Business Report

Project –SMDM

Created by Amit Jain

Table of Contents

List of Figure

1. SALARY DATA ANALYSIS.

Problem Statement:

Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] 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.

Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

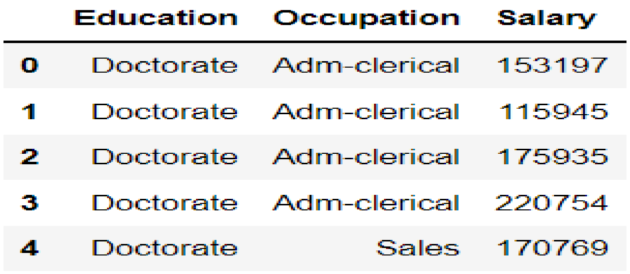
Assumption:

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Step of understanding the data:

- Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:



	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

Insight of the data set:

1. Data Have 3 fields, Education, Occupation and Salary
2. Education have 3 categories: Doctorate, Bachelors and HS-Grad
3. Occupation have 4 Categories: Adm-clerical', ' Sales', ' Prof-specialty', ' Exec-managerial'
4. We need to analyze effects of Education and Occupation Categories over Salary.
5. Shape of the data set is 40 Rows and 3 Columns
6. There are no null values in any Column values
7. Salary ranging between 50103 and 260151 is a good sign, and seems all data is normal in nature, no bad data

1A-1 : State the null and the alternate hypothesis for conducting one-way ANOVA

Hypothesis for Education Qualification

Null Hypothesis Ho : Mean Salary is same across all Education Qualification

Alternate Hypothesis Ha: Mean salary is different for at least one level of Education Qualification.

Hypothesis for Occupation Qualification

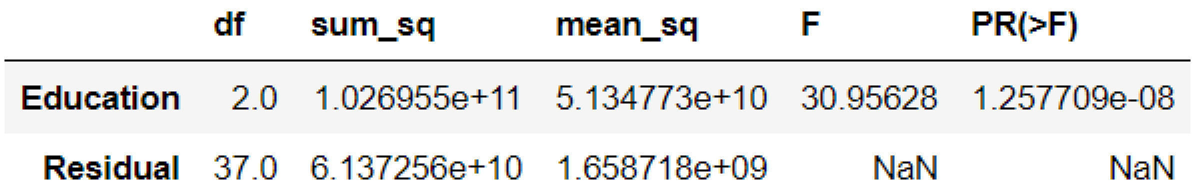
Null Hypothesis Ho : Mean Salary is same accross all Occupation Level

Alternate Hypothesis Ha: Mean salary is different for at least one level of Occupation Level

1A-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.

For performing One way ANOVA Test we have used anova_lm module from statsmodels.stats.anova python Library

After performing one way ANOVA test, following P value was generated:



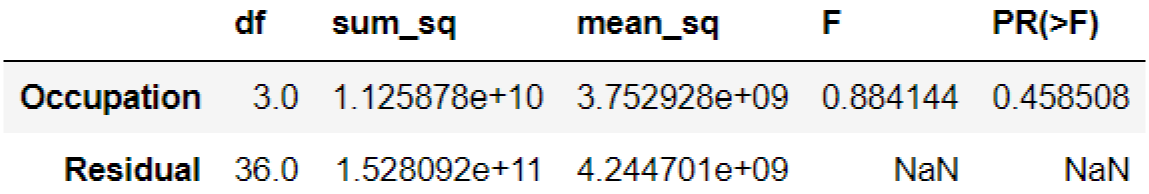
	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	30.95628	1.257709e-08
Residual	37.0	6.137256e+10	1.658718e+09	NaN	NaN

Since the P value is less than alpha value 0.05, so we REJECT the NULL Hypothesis, which says, that mean Salary is same across all Education Qualification level, which means there are differences in mean salary based on Education

1A-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.

For performing One way ANOVA Test we have used anova_lm module from statsmodels.stats.anova python Library

After performing one way ANOVA test, following P value was generated:



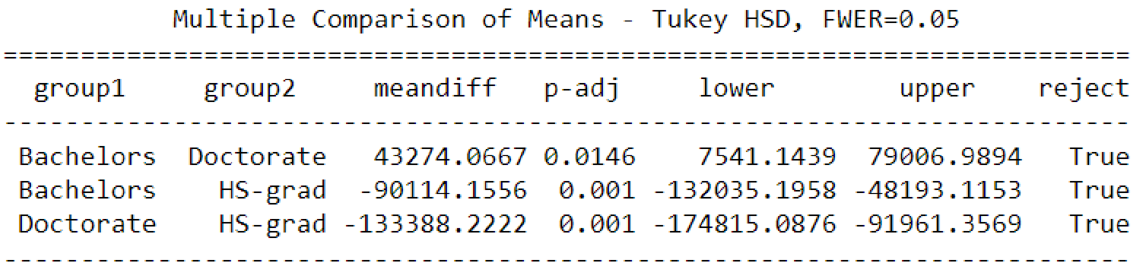
	df	sum_sq	mean_sq	F	PR(>F)
Occupation	3.0	1.125878e+10	3.752928e+09	0.884144	0.458508
Residual	36.0	1.528092e+11	4.244701e+09	NaN	NaN

Since the P value is Greater than alpha value 0.05, so we FAIL TO REJECT the NULL Hypothesis, which says, that Mean Salary is same across all Occupation level Which means , Mean salary is same for all Occupation categories

1A-4 : If the null hypothesis is rejected in either (2) or in (3), find out which class means are significantly different. Interpret the result. (Non-Graded)

As per results from Problem 1A-2 and 1A-3, that There was significant Difference between Mean salary for Educational Categories, so Let’s compare the mean for Educations levels. Education have 3 categories with in it: ' Doctorate', ' Bachelors', ' HS-grad'

For analyzing the mean Salary , based on multiple categories of Education, we would use MultiComparison module from statsmodels.stats.multicomp Python Library, and after analysis, Following P values will be driven, for all categories of Education Column:



```
           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
--------------------------------------------------------------------------
```

From the Output of the MultiCompare Test result we observe that the P values of all the Category of Education Levels are lower than 0.05, So we can say that Mean Salary would be different for all categories of Education levels.

Now check for Mean Salaries for the Given sample data set for all 3 Education levels:

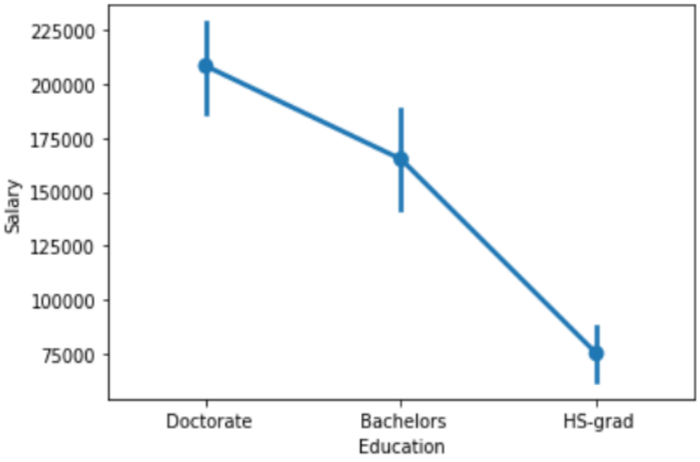


Fig:1 Point Plot

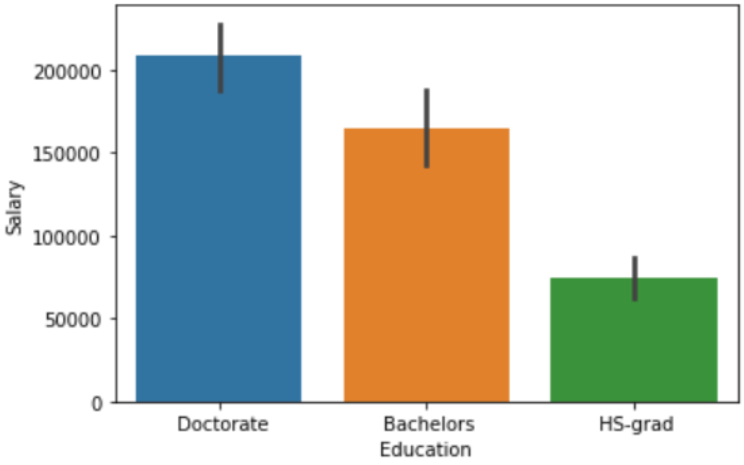
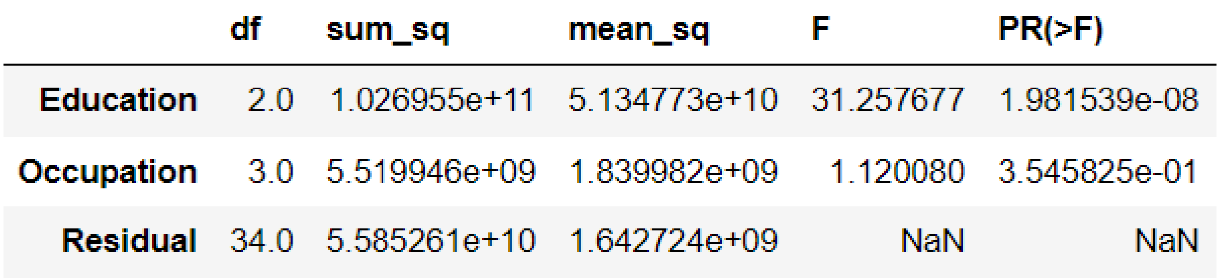


Fig:2 Count Plot

Based on Point Plot and Bar Plot , we can clearly see that mean salary for Doctorate Educational level will be Higher than all and HS-grad Mean salary should be least in all Educational levels.

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.[hint: use the ‘point plot’ function from the ‘seaborn’ function]

Since this problem statement is seeking comparative analysis of both treatments (Education and Occupation) Over salary field, So we would use multi compression and check the P-values for Combined effect of education and Occupation over salary :



	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	31.257677	1.981539e-08
Occupation	3.0	5.519946e+09	1.839982e+09	1.120080	3.545825e-01
Residual	34.0	5.585261e+10	1.642724e+09	NaN	NaN

The p-value in the treatments are Less than α (0.05), so with 95% confident level we can say that Mean Salary are not same for Combined effect of Education and Occupation, and there will be significant differences in Salaries based on combination of Education and Occupation .

Lets Check the interactions using Point Plot method

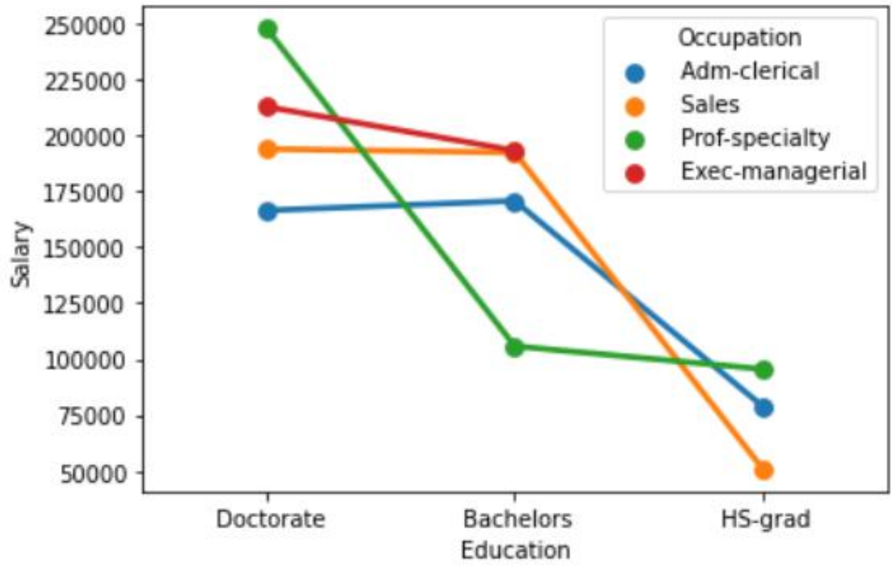


Fig 3: Intersection Plot

Check for the count plot too for each treatment:

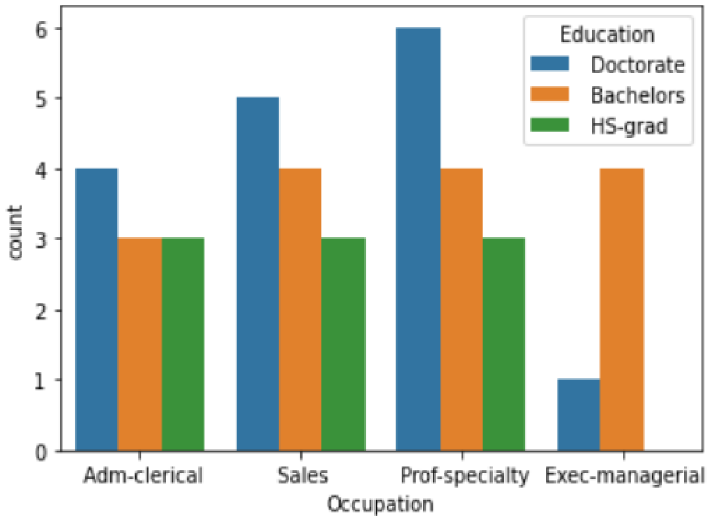


Fig 4: Count Plot

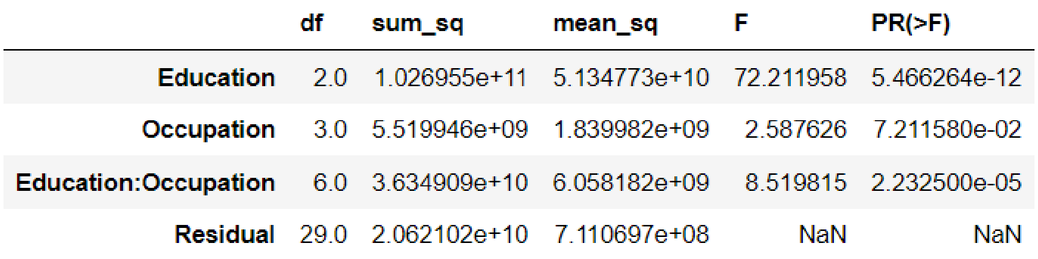
Problem 1B - 2: Perform a two-way ANOVA based on Salary with respect to both Education and Occupation (along with their interaction Education*Occupation). State the null and alternative hypotheses and state your results. How will you interpret this result?

As we confirmed from the interaction plot that there is interactions between education and Occupation levels, So we would conduct a two way anova to check the interaction between Education and occupation.

Null Hypothesis Ho : There is no interaction between Education and Occupation

Alternate Hypothesis Ha : There is an interaction between Education and Occupation

We will use the ANOVA test for Interaction effect of both treatment over Salary . Generate the P value metrics



	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
Occupation	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
Education: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

We will now check Interaction Plot method

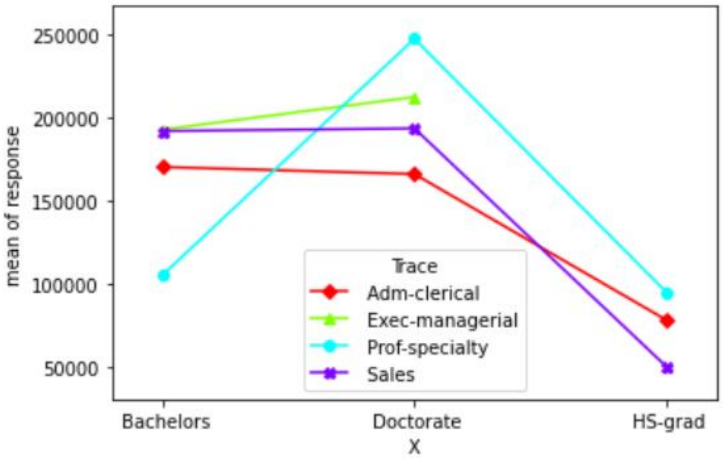


Fig 5: Interaction Plot

We can see that there is some sort of interaction between the two treatments.

Since P value is less than 0.05, So we reject the null hypothesis , which says that There is no interaction between Education and Occupations over Salary and with 95% significance level there is an interaction between Education and Occupation. which means the mean salary is affected by an interaction of Education and Occupation. From the interaction plot, we can say that there is interaction between Education and Occupation levels

Problem 1B - 3: Explain the business implications of performing ANOVA for this particular case study.

By conducting One way ANOVA test : We analyzed Mean salary for any specific Category, So there is no difference in Mean salary for any occupation levels, whereas there are differences in Mean Salary for Educational Levels.

Highest earning Salary have done the Doctorate and lowest earning group is High school graduate.

When we Test for Combined effect of Education and Occupation, we did 2-way ANOVA test, and found that Education and Occupation together are affecting the salary as well and there are differences in mean salary for the combinations of Both categories. and visualized it through intersection plot and observed that there is a interaction between the two categories.

Problem 2: College and Student data analysis

Problem Statement:

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. The data dictionary of the 'Education – Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

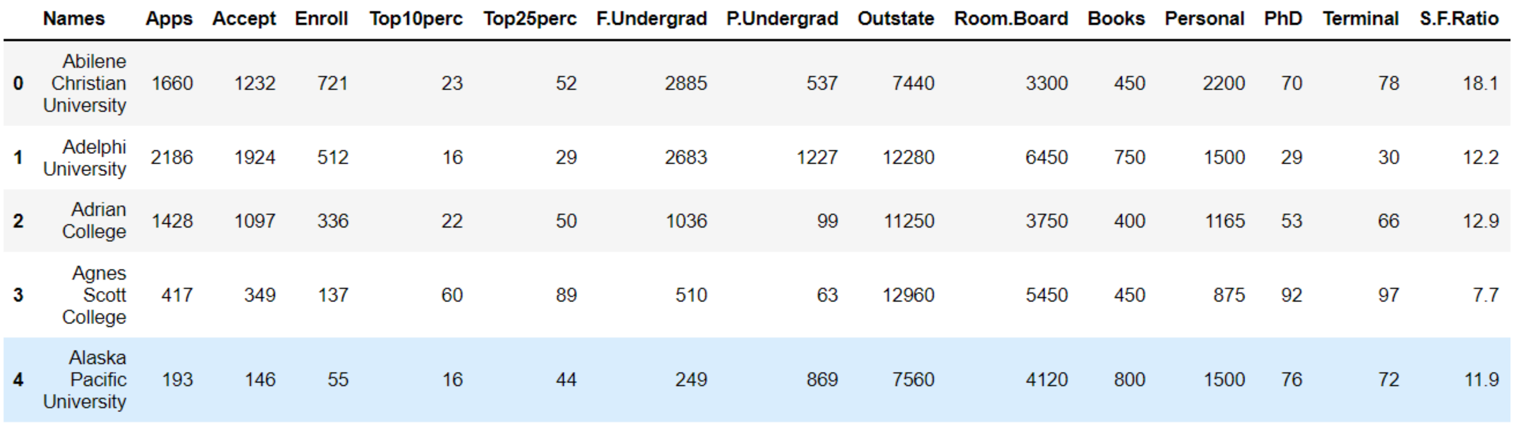
Assumption:

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Step of understanding the data:

- Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:



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

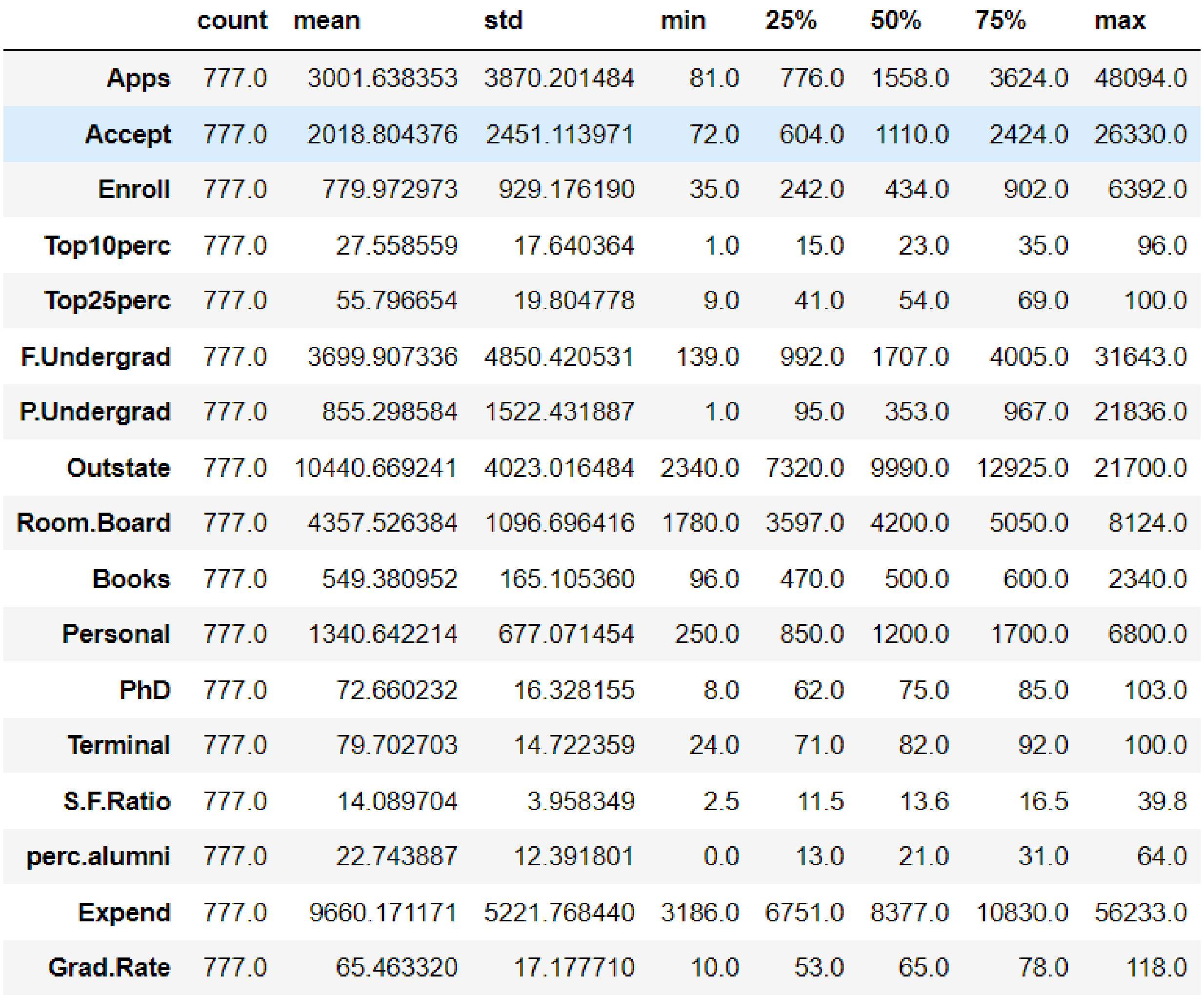
Insights of the data set:

1. As per the data dictionary given, following are the details of all the Columns in the source data set:

    1.1. Names: Names of various university and colleges
    1.2. Apps: Number of applications received
    1.3. Accept: Number of applications accepted
    1.4. Enroll: Number of new students enrolled
    1.5. Top10perc: Percentage of new students from top 10% of Higher Secondary class
    1.6. Top25perc: Percentage of new students from top 25% of Higher Secondary class
    1.7. F.Undergrad: Number of full-time undergraduate students
    1.8. P.Undergrad: Number of part-time undergraduate students
    1.9. Outstate: Number of students for whom the particular college or university is Out-of-state tuition
    1.10. Room.Board: Cost of Room and board
    1.11. Books: Estimated book costs for a student
    1.12. Personal: Estimated personal spending for a student
    1.13. PhD: Percentage of faculties with Ph.D.'s
    1.14. Terminal: Percentage of faculties with terminal degree
    1.15. S.F.Ratio: Student/faculty ratio
    1.16. perc.alumni: Percentage of alumni who donate
    1.17. Expend: The Instructional expenditure per student
    1.18. Grad.Rate: Graduation rate

2. There are total 17 Numeric fields in the data set
3. Data set have one column “Names” , which is of Object data type and it can be ignored for the data analysis
4. There are no null values in the data set
5. S.F.Ratio is float64

6. All other columns are in int64 data type,
7. Describe the data set:



	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

8. When Compare to Describe Output with Given data Dictionary Given , Following are implications:

   8.1 "Names" are the Object data type, which can be Dropped for performing EDA

   8.2 "Apps", "Accepts" , "Enroll" , "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board ", "Books", "Personal" seems normal Columns.

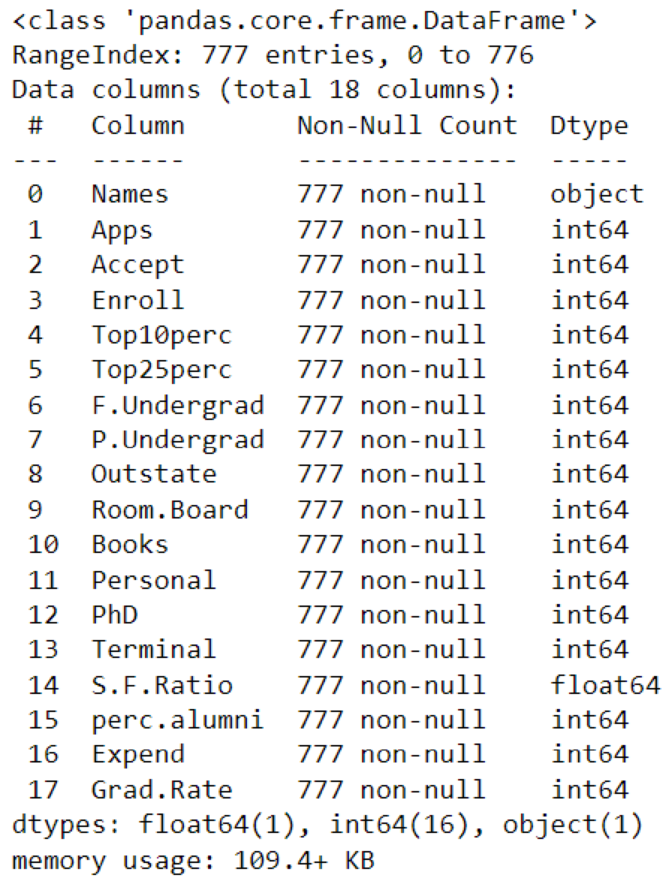
   8.3 "PhD" , "Terminal" and "Grad.Rate" are the percentage values as per data dictionary, but it has maximum values 100 or more than 100, which is not normal, need to change

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

univariate Analysis

in Univariate analysis we need to check Individual columns and its data set distribution , Outlier s, missing values , Duplicates, Special characters etc. everything should be tested against each individual column.

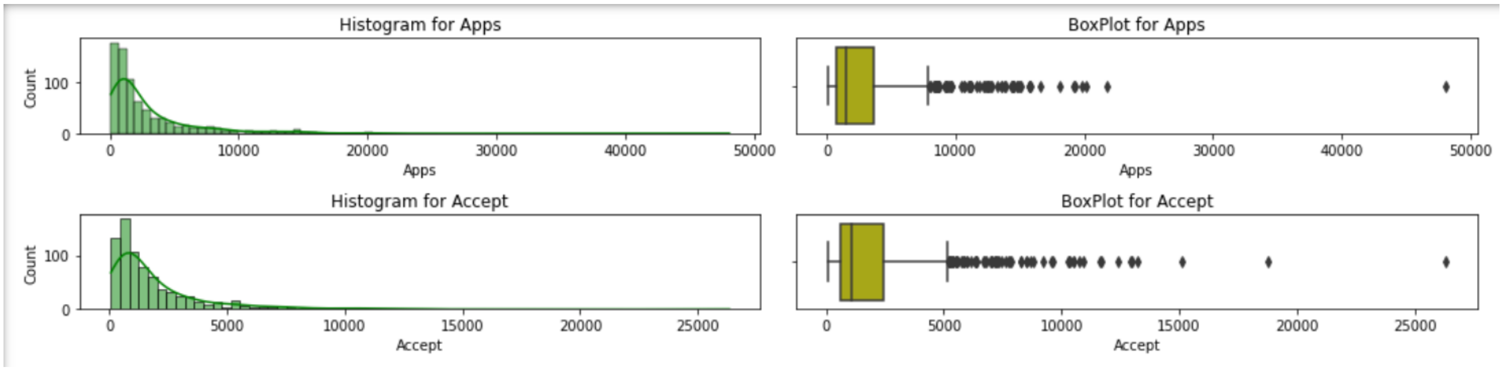
Check for missing values :

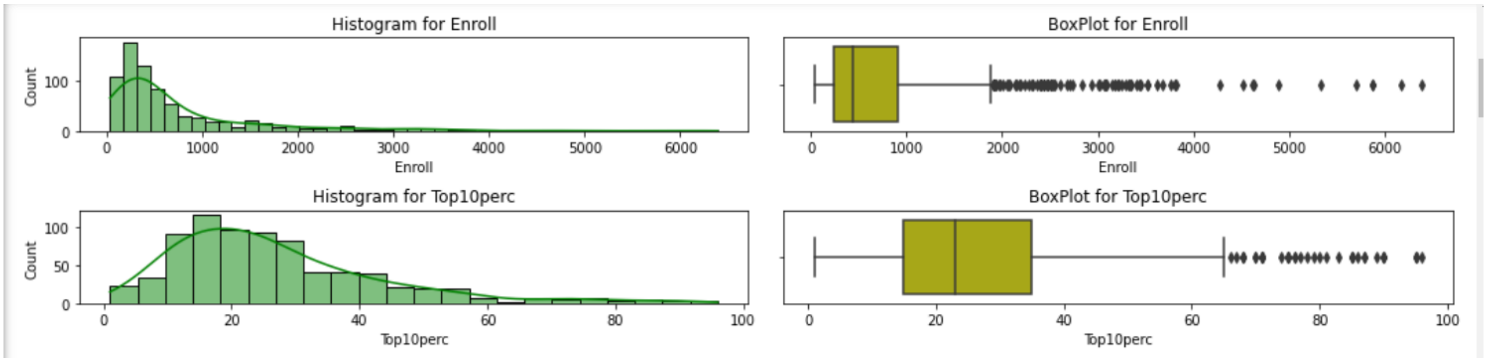


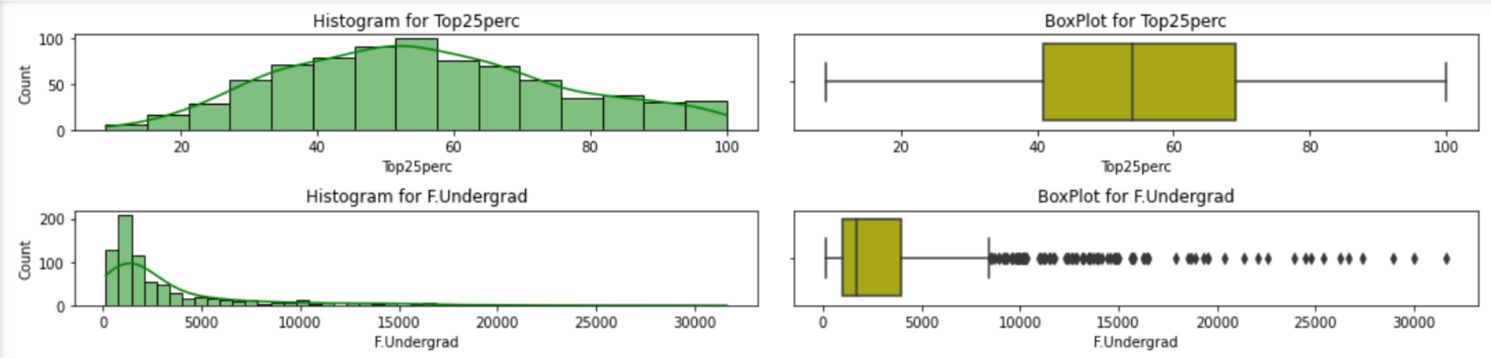
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 777 entries, 0 to 776
Data columns (total 18 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   Names        777 non-null    object 
 1   Apps         777 non-null    int64  
 2   Accept       777 non-null    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
```

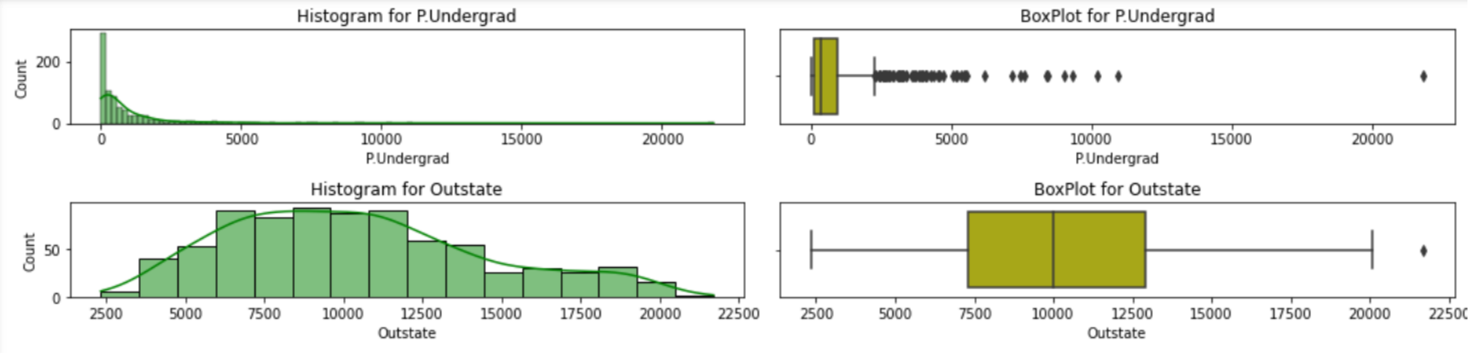
So No Null values in any of the field.

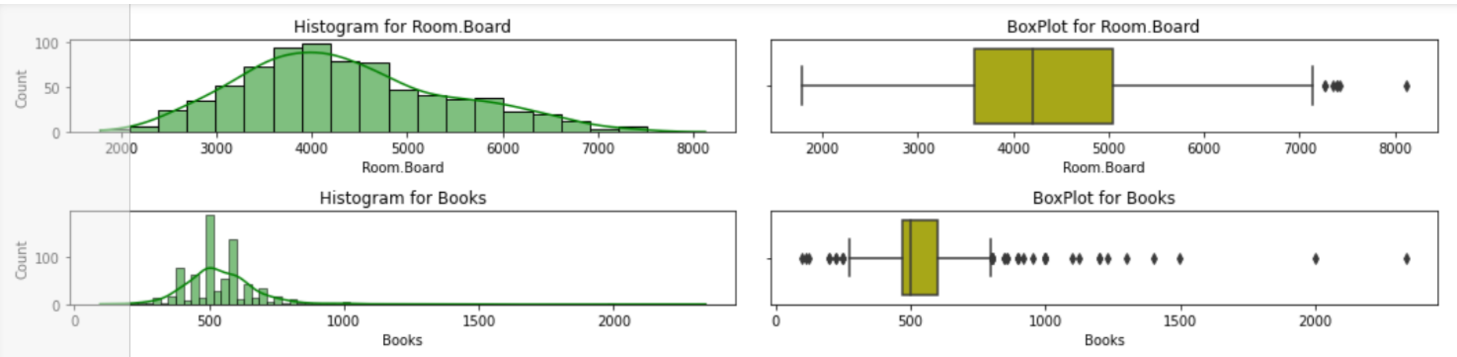
Check for Distribution and Outliers in the data set:

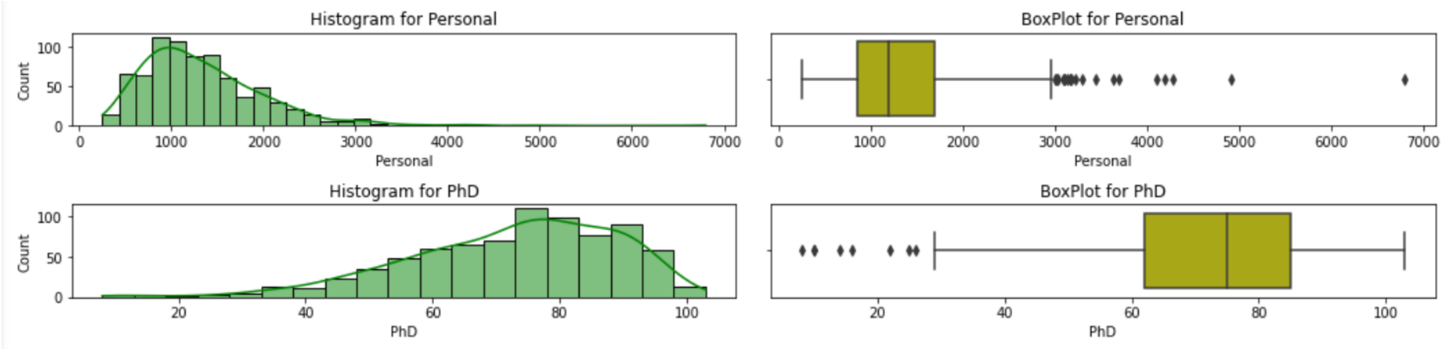


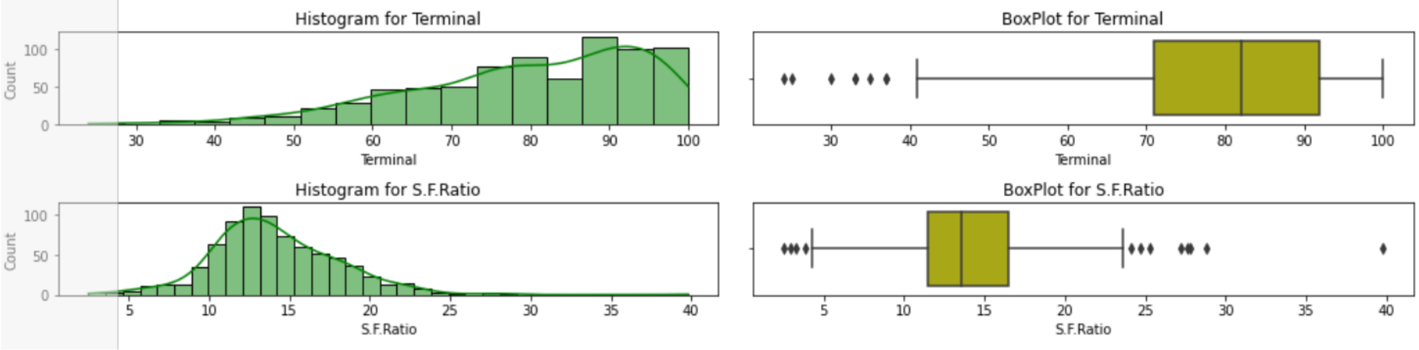


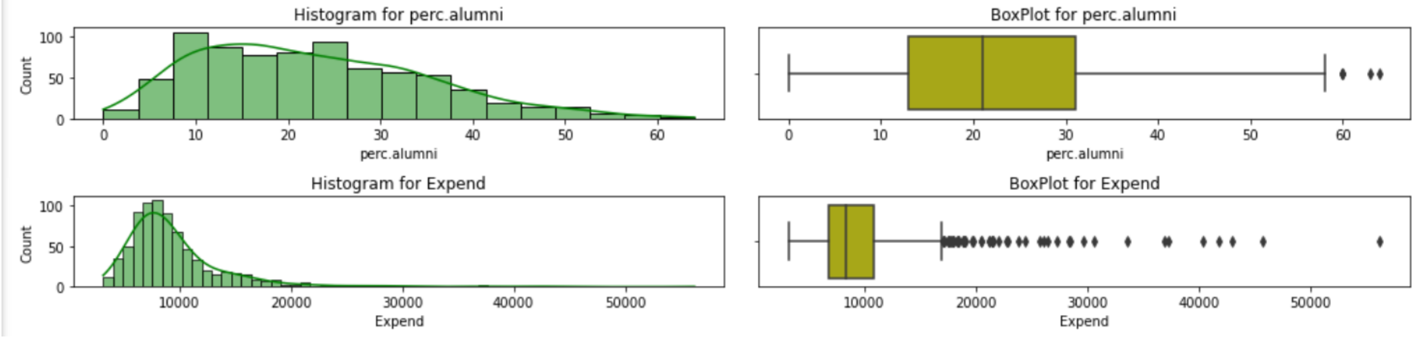





Histogram for Room.Board
Count
100
50
0
2000
3000
4000
5000
6000
7000
8000
Room.Board
BoxPlot for Room.Board
2000
3000
4000
5000
6000
7000
8000
Room.Board
Histogram for Books
Count
100
0
0
500
1000
1500
2000
Books
BoxPlot for Books
0
500
1000
1500
2000
Books


Histogram for Personal
Count
100
50
0
0
1000
2000
3000
4000
5000
6000
7000
Personal
BoxPlot for Personal
0
1000
2000
3000
4000
5000
6000
7000
Personal
Histogram for PhD
Count
100
50
0
20
40
60
80
100
PhD
BoxPlot for PhD
20
40
60
80
100
PhD


Histogram for Terminal
Count
100
50
0
30
40
50
60
70
80
90
100
Terminal
BoxPlot for Terminal
30
40
50
60
70
80
90
100
Terminal
Histogram for S.F.Ratio
Count
100
50
0
5
10
15
20
25
30
35
40
S.F.Ratio
BoxPlot for S.F.Ratio
5
10
15
20
25
30
35
40
S.F.Ratio


Histogram for perc.alumni
Count
100
50
0
0
10
20
30
40
50
60
perc.alumni
BoxPlot for perc.alumni
0
10
20
30
40
50
60
perc.alumni
Histogram for Expend
Count
100
50
0
10000
20000
30000
40000
50000
Expend
BoxPlot for Expend
10000
20000
30000
40000
50000
Expend

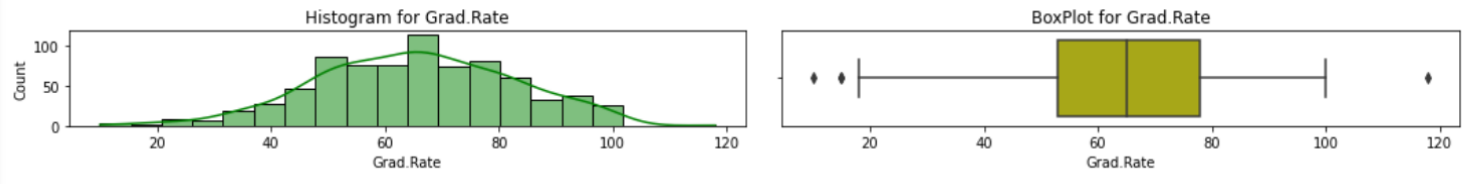
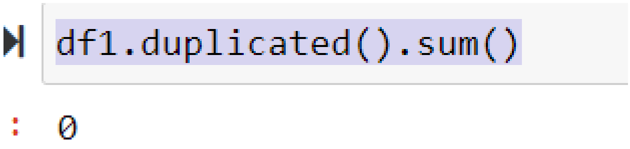


Fig 6: Distribution and Box Plot

Insights: Looking at Box Plot, we can say that : Except Top25perc , remaining all other fields have outliers in upper values or in lower values.

Check for Duplicates:



```
df1.duplicated().sum()
```

```
: 0
```

We don’t have any Duplicate values in the data set.

Multivariate Analysis

In Multivariate analysis , we analyze the combined relations between different fields, and we can consider 2 or more columns and their effects on each other . Directions and strength of the relationships between them. We use Covariance and correlation mechanism for this comparison.

correlations Metrics: We will us heat map for checking the co relation between fields:

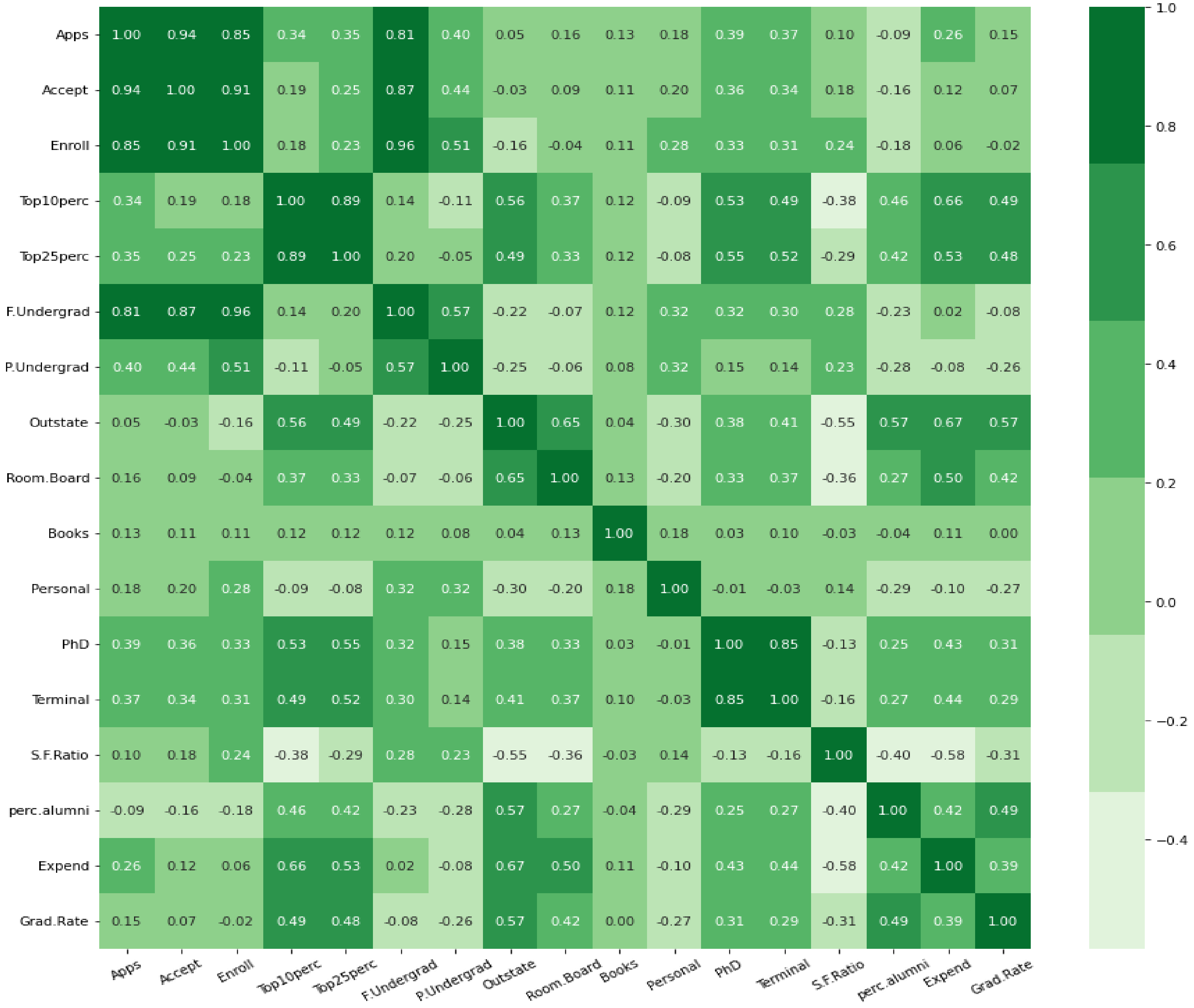


Fig 7 : Heat Map

Check for the Box Plot , with compare to each other , when having single Y – Axis:

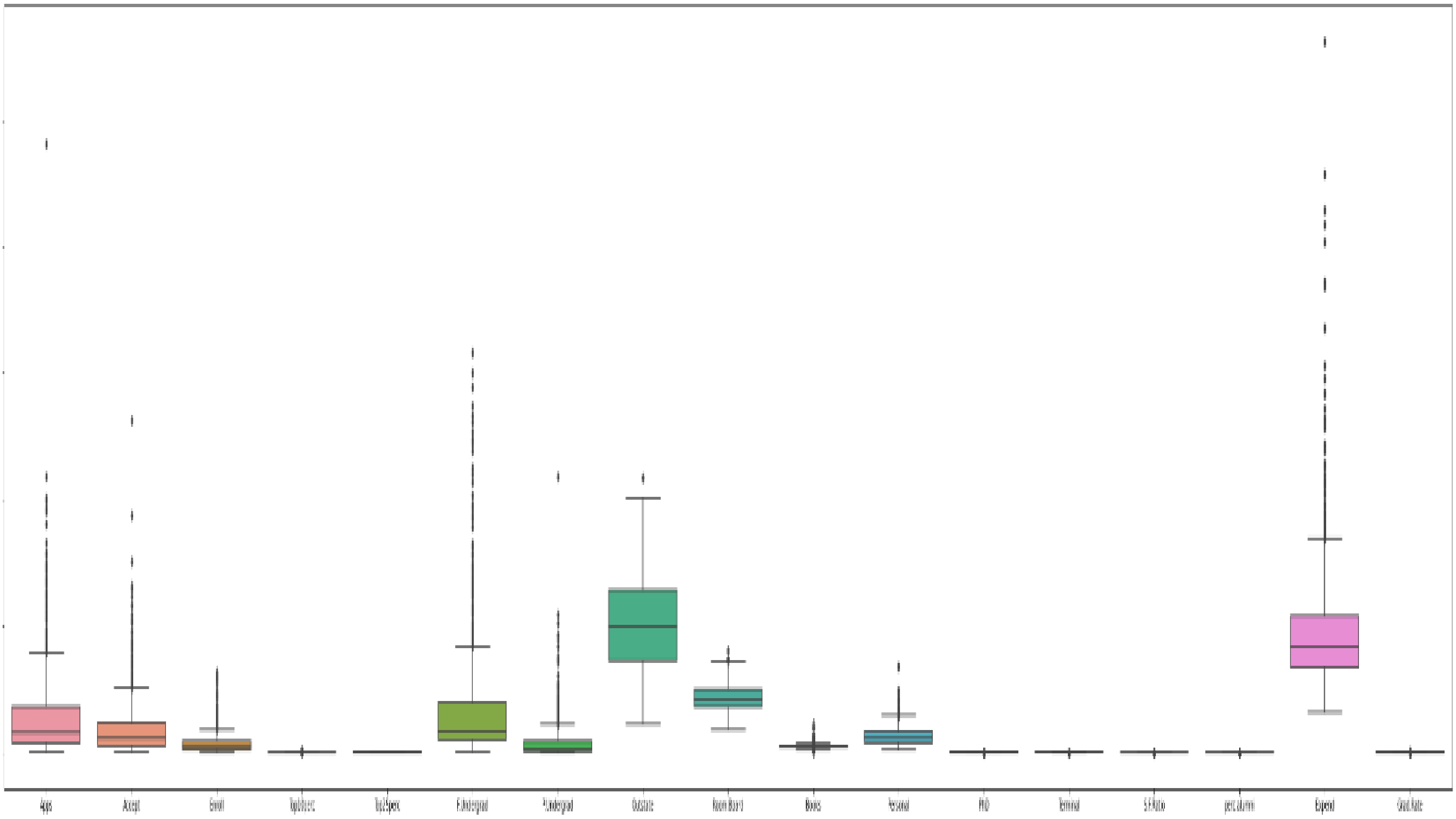


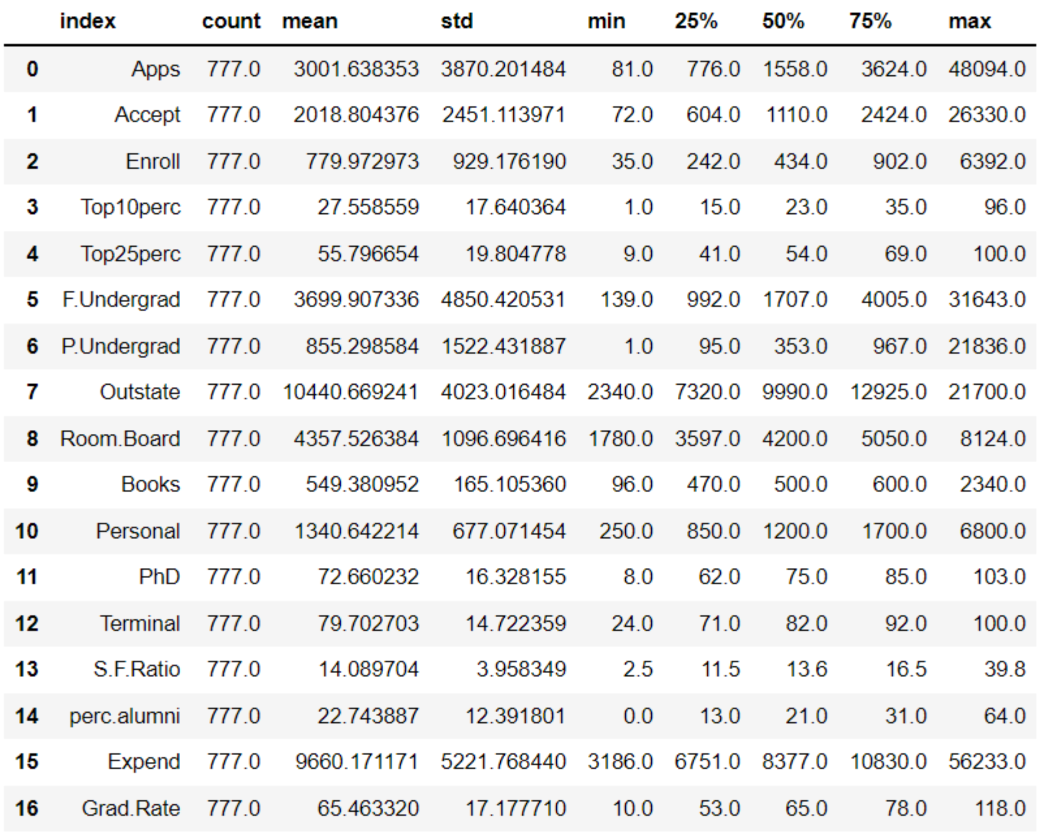
Fig 8: Combined Box Plot

Attached Image for the Box Plot for better visibility :

BoxPlot.png

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

We need to Describe data set, its Minimum and maximum values, Mean, median etc, for evaluating, whether we need scaling or not



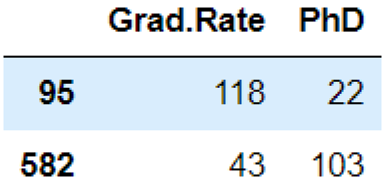
	index	count	mean	std	min	25%	50%	75%	max
0	Apps	777.0	3001.638353	3870.201484	81.0	776.0	1558.0	3624.0	48094.0
1	Accept	777.0	2018.804376	2451.113971	72.0	604.0	1110.0	2424.0	26330.0
2	Enroll	777.0	779.972973	929.176190	35.0	242.0	434.0	902.0	6392.0
3	Top10perc	777.0	27.558559	17.640364	1.0	15.0	23.0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

By looking at the min and max values, we can understand, that Minimum values accross all the column ranging between 1 to 3186 and maximum values ranging between 39 and 56233.0, which means, we have different Scales for all the columns, that's Why "WE NEED TO SCALE THE DATA BEFORE PROCESSING"

Treat BAD Data

Based on the data Describe Implications are: When Compare to Describe Output with Given data Dictionary Given , Following are implications:
1. "Names" are the Object data type, which can be Dropped for perforing EDA
2. "Apps", "Accepts" , "Enroll" , "F.Undergrad", "P.Undergrad", "Outstate", "Room.Board", "Books", "Personal" seems normal Columns.
3. "PhD" and "Grad.Rate" are the percentage values as per data dictionary, but it has maximum values more than 100, which is not normal, need to change.

Check how many rows , we have with percentage value of "PhD" and "Grad.Rate" more than 100.



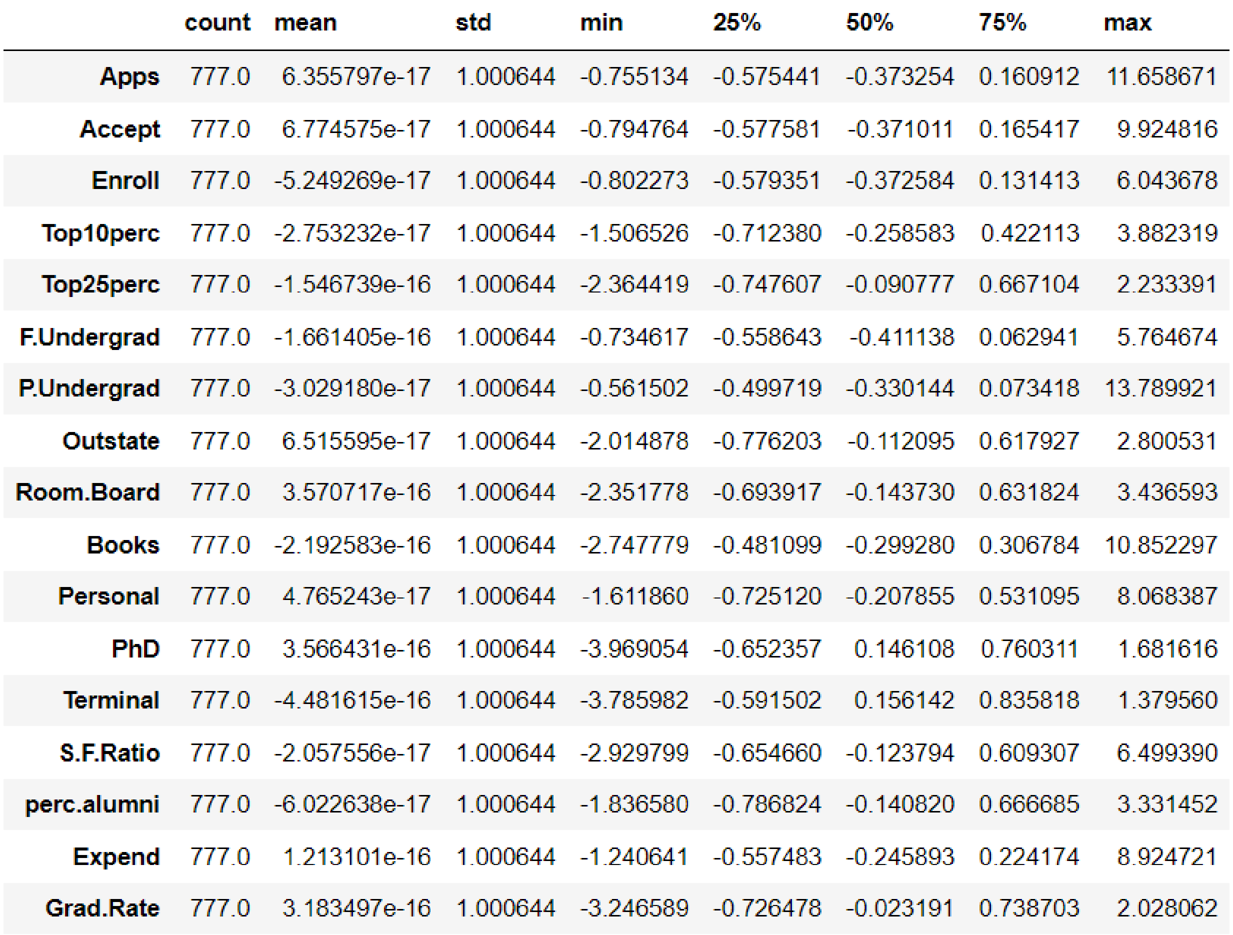
	Grad.Rate	PhD
95	118	22
582	43	103

So we have only 2 rows, in row Index 95, we have Grad.Rate more than 100 and in Index = 582 , we have PhD percent more than 100 Lets Treat both them one by one and replace them with Mean valu e of that Portion

We will replace above values with the MEAN value of that column.

Perform Scaling using z-Score

z-Score is the best method for performing scaling , which suits in most of the cases.
After performing z-Score scaling, Describe the data again :

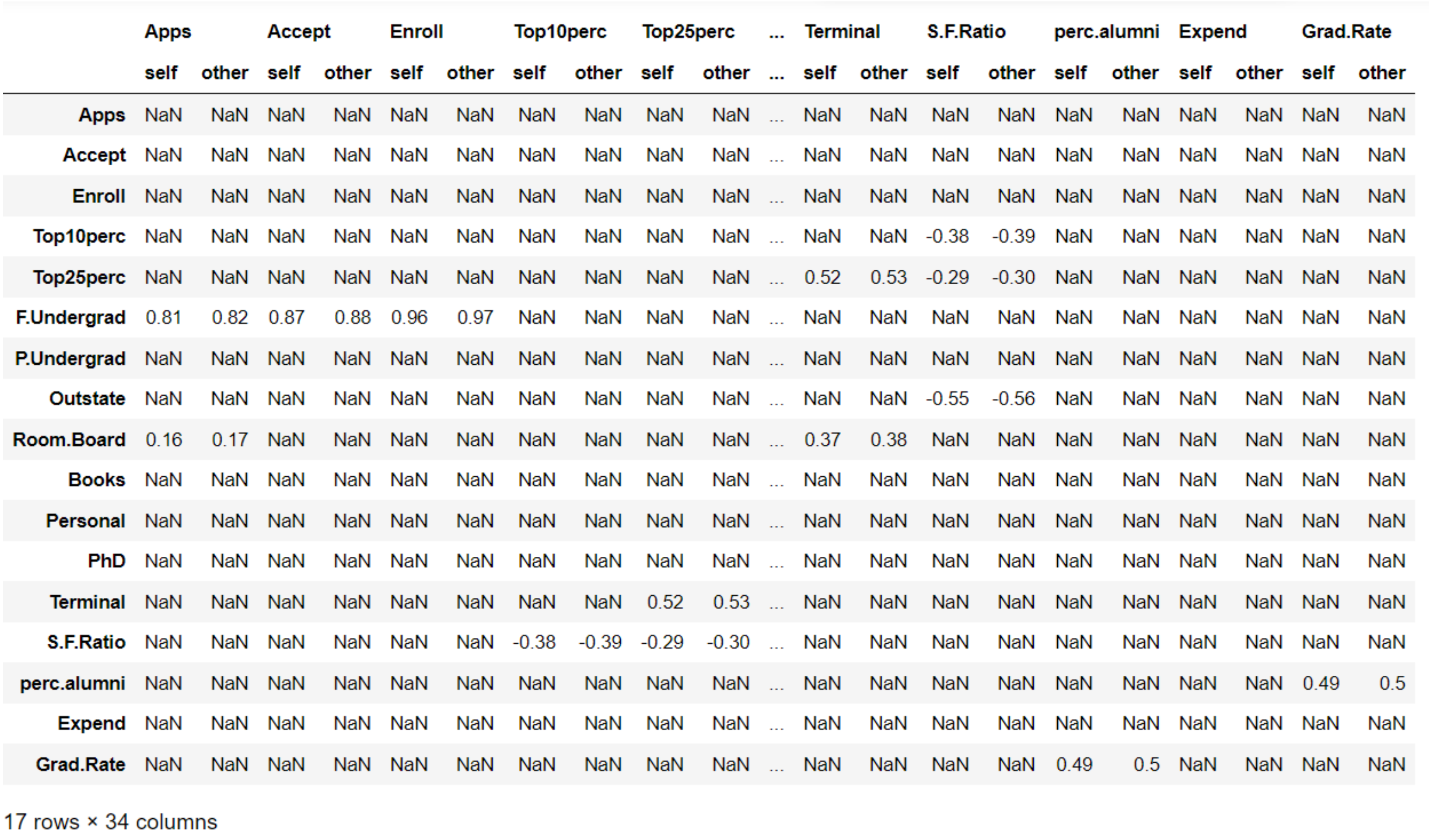


	count	mean	std	min	25%	50%	75%	max
Apps	777.0	6.355797e-17	1.000644	-0.755134	-0.575441	-0.373254	0.160912	11.658671
Accept	777.0	6.774575e-17	1.000644	-0.794764	-0.577581	-0.371011	0.165417	9.924816
Enroll	777.0	-5.249269e-17	1.000644	-0.802273	-0.579351	-0.372584	0.131413	6.043678
Top10perc	777.0	-2.753232e-17	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	3.566431e-16	1.000644	-3.969054	-0.652357	0.146108	0.760311	1.681616
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	3.183497e-16	1.000644	-3.246589	-0.726478	-0.023191	0.738703	2.028062

Now we can see that Minimum and maximum values do not have difference of thousands num bers, as we had earlier before scaling as shown in previous Charts.

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

Following are the correlation and covariance comparison charts :



	Apps		Accept		Enroll		Top10perc		Top25perc		...	Terminal		S.F.Ratio		perc.alumni		Expend		Grad.Rate	
	self	other	self	other	self	other	self	other	self	other	...	self	other	self	other	self	other	self	other	self	other
Apps	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Accept	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Enroll	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Top10perc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	-0.38	-0.39	NaN	NaN	NaN	NaN	NaN	NaN
Top25perc	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.52	0.53	-0.29	-0.30	NaN	NaN	NaN	NaN	NaN	NaN
F.Undergrad	0.81	0.82	0.87	0.88	0.96	0.97	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
P.Undergrad	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Outstate	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	-0.55	-0.56	NaN	NaN	NaN	NaN	NaN	NaN
Room.Board	0.16	0.17	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	0.37	0.38	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Books	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Personal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PhD	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Terminal	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.52	0.53	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
S.F.Ratio	NaN	NaN	NaN	NaN	NaN	NaN	-0.38	-0.39	-0.29	-0.30	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
perc.alumni	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.49	0.5
Expend	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Grad.Rate	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	0.49	0.5	NaN	NaN	NaN	NaN

17 rows × 34 columns

After scaling, when we compare covariance and correlation Metrics, both heat map showing same Graphs. There is very slight difference in the numbers, but those difference numbers are also negligible , if we will round it to ONE decimal point.

CO-Variance Shows the Direction of the relation between 2 variables CO-Relation shows the Direction as well as Strength of the relation between 2 variables

Since both CoVar and CoRel are almost same, we can say that We have good data set after scaling and it is good for processing.

2.4: Check the dataset for outliers before and after scaling. What insight do you derive ? [Do not treat Outliers unless specifically asked to do so]

Check for Outliers Before scaling

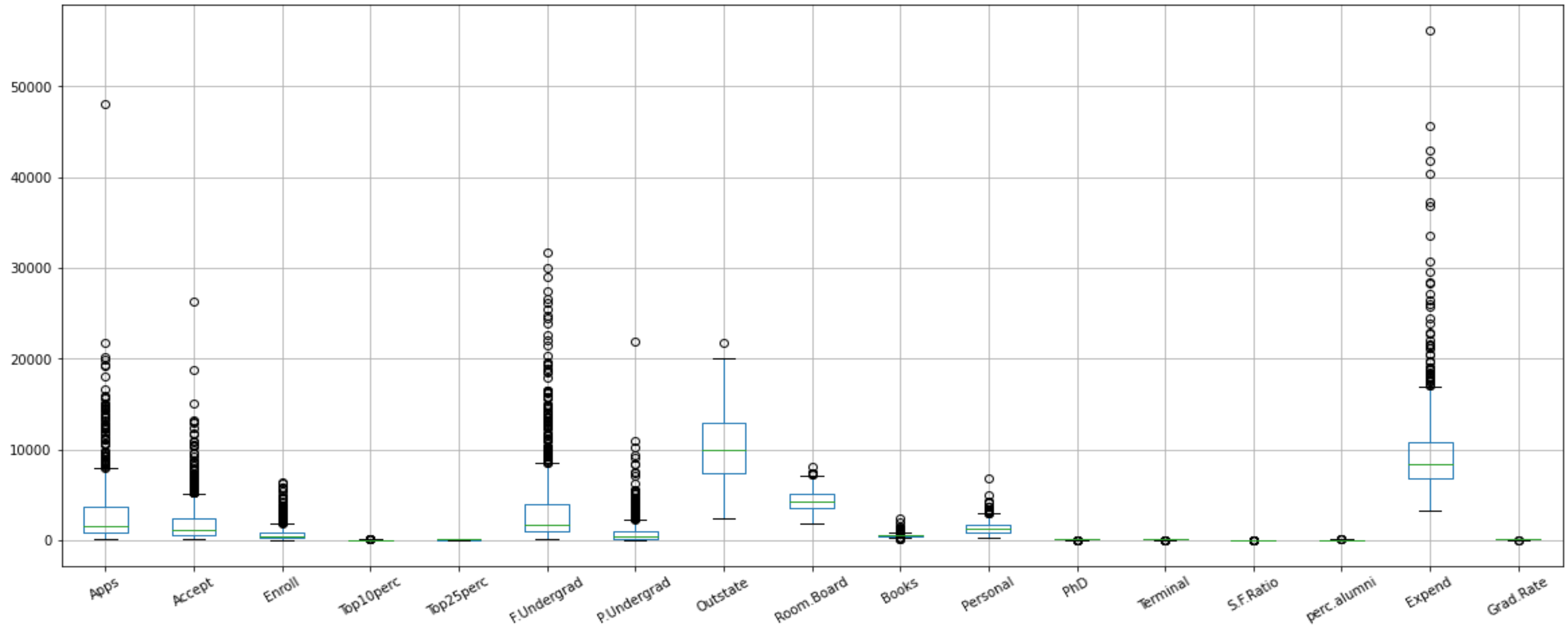


Fig 9 : Box Plot without Scaling after data correction

Check for Outliers After scaling :

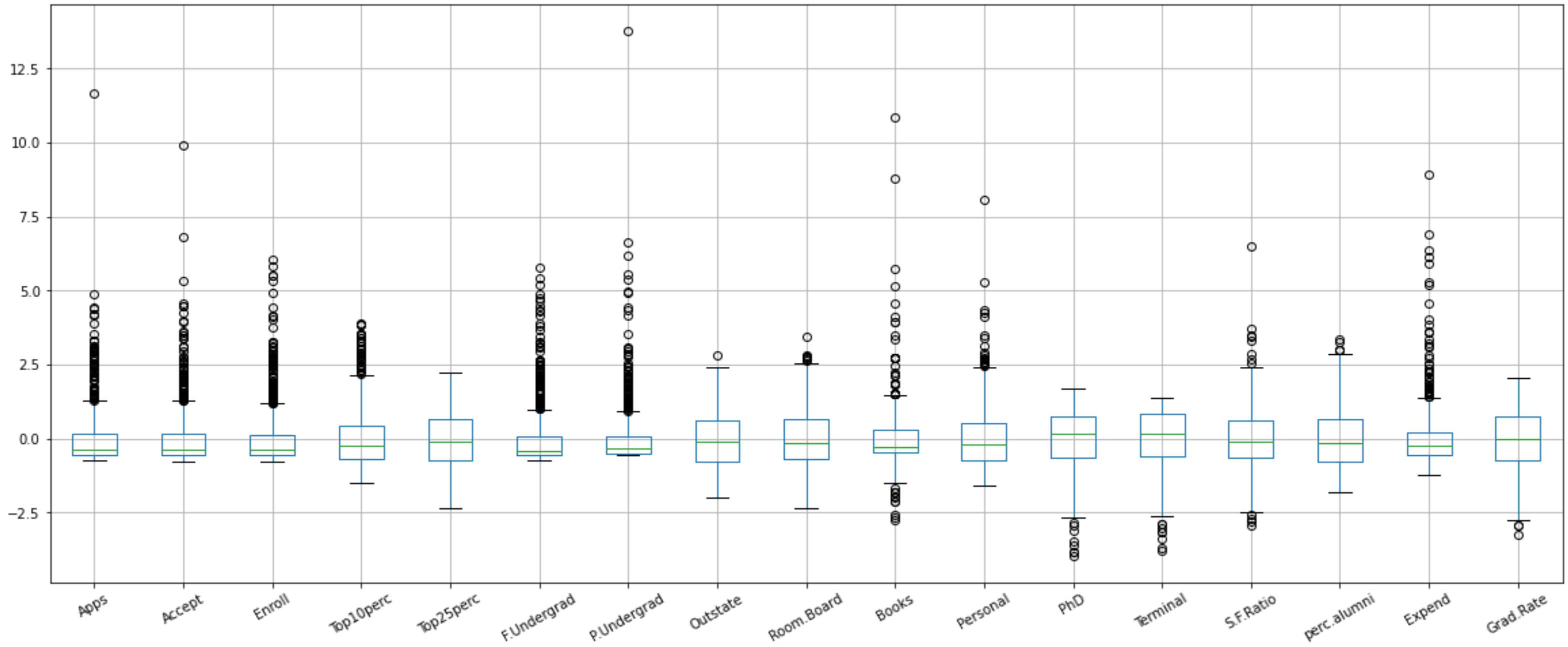


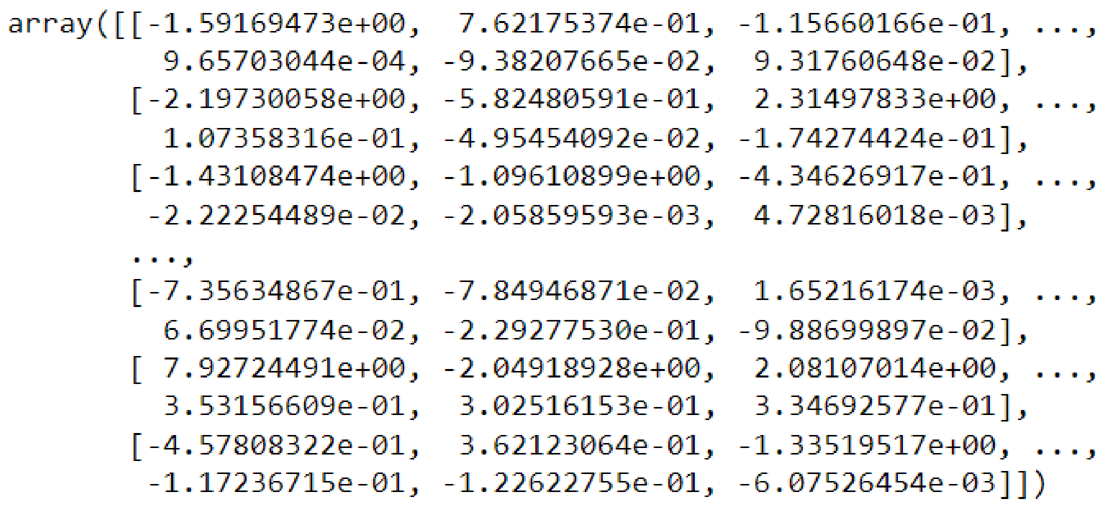
Fig 10: Box Plot after Scaling after data correction

Insights:

1. We can see after scaling Median is closer to 0 for all the Fields.
2. whereas it Median was scattered across all the numbers.
3. We have outliers for about all the fields before and after scaling, as we have not treated the outliers.

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

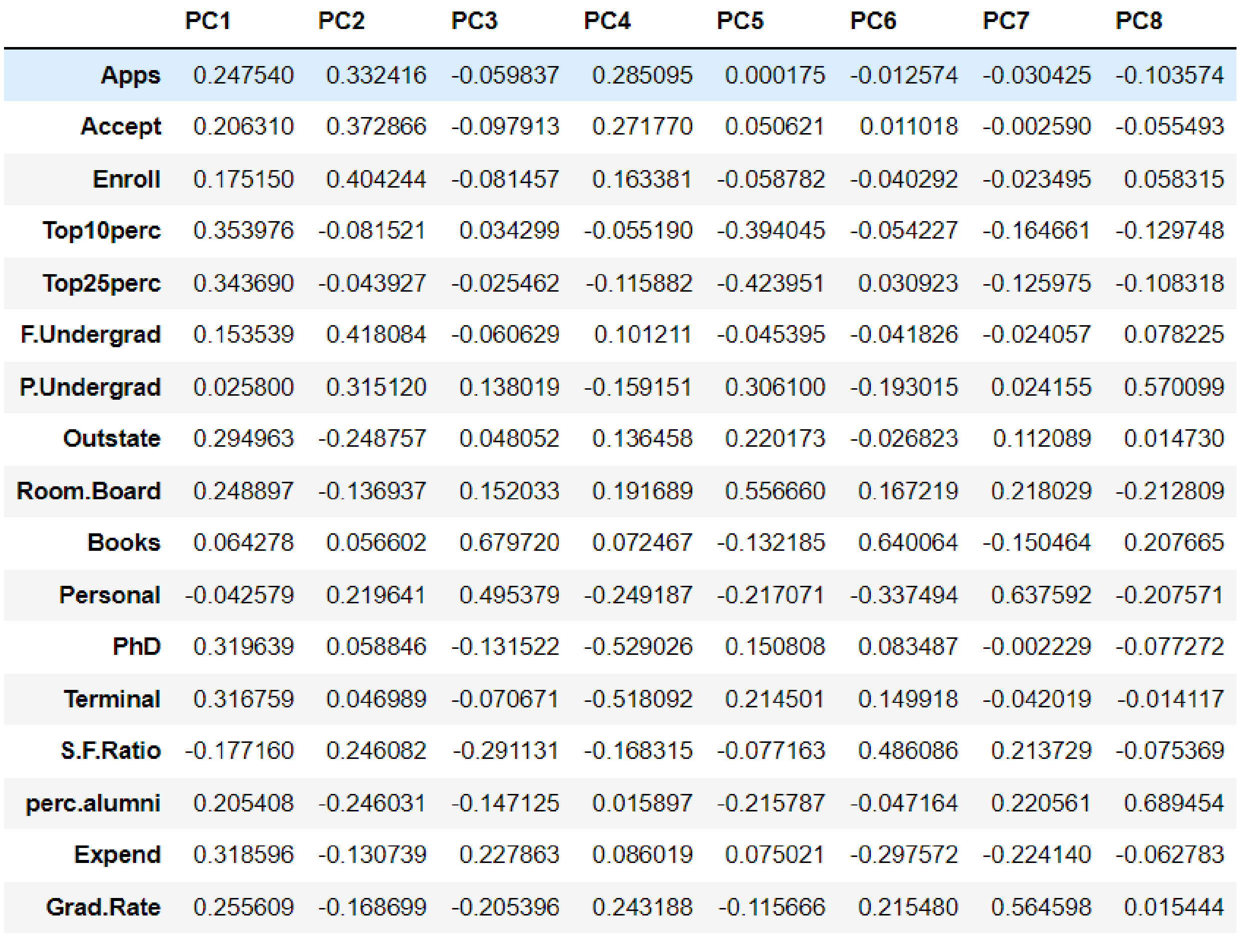
We have generated the eigenvalues and eigen vectors for all the scaled data set.
We have used **PCA** Module from **sklearn.decomposition** Python Library for this analysis



```
array([[-1.59169473e+00,  7.62175374e-01, -1.15660166e-01, ...,
         9.65703044e-04, -9.38207665e-02,  9.31760648e-02],
       [-2.19730058e+00, -5.82480591e-01,  2.31497833e+00, ...,
         1.07358316e-01, -4.95454092e-02, -1.74274424e-01],
       [-1.43108474e+00, -1.09610899e+00, -4.34626917e-01, ...,
        -2.22254489e-02, -2.05859593e-03,  4.72816018e-03],
       ...,
       [-7.35634867e-01, -7.84946871e-02,  1.65216174e-03, ...,
         6.69951774e-02, -2.29277530e-01, -9.88699897e-02],
       [ 7.92724491e+00, -2.04918928e+00,  2.08107014e+00, ...,
         3.53156609e-01,  3.02516153e-01,  3.34692577e-01],
       [-4.57808322e-01,  3.62123064e-01, -1.33519517e+00, ...,
        -1.17236715e-01, -1.22622755e-01, -6.07526454e-03]])
```

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 fram.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Apps	0.247540	0.332416	-0.059837	0.285095	0.000175	-0.012574	-0.030425	-0.103574
Accept	0.206310	0.372866	-0.097913	0.271770	0.050621	0.011018	-0.002590	-0.055493
Enroll	0.175150	0.404244	-0.081457	0.163381	-0.058782	-0.040292	-0.023495	0.058315
Top10perc	0.353976	-0.081521	0.034299	-0.055190	-0.394045	-0.054227	-0.164661	-0.129748
Top25perc	0.343690	-0.043927	-0.025462	-0.115882	-0.423951	0.030923	-0.125975	-0.108318
F.Undergrad	0.153539	0.418084	-0.060629	0.101211	-0.045395	-0.041826	-0.024057	0.078225
P.Undergrad	0.025800	0.315120	0.138019	-0.159151	0.306100	-0.193015	0.024155	0.570099
Outstate	0.294963	-0.248757	0.048052	0.136458	0.220173	-0.026823	0.112089	0.014730
Room.Board	0.248897	-0.136937	0.152033	0.191689	0.556660	0.167219	0.218029	-0.212809
Books	0.064278	0.056602	0.679720	0.072467	-0.132185	0.640064	-0.150464	0.207665
Personal	-0.042579	0.219641	0.495379	-0.249187	-0.217071	-0.337494	0.637592	-0.207571
PhD	0.319639	0.058846	-0.131522	-0.529026	0.150808	0.083487	-0.002229	-0.077272
Terminal	0.316759	0.046989	-0.070671	-0.518092	0.214501	0.149918	-0.042019	-0.014117
S.F.Ratio	-0.177160	0.246082	-0.291131	-0.168315	-0.077163	0.486086	0.213729	-0.075369
perc.alumni	0.205408	-0.246031	-0.147125	0.015897	-0.215787	-0.047164	0.220561	0.689454
Expend	0.318596	-0.130739	0.227863	0.086019	0.075021	-0.297572	-0.224140	-0.062783
Grad.Rate	0.255609	-0.168699	-0.205396	0.243188	-0.115666	0.215480	0.564598	0.015444

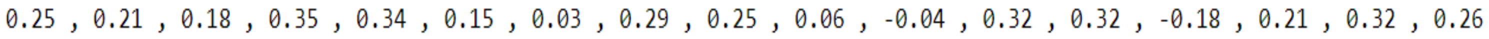
2.7 Write down the explicit form of the first PC (in terms of the eigenvectors. Use values with two places of decimals only). [hint: write the linear equation of PC in terms of eigenvectors and corresponding features]

Linear Equation Example:

Y = mX + C

Similarly, we write Equation in terms of the principal components

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



```
0.25 , 0.21 , 0.18 , 0.35 , 0.34 , 0.15 , 0.03 , 0.29 , 0.25 , 0.06 , -0.04 , 0.32 , 0.32 , -0.18 , 0.21 , 0.32 , 0.26
```

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

```
The Linear Equation for the 1st Component will be
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 +0.26 *
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 sum is as follows:

```
array([0.3210007 , 0.584437  , 0.65343689, 0.71205208, 0.76696205,
       0.81674035, 0.85233271, 0.88686661, 0.91800986, 0.9417166 ,
       0.96011355, 0.97305584, 0.9829117 , 0.99131706, 0.99649123,
       0.99864821, 1.        ])
```

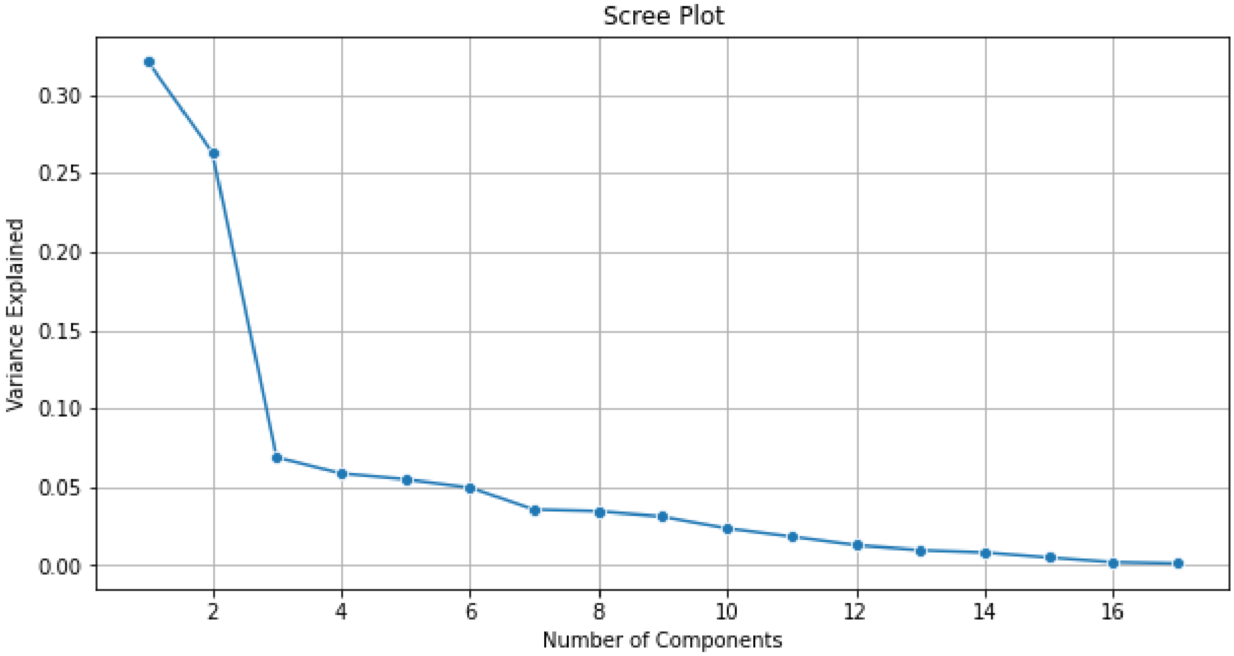


Fig: Weightage chart for each PCA component

We have total 17 columns , but based on cumulative Summation , we can see that initial 7-8 component is contributing approx. 90% of the information, so in our case we will Consider first 8 values of principal components , which will make 88.68% of information. We can take it as optimum number. If we try to take more components it will increase the dimensionality.

2.9: Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

We have performed PCA for making all data points come to the vectors and reduce the dimensionality of the data set. PCA also helps us in understanding the distribution of every ingredient of the principal components extracted out from the data.

We can further disintegrate each principal component and see that which contents is contributing most in each individual PCA component. By performing the PCA we ensure to eliminate the Noise from data , which comes from the degree of the data points away from normal vectors.

When we performed the PCA, we have got following Cumulative summation of percentage contribution of each PCA element:

```
array([0.3210007 , 0.584437  , 0.65343689, 0.71205208, 0.76696205,
       0.81674035, 0.85233271, 0.88686661, 0.91800986, 0.9417166 ,
       0.96011355, 0.97305584, 0.9829117 , 0.99131706, 0.99649123,
       0.99864821, 1.        ])
```

Out of above data points , we have chosen 8 elements and reduced dimensionality by 9 element (there were total 17 element). All 8 components are contributing 88.68% of information, which is enough for data evaluation.

Other Insights are as follows:

1. By using PCA , we could come to know, that 7-8 PCA components are enough for getting more than 85% of the information , Whereas we had 17 Numeric fields to analyze, so we reduce the Dimensionality more than 50%
2. In our analysis we took 8 Components
3. While analyzing individual component, we could come to know, that First Component is covering about 32% of information, and within 1st component itself following fields are playing major role in College "Top10 Percent", "top 25 percent", "Expend", "PhD" and "Terminal" degree,
4. throughout all other components also above listed Fields are playing Vital role in the College information.
5. Since PhD and terminal degree of the teacher are playing good role in each component, we can say that Student prefer to Stick with college with good faculty
6. Though 1st and 2nd Component always have major roles from most of the columns, but when we analyze remaining 6 component out 8 chosen components, we can see that "Books" , "Personal" and "Room Board" also plays major role in information and , might be affecting the reason for choosing the colleges.