Advanced Statistics

Prepared by : CIJITH JOSE

CONTENTS

List of Figures

List of Tables

Problem 1

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High school graduate, Bachelor, and Doctorate. Occupation is at four levels, Administrative and clerical, Sales, Professional or specialty, and Executive or managerial. A different number of observations are in each level of education – occupation combination.

Problem 1A:

1.***State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually***

The Hypothesis for the One Way ANOVA are:

$H0$: The mean salary is the same at 3 levels of Education
Ha: For at least one level of Education , the mean salary is different

H0:The mean salary is the same at 4 levels of occupation.
Ha:For at least one level of occupation, the mean salary is different .

2.***Perform a one-way ANOVA on Salary with respect to Education. State whether the null hypothesis is accepted or rejected based on the ANOVA results.***

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2	1.027E+11	5.13E+10	30.95628009	1.26E-08
Residual	37	6.1373E+10	1.66E+09		

Table1-one way Anova on Education

Since the the P value is very small and lesser than 0.05 , we reject the null hypothesis and state that there is for at least one level of Education, the mean salary is different.

3.***Perform a one-way ANOVA on Salary with respect to Occupation. State whether the null hypothesis is accepted or rejected based on the ANOVA results.***

	df	sum_sq	mean_sq	F	PR(>F)
C(Occupation)	3	1.1259E+10	3.75E+09	0.884144129	0.4585078
Residual	36	1.5281E+11	4.24E+09		

Table 2-one way Anova on Occupation

Since the P value is greater than 0.05 , we fail to reject the null hypothesis and state that the mean salary is same at all levels of occupation.

4. *If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result*

```
\n=========================================================================\n
group1     group2     meandiff      p-adj     lower        upper    reject\n-
 -----------------------------------------------------------------------\n
  Bachelors  Doctorate   43274.0667 0.0146    7541.1439  79006.9894   True\n
   Bachelors   HS-grad  -90114.1556  0.001 -132035.1958 -48193.1153   True\n
Doctorate    HS-grad -133388.2222  0.001 -174815.0876 -91961.3569   True\n---
---------------------------------------------------------------------------
```

Figure 1. Multiple Comparison of Means - Tukey HSD

From the above figure after conducting the Tukey HSD test we can see that the means for all the three combinations are significantly different because the p value is less than 0.05.

Problem 1B:

1.*What is the interaction between two treatments? Analyze the effects of one variable on the other (Education and Occupation) with the help of an interaction plot.*

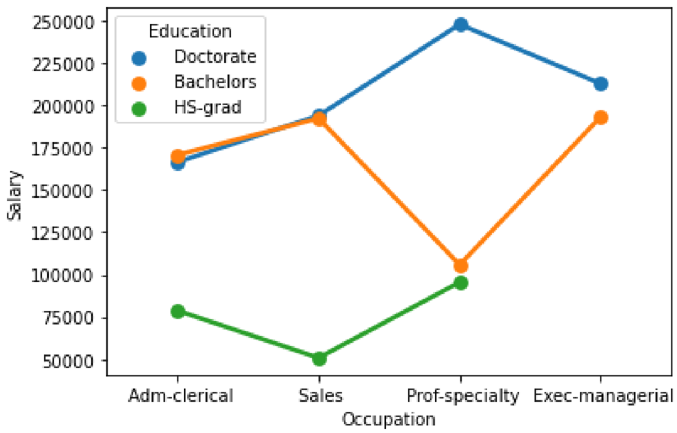


Fig 2: Interaction Plot

From the above plot we can clearly infer that there is significant interaction between Education and Occupation. We can clearly infer that population with doctorate level are on the higher level of occupation like prof-specialty/Managerial with higher income , while the low level Education population like HS-Grad in the above graph are on the low level of occupation such as Adm/Clerical also earning lesser than others .

2.*Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?*

H0: Interaction affect between Education and Occupation does not exist.

Ha: Interaction affect between Education and Occupation exists.

	df	sum_sq	mean_sq	F	PR(>F)
C(Education)	2	1.027E+11	5.13E+10	72.21195806	5.47E-12
C(Occupation)	3	5519946053	1.84E+09	2.587625501	0.0721158
C(Education):C(Occupation)	6	3.6349E+10	6.06E+09	8.51981467	2.23E-05
Residual	29	2.0621E+10	7.11E+08		

Table 3-Two way Anova

After performing the Two way Anova with respect to Education and Occupation along with their Interaction, we can state that there is significant interaction between Education and Occupation by rejecting the null hypothesis since the P-value is much lower than 0.05 .

3.*Explain the business implications of performing ANOVA for this particular case study.*

A two-way Anova was run on a sample of 40 people to examine the effect of Education and Occupation on Salary. There was a significant interaction between the effects of Education and Occupation on salary. Simple main affect analysis showed that Occupation level has no significant effect on the salary, but the Education level has significant impact on the salary.

Problem 2

The dataset Education - Post 12th Standard.csv contains information on various colleges. You are expected to do a Principal Component Analysis for this case study according to the instructions given.

1.*Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?*

In the given data there are total of 777 entries with 18 columns, out of which 16 variables are int64 type, one variable float64 type and one variable of Object type. There are no null values in the dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Names	777	777	Colby College	1							
Apps	777				3001.638	3870.201	81	776	1558	3624	48094
Accept	777				2018.804	2451.114	72	604	1110	2424	26330
Enroll	777				779.973	929.1762	35	242	434	902	6392
Top10perc	777				27.55856	17.64036	1	15	23	35	96
Top25perc	777				55.79665	19.80478	9	41	54	69	100
F.Undergrad	777				3699.907	4850.421	139	992	1707	4005	31643
P.Undergrad	777				855.2986	1522.432	1	95	353	967	21836
Outstate	777				10440.67	4023.016	2340	7320	9990	12925	21700
Room.Board	777				4357.526	1096.696	1780	3597	4200	5050	8124
Books	777				549.381	165.1054	96	470	500	600	2340
Personal	777				1340.642	677.0715	250	850	1200	1700	6800
PhD	777				72.66023	16.32815	8	62	75	85	103
Terminal	777				79.7027	14.72236	24	71	82	92	100
S.F.Ratio	777				14.0897	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777				22.74389	12.3918	0	13	21	31	64

Expend	777				9660.171	5221.768	3186	6751	8377	10830	56233
Grad.Rate	777				65.46332	17.17771	10	53	65	78	118

Table 4 -Description of Dataset

Let’s now perform Univariate analysis on all the variables :

```
Description of Apps -  Number of applications received
---------------------------------------------------------------------------
--
count      777.000000
mean      3001.638353
std       3870.201484
min         81.000000
25%        776.000000
50%       1558.000000
75%       3624.000000
max      48094.000000
Name: Apps, dtype: float64
Distribution of Apps                    Box Plot of Apps
---------------------------------------------------------------------------
--
```

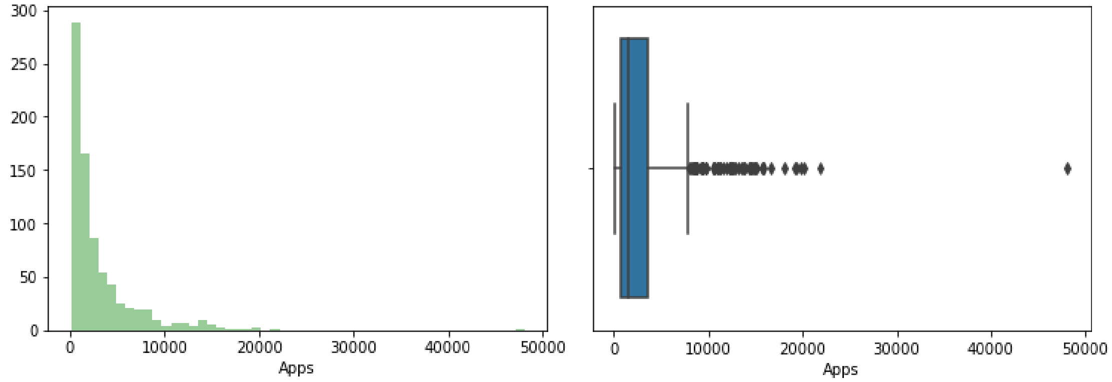


Figure3

```
Description of Accept- Number of applications accepted
---------------------------------------------------------------------------
--
count      777.000000
mean      2018.804376
std       2451.113971
min         72.000000
25%        604.000000
50%       1110.000000
75%       2424.000000
max      26330.000000
```

```
Distribution of Accept                Box Plot of Accept
---------------------------------------------------------------------------
```

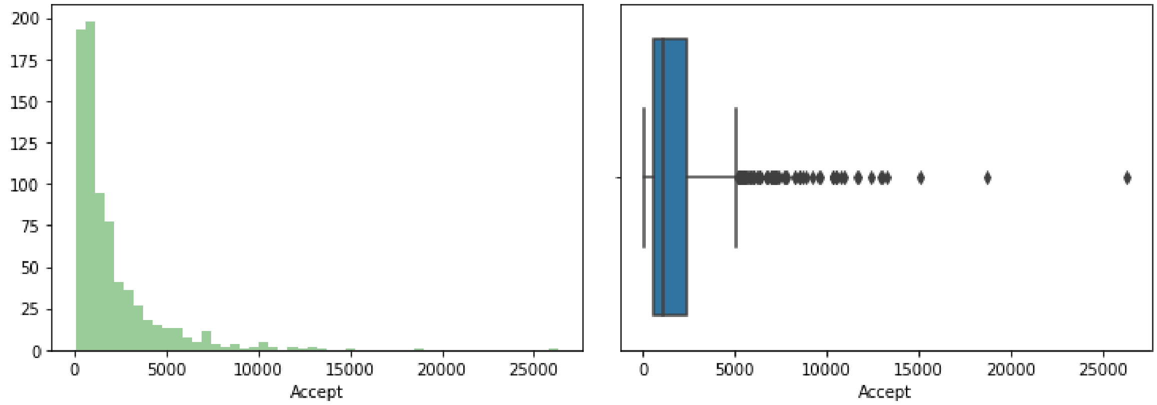


Figure 4

```
Description of Enroll - Number of new students enrolled
---------------------------------------------------------------------------
--
count     777.000000
mean      779.972973
std       929.176190
min        35.000000
25%       242.000000
50%       434.000000
75%       902.000000
max      6392.000000
```

```
      Distribution of Enroll              Box Plot of Enroll
--------------------------------------------------------------------------
```

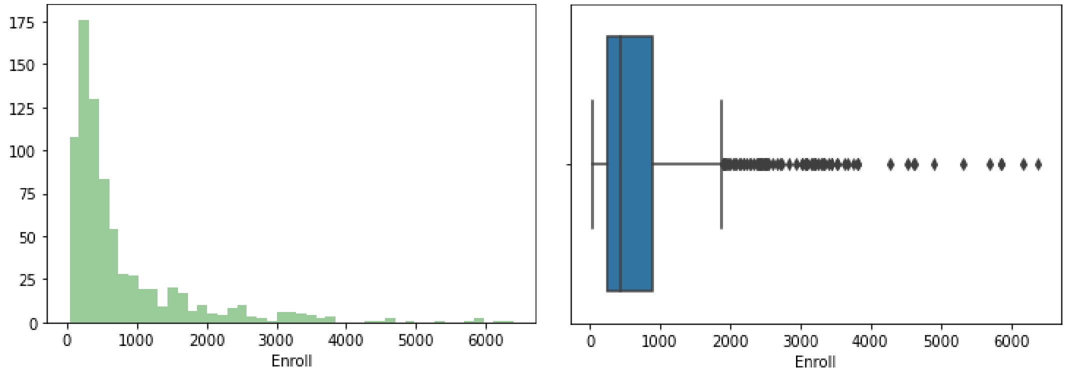


Figure 5

```
Description of Top10perc-%of new students from top 10%of H.Secondary class
---------------------------------------------------------------------------
count    777.000000
mean      27.558559
std       17.640364
```

```
min          1.000000
25%         15.000000
50%         23.000000
75%         35.000000
max         96.000000

Distribution of Top10perc            Box Plot of Top10perc
------------------------------------------------------------------------
```

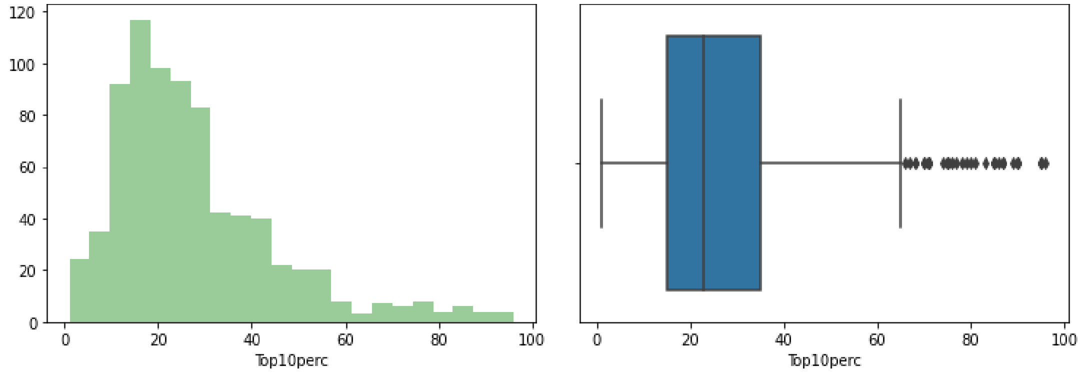


Figure 6

```
Description of Top25perc-%of new students from top 25%of H.Secondary class
--------------------------------------------------------------------------
count      777.000000
mean        55.796654
std         19.804778
min          9.000000
25%         41.000000
50%         54.000000
75%         69.000000
max        100.000000
Distribution of Top25perc            Box Plot Of Top25perc
```

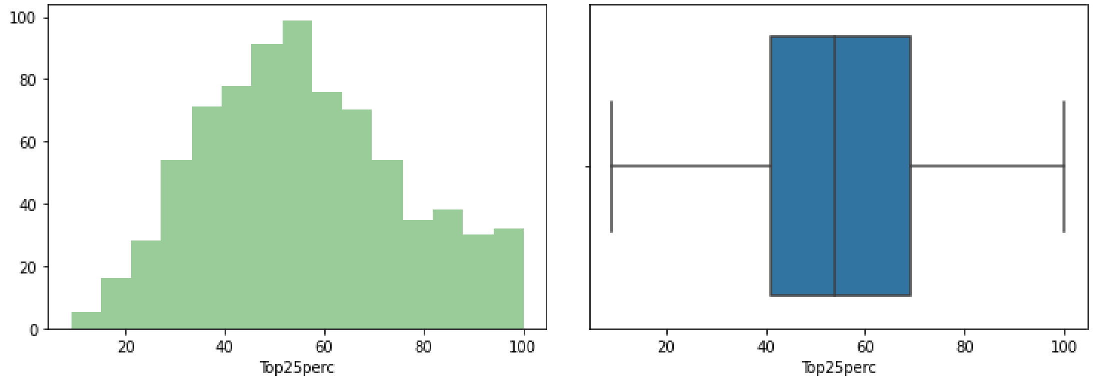


Figure 7

```
Description of F.Undergrad- Number of full-time undergraduate students
---------------------------------------------------------------------------
--
count      777.000000
mean      3699.907336
std       4850.420531
min        139.000000
25%        992.000000
50%       1707.000000
75%       4005.000000
max      31643.000000

Distribution of F.Undergrad            Box Plot of F.Undergrad
-------------------------------------------------------------------------
```

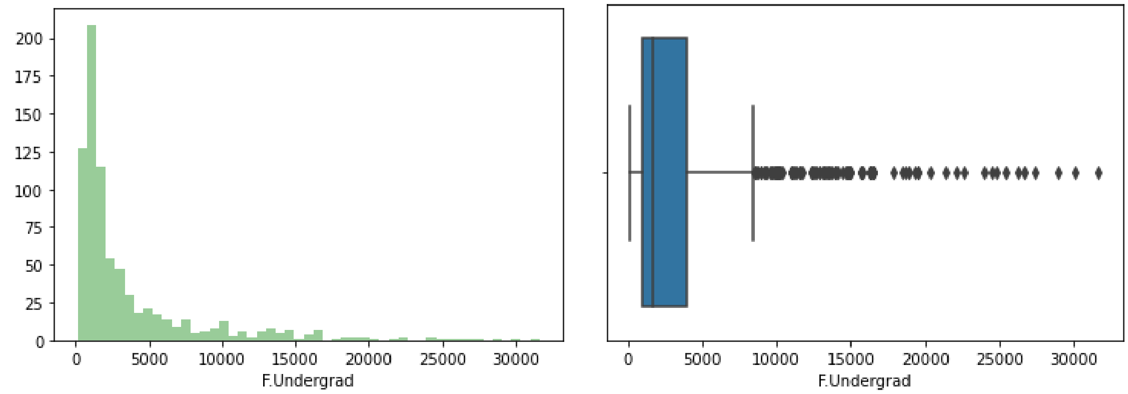
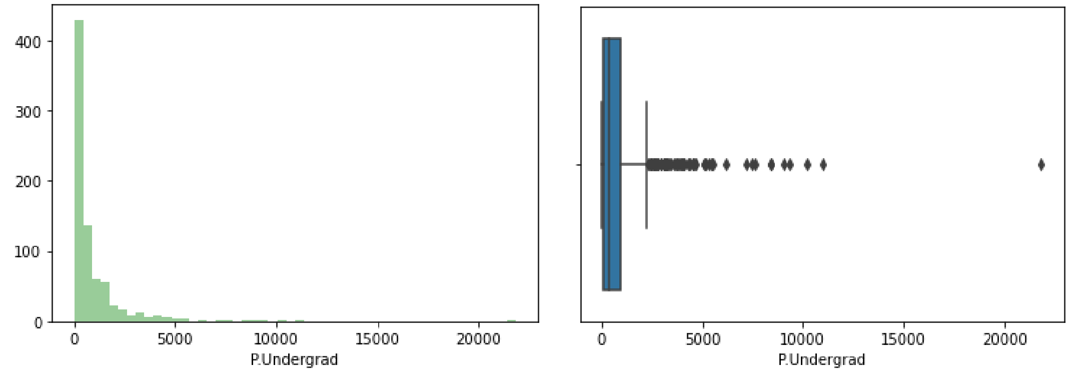


Figure 8

```
Description of P.Undergrad-Number of part-time undergraduate students
---------------------------------------------------------------------------
--
count      777.000000
mean       855.298584
std       1522.431887
min          1.000000
25%         95.000000
50%        353.000000
75%        967.000000
max      21836.000000

Distribution of P.Undergrad              Box Plot of P.Undergrad
-------------------------------------------------------------------------
```

Figure 9



Description of Outstate- outstation students

```
----------------------------------------------------------------------------
count      777.000000
mean     10440.669241
std       4023.016484
min       2340.000000
25%       7320.000000
50%       9990.000000
75%      12925.000000
max      21700.000000

Distribution of Outstate                Box Plot of Outstate
---------------------------------------------------------------------------
```

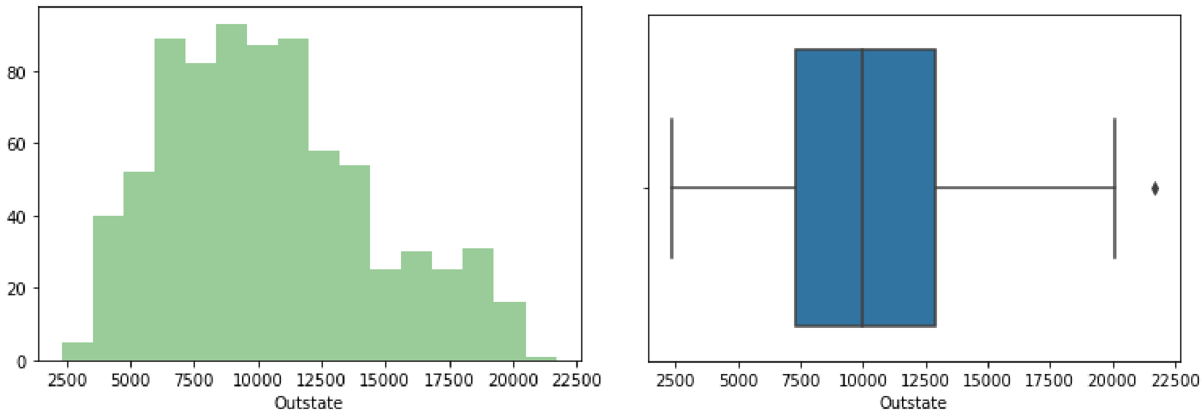


Figure 10

Description of Room.Board -Cost of Room and board

```
-----------------------------------------------------------------------------
--
count     777.000000
mean     4357.526384
std      1096.696416
min      1780.000000
25%      3597.000000
50%      4200.000000
75%      5050.000000
max      8124.000000

Distribution of Room.Board           Box Plot of Room Board
----------------------------------------------------------------------------
```

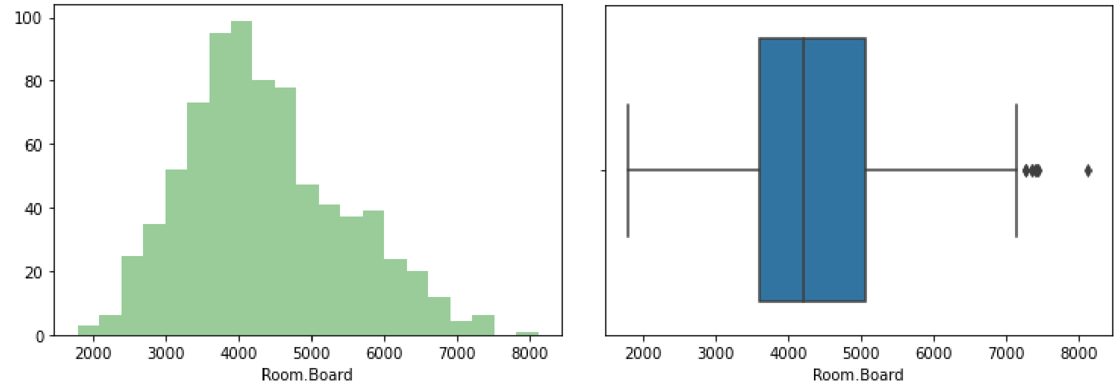


Figure 11

```
Description of Books--Estimated book costs for a student
---------------------------------------------------------------------------
--
count      777.000000
mean       549.380952
std        165.105360
min         96.000000
25%        470.000000
50%        500.000000
75%        600.000000
max       2340.000000


Distribution of Books                Box Plot of Books
-------------------------------------------------------------------------
```

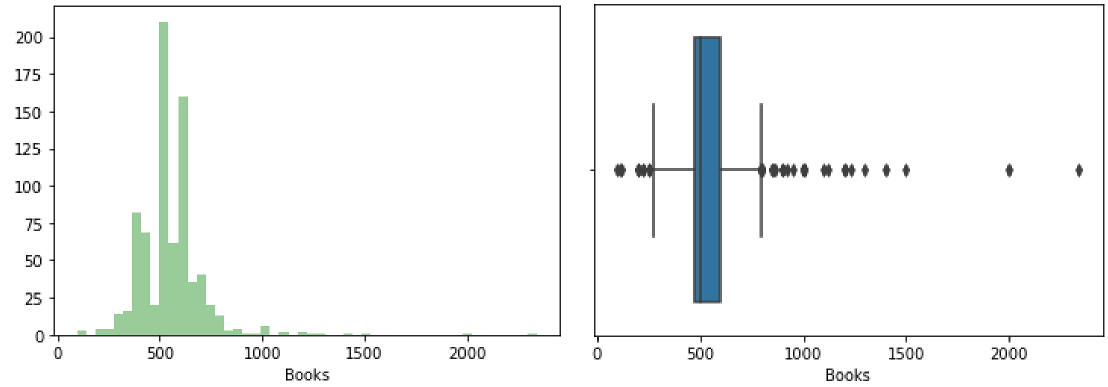


Figure 12

```
Description of Personal--Estimated personal spending for a student
---------------------------------------------------------------------------
--
count      777.000000
mean      1340.642214
std        677.071454
min        250.000000
25%        850.000000
```

```
50%       1200.000000
75%       1700.000000
max       6800.000000


      Distribution of Personal         Box Plot of Personal
------------------------------------------------------------------------
```

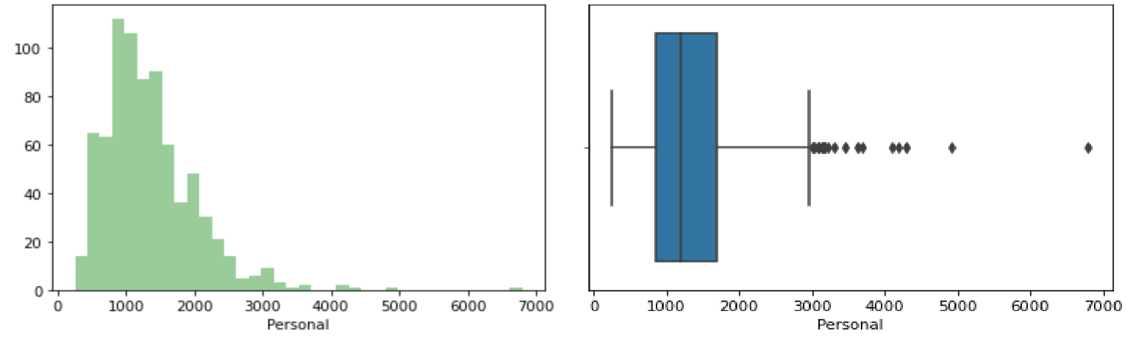


Figure 13

Description of PhD -Percentage of faculties with Ph.D.'s

```
------------------------------------------------------------------------
count    777.000000
mean      72.660232
std       16.328155
min        8.000000
25%       62.000000
50%       75.000000
75%       85.000000
max      103.000000
      Distribution of PhD                   Box Plot of PhD
-------------------------------------------------------------------------
```

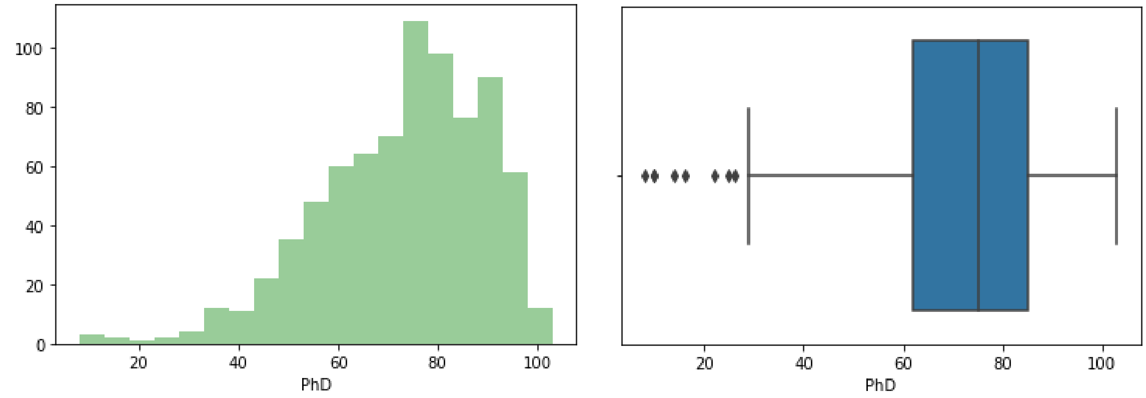


Figure 14

```
Description of Terminal -  Percentage of faculties with terminal degree
---------------------------------------------------------------------------
count    777.000000
mean      79.702703
std       14.722359
min       24.000000
25%       71.000000
50%       82.000000
75%       92.000000
max      100.000000
     Distribution of Terminal            Box Plot Of Terminal
---------------------------------------------------------------------------
```

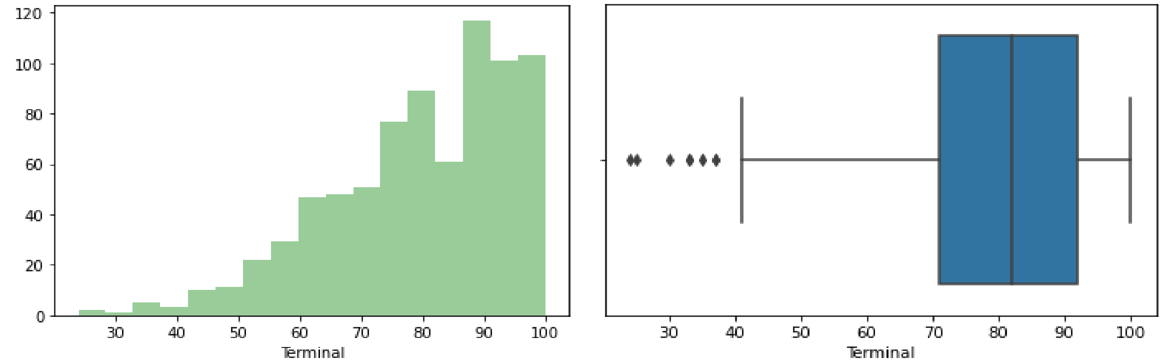


Figure 15

Description of S.F.Ratio --Student/faculty ratio

```
count    777.000000

mean         14.089704
std           3.958349
min           2.500000
25%          11.500000
50%          13.600000
75%          16.500000
max          39.800000

Distribution of S.F.Ratio              Box Plot of S.F Ratio
-----------------------------------------------------------------------
```

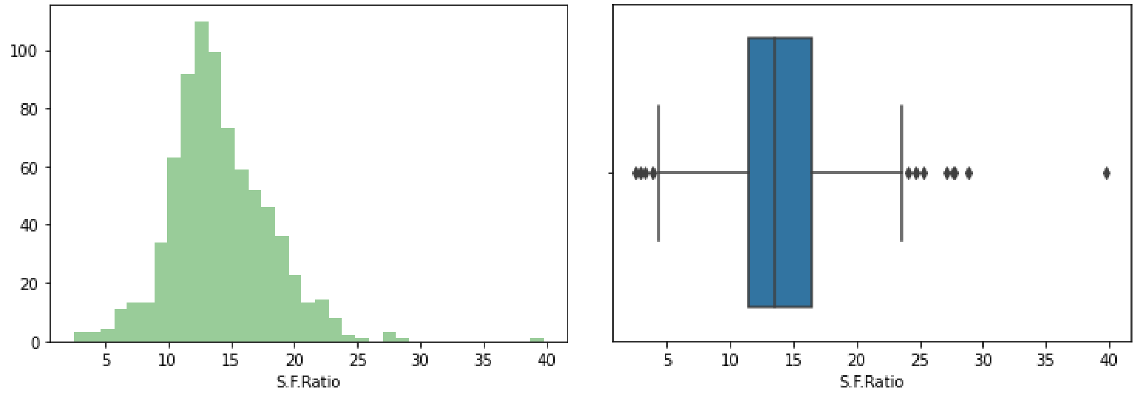


Figure 16

```
Description of perc.alumni--Percentage of alumni who donate
-------------------------------------------------------------------------
count     777.000000
mean       22.743887
std        12.391801
min         0.000000
25%        13.000000
50%        21.000000
75%        31.000000
max        64.000000
Distribution of perc.alumni        Box Plot of Alumni
------------------------------------------------------------------------
```

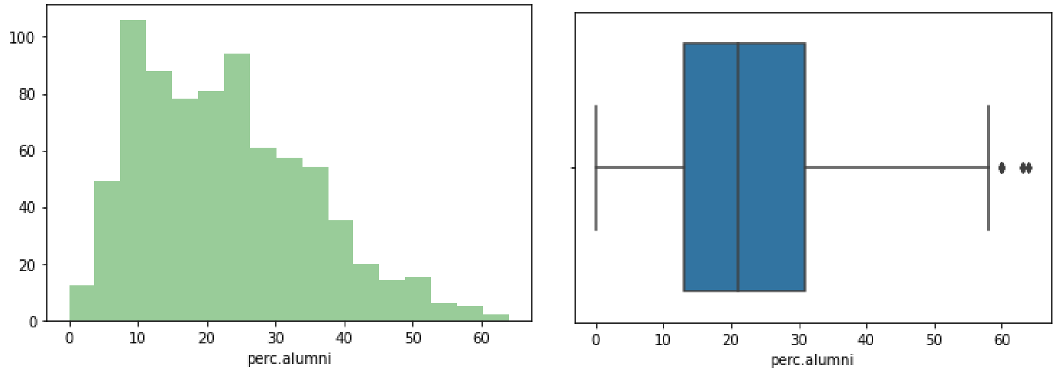


Figure 17

Description of Expend --The Instructional expenditure per student

```
----------------------------------------------------------------------------
--
count      777.000000
mean      9660.171171
std       5221.768440
min       3186.000000
25%       6751.000000
50%       8377.000000
75%      10830.000000
max      56233.000000

Distribution of Expend                       Box Plot of Expend
-------------------------------------------------------------------------
```

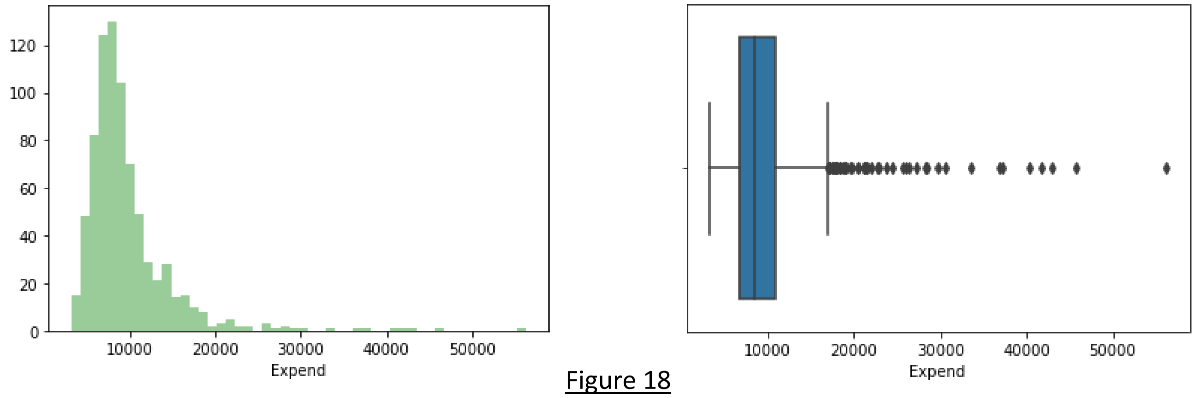


Figure 18

Description of Grad.Rate --Graduation rate

```
----------------------------------------------------------------------------
--
count    777.00000
mean      65.46332
std       17.17771
min       10.00000
25%       53.00000
50%       65.00000
75%       78.00000
max      118.00000
Distribution of Grad.Rate           Box Plot of Grad.Rate
------------------------------------------------------------------------
```

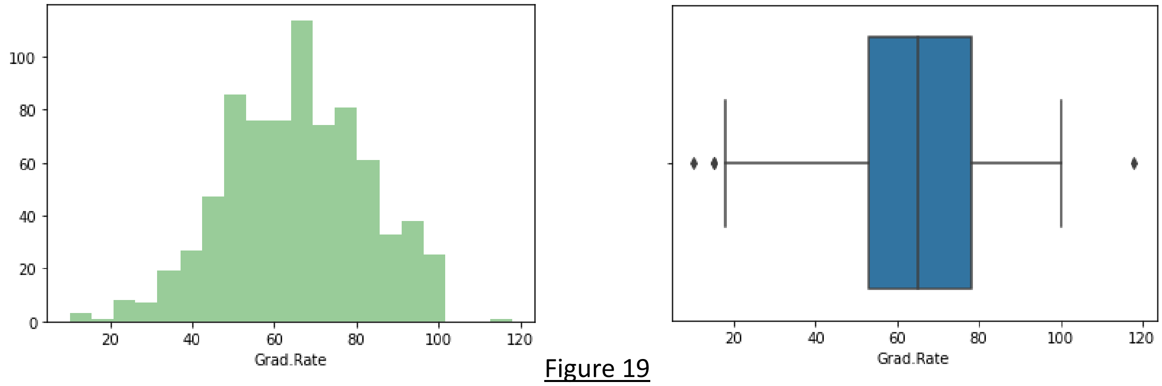


Figure 19

Let’s now perform Multivariate Analysis on all the variables:

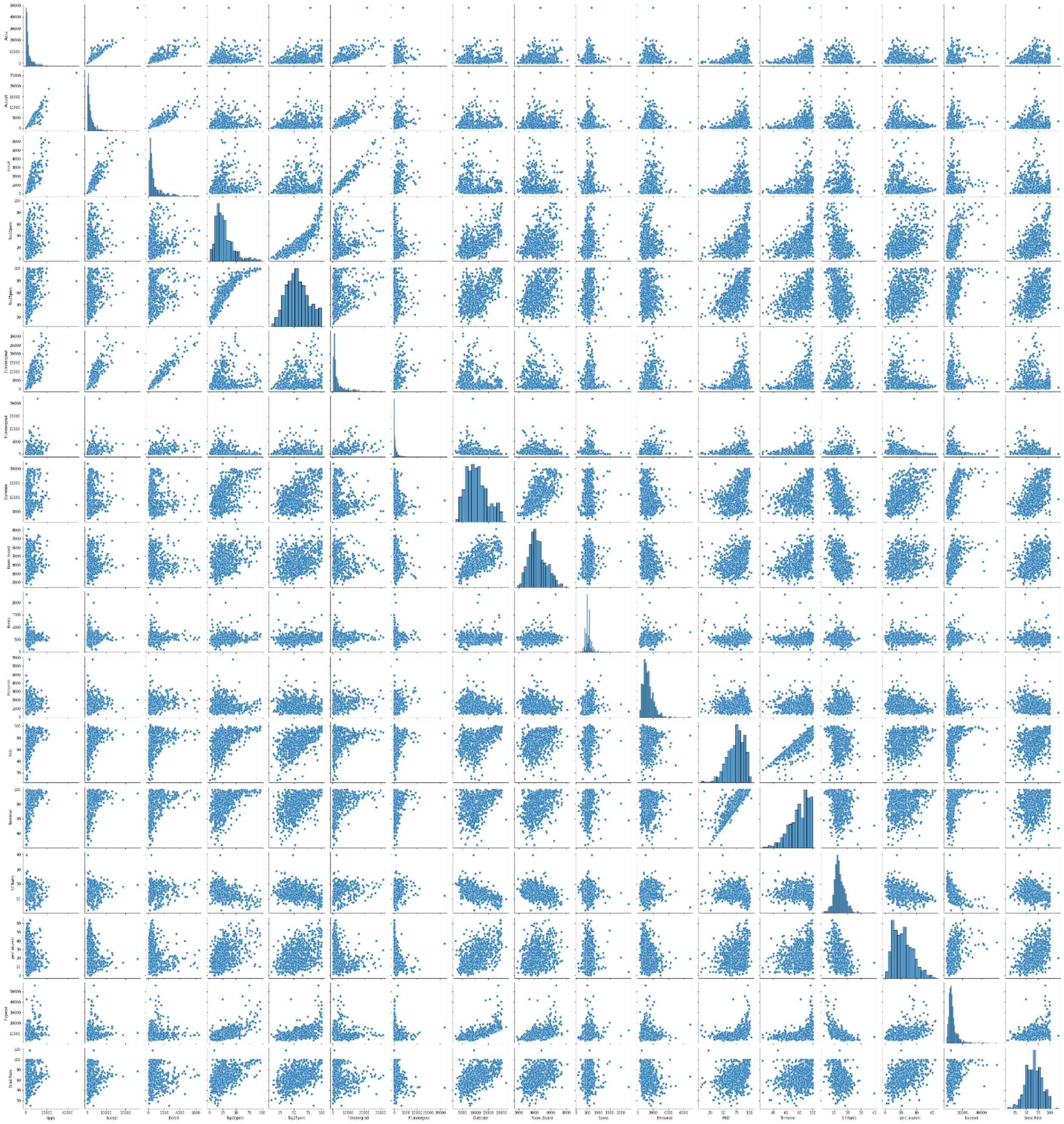


Figure 20– Pair plot of all the variables

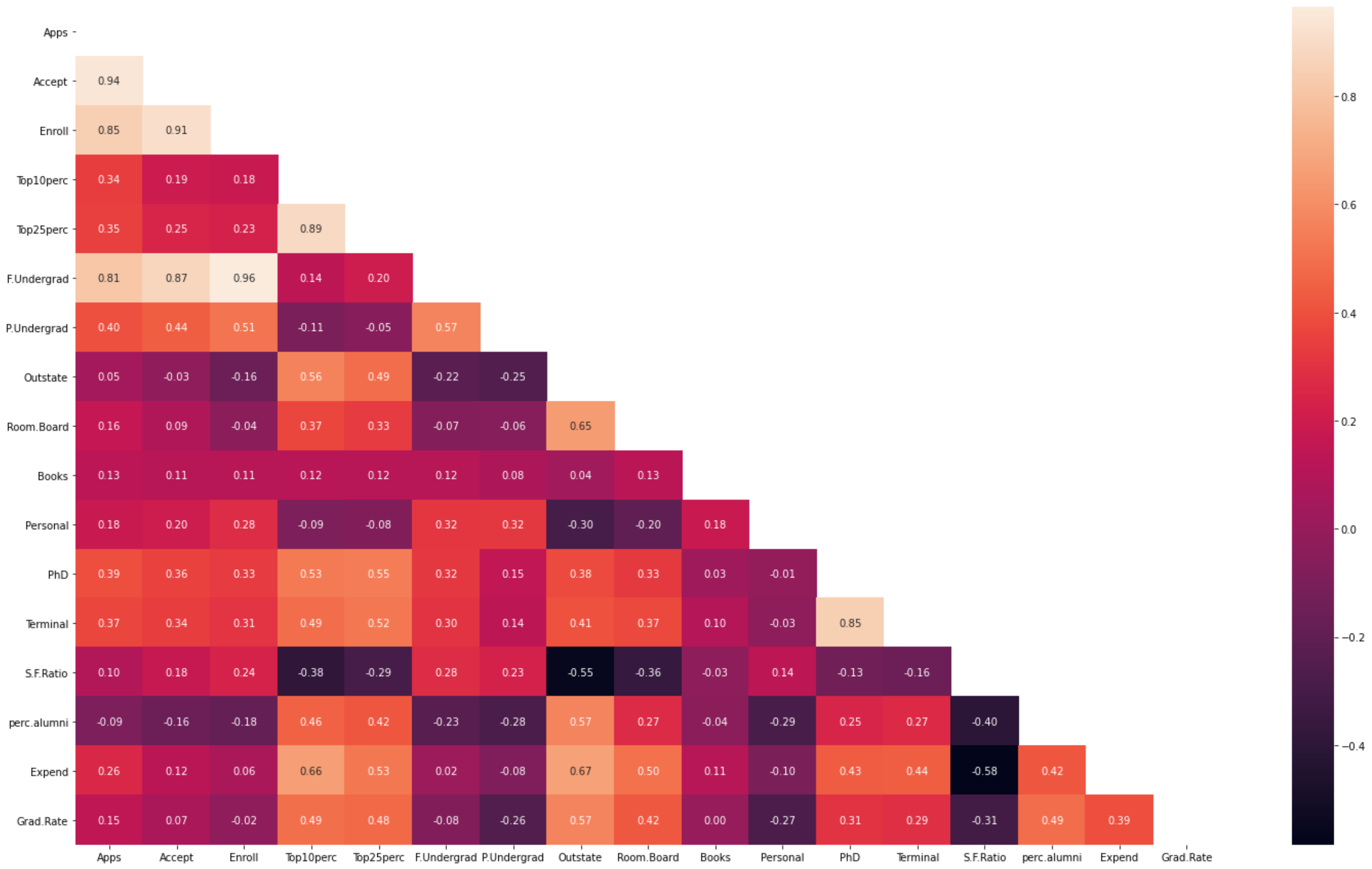


Figure 21- Heat map or correlation map between all the variables

From the Univariate analysis through distribution plot , we have seen that there is skewness in all the data that the distribution is not normal, that is it may be right or left skewed .
This interpretation was further clarified by the box plot as almost all the variables showed significant amount of outliers, except ‘Top 25 Percentage’ which showed no outlier at all .

From the Multivariate analysis through Pair Plot we could see that there are evidence of relationship between certain variables, which was further clarified through the Heat Map where we could see strong relationship between ‘Accepted application and number of application received’ , ‘Fulltime undergraduates & number of new students enrolled’ etc .

2.*Is scaling necessary for PCA in this case? Give justification and perform scaling.*

Scaling is necessary for PCA in this case as the data is not normally distributed. When we work with averages its better to have the data normally distributed since in this data we have seen through the univariate analysis we performed above that there are huge outliers present in the data which would have a negative impact in our analysis .

Refer scaled data in the python file.

3.Comment on the comparison between the covariance and the correlation matrices from this data [on scaled data].

The covariance matrix is mathematical representation of the total variance of the individual dimensions and across dimension. This data shows lot of data present in the mathematical space which could be understood through the covariance matrix.

Covariance and correlation are very closely related to each other. Covariance tell us how much is the variance in the data and correlation helps us understand in which direction is the variance present .

We can see lot of data which are highly corelated in the given data .

4.***Check the dataset for outliers before and after scaling. What insight do you derive her?***

The Figure below shows the outliers before performing the scaling .

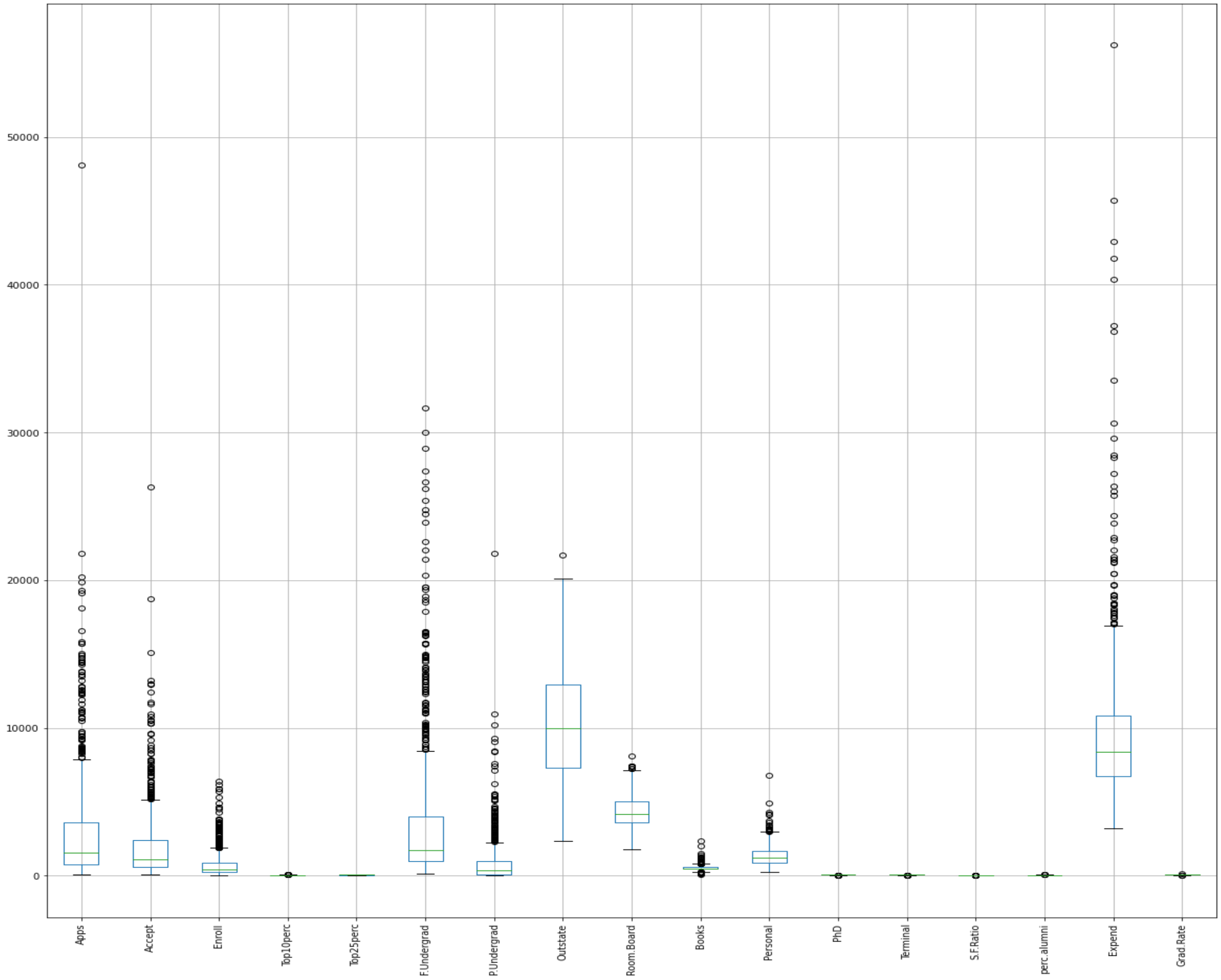


Figure 22– Box plot of unscaled data

The below Figure shows the outliers in a scaled data

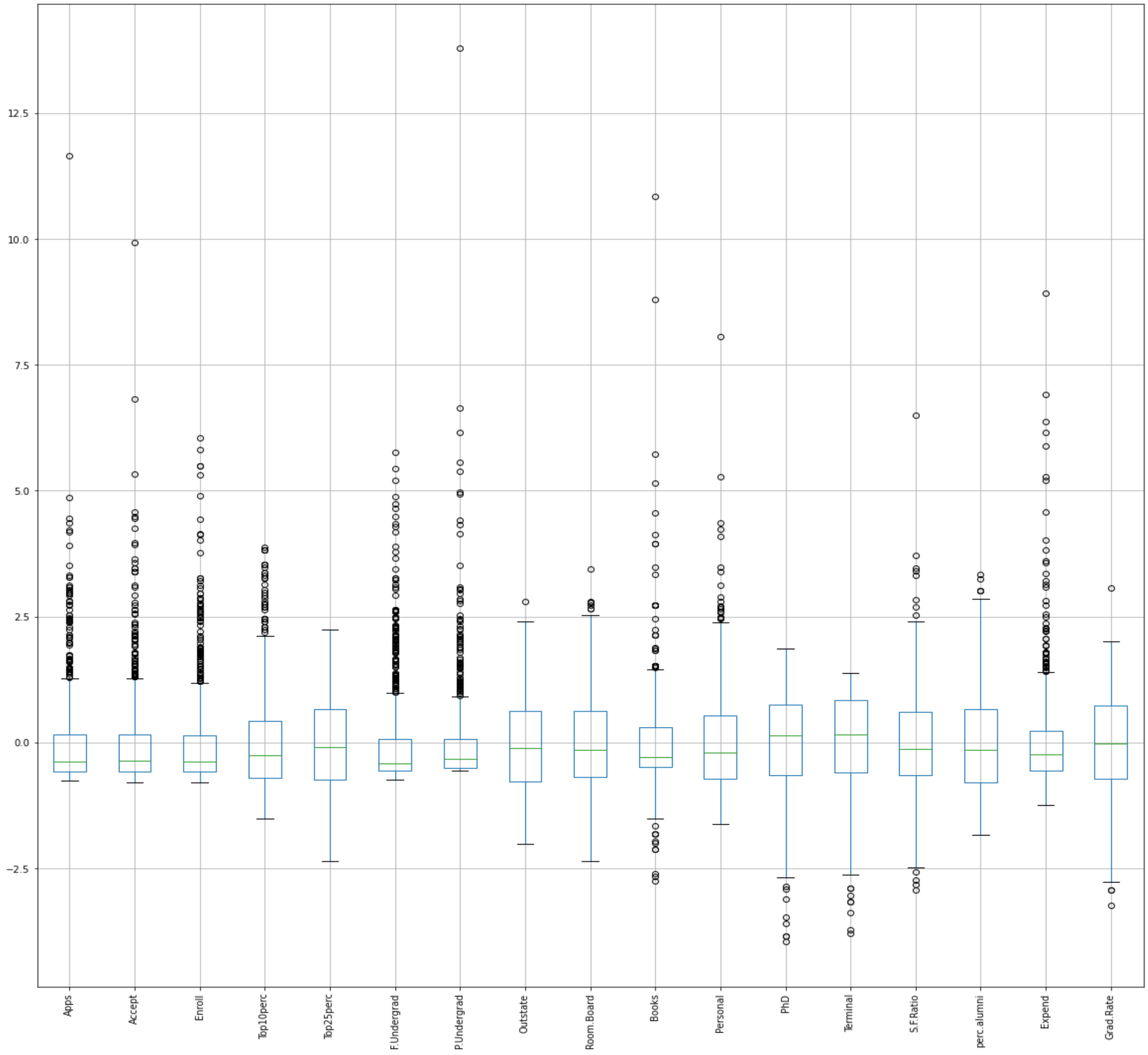


Figure 23- Box plot of scaled data

From both the figures of box plot before and after scaling , we can see that the means of all variables are near to each other after scaling and are normally distributed , in comparison with the unscaled data whe re means are significantly fluctuating .

5.***Extract the eigenvalues and eigenvectors.***

Eigen vectors extracted as below :

```
array([[ 2.48765602e-01,  2.07601502e-01,  1.76303592e-01,
         3.54273947e-01,  3.44001279e-01,  1.54640962e-01,
         2.64425045e-02,  2.94736419e-01,  2.49030449e-01,
         6.47575181e-02, -4.25285386e-02,  3.18312875e-01,
         3.17056016e-01, -1.76957895e-01,  2.05082369e-01,
         3.18908750e-01,  2.52315654e-01],
       [ 3.31598227e-01,  3.72116750e-01,  4.03724252e-01,
        -8.24118211e-02, -4.47786551e-02,  4.17673774e-01,
         3.15087830e-01, -2.49643522e-01, -1.37808883e-01,
         5.63418434e-02,  2.19929218e-01,  5.83113174e-02,
         4.64294477e-02,  2.46665277e-01, -2.46595274e-01,
        -1.31689865e-01, -1.69240532e-01],
       [-6.30921033e-02, -1.01249056e-01, -8.29855709e-02,
         3.50555339e-02, -2.41479376e-02, -6.13929764e-02,
         1.39681716e-01,  4.65988731e-02,  1.48967389e-01,
         6.77411649e-01,  4.99721120e-01, -1.27028371e-01,
        -6.60375454e-02, -2.89848401e-01, -1.46989274e-01,
         2.26743985e-01, -2.08064649e-01],
       [ 2.81310530e-01,  2.67817346e-01,  1.61826771e-01,
        -5.15472524e-02, -1.09766541e-01,  1.00412335e-01,
        -1.58558487e-01,  1.31291364e-01,  1.84995991e-01,
         8.70892205e-02, -2.30710568e-01, -5.34724832e-01,
        -5.19443019e-01, -1.61189487e-01,  1.73142230e-02,
         7.92734946e-02,  2.69129066e-01],
       [ 5.74140964e-03,  5.57860920e-02, -5.56936353e-02,
        -3.95434345e-01, -4.26533594e-01, -4.34543659e-02,
         3.02385408e-01,  2.22532003e-01,  5.60919470e-01,
        -1.27288825e-01, -2.22311021e-01,  1.40166326e-01,
         2.04719730e-01, -7.93882496e-02, -2.16297411e-01,
         7.59581203e-02, -1.09267913e-01],
       [-1.62374420e-02,  7.53468452e-03, -4.25579803e-02,
        -5.26927980e-02,  3.30915896e-02, -4.34542349e-02,
        -1.91198583e-01, -3.00003910e-02,  1.62755446e-01,
         6.41054950e-01, -3.31398003e-01,  9.12555212e-02,
         1.54927646e-01,  4.87045875e-01, -4.73400144e-02,
        -2.98118619e-01,  2.16163313e-01],
       [-4.24863486e-02, -1.29497196e-02, -2.76928937e-02,
        -1.61332069e-01, -1.18485556e-01, -2.50763629e-02,
         6.10423460e-02,  1.08528966e-01,  2.09744235e-01,
        -1.49692034e-01,  6.33790064e-01, -1.09641298e-03,
        -2.84770105e-02,  2.19259358e-01,  2.43321156e-01,
        -2.26584481e-01,  5.59943937e-01],
       [-1.03090398e-01, -5.62709623e-02,  5.86623552e-02,
        -1.22678028e-01, -1.02491967e-01,  7.88896442e-02,
         5.70783816e-01,  9.84599754e-03, -2.21453442e-01,
         2.13293009e-01, -2.32660840e-01, -7.70400002e-02,
        -1.21613297e-02, -8.36048735e-02,  6.78523654e-01,
        -5.41593771e-02, -5.33553891e-03],
       [-9.02270802e-02, -1.77864814e-01, -1.28560713e-01,
         3.41099863e-01,  4.03711989e-01, -5.94419181e-02,
         5.60672902e-01, -4.57332880e-03,  2.75022548e-01,
        -1.33663353e-01, -9.44688900e-02, -1.85181525e-01,
        -2.54938198e-01,  2.74544380e-01, -2.55334907e-01,
        -4.91388809e-02,  4.19043052e-02],
       [ 5.25098025e-02,  4.11400844e-02,  3.44879147e-02,
         6.40257785e-02,  1.45492289e-02,  2.08471834e-02,
        -2.23105808e-01,  1.86675363e-01,  2.98324237e-01,
        -8.20292186e-02,  1.36027616e-01, -1.23452200e-01,
        -8.85784627e-02,  4.72045249e-01,  4.22999706e-01,
         1.32286331e-01, -5.90271067e-01],
       [ 4.30462074e-02, -5.84055850e-02, -6.93988831e-02,
        -8.10481404e-03, -2.73128469e-01, -8.11578181e-02,
         1.00693324e-01,  1.43220673e-01, -3.59321731e-01,
         3.19400370e-02, -1.85784733e-02,  4.03723253e-02,
        -5.89734026e-02,  4.45000727e-01, -1.30727978e-01,
         6.92088870e-01,  2.19839000e-01],
       [ 2.40709086e-02, -1.45102446e-01,  1.11431545e-02,
         3.85543001e-02, -8.93515563e-02,  5.61767721e-02,
        -6.35360730e-02, -8.23443779e-01,  3.54559731e-01,
        -2.81593679e-02, -3.92640266e-02,  2.32224316e-02,
         1.64850420e-02, -1.10262122e-02,  1.82660654e-01,
         3.25982295e-01,  1.22106697e-01],
       [ 5.95830975e-01,  2.92642398e-01, -4.44638207e-01,
         1.02303616e-03,  2.18838802e-02, -5.23622267e-01,
         1.25997650e-01, -1.41856014e-01, -6.97485854e-02,
         1.14379958e-02,  3.94547417e-02,  1.27696382e-01,
        -5.83134662e-02, -1.77152700e-02,  1.04088088e-01,
        -9.37464497e-02, -6.91969778e-02],
       [ 8.06328039e-02,  3.34674281e-02, -8.56967180e-02,
        -1.07828189e-01,  1.51742110e-01, -5.63728817e-02,
         1.92857500e-02, -3.40115407e-02, -5.84289756e-02,
        -6.68494643e-02,  2.75286207e-02, -6.91126145e-01,
         6.71008607e-01,  4.13740967e-02, -2.71542091e-02,
         7.31225166e-02,  3.64767385e-02],
       [ 1.33405806e-01, -1.45497511e-01,  2.95896092e-02,
         6.97722522e-01, -6.17274818e-01,  9.91640992e-03,
         2.09515982e-02,  3.83544794e-02,  3.40197083e-03,
        -9.43887925e-03, -3.09001353e-03, -1.12055599e-01,
         1.58909651e-01, -2.08991284e-02, -8.41789410e-03,
        -2.27742017e-01, -3.39433604e-03],
       [ 4.59139498e-01, -5.18568789e-01, -4.04318439e-01,
        -1.48738723e-01,  5.18683400e-02,  5.60363054e-01,
        -5.27313042e-02,  1.01594830e-01, -2.59293381e-02,
         2.88282896e-03, -1.28904022e-02,  2.98075465e-02,
        -2.70759809e-02, -2.12476294e-02,  3.33406243e-03,
        -4.38803230e-02, -5.00844705e-03],
       [ 3.58970400e-01, -5.43427250e-01,  6.09651110e-01,
        -1.44986329e-01,  8.03478445e-02, -4.14705279e-01,
         9.01788964e-03,  5.08995918e-02,  1.14639620e-03,
         7.72631963e-04, -1.11433396e-03,  1.38133366e-02,
         6.20932749e-03, -2.22215182e-03, -1.91869743e-02,
        -3.53098218e-02, -1.30710024e-02]])
```

Eigen Values extracted as below:

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222, 0.53061262, 0.4043029 ,
       0.31344588, 0.22061096, 0.16779415, 0.1439785 , 0.08802464,
       0.03672545, 0.02302787])
```

6.*Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features*

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**
Apps	0.248766	0.331598	-0.06309	0.281311	0.005741	-0.01624	-0.04249
Accept	0.207602	0.372117	-0.10125	0.267817	0.055786	0.007535	-0.01295
Enroll	0.176304	0.403724	-0.08299	0.161827	-0.05569	-0.04256	-0.02769
Top10perc	0.354274	-0.08241	0.035056	-0.05155	-0.39543	-0.05269	-0.16133
Top25perc	0.344001	-0.04478	-0.02415	-0.10977	-0.42653	0.033092	-0.11849
F.Undergrad	0.154641	0.417674	-0.06139	0.100412	-0.04345	-0.04345	-0.02508
P.Undergrad	0.026443	0.315088	0.139682	-0.15856	0.302385	-0.1912	0.061042
Outstate	0.294736	-0.24964	0.046599	0.131291	0.222532	-0.03	0.108529
Room.Board	0.24903	-0.13781	0.148967	0.184996	0.560919	0.162755	0.209744
Books	0.064758	0.056342	0.677412	0.087089	-0.12729	0.641055	-0.14969
Personal	-0.04253	0.219929	0.499721	-0.23071	-0.22231	-0.3314	0.63379
PhD	0.318313	0.058311	-0.12703	-0.53472	0.140166	0.091256	-0.0011
Terminal	0.317056	0.046429	-0.06604	-0.51944	0.20472	0.154928	-0.02848
S.F.Ratio	-0.17696	0.246665	-0.28985	-0.16119	-0.07939	0.487046	0.219259
perc.alumni	0.205082	-0.2466	-0.14699	0.017314	-0.2163	-0.04734	0.243321
Expend	0.318909	-0.13169	0.226744	0.079273	0.075958	-0.29812	-0.22658
Grad.Rate	0.252316	-0.16924	-0.20806	0.269129	-0.10927	0.216163	0.559944

Table 5- DATAFRAME OF PC COMPONENT EIGENVECTORS WITH ORIGINAL FEATURES

7.*Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only).*

The Explicit form of the first PC in terms of the eigenvectors are:

PC1=a1x1+a1x2+…anxn

	PC1
0	-1.59
1	-2.19
2	-1.43
3	2.86
4	-2.21
5	-0.57
6	0.24

8. *Consider 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?*

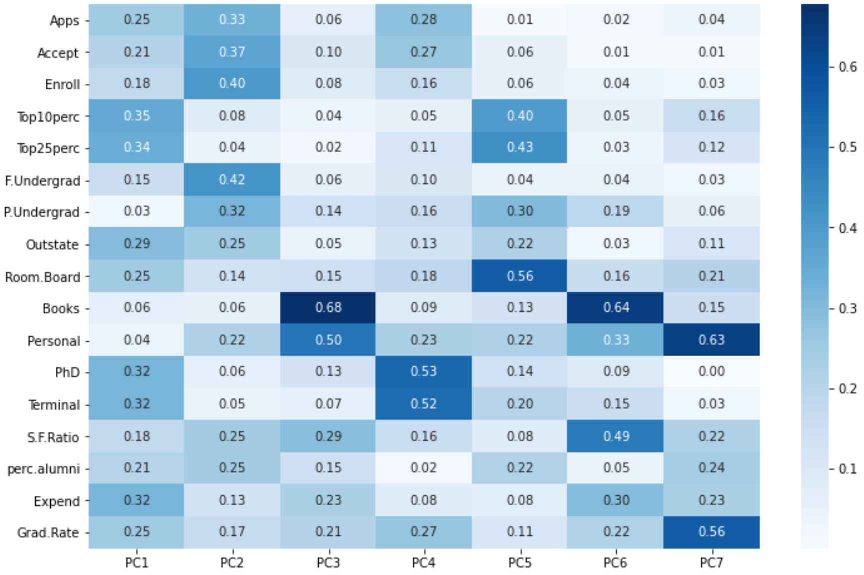
```
array([0.32020628, 0.58360843, 0.65261759, 0.71184748, 0.76673154,
       0.81657854, 0.85216726, 0.88670347, 0.91787581, 0.94162773,
       0.96004199, 0.9730024 , 0.98285994, 0.99131837, 0.99648962,
       0.99864716, 1.        ])
```

As we can above value are the cumulative form of eigen values , which helps us determine how many principal components to filter for further analysis. Each value represent PC's extracted from PC1 TO PC17 equivalent to the number of features .As we can see the first PC shows 32% of the data content in it , and cumulatively PC2 contains 58% of the data and so on until PC17 where we can get 100% of the data if we choose to work on all the PC's. But it contradicts our choice of performing PCA , because the main reason was to reduce the dimensionality Hence for this case study I choose to take 85% of the data for my analysis and go ahead with 7 dimensions PC1,PC2,PC3,PC4,PC5,PC6 & PC7,hene reducing 10 dimension and working with only 7 to ease our analysis .

Eigen vectors indicates the direction of the principal component, which helps in understanding the linear transformation .

9. *Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis?*

After performing Principal component analysis we can conclude that , there is significant corelation in the given data set , and to analyse it further we performed scaling in the data and went ahead with PCA which helped us in reducing the dimensions to 7 PC from 17 PCs, having 85% of the information still availbale to us in the 7 PC's.
Below are some further information on the 7 PC's extracted as to which feature has more information in which PC.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Apps	0.25	0.33	0.06	0.28	0.01	0.02	0.04
Accept	0.21	0.37	0.10	0.27	0.06	0.01	0.01
Enroll	0.18	0.40	0.08	0.16	0.06	0.04	0.03
Top10perc	0.35	0.08	0.04	0.05	0.40	0.05	0.16
Top25perc	0.34	0.04	0.02	0.11	0.43	0.03	0.12
F.Undergrad	0.15	0.42	0.06	0.10	0.04	0.04	0.03
P.Undergrad	0.03	0.32	0.14	0.16	0.30	0.19	0.06
Outstate	0.29	0.25	0.05	0.13	0.22	0.03	0.11
Room.Board	0.25	0.14	0.15	0.18	0.56	0.16	0.21
Books	0.06	0.06	0.68	0.09	0.13	0.64	0.15
Personal	0.04	0.22	0.50	0.23	0.22	0.33	0.63
PhD	0.32	0.06	0.13	0.53	0.14	0.09	0.00
Terminal	0.32	0.05	0.07	0.52	0.20	0.15	0.03
S.F.Ratio	0.18	0.25	0.29	0.16	0.08	0.49	0.22
perc.alumni	0.21	0.25	0.15	0.02	0.22	0.05	0.24
Expend	0.32	0.13	0.23	0.08	0.08	0.30	0.23
Grad.Rate	0.25	0.17	0.21	0.27	0.11	0.22	0.56

Figure 24-Heat Map

Form the above figure 24 we can infer that

- PC1 contains more information on ‘Top10perc’- Percentage of new students from top 10% of H igher Secondary class , ‘Top25perc’ , PhD which is giving more insights on the toppers admitted i n colleges and highly qualified faculty in colleges.
- PC2 contains more Information on application received , application accepted and new students enrolled, hence providing insights on the applied rate and acceptance rate of colleges.
- PC3 is providing us with an data related to costs and funding as we can see from the figure more information is from Books cost and personal spending of student.
- In PC4 we can more information on the faculty staff qualification .
- PC5 provides information on Room Boarding expenses and so on .

Hence the reduced dimension can be further used and analysed with ease to gain significant insights.

-----------x-x----------