Advanced Statistics

13th June 2021

OVERVIEW

The Advanced Statistics module covers important concepts like ANOVA, EDA and PCA which are used to solve the given project problem statements. This report provides valuable outputs, tables, graphs and proposed business solutions derived from the given data sets and dictionaries. Python and it's variarious libraries are used to help do calculations and visualize data.

GOALS

- To strengthen the concepts of Advanced Statistics by implementing the methods to solve the given problems.
- To take a structured approach to treat the data with appropriate code and arithmetics.
- To find patterns in the data through descriptive and exploratory analysis.
- To visualize the findings and interpret the results.
- To state the hypothesis of the problem statement and apply statistical methods to draw conclusions.
- To list the insights from the treatments and propose appropriate business solutions.

Vinay Santosh
PGP Data Science and Business Analytics
March_B 2021 (Online)

CONTENTS:

TABLES:

FIGURES:

Problem 1 A

1.1 Hypothesis Statements for One-Way ANOVA:

For Education:

Null Hypothesis: H0 : The mean Salary is equal for all levels of Education.

Alternate Hypothesis: H1 : At least one level of Education has different mean Salary.

For Occupation:

Null Hypothesis: H0 : The mean Salary is equal for all levels of Occupation.

Alternate Hypothesis: H1 : At least one level of Occupation has a different mean Salary.

1.2 One-way ANOVA on Salary with respect to Education:

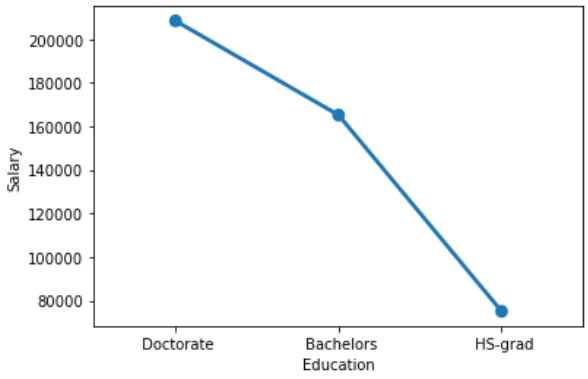


Figure 1.2a

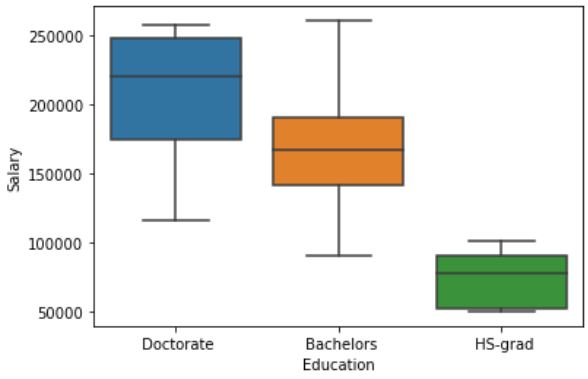


Figure 1.2b

From the above plots, we can observe that employees with a Doctorate level of Education have a higher Salary as compared to High School Graduates.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Table 1.2

P-value = **1.257709e-08** is smaller than the level of significance **α, 0.05**

Thus, the null hypothesis is rejected based on the above treatment and it is concluded that at least one level of Education has different mean Salary.

1.3 One-way ANOVA on Salary with respect to Occupation:

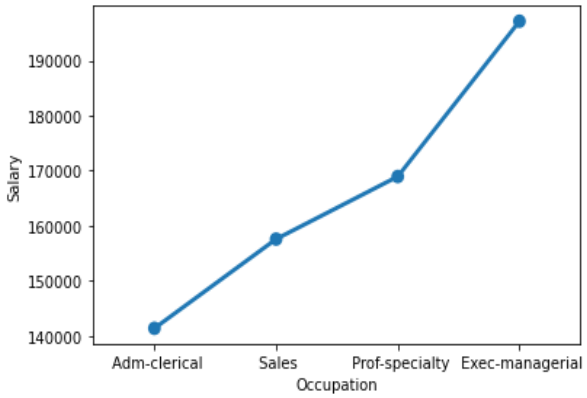


Figure 1.3a

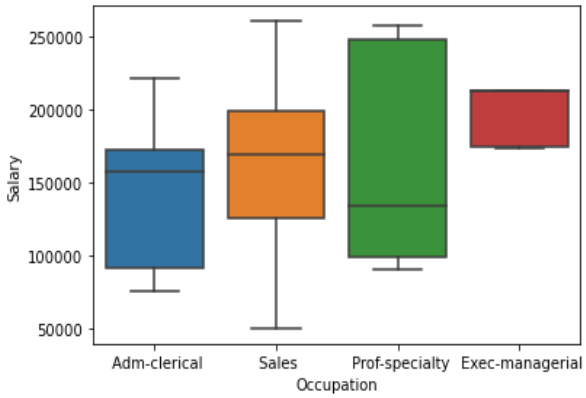


Figure 1.3b

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Table 1.3

p- value = **0.458508** is greater than the level of significance **α, 0.05**

Hence, we fail to reject the null hypothesis based on the above treatment and it is concluded that all levels of Occupations have an identical mean Salary as displayed in Figure 1.3b.

1.4 Interpreting the results of One-way ANOVA:

The null hypothesis is rejected in (2) and we can see, from **figure 1.2a** and **figure 1.2b,** that there is a significant difference in the class means of HS-Grads from Doctate. This shows that the level of Education is a factor that has a significant level of impact on Salary.

Problem 1B :

1.5 Interaction between two treatments:

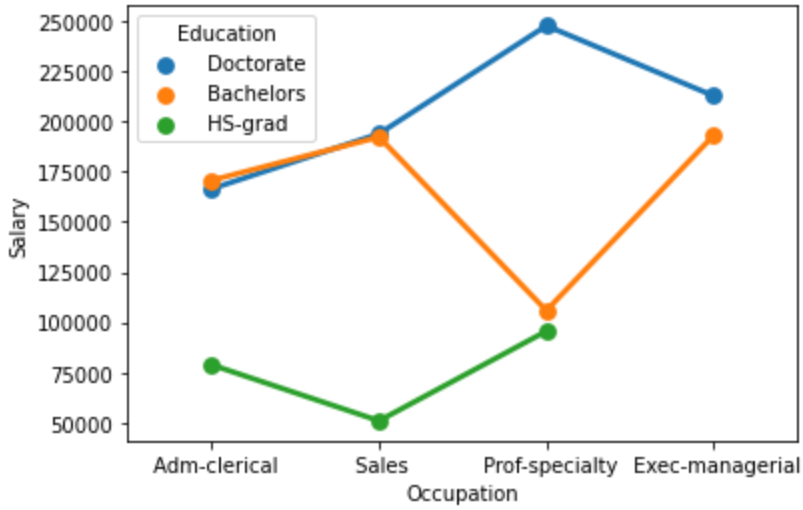


Figure 1.2.1

- We can clearly see, from the figure (figure 1.2.1) above, that there is an interaction between the treatments.
- The positions Adm-clerical and Sales have a similar Salary range for both Doctorates and Bachelors.
- There is a drop in Salary for employees with a Bachelors who take up Prof-speciality roles whereas there is an increase in pay for Doctorates in this field, which is significantly different.
- The Exec- managerial positions also pay a higher mean Salary to employees with Doctorates.
- The HS-Grad does not show any interaction with the other variables in this dataset and we can see that the mean Salary does not show a significant variance.

1.6 Two-way ANOVA based on Salary with respect to both Education and Occupation:

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
C(Occupation)	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
C(Education):C(Occupation)	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

Table 1.6

From the table (Table 2b) we can observe that the F value of the first treatment (Education) is significantly different and the p-values are used to reject or fail to reject the null hypotheses.

Treatment 1 Hypothesis:

Null hypothesis: All levels of Education have equal mean Salary

Alternate hypothesis: At least one level of Education has a different mean Salary

Result:

The corresponding p-value for this treatment (**from Table 2b**) is **lower than** the level of significance **0.05** which proves that the level of Education has an impact on Salary.

Treatment 2 Hypothesis:

Null hypothesis: All levels of Occupation have equal mean Salary

Alternate hypothesis: At least one level of Occupation has a different mean Salary

Result:

The corresponding p-value for this treatment (**from Table 2b**) is **greater than** the level of significance **0.05** which proves that the Occupation level of the employee does not have a significant impact on Salary.

Interaction Hypothesis:

Null hypothesis: There is no significant interaction between Education and Occupation

Alternate hypothesis: There is a significant level of interaction between both factors

Result:

The corresponding p-value for the interaction (**from Table 2b**) is **lower than** the level of significance **0.05** which proves that this interaction does have a significant impact on Salary which is to be considered.

1.7 Business implications of ANOVA:

In this problem we check to see if the dependent variable Salary is affected by the Education level or Occupation (role) of a person. By performing ANOVA techniques, we are able

to distinguish the effects of the independent variables, individually and as an interaction, on the dependent.

In this case, we observe that a person with an educational level higher than HS-Grad is more likely to have a higher Salary. Also the current occupation of a person, does not prove to be a determining factor in influencing Salary. The mean Salary for Exec-managerial positions is higher than all other positions. This statistical analysis can be used by hiring managers to recruit the ideal candidates for different roles in their company.

Problem 2:

2.1. Exploratory Data Analysis and Insights:

Table 2.1

```
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)
```

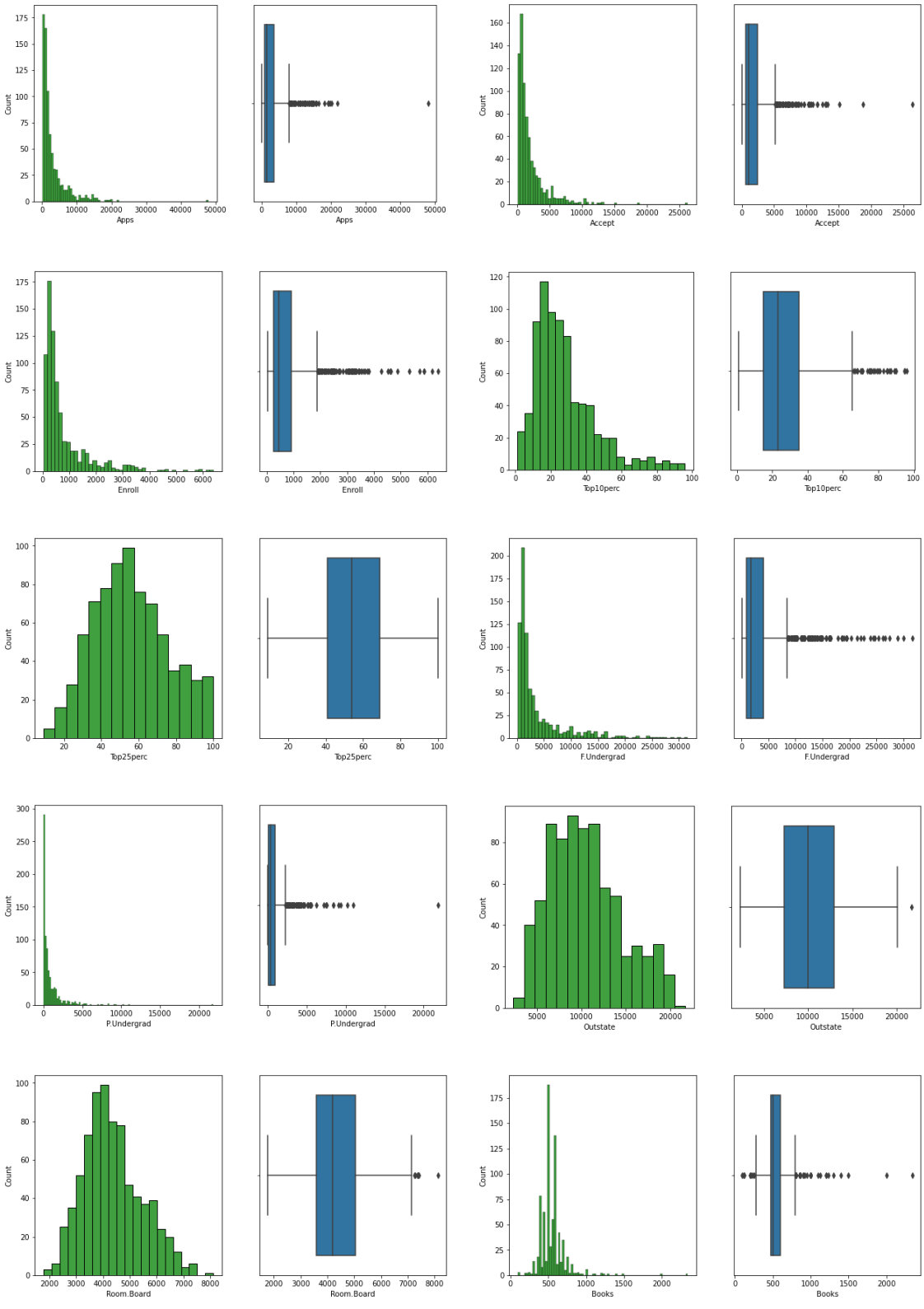
- The provided dataset consists of information about colleges, past and present students, faculty, expenditures, graduation rate etc.
- The dataset has information ranging from the colleges accepting students to the estimated expenses each student will have to spend in college.
- By performing EDA, we can take a closer look at the given dataset.
- From the information given in table 2.1, we observe that there are 18 factors that are taken into consideration with 777 non- null entries.
- All features except the college names are numeric in nature (either int or float)

- The description of the given dataset, provides us the values of central tendencies across all the different features.
- From the above table, Table 2.1.1, we observe the summary of the data which gives us the range, interquartile values of each feature individually.
- On average, there are more full-time students as compared to part-time students.

	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.796654	19.804778	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.907336	4850.420531	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.298584	1522.431887	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.669241	4023.016484	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.526384	1096.696416	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.380952	165.105360	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.642214	677.071454	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.660232	16.328155	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.702703	14.722359	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.089704	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.743887	12.391801	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.171171	5221.768440	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.463320	17.177710	10.0	53.0	65.0	78.0	118.0

Table 2.1.1

- We observe that the acceptance rate decreases with the increase in applications.
- The alumni contribution percentage is higher for colleges with a good graduation rate .
- From the table of figures Figure 2.1.1 given below, we observe the distribution of each feature individually.
- We observe that features such as “Top25perc” “Grad Rate” “perc.alumni”, “outstate” has less skewness in them.
- The distribution of top students from various high schools and outstate students, seem to be spread across most of the colleges. Which means, students who perform well in high school are more likely to get accepted into colleges.
- Colleges that spend more on instructional expenses, have a lower student/faculty ratio and a higher graduation rate.
- Colleges with more students from other states, have a higher percentage of top performing students in high school.


Count
Apps
Apps
Count
Accept
Accept
Count
Enroll
Enroll
Count
Top10perc
Top10perc
Count
Top25perc
Top25perc
Count
F.Undergrad
F.Undergrad
Count
P.Undergrad
P.Undergrad
Count
Outstate
Outstate
Count
Room.Board
Room.Board
Count
Books
Books

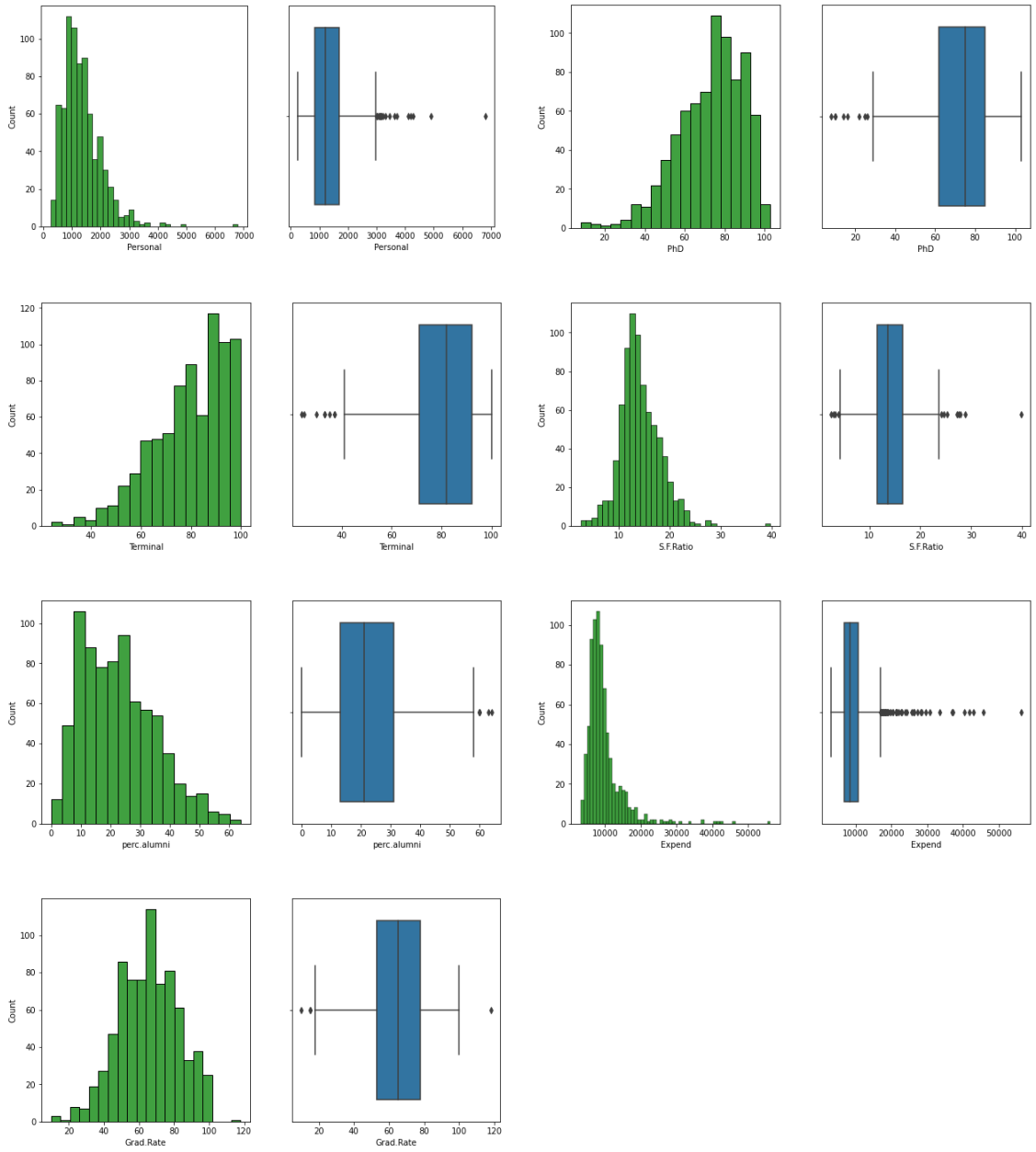


Figure 2.1.1

- Figure 2.1.0 displays a heatmap from which we can observe the correlation between the features. By performing multivariate analysis, we are able to visualise the correlation between the features.
- The features with darker colors have a lower correlation and lighter colors have higher correlation.
- The number of full-time undergraduates (“F.Undergrad”) are heavily correlated with the number of students enrolled in college (“Enroll”).

- We observe a negative correlation between student/faculty ratio (“S/F Ratio”) and the instructional expenses per student (“Expend”)

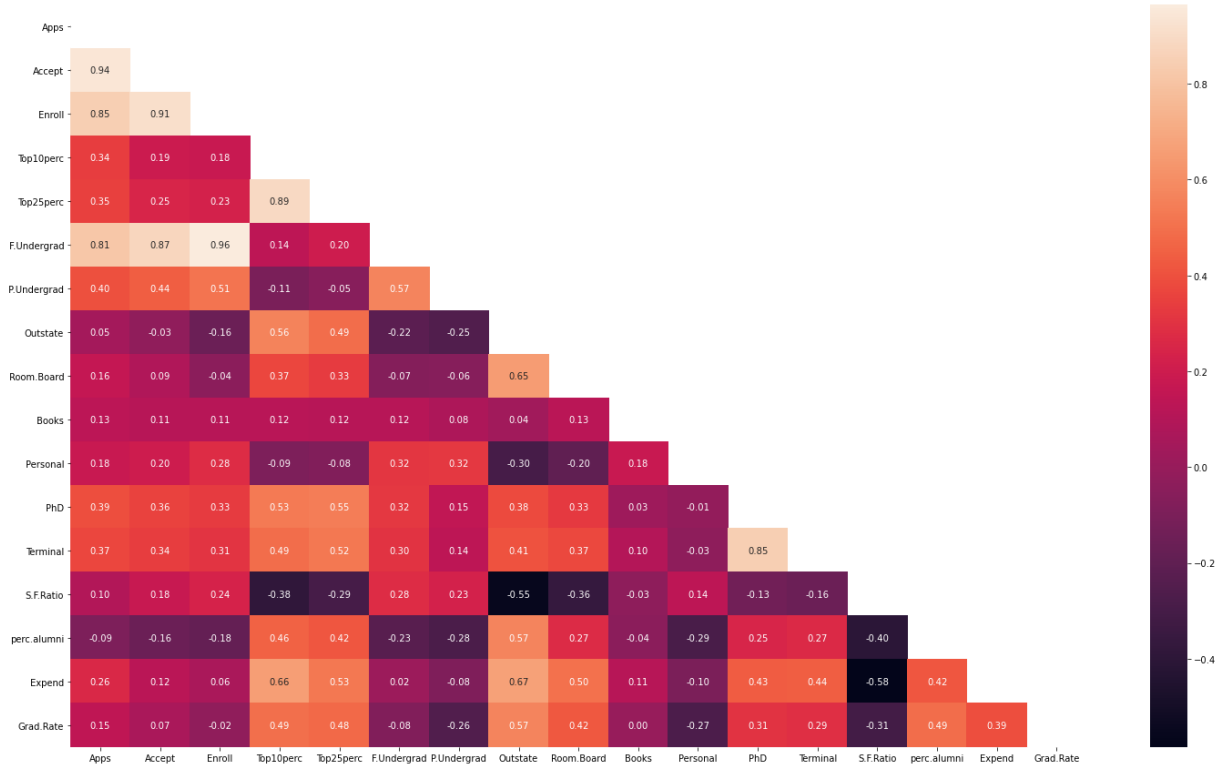


Figure 2.1.2

2.2 Data Scaling:

- In order to perform PCA on any data, we require a standardized set of data. This refers to centralizing the data to its origin.
- In the given data, from the EDA, we can see that there is a significant difference in the mean from its standard deviation in all of the features.
- By scaling data, we centralize it to zero and reduce the deviation from the mean.
- Upon performing the Z-score technique of scaling we derive the following table (Table 2.2.1).
- The scaled data now has an even standard deviation which is close to the means of all the features. We can now proceed with the remaining steps of PCA.

	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

Table 2.2.1

2.3. Covariance and Correlation of Scaled Data:

- In a covariance matrix the diagonals correspond to the variances and the other values correspond to the covariance between the features.
- In a correlation matrix scaled form of covariance and the diagonals are 1 because the features are 100% correlated with each other. Thus the covariances (which are off diagonal values) become correlations and we get similar matrices.
- The scaling of data is used to center the data in order for us to perform PCA and proceed with decomposition.
- The tables below (Table 2.3.1a, 2.3.1b) is the covariance/correlation matrix for the given dataset.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026

Table 2.3.1a

Covariance Matrix

%s [[1.00128866 0.94466636 0.84791332 0.33927032 0.35209304 0.81554018 0.3987775 0.05022367 0.16515151 0.13272942 0.17896117 0.39120081 0.36996762 0.09575627 -0.09034216 0.2599265 0.14694372]
[0.94466636 1.00128866 0.91281145 0.19269493 0.24779465 0.87534985 0.44183938 -0.02578774 0.09101577 0.11367165 0.20124767 0.35621633 0.3380184 0.17645611 -0.16019604 0.12487773 0.06739929]
[0.84791332 0.91281145 1.00128866 0.18152715 0.2270373 0.96588274 0.51372977 -0.1556777 -0.04028353 0.11285614 0.28129148 0.33189629 0.30867133 0.23757707 -0.18102711 0.06425192 -0.02236983]
[0.33927032 0.19269493 0.18152715 1.00128866 0.89314445 0.1414708 -0.10549205 0.5630552 0.37195909 0.1190116 -0.09343665 0.53251337 0.49176793 -0.38537048 0.45607223 0.6617651 0.49562711]
[0.35209304 0.24779465 0.2270373 0.89314445 1.00128866 0.19970167 -0.05364569 0.49002449 0.33191707 0.115676 -0.08091441 0.54656564 0.52542506 -0.29500852 0.41840277 0.52812713 0.47789622]
[0.81554018 0.87534985 0.96588274 0.1414708 0.19970167 1.00128866 0.57124738 -0.21602002 -0.06897917 0.11569867 0.31760831 0.3187472 0.30040557 0.28006379 -0.22975792 0.01867565 -0.07887464]
[0.3987775 0.44183938 0.51372977 -0.10549205 -0.05364569 0.57124738 1.00128866 -0.25383901 -0.06140453 0.08130416 0.32029384 0.14930637 0.14208644 0.23283016 -0.28115421 -0.08367612 -0.25733218]

Table 2.3.1b

2.4. Outliers, pre and post scaling:

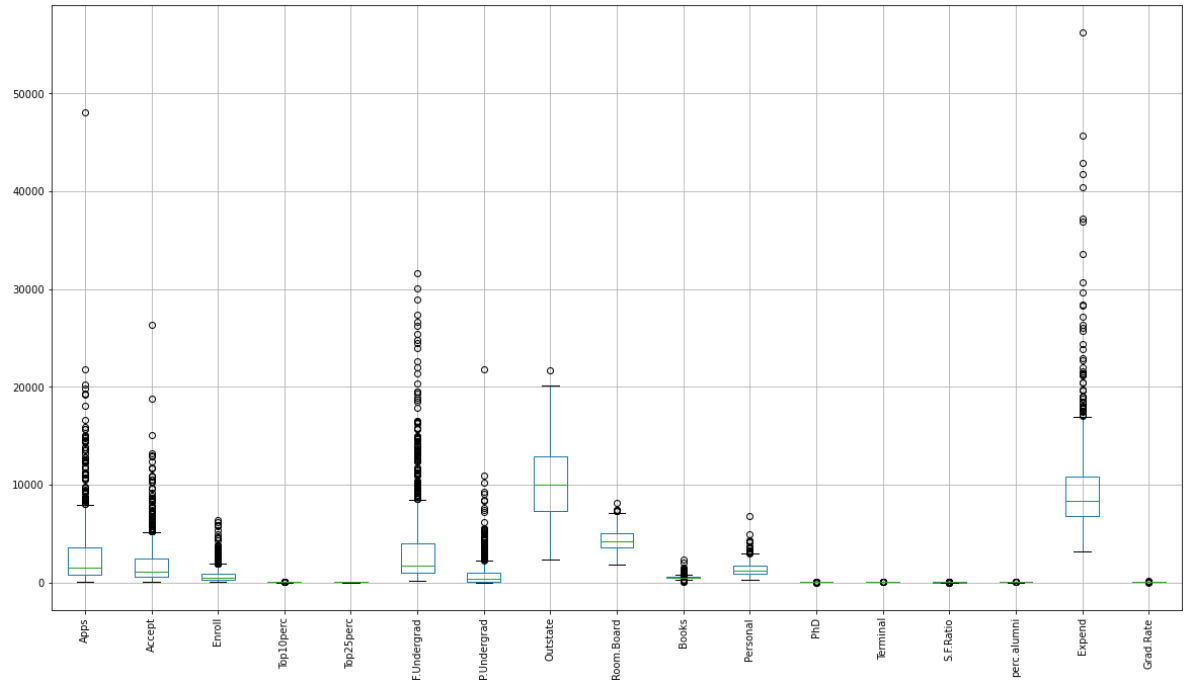


Figure 2.4.1 a (before)

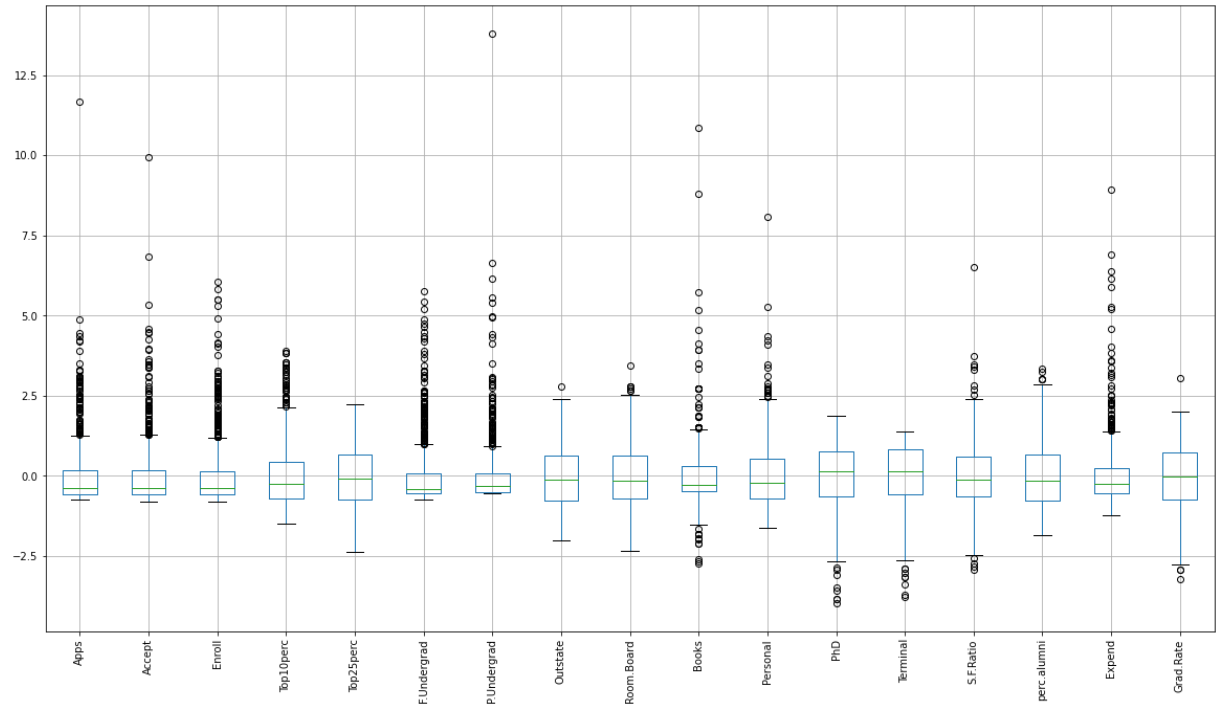


Figure 2.4.1 b (after)

- The above two figures show the presence of outliers before and after scaling.
- However, the scaled dataset shows us where the outliers lie in the data.
- Features like “PhD” and “Terminal” have outliers below the lower quartile range whereas “Outstate” and “Room.Board” have outliers above the higher quartile range.
- We can observe that “Top25perc” has no outliers present.
- The scaled boxplot has a much smaller range from -5 to 15.

2.5 EigenValues and EigenVectors:

Eigen Vectors:

```
%s [[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
```

```
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
  2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
  4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
  1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
 -8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
  5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
 -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
 -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

EigenValues:
%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]

2.6 Principal Component Analysis:

With the scaled data, we perform PCA in order to reduce the dimensions of the data set while retaining the original features. The PCA components, which are a linear combination of the eigenvectors and the original value of the data, can be fitted to the original dataset. The table below is the head of the dataset with the extracted components fitted to it's feature.

	0	1	2	3	4
Apps	0.248766	0.331598	-0.063092	0.281311	0.005741
Accept	0.207602	0.372117	-0.101249	0.267817	0.055786
Enroll	0.176304	0.403724	-0.082986	0.161827	-0.055694
Top10perc	0.354274	-0.082412	0.035056	-0.051547	-0.395434
Top25perc	0.344001	-0.044779	-0.024148	-0.109767	-0.426534
F.Undergrad	0.154641	0.417674	-0.061393	0.100412	-0.043454
P.Undergrad	0.026443	0.315088	0.139682	-0.158558	0.302385
Outstate	0.294736	-0.249644	0.046599	0.131291	0.222532
Room.Board	0.249030	-0.137809	0.148967	0.184996	0.560919
Books	0.064758	0.056342	0.677412	0.087089	-0.127289
Personal	-0.042529	0.219929	0.499721	-0.230711	-0.222311
PhD	0.318313	0.058311	-0.127028	-0.534725	0.140166
Terminal	0.317056	0.046429	-0.066038	-0.519443	0.204720
S.F.Ratio	-0.176958	0.246665	-0.289848	-0.161189	-0.079388
perc.alumni	0.205082	-0.246595	-0.146989	0.017314	-0.216297
Expend	0.318909	-0.131690	0.226744	0.079273	0.075958
Grad.Rate	0.252316	-0.169241	-0.208065	0.269129	-0.109268

Table 2.6

2.7 Linear equation of first PC in terms of eigenvectors and corresponding features:

0.25 ***Apps*** + 0.21 ***Accept*** + 0.18 ***Enroll*** + 0.35 ***Top10perc*** + 0.34 ***Top25perc*** + 0.15 ***F.Undergrad*** + 0.03 ***P.Undergrad*** + 0.29 ***Outstate*** + 0.25 ***Room.Board*** + 0.06 ***Books*** + -0.04 ***Personal*** + 0.32 ***PhD*** + 0.32 ***Terminal*** + -0.18 ***S.F.Ratio*** + 0.21 ***perc.alumni*** + 0.32 ***Expend*** + 0.25 ***Grad.Rate***

2.8 Understanding Cumulative Variances:

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.]

The cumulative eigenvalues for this case are represented above. This shows the proportion of variance captured in each of the features. Here, the first component captures 32% of the variance and the first two components capture 58% and so on. All of the components together capture 100% of the variance in the dataset. To capture approximately 80% of the variance, we can use 8 of the 17 components.

The components indicate the linear combination of the features with its respective weights or loadings. These loadings are the eigenvectors themselves. The Scree plot below provides us a visual representation of the number of components required to capture the percentage of variance required for our study.

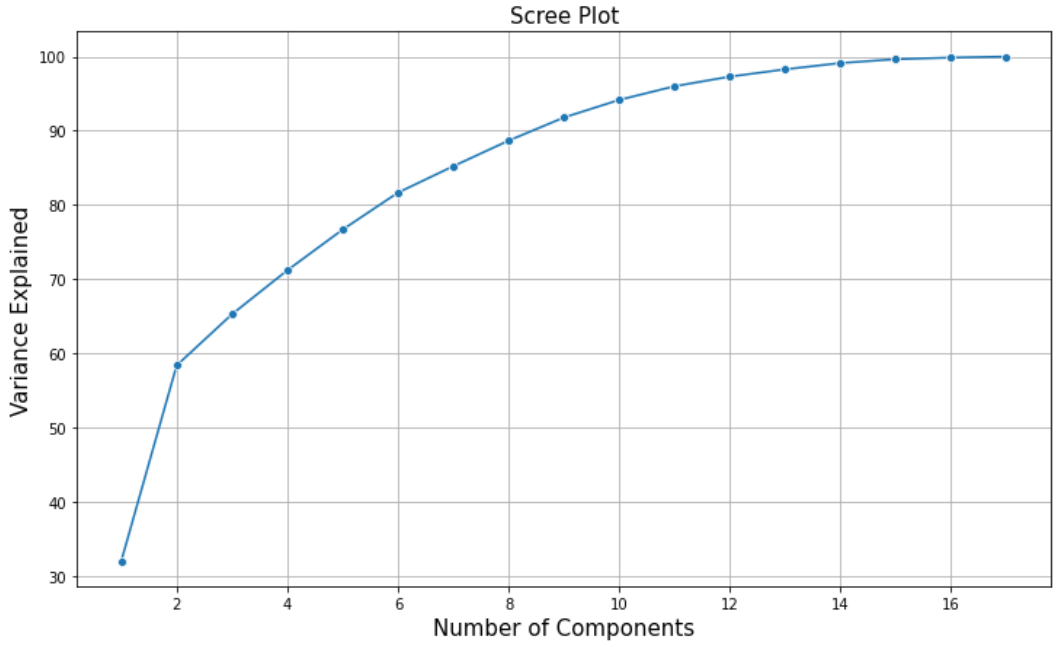


Figure 2.8

2.9 Business Implications of PCA and Insights:

- Principal component analysis is used for the compression and reduction of dimensions in a given dataset.
- It helps us identify patterns and deduce insights. By implementing PCA, we are able to compress large amounts of data into values which correspond to its weight.

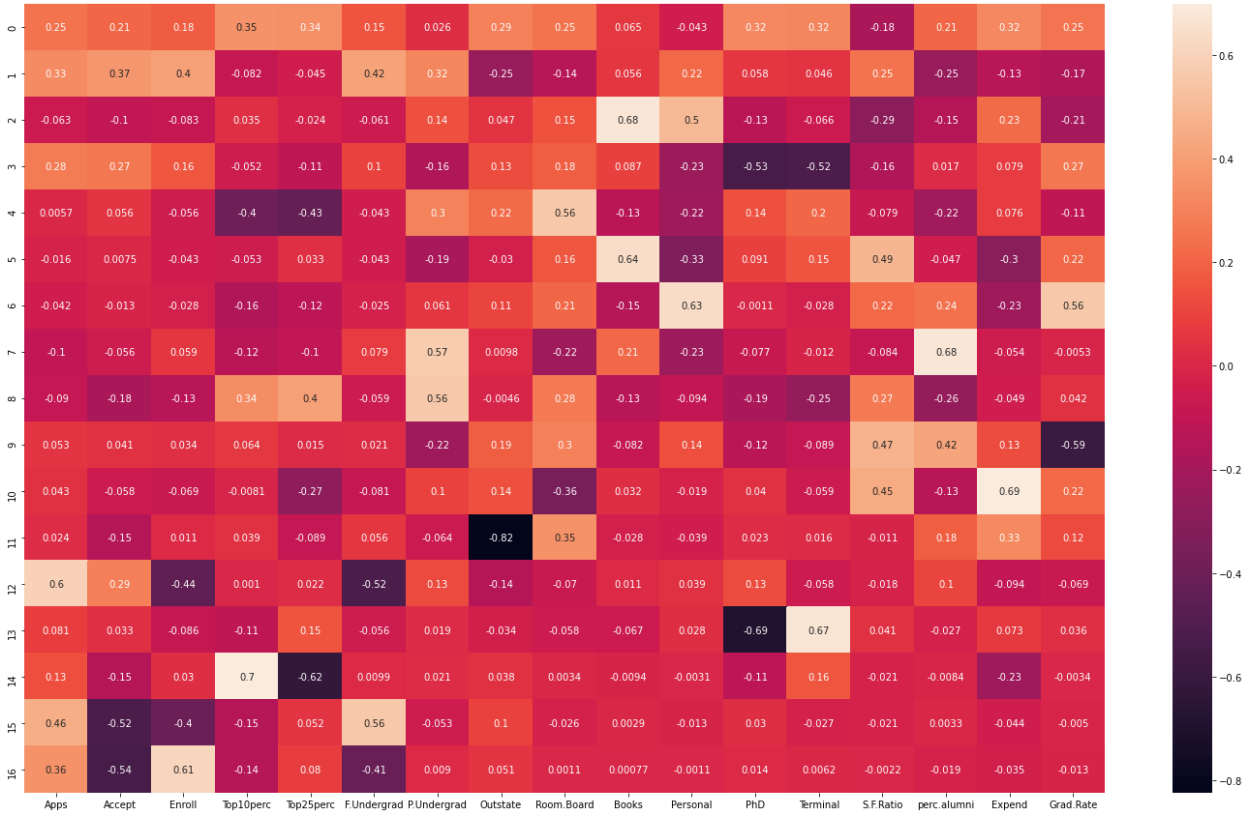


Figure 2.9.1

- On plotting the compressed values, we can visually notice the variance captured by each principal component.
- This displays the correlation of the original variable with the compressed data and the combination of these components can be used to derive further information.
- Based on this we can notice that the first two components have a high percentage of data on variables such as applications received, number of students accepted and enrolled (application related information).
- Principal component 1 captures the majority of the variations in the data,
- Components 3, 5, 6 and 7 collect most of the information on student expenses and can be used to predict spending patterns in students.

- The graduation rate of the colleges are captured most by the 7th principal component which also has a positive correlation with student/faculty ratio.
- This indicates that colleges with lower ratios of students to faculty, tend to have a higher graduation rate.
- Grad rate also displays a negative correlation with student expenses.
- The mean of all the variables is centralized to 0 due to scaling. From the boxplot given below, we can visibly notice the reduction of outliers and the advantages of implementing compressed data for further analysis and model building.
- The accommodation expenses for all colleges have a higher median value.
- PCA is one of the techniques used to reduce data dimensions without dropping original features.

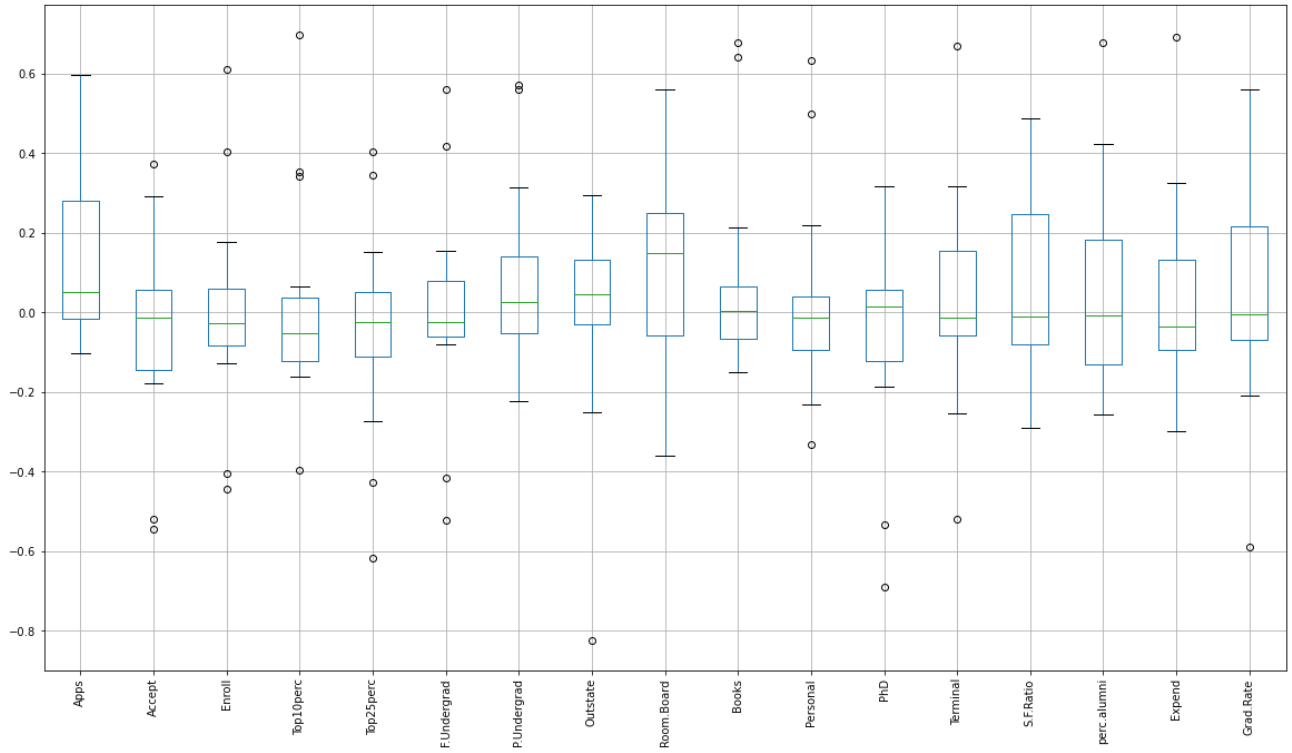


Figure 2.9.2