253	-0.912539	-0.888017	-0.695834	-0.596031	-0.777479	0.132410	-0.717252	-0.212377	2.054112	0.325766	0.196665	-0.560031	-0.584624	-1.701629	0.472798	-1.690169

Before scaling and after scaling comparision of the data without outliers

a.) Box plot of data before scaling without outliers



b.)Box plot of data after scaling without outliers



Inference 2: comparison between scaled and non-scaled Data without outliers :

The outliers are reduced / replaced from the data set using the IQR and the percentiles .

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

a.) Covariance matrix:

Built the covariance matrix for scaled data without outliers.

Covariance matrix from the output.

```
Covariance Matrix
%s [[ 1.00128866e+00  9.34924522e-01  8.71047851e-01  3.24267498e-01
   3.62868715e-01  8.17850066e-01  5.01757457e-01  6.32865674e-02
   1.82959147e-01  2.33291015e-01  2.29533091e-01  4.46769909e-01
   4.17266656e-01  1.14213807e-01 -1.00610235e-01  2.54533823e-01
   1.47156965e-01]
 [ 9.34924522e-01  1.00128866e+00  9.22355696e-01  2.20622220e-01
   2.68343861e-01  8.71549611e-01  5.56937826e-01 -1.42554840e-02
   1.10325234e-01  2.13905650e-01  2.56349672e-01  4.07561902e-01
   3.84055661e-01  1.82533467e-01 -1.62609751e-01  1.66268254e-01
   6.95416281e-02]
 [ 8.71047851e-01  9.22355696e-01  1.00128866e+00  1.67564256e-01
   2.24859762e-01  9.48490114e-01  6.42498582e-01 -1.60758239e-01
  -3.72502918e-02  2.13398113e-01  3.47533881e-01  3.61278980e-01
   3.36082067e-01  2.67904002e-01 -2.13741732e-01  5.73414255e-02
  -4.05441513e-02]
 [ 3.24267498e-01  2.20622220e-01  1.67564256e-01  1.00128866e+00
   9.14894801e-01  1.03808804e-01 -1.46595772e-01  5.64307196e-01
   3.55518043e-01  1.54141858e-01 -1.20330447e-01  5.47194434e-01
   5.10761922e-01 -3.83391995e-01  4.50626637e-01  6.63757791e-01
   4.99596967e-01]
 [ 3.62868715e-01  2.68343861e-01  2.24859762e-01  9.14894801e-01
   1.00128866e+00  1.68331443e-01 -6.90652359e-02  4.91381245e-01
   3.30006448e-01  1.71971491e-01 -8.81163647e-02  5.55515750e-01
   5.28486082e-01 -2.91217587e-01  4.12790568e-01  5.76866972e-01
   4.87365165e-01]
 [ 8.17850066e-01  8.71549611e-01  9.48490114e-01  1.03808804e-01
   1.68331443e-01  1.00128866e+00  6.83706040e-01 -2.33426909e-01
  -7.97104222e-02  2.02639704e-01  3.64907776e-01  3.31586000e-01
   3.11509022e-01  3.08545857e-01 -2.70986793e-01  2.20793985e-04
  -1.03328032e-01]
 [ 5.01757457e-01  5.56937826e-01  6.42498582e-01 -1.46595772e-01
  -6.90652359e-02  6.83706040e-01  1.00128866e+00 -3.36167066e-01
  -7.49062347e-02  1.36212362e-01  3.31882174e-01  1.50869880e-01
   1.42681744e-01  3.50967077e-01 -3.94242323e-01 -1.67158835e-01
  -2.76670307e-01]
 [ 6.32865674e-02 -1.42554840e-02 -1.60758239e-01  5.64307196e-01
   4.91381245e-01 -2.33426909e-01 -3.36167066e-01  1.00128866e+00
   6.60163530e-01 -4.34567796e-03 -3.30937753e-01  4.02146639e-01
   4.19751257e-01 -5.79714939e-01  5.63482447e-01  7.74350317e-01
   5.81048783e-01]
 [ 1.82959147e-01  1.10325234e-01 -3.72502918e-02  3.55518043e-01
   3.30006448e-01 -7.97104222e-02 -7.49062347e-02  6.60163530e-01
   1.00128866e+00  1.07189891e-01 -2.26990151e-01  3.53391168e-01
   3.83940544e-01 -3.82746091e-01  2.72471331e-01  5.81799704e-01
   4.30742882e-01]
 [ 2.33291015e-01  2.13905650e-01  2.13398113e-01  1.54141858e-01
   1.71971491e-01  2.02639704e-01  1.36212362e-01 -4.34567796e-03
   1.07189891e-01  1.00128866e+00  2.37673348e-01  1.52781255e-01
   1.69157441e-01 -3.63949917e-03 -4.88870781e-02  1.46001071e-01
  -3.49849712e-03]
 [ 2.29533091e-01  2.56349672e-01  3.47533881e-01 -1.20330447e-01
  -8.81163647e-02  3.64907776e-01  3.31882174e-01 -3.30937753e-01
  -2.26990151e-01  2.37673348e-01  1.00128866e+00 -1.80160909e-02
  -3.45272440e-02  1.86768734e-01 -3.09448236e-01 -1.67436256e-01
  -2.89388165e-01]
 [ 4.46769909e-01  4.07561902e-01  3.61278980e-01  5.47194434e-01
   5.55515750e-01  3.31586000e-01  1.50869880e-01  4.02146639e-01
   3.53391168e-01  1.52781255e-01 -1.80160909e-02  1.00128866e+00
   8.67651517e-01 -1.32887210e-01  2.43585240e-01  5.24223727e-01
   3.19192258e-01]
[ 4.17266656e-01  3.84055661e-01  3.36082067e-01  5.10761922e-01
  5.28486082e-01  3.11509022e-01  1.42681744e-01  4.19751257e-01
  3.83940544e-01  1.69157441e-01 -3.45272440e-02  8.67651517e-01
  1.00128866e+00 -1.52328301e-01  2.63442196e-01  5.28106224e-01
  2.94336635e-01]
[ 1.14213807e-01  1.82533467e-01  2.67904002e-01 -3.83391995e-01
 -2.91217587e-01  3.08545857e-01  3.50967077e-01 -5.79714939e-01
 -3.82746091e-01 -3.63949917e-03  1.86768734e-01 -1.32887210e-01
 -1.52328301e-01  1.00128866e+00 -4.13884455e-01 -6.56336852e-01
 -3.17442424e-01]
[-1.00610235e-01 -1.62609751e-01 -2.13741732e-01  4.50626637e-01
  4.12790568e-01 -2.70986793e-01 -3.94242323e-01  5.63482447e-01
  2.72471331e-01 -4.88870781e-02 -3.09448236e-01  2.43585240e-01
  2.63442196e-01 -4.13884455e-01  1.00128866e+00  4.60176406e-01
  4.92274370e-01]
[ 2.54533823e-01  1.66268254e-01  5.73414255e-02  6.63757791e-01
  5.76866972e-01  2.20793985e-04 -1.67158835e-01  7.74350317e-01
  5.81799704e-01  1.46001071e-01 -1.67436256e-01  5.24223727e-01
  5.28106224e-01 -6.56336852e-01  4.60176406e-01  1.00128866e+00
  4.25256195e-01]
[ 1.47156965e-01  6.95416281e-02 -4.05441513e-02  4.99596967e-01
  4.87365165e-01 -1.03328032e-01 -2.76670307e-01  5.81048783e-01
  4.30742882e-01 -3.49849712e-03 -2.89388165e-01  3.19192258e-01
  2.94336635e-01 -3.17442424e-01  4.92274370e-01  4.25256195e-01
  1.00128866e+00]]
```

b.) Eigenvalues and Eigen vectors:

Eigen Vector from output :

```
Eigen Vectors
%s [[-1.51051724e-01  5.73869368e-01  2.54721171e-02  3.50002377e-01
   -4.76265776e-01 -2.73993248e-02 -6.79408155e-02 -1.34049566e-01
    1.84655032e-01  3.41030317e-02  1.23782113e-02  4.76414519e-02
   -2.28743180e-01 -1.02559773e-01  9.77100175e-02 -3.24930495e-01
   -2.42671239e-01]
  [ 4.52766958e-01 -6.43625404e-01 -4.08143058e-02  1.12837998e-01
   -2.08677137e-01 -1.27528369e-01 -2.86891699e-02 -1.23207526e-01
    1.89697047e-01  1.02521665e-01 -1.41529768e-03  3.31338141e-02
   -2.02792107e-01 -1.21914245e-01  1.25144023e-01 -3.57755851e-01
   -2.08095876e-01]
  [-7.50067816e-01 -2.58381892e-01  3.37484396e-02 -2.25003975e-01
    2.65981931e-01 -1.80558174e-02 -2.29745788e-02 -4.79563882e-02
    5.20184210e-02  1.34762063e-01  7.92830517e-03 -3.89761143e-02
   -1.72168365e-01 -1.42497171e-02  9.44419384e-02 -3.95824297e-01
   -1.64564266e-01]
  [ 5.89947774e-02 -5.31897461e-02  7.23553559e-01 -3.22924466e-02
   -1.62488072e-02  4.57358763e-02 -6.57319491e-03 -7.14429611e-02
   -1.10851590e-01 -2.89094711e-01  2.58267694e-01 -8.37673857e-02
   -1.45905144e-01  3.75563233e-01 -7.23866450e-02  7.53900839e-02
   -3.44633526e-01]
  [-1.47356588e-02 -3.70257583e-03 -6.58266244e-01  2.64582929e-02
    3.47432747e-02 -1.58456329e-01 -1.32078205e-01 -4.53255044e-02
   -1.89924670e-01 -3.36249057e-01  2.34717438e-01 -2.14918233e-02
   -1.20536687e-01  4.27876370e-01 -4.63368319e-02  3.67211412e-02
   -3.37858398e-01]
  [ 4.51829780e-01  4.13880625e-01 -1.05340454e-02 -3.50240396e-01
    5.20754661e-01  7.88826178e-02 -3.63762678e-02  1.12660606e-02
   -1.41252801e-04  1.22385171e-01  2.79162755e-02 -5.49956869e-02
   -1.15073146e-01 -1.46165800e-02  8.72397333e-02 -4.06243667e-01
   -1.34287678e-01]
  [ 4.97285352e-03 -3.24986907e-02  3.82640339e-02  1.01785907e-01
   -1.61437628e-01 -3.58599650e-02  1.89391557e-01  4.23776360e-01
   -7.36103386e-01 -5.41905661e-02  9.36586774e-02 -5.16448338e-02
    1.32038801e-01 -2.07265372e-01  3.86964803e-02 -3.54916637e-01
   -1.45128920e-02]
  [-4.74030188e-03  9.51907262e-02  2.55089170e-03 -2.23348292e-01
    7.53206365e-03 -5.57302446e-01  6.09931131e-01 -1.87206448e-01
   -1.46112057e-02 -2.38893511e-02 -1.04399025e-01 -1.39668813e-02
   -4.29684243e-02 -2.53851713e-01  2.05908405e-02  2.37362415e-01
   -2.97304568e-01]
  [-1.79855036e-02 -2.24494212e-02  3.34522167e-02 -9.10703410e-02
    8.88393882e-02  1.05909326e-01 -4.62002002e-01 -3.04566995e-01
   -2.17093330e-01 -3.55685905e-01 -1.25975104e-01  2.57756666e-01
    9.02072957e-02 -5.66793784e-01 -2.60693995e-02  1.23789047e-01
   -2.51192093e-01]
  [ 2.61824511e-03  2.76914586e-03  8.27989701e-03 -4.23639027e-02
   -9.11163317e-03 -4.85902891e-02  5.14384110e-02  7.45947123e-02
    1.62931047e-02  2.56097327e-01  1.39285992e-01  6.08723983e-01
    1.66299240e-02  4.72789590e-02 -7.13557985e-01 -1.06015391e-01
   -9.35681745e-02]
  [ 1.80670889e-02 -1.08413791e-02  1.34656529e-03  2.94677640e-02
   -2.34506560e-02 -9.33480418e-03  1.75508664e-02 -9.21496211e-02
   -3.28224085e-02 -2.51641980e-01 -6.56948779e-01 -3.84138124e-01
   -6.29349774e-02  1.07878431e-01 -5.21834336e-01 -2.35469217e-01
    4.84668755e-02]
  [ 3.31773567e-04  5.14472664e-03 -5.66570968e-02 -5.33328919e-01
   -4.34191485e-01  1.86806043e-01 -3.61407075e-02  1.25222037e-01
    1.67560289e-01  4.70574563e-02 -9.61136496e-02 -6.21741995e-02
    5.47356939e-01  1.23470983e-01  5.72580979e-02 -7.06517594e-02
   -3.24667558e-01]
 [-1.52860888e-02 -1.06241335e-03  8.90053732e-02  5.22450951e-01
   3.67220609e-01 -2.64231329e-01 -1.04786025e-01  7.52852907e-02
   1.29240181e-01  1.16057935e-01 -9.84467080e-02 -4.79218101e-02
   5.85124026e-01  7.31469402e-02  3.74577785e-02 -5.96664001e-02
  -3.20509921e-01]
 [ 1.26724133e-03 -1.49408711e-02 -8.57331081e-03  8.61575853e-02
   4.40006223e-02  2.33516509e-01  3.83414110e-01 -4.58497282e-01
  -1.21020302e-01 -2.15537935e-01 -1.74587187e-01  4.42093906e-01
   2.26758818e-01  2.83024041e-01  2.58375559e-01 -2.47834896e-01
   1.78476677e-01]
 [ 2.63795357e-02  3.68623907e-03  8.85282560e-03  1.17084626e-02
  -6.97525565e-02  5.84510444e-02 -1.69270316e-01 -2.50492792e-01
  -4.57694691e-01  6.35277046e-01 -3.21857779e-01 -5.59562277e-03
  -1.38310455e-01  2.29944433e-01  1.09906654e-01  2.43261851e-01
  -1.98617542e-01]
 [-9.78083104e-03 -6.46344811e-02 -1.59986586e-01  2.25352711e-01
   1.17933520e-01  6.64009736e-01  3.96898901e-01 -6.52884786e-02
   5.05354371e-02  8.64143063e-02  1.50523749e-01 -2.38206017e-01
  -2.83072831e-02 -2.20176259e-01 -1.72929690e-01  1.35747859e-01
  -3.40157000e-01]
 [-2.27570853e-03 -1.81770778e-02  7.18415214e-03  5.27965354e-02
   8.65713354e-02  1.41880695e-01  7.39141063e-02  5.80722250e-01
   1.29825028e-01 -1.62585522e-01 -4.63707778e-01  3.72585008e-01
  -2.87880353e-01  7.41465533e-02  2.31028150e-01  1.60607758e-01
  -2.48644778e-01]]
```

Eigen Value from the output:

```
 Eigen Values
%s [5.64307841 4.82973672 1.10030644 0.9966849  0.8977433  0.76549205
 0.58709565 0.55450358 0.44319291 0.38222641 0.24563729 0.03891348
 0.05597992 0.07466871 0.12376406 0.13603844 0.14684496]
```

Inference

Successfully built the Covariance matrix, Eigen value and Eigen vector for the given educational dataset.

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

```
he explicit form of the first PC :
([[ 1.73690056,  1.59813592,  1.54279982, -3.18198787,  1.78588136,
    0.54961821, -0.23204615, -1.90442505, -0.79778763,  2.83704769,
   -1.92917206, -2.19751868,  0.08621937, -0.88461566,  2.20172327,
    1.50532369, -5.22639492,  2.21614773,  2.02850424,  2.98218645,
   -0.16647671, -0.42486665,  1.79587893, -1.14116234, -0.69064956,
    3.45608305, -1.26658595, -1.35107274,  1.65014044, -1.02727695,
    0.94417962, -1.16286652,  2.65509454,  1.96644492,  0.03207601,
    1.15445763, -3.82111755, -3.73537582,  0.47199807, -1.35922637,
   -0.45281261,  0.05659088,  1.70842694,  1.34404146, -0.93545893,
    3.64699493,  2.20110218,  0.40224536, -1.26850553,  0.34056911,
    2.04117675,  1.46787903,  3.21901363,  3.46314232, -1.14836404,
    1.16531024,  1.04781343,  3.13600604,  1.27843908, -4.91626359,
   -5.31141129, -0.72430214,  0.56153994, -0.81583818, -4.20667789,
    1.60678943,  3.93267614,  2.29319743,  0.51838451, -1.32546715,
   -6.59793242, -4.93923926, -4.26581836,  0.57161781, -1.283442  ,
    1.48631412,  0.64358518,  0.4265354 , -1.1451771 ,  0.09866348,
   -0.08561853,  1.52918759,  2.6049984 , -0.05809533, -0.20807246,
    3.38106163, -4.18769947, -5.28607588,  0.62075597,  1.72104422,
    0.72967797, -4.16516146,  1.31424779,  1.30291269, -1.66424398,
    1.62913274, -0.42145742,  1.74463892,  1.72581055, -0.07005301,
    1.86031257, -0.28250191,  2.33118364,  2.26897182,  1.40699704,
    3.28925711, -2.69693384, -0.07606795, -1.54185602,  0.08081483,
   -0.19885417,  1.01382209,  2.32049143,  2.77772781, -4.7839659 ,
   -2.15479279,  1.41194604, -2.35309793, -2.17591392,  3.18717336,
   -0.28956555,  0.83495321, -3.92186372, -4.30582766,  0.21415541,
    1.04757479,  1.84894822, -0.28811643,  1.03895758,  0.12310671,
   -0.18191919,  0.37466172,  0.32386196,  0.65390341,  1.8569272 ,
    2.63498017,  0.60430396, -4.51416199, -3.44208885, -2.16700461,
   -3.46988023, -1.67700432,  2.58019343,  0.63734249, -5.70019554,
    2.27036761,  2.66326344,  1.26870982,  1.69484144, -3.96747211,
   -0.14467483, -0.59535398, -2.44370953,  2.38696944,  3.00668178,
    1.44918346,  3.02853597,  2.25849728, -6.53048149, -4.59684469,
    2.73892451,  3.42540434, -2.41609599, -2.36306709, -2.85470229,
    3.90636028,  1.39094928,  1.60518149,  3.81350428,  1.87077896,
    1.93433894, -1.47345134, -3.66810547, -0.25206698, -6.64135243,
   -2.2066802 , -0.71046121,  1.53200204,  3.78558501,  1.93091022,
    1.69064966,  1.52426423,  1.32334612,  2.64279982, -1.279083  ,
   -0.51041586,  0.0767806 ,  0.64233941,  1.34205088,  2.76880035,
   -0.53228367, -6.31513317,  1.47153133, -0.39317967,  0.58670513,
    1.22839897, -2.64230727,  3.50939423,  3.04251373,  3.20406126,
   -1.07238558, -1.02610389,  1.73657237, -2.28301602,  3.32667337,
   -2.13016508,  1.61403804,  2.80610189,  1.60146762,  0.22913626,
    2.39809165,  2.0405452 ,  0.71167465, -2.63915891,  1.5271081 ,
    2.85790204,  1.71064658,  1.93513291, -0.31802421, -4.11238204,
    0.90790878, -6.06208616, -3.80617907,  1.72399962,  1.84623546,
   -3.27830834,  4.66919982, -0.78664816,  0.41663239,  1.04041768,
   -2.39604654,  2.23913772,  3.32389793,  1.36770308,  1.87242734,
    2.13662206,  1.98598829, -3.80424423,  0.30990106,  0.15918291,
   -1.54935099,  1.24251809, -3.85468208, -1.12739263, -1.28541486,
    0.67691215,  0.01785441,  2.36006251,  0.27844912, -0.97215011,
   -6.45122619, -5.37442309,  2.21116955, -0.81447022, -0.87435449,
   -1.36918636, -2.7611951 , -2.0723166 , -1.44686632, -1.38892581,
   -0.79635468,  0.03096386,  1.96725289,  1.37185965,  4.84645824,
    4.15377537, -0.01913202,  0.94460077, -1.50119179,  0.09707688,
   -2.0244582 ,  1.38855617,  0.85365374,  1.02774593, -1.98854423,
    3.13077595,  0.09545367, -1.53320528, -1.77309303, -2.32071851,
    3.2129836 ,  2.78070363,  1.58305584, -1.33383394, -6.11333944,
    2.26673242,  2.88469844, -1.77230155, -0.14312723,  2.567682  ,
    2.05156724,  1.72425841, -3.33349817,  1.11205935,  0.65002374,
```

```
 0.97045388, -2.1047155 ,  2.19886724, -0.65600216, -3.50575677,
 1.27303784, -2.02361073,  2.47059268,  3.30227085,  3.12608606,
 2.92524166, -2.58822788, -1.02373212, -0.60670518, -4.69793359,
 0.28279537,  1.39702386,  1.84812107, -2.04456229,  2.98999645,
 1.94127339,  2.21015667,  1.85084855, -1.39696214,  0.88298054,
 2.5713972 ,  1.46204944,  1.50148796,  1.49917705, -0.18137951,
 2.35861226, -1.58310036, -1.92762182, -0.79730189, -2.0051273 ,
-0.9490778 ,  0.49512404,  1.46340032,  3.07567792, -3.00007994,
 2.18045408,  3.27606403,  1.03714415, -1.49054947, -1.49839459,
 2.52911556,  2.87261975, -1.07761892,  0.02164143, -2.30595358,
 0.32207042, -1.09878367, -0.53789135, -0.11957169,  0.67105815,
 0.98026727,  0.0132106 ,  2.19304619,  0.65552666, -5.82708863,
 3.78793178,  2.15461899,  2.79677098,  2.38247633, -1.48338622,
 2.04036737,  0.7254916 ,  0.35954834,  4.20634029, -0.20437783,
-1.89078517, -2.55309641, -0.72526303,  3.64298877,  0.33552459,
 1.60440702,  0.22132666, -0.63029784,  1.81162293,  1.37773269,
 0.38907535,  3.52386607,  2.96595648,  3.6936194 , -0.70313631,
 0.67814363,  0.75021703,  1.38487562,  0.68061644,  3.73909212,
 1.92636339, -0.7256306 ,  0.01208241,  2.09537304,  3.67962898,
-3.97897212,  3.19711926,  2.63981286,  2.43883811,  3.50882584,
-0.2388621 ,  1.23585411, -0.46543762, -0.07628002,  3.12154995,
-1.56615286,  1.24534686,  0.10200342,  1.76940474, -0.54849479,
-1.30292282,  0.12367594, -6.12157318,  2.11955304, -1.62004837,
 2.80973328,  2.14204182, -3.63665797,  1.85357566, -0.58927983,
 1.38363199,  0.56569748,  0.08461451, -1.96154826,  0.41515347,
 0.06880922,  3.13241131,  2.18678501,  1.60085837, -6.71256985,
 0.75915552,  3.23989383,  0.49824259, -4.33928372, -3.22968929,
-1.59404003, -1.11588938, -1.19667536, -2.36689458,  1.9775436 ,
 2.99279362, -0.43197194,  0.26272963,  1.68627186,  2.15310943,
-2.44923898, -0.82615818,  1.118989  , -0.21933072,  2.56606863,
-3.02043142, -4.2340279 ,  3.1704897 ,  2.96319226,  1.08469245,
 1.86552065,  1.60314325,  2.86207426, -2.71988612,  1.31027787,
 1.98332798, -2.09709989,  2.9469816 , -0.59720461, -6.34326418,
-0.94664993, -2.03336878,  0.1410854 ,  0.77513321, -0.6367393 ,
 0.96435284,  1.98871559, -0.19855341, -2.06414288, -3.17746669,
-0.6194946 , -4.02487431, -3.2149828 , -0.78045575, -1.07487425,
 2.09913114, -0.61628558,  0.93140836,  1.99060892,  2.07272134,
-1.57362333, -0.1056797 ,  0.87111529, -3.16966393, -0.48819806,
-0.87915935,  0.37372154,  2.15763886,  0.08241169,  1.97437423,
 2.91584624,  1.5759083 ,  0.03230475,  1.75339995,  1.99075863,
-1.23625547, -0.39882772, -2.8674525 , -1.14144808,  2.04593782,
 1.45156099, -1.03541094, -1.93067397,  0.58604876,  1.1958066 ,
-0.08891941,  1.77529858, -0.11475757,  1.23310808,  0.2548118 ,
-1.64165919, -2.67610943, -3.72188761,  0.86353768,  1.77586486,
-4.19228994, -1.03794776, -1.04055563, -0.73132522,  0.45077953,
 0.65348423,  1.79053232, -1.07610296,  3.22672988, -1.55877038,
 0.47493077,  2.53726202, -2.6941781 , -4.4406798 ,  2.22305299,
 2.13258709,  3.56726517,  2.54993974,  1.95715039, -1.83756084,
 3.43832836,  1.0979323 ,  2.73709154,  3.51847869,  2.177698  ,
-1.72513354,  2.05963418, -0.88441556,  3.23467745, -0.15833685,
-0.4841842 ,  0.12202023, -2.91756179,  3.0516152 , -1.75204999,
-0.7726868 ,  0.44021456, -0.26693552,  4.05342754,  1.97018594,
 1.75400468, -1.1841815 , -2.62462076,  0.76176321, -1.32572879,
-2.04612023, -3.48540179, -3.58893334, -2.72092591,  1.10850954,
-0.17029583,  1.65435743,  0.7113155 ,  0.54526005, -1.92148215,
-0.11030747,  0.72357719,  1.15246721,  1.17434865, -1.32612427,
-2.90528195, -3.76059845,  2.51052704,  2.23611059, -0.13905325,
 3.36929211, -2.55658317,  1.44209222, -0.6524969 ,  2.31000768,
 0.85587976,  1.84301035,  1.03388032,  1.06127056,  3.01158075,
-1.62812704, -1.80855431,  2.19965353, -4.28523067, -1.48358266,
 0.70776836, -3.54141075, -4.49587568,  2.91161114,  0.88561431,
 2.21195798, -4.16903425,  1.5476178 , -0.75962467,  0.36012379,
-5.26580998, -5.25268182, -0.46661987,  2.48633709, -5.79885336,
-2.40287957, -2.94233654, -1.17059355, -1.63781801, -2.12734837,
-1.00847019,  0.79598641,  1.37752308, -0.2469104 , -4.06624932,
-2.59582747, -0.01261397, -0.91234316, -3.15458577, -1.86755044,
 1.39624264, -1.82061742,  1.12008256,  0.07496428,  1.8251573 ,
 3.31799029,  3.17429498,  0.32552589, -2.04568808, -1.52118317,
 0.88830922, -4.28355457, -5.59434025,  0.71013645,  0.05547099,
-2.92855077,  0.58591586, -1.62626613, -0.92048156,  1.66665929,
 2.90944782,  2.23565566, -1.6139704 ,  0.46192509, -2.15043663,
 0.95094954, -4.99971911, -0.1288452 ,  0.19020783,  0.3208212 ,
 0.15345819,  0.20035019,  0.81515632,  1.52200415,  0.13214566,
-5.26884862, -1.11857138, -1.41987572, -7.13975733, -3.23402806,
-0.69976082, -1.91553636, -1.63303678, -2.48695727, -5.674443  ,
-2.07476957, -0.69273266,  3.80249379, -1.94681834,  3.60030886,
-1.47764749, -0.86184235, -5.0554999 ,  2.79777338,  3.5100708 ,
 0.97454105, -0.20602572,  0.39227363, -1.29952046,  2.52378432,
-2.49816651,  0.34575892,  2.12075869, -2.07854364, -3.14069714,
-1.20110592, -0.98959544, -2.52090952, -5.12706154, -4.09887053,
 1.23990814,  2.21426918,  2.22829043,  1.09682608,  1.98279776,
-3.56343407,  0.50724973,  0.53681439,  2.42216411, -2.05899455,
 2.08192528,  4.80398277, -1.57542076, -5.89778682, -4.62546932,
-3.50729574, -0.58909167,  2.94008499, -2.24685904,  2.60908294,
 0.78434249,  1.88434331,  4.78455853, -1.91851702,  0.36438053,
-4.6529743 ,  1.12847435,  1.40815424, -0.07971701, -1.70432339,
-3.83015724, -1.13967762, -2.19308103, -5.97858449,  3.65421649,
 2.65912633,  3.49484069,  2.27674014, -5.05929865, -2.8291165 ,
 1.04086912,  2.45299464, -4.57831608,  0.69702515,  4.23938746,
 0.36370993,  2.35453756, -0.88773896, -0.72152928,  1.19080985,
 2.8432703 , -0.50895166,  2.20644326,  1.45056972,  0.16666725,
 0.83241466, -0.78903614, -1.96816746, -2.73144128,  1.04588602,
-2.56085188, -1.28062868, -0.51339419, -0.7049714 ,  0.23122233,
-2.06283521,  0.802929  ,  2.58288919, -5.28184487,  0.23238862,
 1.66030907,  2.87573106,  0.8529374 ,  2.59607281, -1.6959675 ,
-1.67594748, -2.90984372,  3.58454345, -0.26911621,  0.67443642,
-6.6604879 ,  0.62211979]])
```

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.

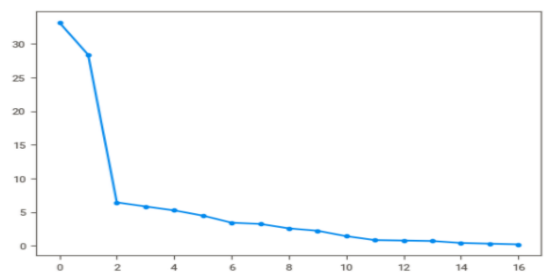
a.) Cumulative Value, optimum number of Principle components and Eigen vectors.

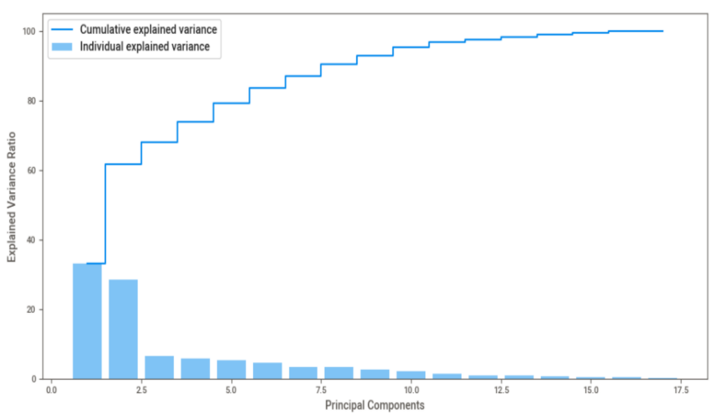
Cumulative values is the sum of all eigen values. To find the variance explained by each component you should divide each component's eigenvalue by the cumulative values.
By knowing the cumulative values, we can know how many percent of information is captured in each principle components and decide the number of optimum numbers.
Eigen vector indicate the coefficient of the features or numerical columns

Cumulative Value:

```
Cumulative Variance  [ 33.15185743  61.52550945  67.98957029  73.84487717  79.11892366
  83.61602283  87.06508197  90.32266981  92.9263317   95.17182864
  96.61489423  97.47757639  98.27677262  99.00385952  99.44252192
  99.77139178 100.        ]
```

Plotting the Value for Scree Test





Insights

From the above cumulative graph, we can find that first two PCA components is picking up around 60 percent. For 3 components its around 68 percent.
As per the Kaiser's Stopping Rule the optimum number of PCA is three. (choose all components whose eigenvalues are greater than 1)

b.) Perform PCA

PCA using sklearn PCA Function.

```
array([[-1.73690056, -1.59813582, -1.54279984, ..., -0.67443637,
         6.66048788, -0.62211975],
       [ 0.78652264, -0.33203968, -1.37926808, ..., -0.14322337,
        -1.08947694,  0.63056959],
       [ 0.09135343,  2.12904388, -0.60247563, ...,  0.37330976,
         1.41458627, -1.315047  ]])
```

PCA components:

```
array([[ 0.24267124,  0.20809588,  0.16456427,  0.34463353,  0.3378584 ,
         0.13428768,  0.01451289,  0.29730457,  0.25119209,  0.09356817,
        -0.04846688,  0.32466756,  0.32050992, -0.17847668,  0.19861754,
         0.340157  ,  0.24864478],
       [ 0.3249305 ,  0.35775585,  0.3958243 , -0.07539009, -0.03672114,
         0.40624367,  0.35491664, -0.23736241, -0.12378905,  0.10601539,
         0.23546922,  0.07065176,  0.0596664 ,  0.2478349 , -0.24326185,
        -0.13574786, -0.16060776],
       [-0.09770807, -0.12514405, -0.09444334,  0.07238641,  0.04633681,
        -0.08724118, -0.03869596, -0.02059314,  0.0260686 ,  0.71355752,
         0.52183455, -0.05726439, -0.03745204, -0.25837397, -0.10990679,
         0.17293335, -0.23102718]])
```

c.)Exporting the data of the Principal Component scores into a data frame.

PCA SCORE IN DATAFRAME

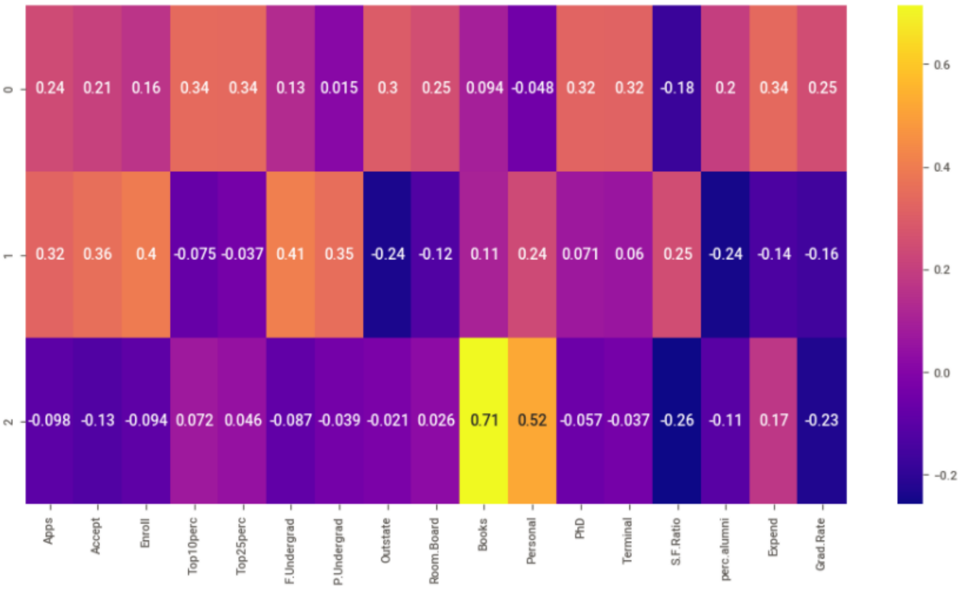
PCA SCORE DATAFRAME HEAD

	principal component 1	principal component 2	principal component 3
0	-1.736901	0.786523	0.091336
1	-1.598136	-0.332040	2.129007
2	-1.542800	-1.379268	-0.602485
3	3.181988	-2.993983	0.335527
4	-1.785882	-0.202226	2.731232

Principle components data frame

	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	0.242671	0.208096	0.164564	0.344634	0.337858	0.134288	0.014513	0.297305	0.251192	0.093568	-0.048467	0.324668	0.320510	-0.178477	0.198618	0.340157	0.248645
1	0.324930	0.357756	0.395824	-0.075390	-0.036721	0.406244	0.354917	-0.237362	-0.123789	0.106015	0.235469	0.070652	0.059666	0.247835	-0.243262	-0.135748	-0.160608
2	-0.097708	-0.125144	-0.094443	0.072386	0.046337	-0.087241	-0.038696	-0.020593	0.026069	0.713558	0.521835	-0.057264	-0.037452	-0.258374	-0.109907	0.172933	-0.231027

Heatmap



Communality of the data. (It is the extent to which an item correlates with all other items.)

```
     The communality is
5]:  Apps           0.665310
     Accept         0.690996
     Enroll         0.654832
     Top10perc      0.492410
     Top25perc      0.420916
     F.Undergrad    0.627773
     P.Undergrad    0.408125
     Outstate       0.555260
     Room.Board     0.401050
     Books          0.913141
     Personal       0.805771
     PhD            0.452584
     Terminal       0.417628
     S.F.Ratio      0.684686
     perc.alumni    0.551786
     Expend         0.648838
     Grad.Rate      0.640280
     dtype: float64
```

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

The three Principle components (PC0, PC1 and PC2) created are free from multicollinearity
Just three PCA (out of 17) components is picking up around 68 % of variability.
PC0 explains most of variables at average level of .22 with good explanatory for Top10perc, top20perct, Expend, PhD, terminal, outstate variables.
pc1 has good explanatory for F. undergrad, Enroll, Accept, P. undergrad, accept and apps
PC2 has highest explanatory for Books and personal.
The highest communality variable is Personal with 81% communality
The lowest communality variable is with 40 % communality.