Advanced statistics PROJECT

Business report

1.Salary Data analysis.

2. Education post 12th standard PCA analysis.

Student Name: Vivek Bhatia

PGP-DSBA Online Jan_C 2022

Date: 22/04/2022

Table of Contents

List of figures

List of Pictures

Salary Data

1.Executive summary

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.

2.Introduction

The data set provides salaries of 40 individuals along with their educational qualification. This report provides detailed explanation of various problems mentioned in the assignment and its related solution and the inferences made out of that analysis.

3.Data set 1 description

Salary data was imported into jupyter note book to perform one way and two-way anova on salary with respect to education and occupation and explain the business implications of performing ANOVA for this particular case study.

4.Exploratory data analysis

4a. Sample of data set

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769

4b. Check for the type of variables in the data frame

The data set consists of 40 rows and 3 columns with salary as int64 and Education and Occupation as Object data type. Also, we can see there are no null values in the data set.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Education   40 non-null     object
 1   Occupation  40 non-null     object
 2   Salary      40 non-null     int64 
dtypes: int64(1), object(2)
memory usage: 1.1+ KB
```

5.Question Problem set 1

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

Hypothesis testing for educational qualification.

Ho: The mean salary is same across all categories of educational qualification. Ha: The mean salary is not same across atleast one category of educational qualification.

Hypothesis testing for Occupation.

Ho: The mean salary is same across all categories of Occupation. Ha: The mean salary is not same across atleast one category of Occupation.

1.2 Perform one-way ANOVA for Education with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

	df	sum_sq	mean_sq	F	PR(>F)
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

As P value is less than 0.05, at 95% significance level, **we reject the null hypothesis** that the mean salary is same across all categories of educational qualification. The mean salary for different categories of education is different for at least one of the categories.

1.3 Perform one-way ANOVA for variable Occupation with respect to the variable 'Salary'. State whether the null hypothesis is accepted or rejected based on the ANOVA results.

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

As P value is more than 0.05, at 95% significance level, **we fail to reject the null hypothesis or we accept the null hypothesis** that the mean salary is same across all categories of Occupation.

1.4 If the null hypothesis is rejected in either (1.2) or in (1.3), find out which class means are significantly different. Interpret the result.

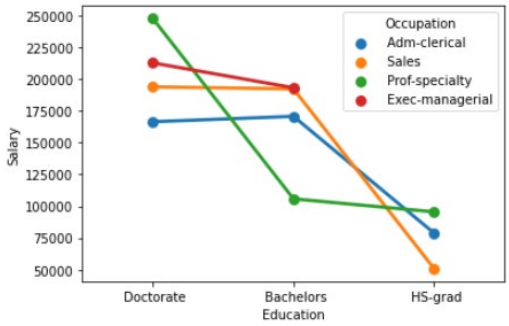
As observed the null hypothesis is rejected for education, we do a tukeyhsd test to determine which category of education has a different mean for salary and do a multi comparison test to find out the results.

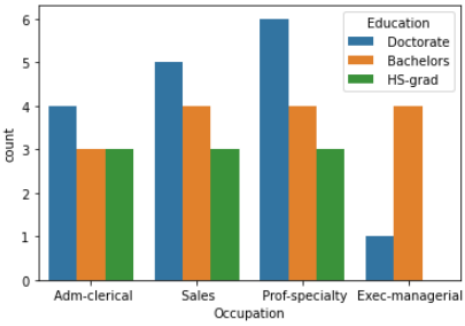
Multiple Comparison of Means - Tukey HSD, FWER=0.05

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

The result indicates that P-adj of all the groups falls below 0.05 and hence all the categories of education have different means for salaries. The reject column also highlight that we reject all the groups in terms of their mean salaries being same.

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





Observations:

To observe education w.r.t occupation let us analyse each educational qualification and compare their earnings, dominance and value addition and determine weather or not a particular occupation is suitable for a person with a particular educational qualification.

Doctorate:

- From the interaction plot we can observe that doctorate people are found in all the domains of occupation categories and from the count plot we can see that Doctorate degree holders are dominant in Prof-Speciality with a count of 6 and draw the highest salaries of approximately 250000.
- 4 people with doctorate degree holders also work in Adm-clerical position and have lowest salaries of around 160000.
- Around 5 people work in sales drawing a salary of around 190000.
- Lowest count of 1 is observed in Exec-managerial however the salaries is high around 210000.
- From the above analysis it is evident that with doctorate degree one should work in Prof-Speciality or Exec-managerial roles to draw maximum salary. There is no value addition for a doctorate personnel to work in Adm-clerical position.

Bachelors Education:

- People with Bachelor education are having identical count of 4 in sales, Prof-Speciality and Exec-managerial domains and have the lowest count of 3 in Adm-clerical roles.
- People working in Exec-managerial and sales earn highest amount of approximately 200000 followed by Adm-clerical with earnings of 175000 and the lowest earning domain for bachelors' education is Prof-Speciality.
- From the interaction plot and count plot analysis we can conclude that with Bachelor's education working in Prof-Speciality is not an ideal choice. However, sales, Prof-Speciality and Exec-managerial offer them with wide variety of options and almost at par salary.

HS-grad:

- HS-grad are working in sales, Prof-Speciality and Adm-clerical and have identical count of 3 in each of the domain.
- No HS-grad are observed to be working in Exec-managerial roles.
- HS-grad is the lowest earning group and the maximum average salary drawn by them is in Prof-Speciality with an earning of around 100000.
- The earning reduces to 75000 for Adm-clerical jobs and 50000 for sales.
- Hence, for HS-grad it can be concluded that they should work in Prof-Speciality to earn the most and doesn't have much value addition in sales department.

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

The null and alternative hypothesis can be formulated as:

Ho:There is no interaction between Education and Occupation Ha: There is an interaction between Education and Occupation

Two-way anova result:

	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

As P value is less than 0.05, we reject the null hypothesis and conclude that at 95% significance level there is an interaction between Education and Occupation. In other words, the mean salary is affected by an interaction of Education and Occupation. Refer to 1.5 for the interaction plot, it is evident that people with Doctorate degree falls into the highest earning groups followed by Bachelor's education and High school graduates.

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

By conducting ANOVA analysis, we can see the effect of two categorical variables namely Education and Occupation on the numeric variable salary. It becomes very evident on how Education and Occupation are related to salary and also if the mean salary is same for both or not.

Our hypothesis concluded that mean salary is same for Occupation but different for different education levels.

Highest earning group consisted of Doctorate and lowest earning group is High school graduate.

We also determined the most value additive job as per the educational qualifications and summarised it in section 1.5 of this report.

We also conducted TWO-WAY ANOVA to determine the combined effect of Education and Occupation and tried to visualise it through intersection plot and observed that there is a definite interaction between the two categories and to finally conclude we can infer that the higher the education the more the earning.

End of problem set 1-Salary data analysis.

Education Post 12th Standard data

1.Executive summary

The dataset contains information on various colleges to conduct a PCA analysis and explain the business implications for this case study.

2.Introduction

The data set provides detailed information on the following:

Table 1	
Abbreviation	Description
Names	Names of various university and colleges
Apps	Number of applications received
Accept	Number of applications accepted
Enroll	Number of new students enrolled
Top10perc	Percentage of new students from top 10% of Higher Secondary class
Top25perc	Percentage of new students from top 25% of Higher Secondary class
F.Undergrad	Number of full-time undergraduate students
P.Undergrad	Number of part-time undergraduate students
Outstate	Number of students for whom the particular college or university is Out-of-state tuition
Room.Board	Cost of Room and board
Books	Estimated book costs for a student
Personal	Estimated personal spending for a student
PhD	Percentage of faculties with Ph.D.'s
Terminal	Percentage of faculties with terminal degree
S.F.Ratio	Student/faculty ratio
perc.alumni	Percentage of alumni who donate
Expend	The Instructional expenditure per student
Grad.Rate	Graduation rate

3.Data set 1 description

Education Post 12th Standard data was imported into jupyter note book to PCA to explain the business implications for this particular case study.

4.Exploratory data analysis

4a. Sample of data set

Names	Abilene Christian University	Adelphi University	Adrian College	Agnes Scott College	Alaska Pacific University
Apps	1660	2186	1428	417	193
Accept	1232	1924	1097	349	146
Enroll	721	512	336	137	55
Top10perc	23	16	22	60	16
Top25perc	52	29	50	89	44
F.Undergrad	2885	2683	1036	510	249
P.Undergrad	537	1227	99	63	869
Outstate	7440	12280	11250	12960	7560
Room.Board	3300	6450	3750	5450	4120
Books	450	750	400	450	800
Personal	2200	1500	1165	875	1500
PhD	70	29	53	92	76
Terminal	78	30	66	97	72
S.F.Ratio	18.1	12.2	12.9	7.7	11.9
perc.alumni	12	16	30	37	2
Expend	7041	10527	8735	19016	10922
Grad.Rate	60	56	54	59	15

4b. Check for the type of variables in the data frame

The data set consists of 777 rows and 18 columns with different variables and the types are shown as below:

```
<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
```

There are no null values in the data set and the data type of name is Object, S.F.Ratio is float64 and remaining 16 columns are int64.

4c. Describe the dataset.

	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

From the above statistics we can observe the following:

- Minimum number of applications received is 81 and the maximum is 48094.
- Minimum number of applications accepted is 72 and the maximum is 26330.From this we can conclude that almost 54% of the applications are accepted.
- Minimum number of students enrolled is 35 and maximum is 6392. This further implies that only 13.2% of students who actually apply enroll for the course. Also, we can say that 24.27% of students whose applications are accepted finally enroll for the college studies. From this we can conclude that there is a large gap between students applying for a university and then finally getting enrolled for their subject. This area needs to be worked upon from business perspective.
- Minimum number of Percentage of new students from top 10% of Higher Secondary class is 1 and maximum is 96. This signifies the admission criteria and the level of student intake in a particular university.
- Minimum number of Percentage of new students from top 25% of Higher Secondary class is 9 and maximum is 100. This signifies the admission criteria and the level of student intake in a particular university.
- Minimum number of full-time undergraduate students is 139 with a maximum count of 31643.
- Minimum number of part-time undergraduate students is 1 with a maximum count of 21836.
- Minimum Number of students for whom the particular college or university is Out-of-state tuition is 2340 and the maximum count is 21700. From this we can conclude that a lot of students come from outstation to study and hence their fee structure and living costs will be different from students who are locally available in town.
- Minimum Cost of Room and board is 1780 and maximum cost is 8124. This implies that the universities offer a variety of options for stay and depending on the location and university campus the cost varies greatly.
- Minimum estimated personal spending for a student is 250 and can go up to a maximum of 6800.
- Minimum percentage of faculties with Ph.D.'s is 8 and maximum is 103. As the percentage cannot be greater than 100, we need to treat this data before performing PCA.
- Minimum percentage of faculties with terminal degree is 24 and maximum is 100. This data indicates the qualifications of faculty in a particular university as terminal degree refers to the highest degree that can be awarded in a particular field.
- Minimum student/faculty ratio is 2.5 and the maximum ratio is 39.8. This parameter is crucial to decide the number of students per teacher and hence can be a measure of personal attention per student in a class.
- Minimum percentage of alumni who donate is 0 and the maximum count for the same is 64.
- Minimum Instructional expenditure per student ranges from 3186 to 56233.
- Minimum graduation rate is 10 and maximum is 118. As the percentage cannot be greater than 100, we need to treat this data before performing PCA.

5.Question Problem set 2

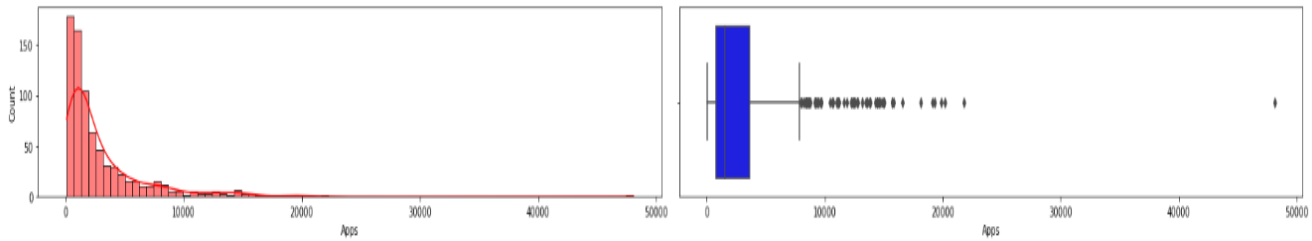
2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA?

Univariate analysis:

To understand and draw conclusions from Univariate analysis we plot histplot and boxplot for all the numeric columns and observe the data.

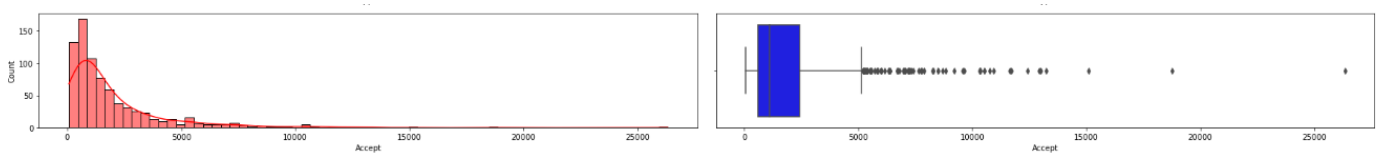
Inferences:

1. **Apps:**



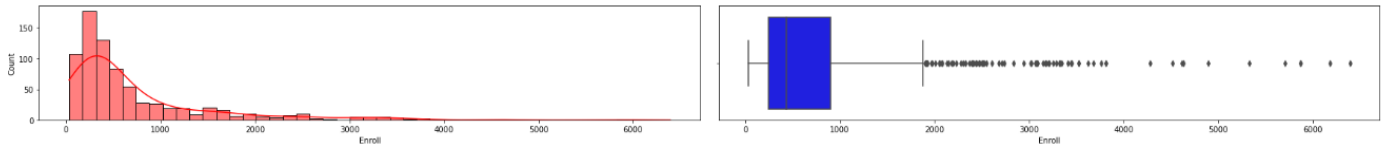
From the histplot and box plot we can see that the data is positively skewed with a lot of outliers. The histplot shows that the majority of the applications received are in the range of 3000 to 5000. However, the maximum number of applications goes up to 50000.

2. **Accept:**



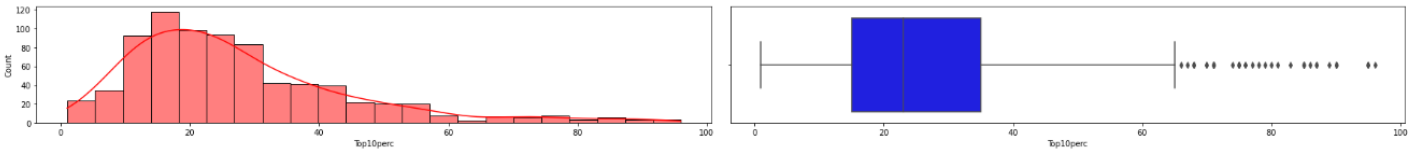
The data is positively skewed and have outliers. The histplot shows that majority of applications accepted by university are in the range of 800 to 2500 with maximum number going to around 27000.

3. **Enroll:**



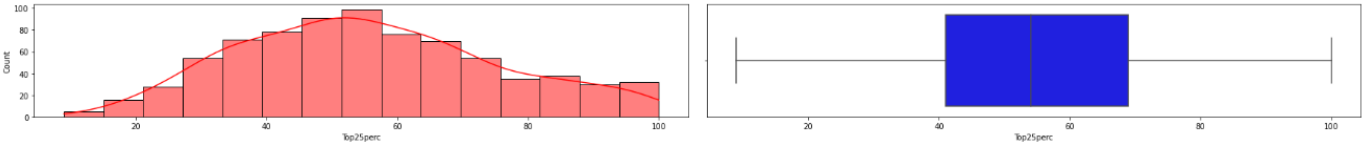
Data shows positive skewness for enroll with outliers. From the plots we can conclude that majority of students enrolling is between 200 to 1000.

4. Top 10 perc:



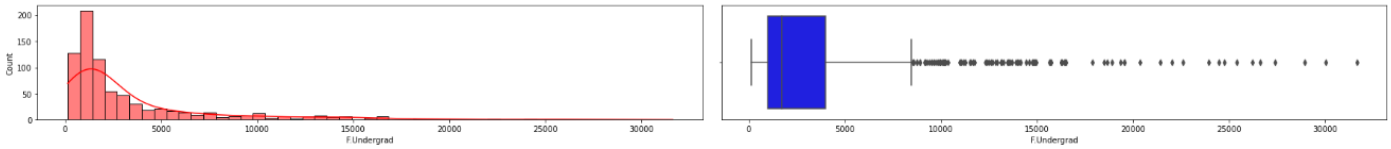
The data is slightly positive skewed and has outliers. The number of students who are in top 10 percent in higher secondary falls mostly in the range of 12 to 24.

5. Top 25 perc:



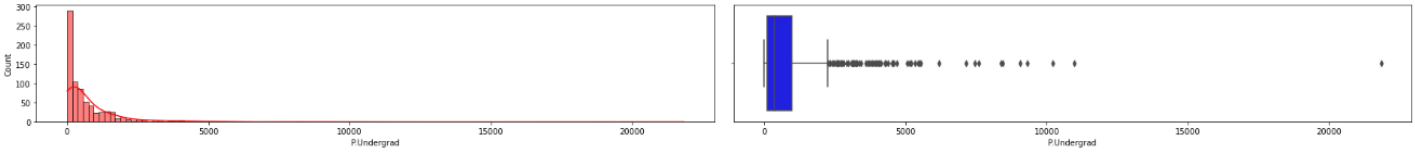
The data is almost normally distributed and has no outliers. The majority of the students who are in top 25 percentage in a university is around 50.

6. F.Undergraduate:



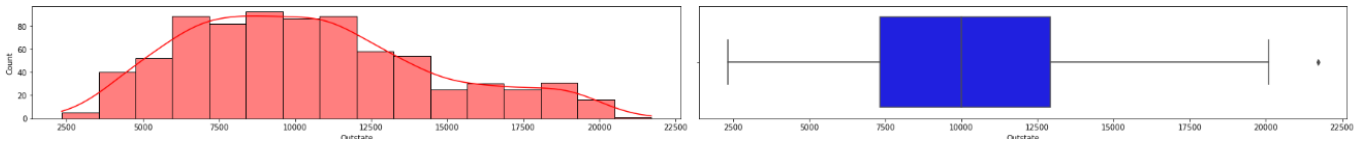
The data is highly skewed on the right and has outliers. Almost 3000 to 4500 students are full time undergraduate in the university.

7. P.Undergraduate:



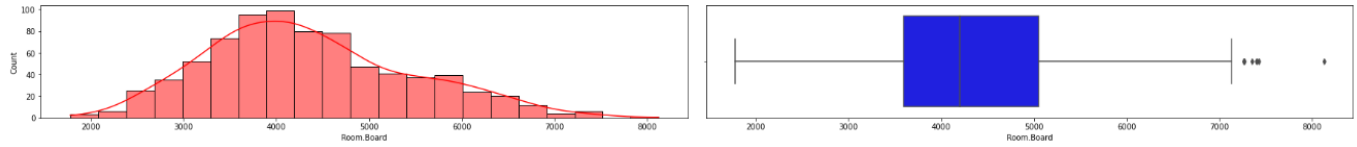
The data is highly skewed on the right and has outliers. Almost 1000 students are part time undergraduate in the university.

8. Outstate:



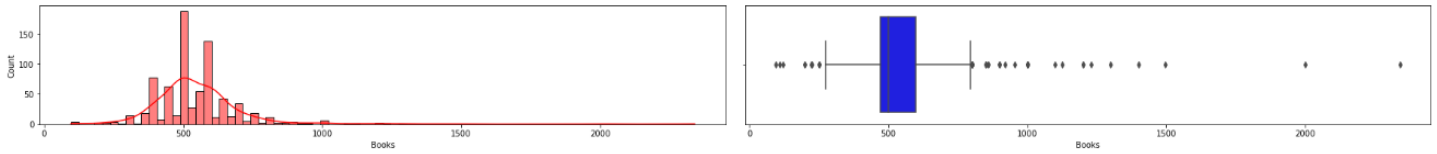
The data looks almost normal in distribution and has only one outlier. The number of students vary between 7000 to 13000.

9. Room.Board:



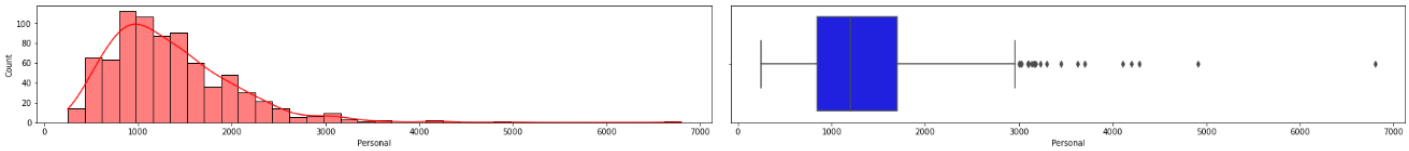
The data looks almost normal in distribution and has outliers. The cost of room and board of students vary between 3500 to 5000.

10. Books:



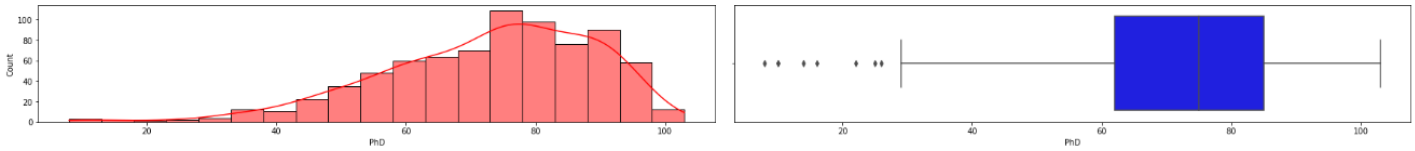
The data looks normal in distribution and has outliers. The cost of books varies from 500 to 650 for majority of the students.

11. Personal:



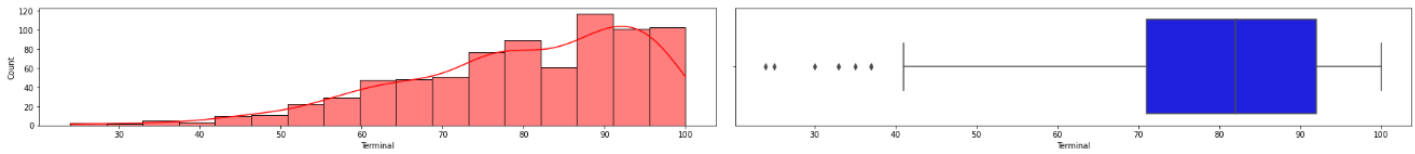
Personal expense also shows positive skewness and has outliers. The personal expenses vary from 900 to 1700 for most of the students.

12. PhD:



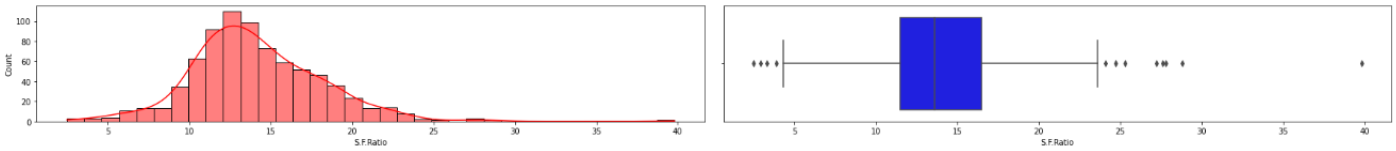
This is a negatively skewed data with outliers. Majority of the universities have PhD faculty between 62-85.

13. Terminal:



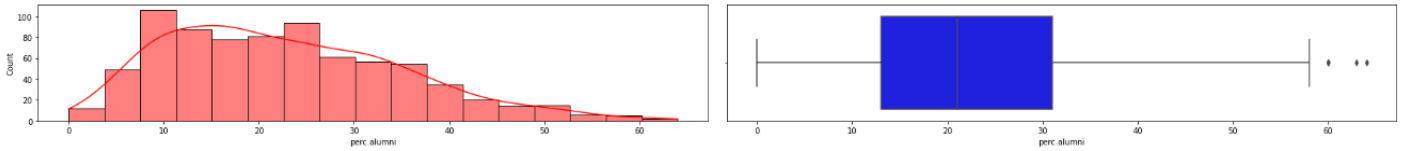
The data is negatively skewed with outliers and the terminal faculty numbers lie between 72-93 for most of the universities.

14. S.F Ratio:



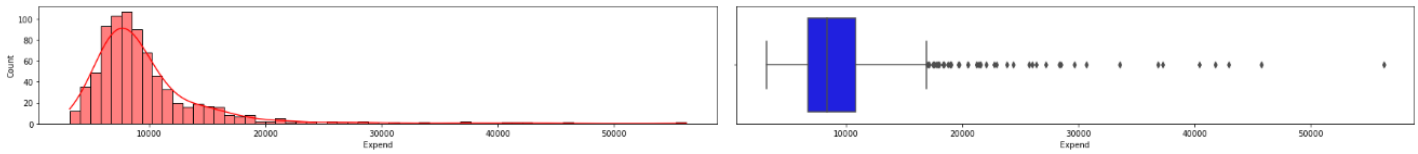
The student faculty ratio data seems fairly normal in distribution and has outliers. The data varies between 12-17 for most of the universities.

15. Perc.alumni:



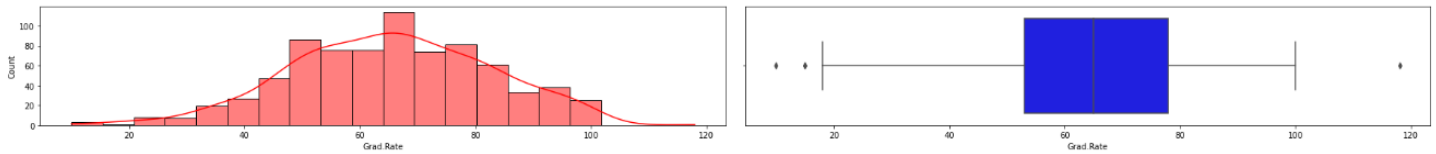
The data is right skewed and has outliers. The percentage of alumni who donate lies between 14%-32% for most of the universities.

16. Expend:



The instructional expenditure per student is fairy normally distributed and has outliers. The expenditure varies between 6000 to 11000 for most of the universities.

17. Grad rate:



The graduation rate data also looks normally distributed, has few outliers and varies between 55-78 for most of the universities.

Multivariate analysis

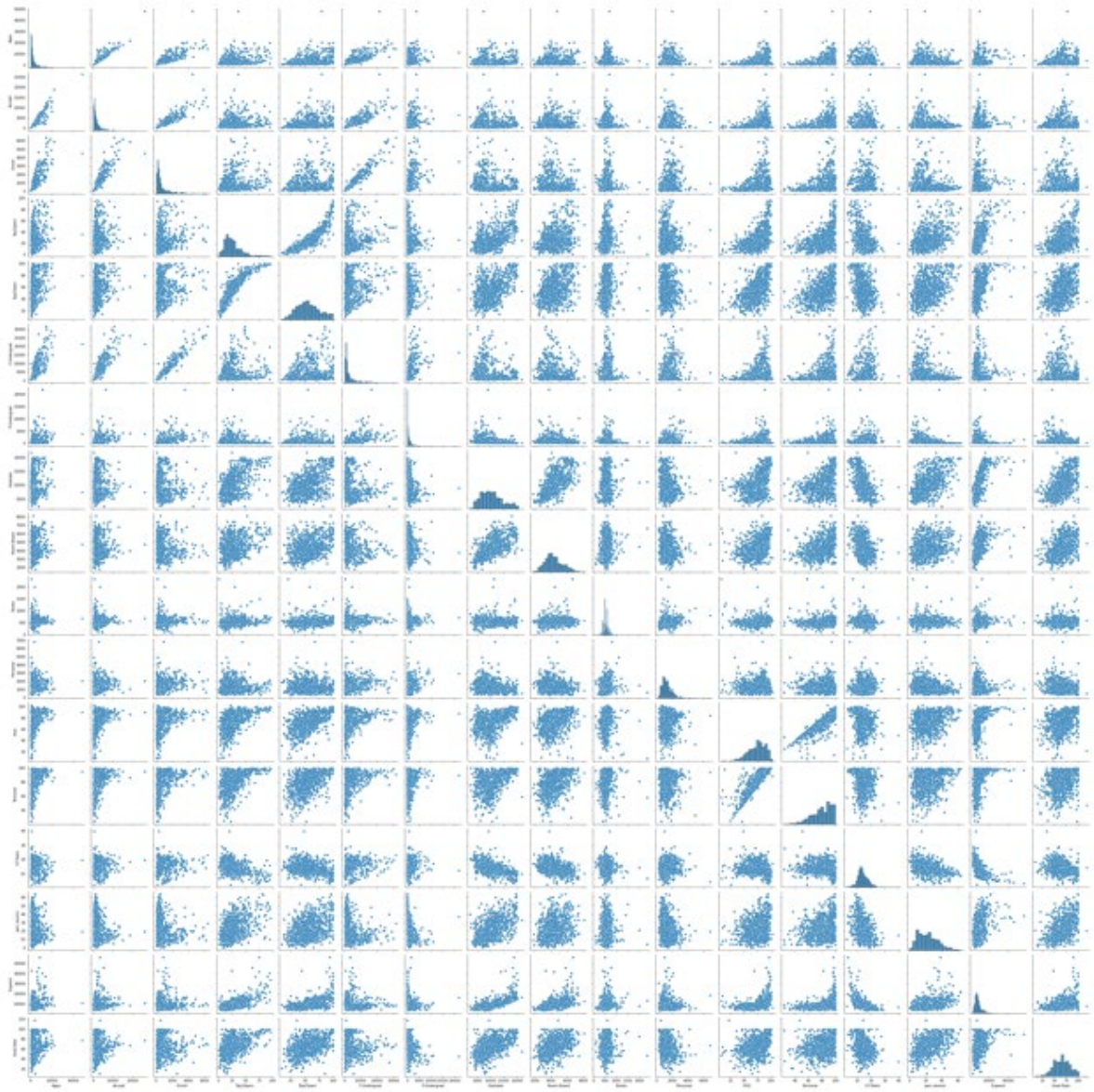
To understand and draw conclusions from Univariate analysis we plot pairplot and heatmap for all the numeric columns and observe the data. Pair plot gives a clear representation of the pairwise correlation between various variables in a data set. Heat map is a value showing chart representing the correlation matrix between different variables. Lighter shade shows higher correlation and darker the shade lower or even negative correlation can be seen between variables.

Inferences:

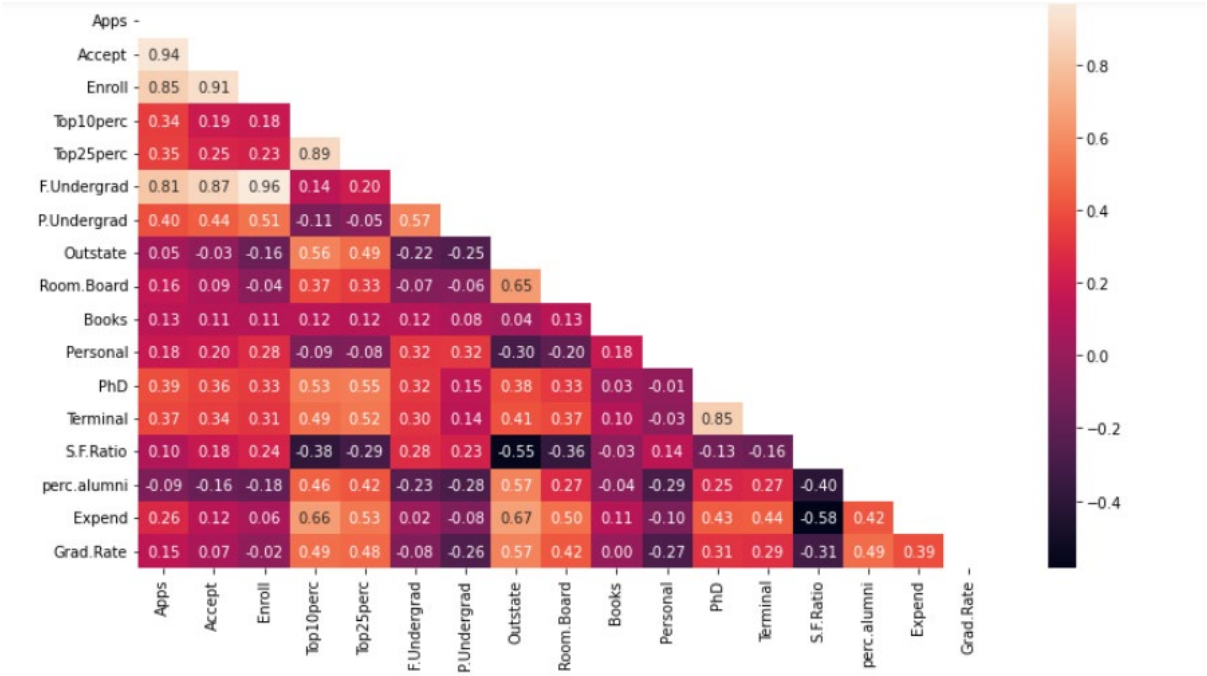
From the pair plot it is evident that Apps shares a very high correlation with Accept, Enroll and Full-time graduate. It also shows a negative correlation with perc. Alumni.

Similarly, S.F Ratio shares a very high negative correlation with Top 10 perc, top 25 perc, outstate, room board and expend.

Pairplot



Heat map



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

Yes, as elaborated in Table 1, scaling is necessary because there are different data corresponding to numbers, percentage, cost, expenditure per student and graduation rate. To pre-process the data, normalize it and bring it into a common scale we need to do scaling.

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

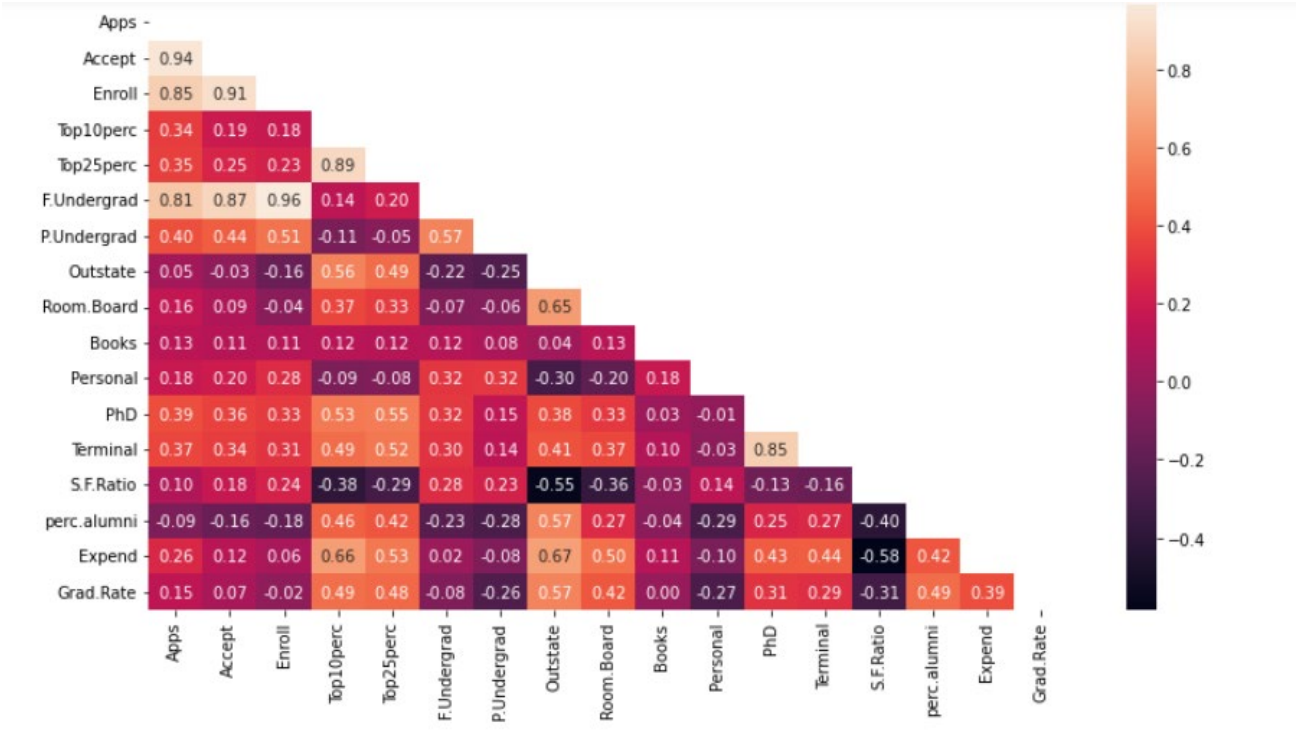
Covariance is measure of how the variables vary together with respect to the data. It gives the direction of linear relationship between the variables. Positive covariance means the variables are directly proportional whereas negative covariance means they share an inverse relationship.

Correlation is simply the strength of the of how the variable is related to one another. Correlation can be highly positive or negative and always lies between 0 & 1.

Covariance

```
cov
[[ 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]
 [ 0.05022367 -0.02578774 -0.1556777   0.5630552   0.49002449 -0.21602002
  -0.25383901  1.00128866  0.65509951  0.03890494 -0.29947232  0.38347594
   0.40850895 -0.55553625  0.56699214  0.6736456   0.57202613]
 [ 0.16515151  0.09101577 -0.04028353  0.37195909  0.33191707 -0.06897917
  -0.06140453  0.65509951  1.00128866  0.12812787 -0.19968518  0.32962651
   0.3750222  -0.36309504  0.27271444  0.50238599  0.42548915]
 [ 0.13272942  0.11367165  0.11285614  0.1190116   0.115676    0.11569867
   0.08130416  0.03890494  0.12812787  1.00128866  0.17952581  0.0269404
   0.10008351 -0.03197042 -0.04025955  0.11255393  0.00106226]
 [ 0.17896117  0.20124767  0.28129148 -0.09343665 -0.08091441  0.31760831
   0.32029384 -0.29947232 -0.19968518  0.17952581  1.00128866 -0.01094989
  -0.03065256  0.13652054 -0.2863366  -0.09801804 -0.26969106]
 [ 0.39120081  0.35621633  0.33189629  0.53251337  0.54656564  0.3187472
   0.14930637  0.38347594  0.32962651  0.0269404  -0.01094989  1.00128866
   0.85068186 -0.13069832  0.24932955  0.43331936  0.30543094]
 [ 0.36996762  0.3380184   0.30867133  0.49176793  0.52542506  0.30040557
   0.14208644  0.40850895  0.3750222   0.10008351 -0.03065256  0.85068186
   1.00128866 -0.16031027  0.26747453  0.43936469  0.28990033]
 [ 0.09575627  0.17645611  0.23757707 -0.38537048 -0.29500852  0.28006379
   0.23283016 -0.55553625 -0.36309504 -0.03197042  0.13652054 -0.13069832
  -0.16031027  1.00128866 -0.4034484  -0.5845844  -0.30710565]
 [-0.09034216 -0.16019604 -0.18102711  0.45607223  0.41840277 -0.22975792
  -0.28115421  0.56699214  0.27271444 -0.04025955 -0.2863366   0.24932955
   0.26747453 -0.4034484   1.00128866  0.41825001  0.49153016]
 [ 0.2599265   0.12487773  0.06425192  0.6617651   0.52812713  0.01867565
  -0.08367612  0.6736456   0.50238599  0.11255393 -0.09801804  0.43331936
   0.43936469 -0.5845844   0.41825001  1.00128866  0.39084571]
 [ 0.14694372  0.06739929 -0.02236983  0.49562711  0.47789622 -0.07887464
  -0.25733218  0.57202613  0.42548915  0.00106226 -0.26969106  0.30543094
   0.28990033 -0.30710565  0.49153016  0.39084571  1.00128866]]
```

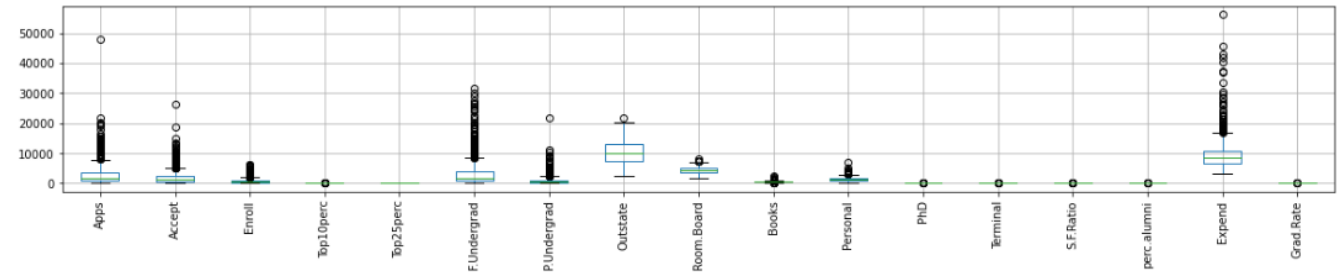
Correlation



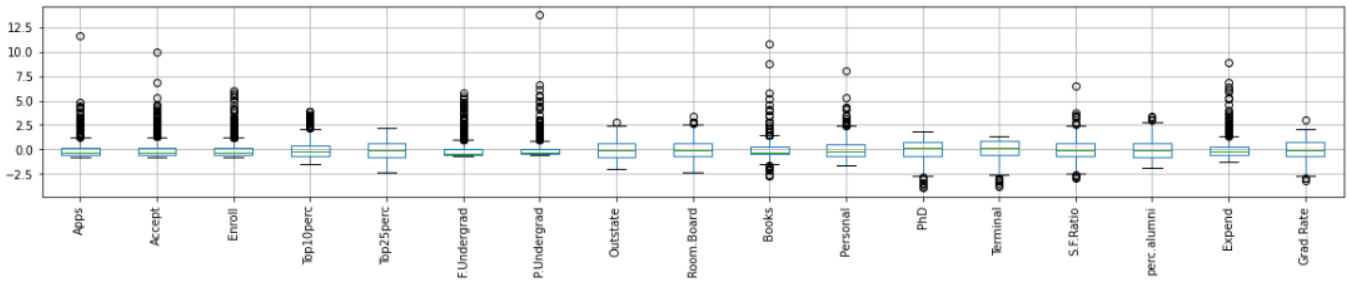
2.4 Check the dataset for outliers before and after scaling. What insight do you derive here?

To check for outliers, we need to plot box plot.

Before scaling:



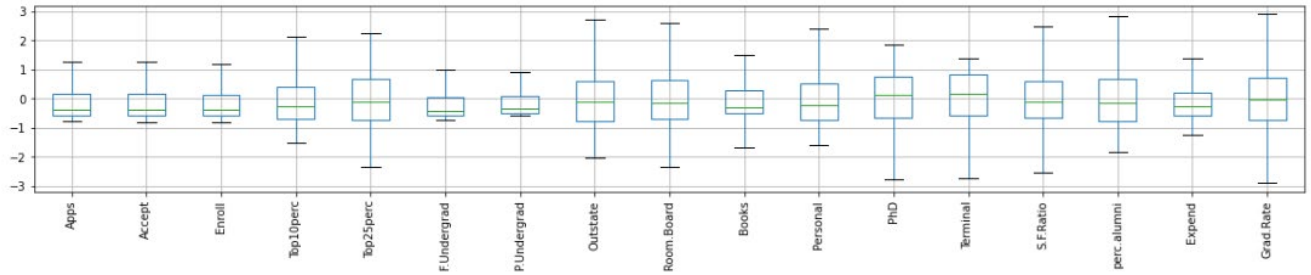
After scaling:



As we can observe the outliers are still present in the data. Scaling does not remove outliers. It only processes the data and bring them into a common scale with mean as zero and standard deviation as 1.

We need to treat these outliers before we can do the PCA analysis.

Post outlier treatment we get the boxplot as follows:



We can clearly see that there are no outliers now and we are ready to proceed for PCA analysis.

2.5 Extract the eigenvalues and eigenvectors. [Using Sklearn PCA Print Both]

Eigenvalues are the special set of scalars associated with the system of linear equations. It is mostly used in matrix equations. 'Eigen' is a German word that means 'proper' or 'characteristic'. Therefore, the term eigenvalue can be termed as characteristic value, characteristic root, proper values or latent roots as well. In simple words, the eigenvalue is a scalar that is used to transform the eigenvector.

The basic equation is $Ax = \lambda x$

The number or scalar value "λ" is an eigenvalue of A.

In Mathematics, an eigenvector corresponds to the real non zero eigenvalues which point in the direction stretched by the transformation whereas eigenvalue is considered as a factor by which it is stretched. In case, if the eigenvalue is negative, the direction of the transformation is negative.

Using PCA fit transform function we calculated the "pca components" and "pca explained variance" to obtain eigen vectors and eigen values. Each eigen vector represents a PC value. Here we have 13 vectors so we will have 13 principal components which we will treat later for dimensionality.

Eigen vectors

```
eigen vectors
[[ 0.0929684   0.06592707  0.03166929  0.33452816  0.36427546  0.01149875
  -0.04622402  0.37830181  0.29777508  0.0401252  -0.10944135  0.31241468
   0.31563577 -0.238936    0.28566756  0.24699418  0.31117828]
 [ 0.32104652  0.3319699   0.35033549  0.06754279  0.13142781  0.32452837
   0.20972858 -0.20665209 -0.07383062  0.13395669  0.29386253  0.30728219
   0.28923834  0.27738993 -0.26160327 -0.02920582 -0.12680266]
 [ 0.06660652  0.07883241  0.01381154 -0.32328505 -0.41399578  0.02193807
   0.1038968   0.2446006   0.65435548  0.06932308  0.02906196  0.01051885
   0.07534077 -0.19726187 -0.35032544  0.14253472 -0.14343185]
 [-0.0129432  -0.03420729 -0.01122623  0.21141739  0.19443136 -0.01744947
  -0.02676078  0.02000679 -0.07736692  0.29066279  0.60616613 -0.21336602
  -0.22034156 -0.50833033 -0.05443422  0.17754101 -0.26409767]
 [ 0.24674827  0.22877472  0.19114851  0.07741651  0.11797053  0.15029942
   0.06989958  0.04640648  0.20589468  0.05440207 -0.01326297 -0.44256735
  -0.48498252  0.128203   -0.08731305 -0.0657708   0.54379453]
 [ 0.00650339  0.02487384  0.03094669 -0.32096376 -0.37686172  0.01698553
   0.00474311  0.05172404 -0.0558972   0.08091173  0.52880595  0.07682269
   0.10434108  0.06989583  0.56468516 -0.05988252  0.34192987]
 [-0.2400326  -0.27288698 -0.26523924  0.09974811  0.18713202 -0.21021832
  -0.08937877 -0.05471221  0.31550426  0.50411092  0.19661216  0.04689621
   0.06899073  0.48926256 -0.1645982  -0.1378192   0.11257333]
 [ 0.13180129  0.12917757  0.12023338 -0.02170441 -0.01531995  0.10021068
   0.0537958   0.02283038  0.11989134  0.48393685 -0.33142515 -0.18837226
  -0.09780965  0.16012666  0.52854784  0.05801938 -0.47485834]
 [ 0.01773119  0.0195307   0.00760842 -0.13379303 -0.18486533  0.01175077
  -0.04181848 -0.17673816 -0.33073591  0.6116081  -0.31893212  0.0983305
   0.12229609 -0.36247736 -0.21127982  0.02433122  0.35725708]
 [ 0.03400089  0.06133927  0.00639234  0.00700079 -0.12749803 -0.02574847
  -0.13366948  0.75697305 -0.42976697  0.10870293  0.03384734 -0.02010444
  -0.06046494  0.31211523 -0.21023249  0.20131984 -0.05772333]
 [-0.03784437 -0.00907891  0.01188448 -0.08516089  0.13131824  0.01666731
   0.08576745  0.0571668  -0.05833278 -0.0378997   0.01497856 -0.70146927
   0.67932403 -0.01040467 -0.03469458 -0.04429777  0.03267103]
 [-0.22445438 -0.17576693 -0.04255122 -0.14064341  0.18030241  0.06116207
   0.8706629   0.21583357 -0.09487917  0.05038066 -0.03722249  0.09315117
  -0.11394885 -0.06491485  0.02561499 -0.12534642  0.03283629]
 [-0.11116423 -0.15058739 -0.01589244  0.29511711 -0.29567976  0.05081468
   0.24738219 -0.27224079 -0.01923638 -0.04789243 -0.00703647 -0.10037346
   0.01716347  0.20799672  0.01145122  0.76406889  0.10963549]
 [ 0.00914552 -0.00281244 -0.02985794 -0.69375696  0.511526    0.01136756
  -0.14019587 -0.09997777 -0.02476745 -0.02826868  0.0047994   0.0434578
  -0.08863951  0.07241455 -0.00883861  0.45737401  0.02419997]
 [ 0.5569348   0.26755069 -0.49315933  0.00901027 -0.00267915 -0.55093616
   0.23891433 -0.07289772 -0.06810512 -0.01670698  0.02161283 -0.00454371
   0.02707478  0.02278083  0.01480452  0.04412393 -0.01406045]
 [ 0.59269472 -0.70710813 -0.13493584 -0.02159112 -0.01682207  0.35140689
  -0.03885668  0.04554559 -0.01161555 -0.00902118 -0.0047103  -0.00548016
   0.00898686 -0.01073745 -0.00137953 -0.04641723 -0.00788293]
 [-0.14346037  0.32336365 -0.69930928  0.0310452  -0.00890279  0.61814444
  -0.04589764  0.00367566 -0.00556533 -0.00416292  0.00630645 -0.00804
  -0.0048878  -0.00662521  0.01383698 -0.01193343  0.00123347]]
```

Eigen Values:

```
eigen_values
[4.75579369 2.3800885  0.88497491 0.81453646 0.72423975 0.52688069
 0.47958062 0.41127635 0.36620193 0.23942458 0.12943793 0.09751277
 0.08189987 0.06059116 0.03582106 0.01435481 0.00793972]
```

Dividing the eigen values by the number of components gives us the pca explained variance ratio. It gives us a very clear understanding of the strength of each principal component in the data.

For e.g., PC1 alone captures 39.5 % of the variability of the data, PC2 captures 19.81% and so on.

```
exp_var
[0.39596786 0.19816641 0.0736831  0.06781839 0.06030027 0.04386814
 0.03992993 0.03424291 0.03049001 0.01993451 0.01077702 0.00811892
 0.00681899 0.00504483 0.00298246 0.00119518 0.00066106]
```

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

The PCA analysis was performed and the principal components were exported into a 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	0.092968	0.321047	0.066607	-0.012943	0.246748	0.006503	-0.240033	0.131801	0.017731	0.034001	-0.037844	-0.224454	-0.111164	0.009146	0.556935	0.592695	-0.143460
1	0.065927	0.331970	0.078832	-0.034207	0.228775	0.024874	-0.272887	0.129178	0.019531	0.061339	-0.009079	-0.175767	-0.150587	-0.002812	0.267551	-0.707108	0.323364
2	0.031669	0.350335	0.013812	-0.011226	0.191149	0.030947	-0.265239	0.120233	0.007608	0.006392	0.011884	-0.042551	-0.015892	-0.029858	-0.493159	-0.134936	-0.699309
3	0.334528	0.067543	-0.323285	0.211417	0.077417	-0.320964	0.099748	-0.021704	-0.133793	0.007001	-0.085161	-0.140643	0.295117	-0.693757	0.009010	-0.021591	0.031045
4	0.364275	0.131428	-0.413996	0.194431	0.117971	-0.376862	0.187132	-0.015320	-0.184865	-0.127498	0.131318	0.180302	-0.295680	0.511526	-0.002679	-0.016822	-0.008903

2.7 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 first PC in terms of eigen vectors up to two places of decimals is obtained as follows:

```
The explicit form of the first PC upto two places of decimals are
[ 0.09  0.07  0.03  0.33  0.36  0.01 -0.05  0.38  0.3   0.04 -0.11  0.31
  0.32 -0.24  0.29  0.25  0.31]
```

The linear equation of PC in terms of eigenvectors and corresponding features is shown below:

```
The linear equation of PC in terms of eigenvectors and corresponding features is given as
0.09 Apps+0.07 Accept+0.03 Enroll+0.33 Top10perc+0.36 Top25perc+0.01 F.Undergrad+-0.05 P.Undergrad+0.38 Outstate+0.3 Room.Board
+0.04 Books+-0.11 Personal+0.31 PhD+0.32 Terminal+-0.24 S.F.Ratio+0.29 perc.alumni+0.25 Expend+0.31 Grad.Rate+
```

2.8 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?

```
cumulative values of the eigenvalues
[0.39596786 0.59413427 0.66781737 0.73563576 0.79593603 0.83980417
 0.8797341  0.91397701 0.94446702 0.96440153 0.97517855 0.98329747
 0.99011646 0.99516129 0.99814376 0.99933894 1.        ]
```

To decide on the optimum number of principal components we need to look at the percentage of information covered by the components and also the incremental value between the components. As we can see 83.98 % of information is covered up to PC6 and beyond that the incremental value is less than 5%, so we can take it as optimum number. If we try to take more components it will increase the dimensionality without much value addition to the information required.

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

PCA helps us in understanding the distribution of each and every ingredient of the principal components extracted out from the data. As we have seen post scaling and treating the data, we perform a principal component analysis which in essence gives us the principal components the whole data set is made up of. To further understand we dive deep, reduce the dimensionality and try to disintegrate each principal component and see the contents it's made up of and how much they add value to each principal component. We also ensure that post our analysis the variables don't share any correlation with each other which would mean we have eliminated all the noise from our information which would have otherwise occupied mathematical space between the dimensions of principal components.

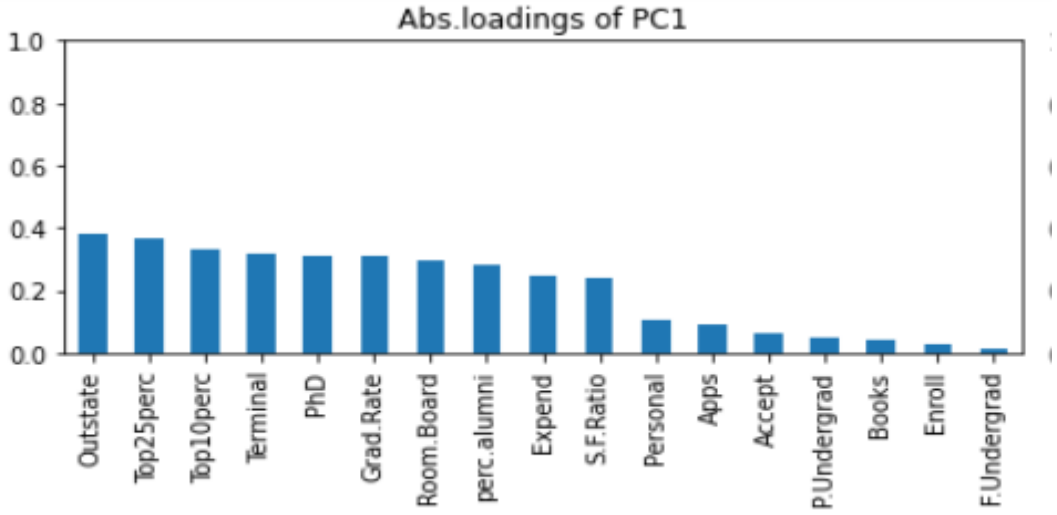
To elaborate further let us first see the principal component weights in the dataset.

```
The eigen values of the final optimum number of principal components are
[0.39596786 0.19816641 0.0736831  0.06781839 0.06030027 0.04386814]
```

From the above solution we can see that PC1 adds 39.59% value to the entire dataset. Similarly, PC2 adds a value of 19.81 %, PC3 7.36%, PC4 6.78%, PC5 6.03% and PC6 brings a value addition of 4.38%. The reason for selecting this optimum number is explained in section 2.8 of the business report.

Once the principal components are identified we can dive deeper into our analysis and observed the individual sub-components these PCs are made up of.

Let us analyse PC1:



PC1 carries a maximum weight of 39.59% of the data set. From the above plot we can see that the most important contributing factors for PC1 are outstate, Top 25 perc, Top 10 perc, Terminal, PhD, Grad rate. This business model is the most successful model and attracts the highest number of students with an excellent academic record.

In terms of students, it would mean that a good university will be a choice of most of the students from outstation and who have performed academically well in their high school and are among top 10 % and top 25 % students. In other words, these universities will set the highest standards for their admission criteria and selects only the best of the best students.

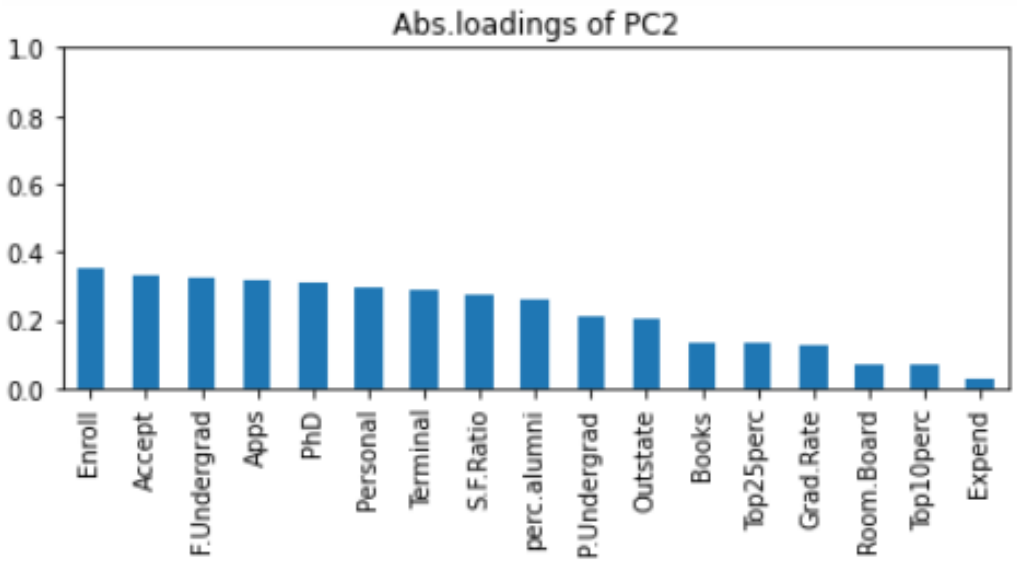
The students will also choose universities which have a good number of Terminal degree holder and PhD's in their faculty and same can be observed to contribute highly from the above plot. University management team needs to ensure they have the best faculty to make their university the obvious choice for the students applying for higher studies.

Moreover, the graduation rate will also determine the decision to enroll in the university as the students expects to complete their degree and this percentage gives a fair amount of idea on how the university will assist a candidate to successfully complete his/ her studies and pass out of the university with a degree in the defined time frame of the course enrolled.

Last but not the least this university has very low expenses in terms of personal and books making it easier on the pocket of students and help them in enrolling for the university.

The above parameters make an excellent business decisive factor for the university management to ensure they attract a good number of students from across the globe and in turn make it a profitable venture.

Let us analyse PC2 and see what similarities and differences we can conclude with respect to PC1.



Let us remember that PC2 only carries around 19.81 % weight of the complete data set. However, in those 19.81 % we can see a significant contribution is from Enroll, Accept, F. Undergraduate, Apps, PhD, Terminal.

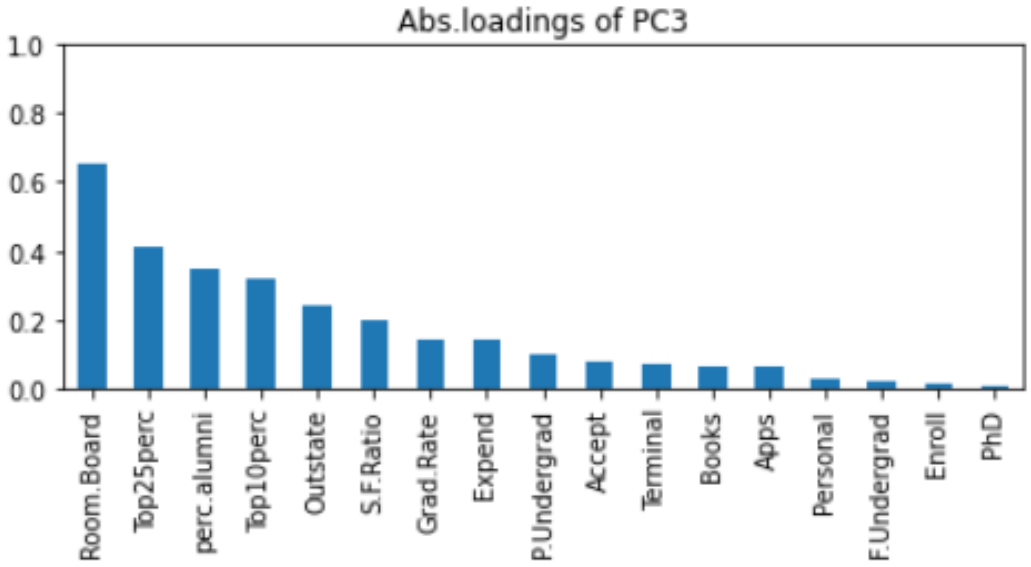
We can again conclude that maximum students will apply and enroll in universities having good faculty. As PC2 doesn't offer a very strict regime of taking only top performing students we can see that maximum applications and enrolment happen in these universities.

Personal expenses also seem to contribute in this kind of universities and university management can work on these areas to improve their business significantly.

PC1 and PC2 can be regarded as the top universities which are the best in demand colleges for students and they both have very high standards faculty in common.

From overall business perspective Universities need to continuously maintain and improvise on the overall standards of their faculty with changing time and technological advances.

Let us continue our analysis for PC3.

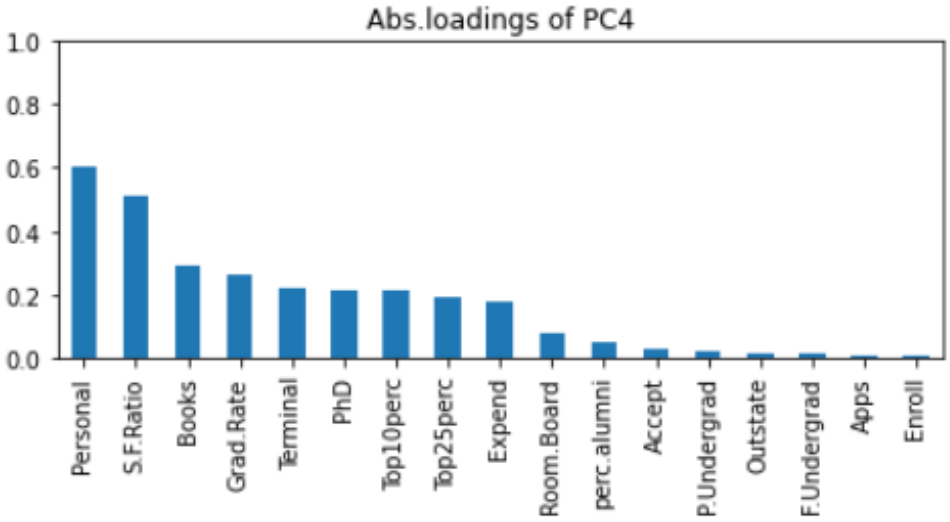


Here we have a sightly different picture with respect to PC1 and PC2.

PC3 contributes a total of 7.36 % weight in the entire data set. As seen cost of room and board plays an important role in budgeting the entire plan of study and therefore a lot of top 25 % of students tend to choose this as their choice of university along with a significant number of top 10 % of high school graduates. This business model also offers a decent graduation rate and S.F ratio which will be a deterministic factor for students opting to choose between various universities.

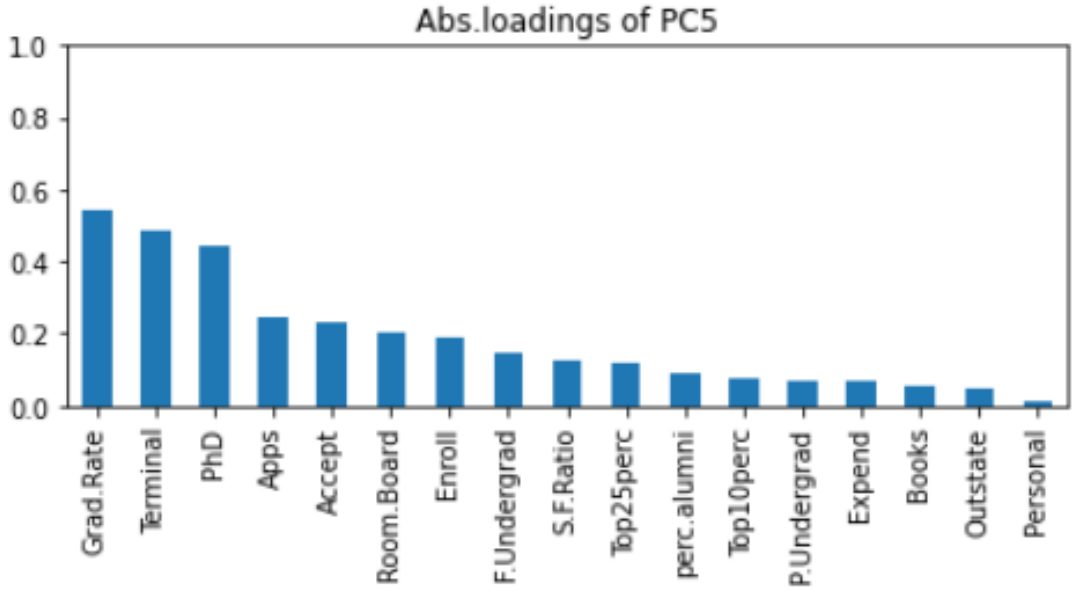
However, overall faculty standard is very low compared to PC1 and PC2 and we can confidently conclude that the education standard will be considerably low and students choose this university mostly out of their budget constraints.

Analysis of PC4



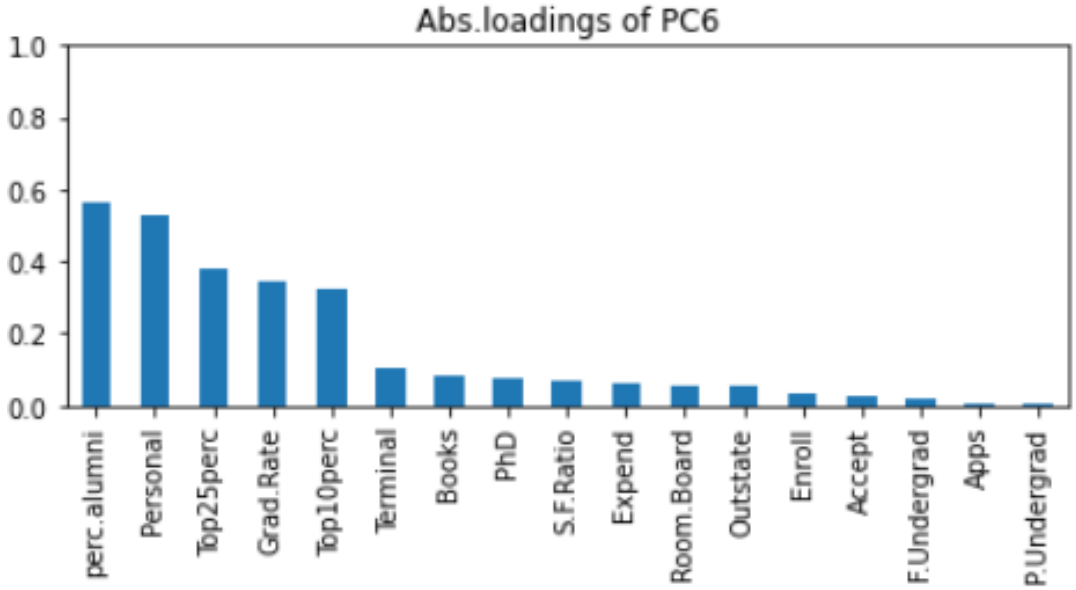
PC4 carries a weight of only 6.78 % and we can easily see that the application and enroll values are the minimum in this universities. Inspite of having a decent faculty in terms of PhD and terminal degree, a good S.F Ratio and graduation rate, the reason for such low admission can be attributed to personal cost being very high either due to location or due to the fee structure of the university.

Analysis of PC5



PC5 contributes around 6.03% and can be seen as a very good business model with good graduation rate, decent faculty and hence a decent number of applications and enrolment. However, we should observe that this university is the choice for students who are below top 25% in high school or average performers. But at the same time this university seems to be an ideal choice for those who could not get enrolled into the top colleges due to stringent admission criteria as it offers a good quality education with lowest personal, books and expend costs.

Analysis of PC6



PC6 carries the least weight of only 4.38 % and shows a very less application, acceptance and enrolment record. This university model relies heavily on the donation of alumni and again similar to PC4 has a very high personal expense sub-component. From comparison of PC4 and PC6 it is evident that students don't wish to enroll in universities with a high personal expense component and under those circumstances PC5 model suits them best.

Overall conclusion:

The overall conclusion is to have a business model with excellent faculty, good intake and acceptance ratio, low personal and other expenses and a decent graduation rate. The university may choose to have very high standards of admission criteria as depicted in PC1 and PC2 or may chose to go moderate way as reflected in PC5. However, the expense of books, room & board and personal needs to be kept under check as it hugely impacts the overall business. Hence, higher management of universities should focus on improving the education standards by hiring the best and most experienced faculty and at the same time keep in mind the budget constraints of the students so that they optimize the admission process and manage to secure maximum number of intakes every academic year.

End of problem set 2-Education Post 12th standard data.