Business Report

DSBA

Advanced Statistics Module Project

Name: Sandeep Immadi

Date: 13/06/2021

CONTENTS:

Problem 1:

1.1 State the null and the alternate hypothesis for conducting one-way ANOVA for both Education and Occupation individually…… 5

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…...6

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……………………………………………………..8

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……9

1.5 What is the interaction between the two treatments? Analyse the effects of one variable on the other (Education and Occupation) with the help of an interaction plot……………………………………………………………10

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?...12

1.7 Explain the business implications of performing ANOVA for this particular case study……………………………13

Problem 2:

2.1 Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. What insight do you draw from the EDA? ………16

2.2 Is scaling necessary for PCA in this case? Give justification and perform scaling………………………………………23

2.3 Comment on the comparison between the covariance and the correlation matrices from this data. [on scaled data]……… 26

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

2.5 Extract the eigenvalues and eigenvectors. [print both]………………………………………………………………………… 29

2.6 Perform PCA and export the data of the Principal Component (eigenvectors) into a data frame with the original features…… 31

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]……… 34

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?... 35

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]………36

List of Figures

- Fig.1 – Salary-education plot……4
- Fig.2 – Salary-occupation plot ……7
- Fig.3 – Inter plot salary-occupation ……8
- Fig.4 – Inter plot salary-education ……10
- Fig.5 – Two way ANOVA…… 10
- Fig.6 – Dataset info ……11
- Fig.7 – Histogram……13
- Fig.8 – Boxplot…… 16
- Fig.9 – Heatmap…… 17
- Fig.10 – Before scaling……19
- Fig.11 – After scaling……19
- Fig.12 – Scatter plot……21
- Fig.13 – PCA exp var…… 22
- Fig.14 – Linear Equation……27
- Fig.15 – Cum values……28
- Fig.16 – After scaling…… 29
- Fig.17 – Scatter plot…… 33
- Fig.18 – Eigen…… 35
- Fig.19 – cum_var……35
- Fig.20 – PCA PLOT……36

List of Tables

- Tab.1 – Education one way ANOVA……7
- Tab.2 – Occupation one way ANOVA ……8
- Tab.3 – Tukey HSD……10
- Tab.4 – Summary dataset……16
- Tab.5 – Skewness……17
- Tab.6 – Kurtosis……17
- Tab.7 – Scaled data…… 24
- Tab.8 – Scaled data to data frame…… 24
- Tab.9 – Stats summary Scaled data……25
- Tab.10 – Before scaling…… 26
- Tab.11 – After scaling…… 29
- Tab.12 – PCA exp var…… 31
- Tab.13 – Linear equation…… 32
- Tab.14 – Cumulative variance ……32
- Tab.15 – Linear equation…… 33
- Tab.16 – Cumulative variance ……34

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.

Variables:

Education: Indicates education qualification of individuals – Doctorate, Bachelors and HS-grad.

Occupation: indicates the occupation of individuals – Adm-clerical, Sales, Prof-specialty and Exec-managerial.

Salary: indicates the salary of an individual.

- Basic Parameters:

```
<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
```

```
 Doctorate    16
 Bachelors    15
 HS-grad       9
Name: Education, dtype: int64
```

```
 Prof-specialty     13
 Sales              12
 Adm-clerical       10
 Exec-managerial     5
Name: Occupation, dtype: int64
```

Fig:1

- Inferences:

- Total number of entries = 40
- Total number of columns = 3
- Number of null values = 0
- Data types encountered = int64(1), object (2)
- The data is normally distributed.

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

- One-way refers to the number of independent variables in our analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable.
- It determines whether the observed differences between the means of independent (unrelated) groups are explainable by chance alone, or whether there are any statistically significant differences between groups.
- Analyse of all the sample means at one time and thus precludes the build-up of error rate.
- A completely randomized design is analysed by a one-way analysis of variance. If k samples are being analysed, the following hypotheses are being tested in a one-way ANOVA.
- H0: $\mu 1 = \mu 2 = \mu 3 = \ldots\ldots = \mu k$
- Ha: At least one of the means is different from the others
- The null hypothesis states that the population means for all treatment levels are equal.
- Because of the way the alternative hypothesis is stated, if even one of the population means is different from the others, the null hypothesis is rejected.
- Testing these hypotheses by using one-way ANOVA is accomplished by partitioning the total variance of the data into the following two variances.
- The variance resulting from the treatment (columns)
- The error variance, or that portion of the total variance unexplained by the treatment
- For ANOVA, the null hypothesis assumes equality of population means and the alternate hypothesis assumes that the population means differ.
- To disprove the null hypothesis, it is sufficient to prove that at least one population mean differs.
- Hence, the hypothesis for this case are:

❖ For Education:

✓ H0 = Mean salary is the same for all levels of education.
✓ H1= Mean salary differs for at least one level of education.

❖ For Occupation:

✓ H0 = Mean salary is the same for all levels of occupation.
✓ H1= Mean salary differs for at least one level 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.

- Analysis of variance is used to determine statistically whether the variance between the treatment level means is greater than the variances within levels (error variance).

❖ Assumptions:

- Observations are drawn from normally distributed populations.
- Observations represent random samples from the populations.
- Variances of the populations are equal.
- As part of this process, the total sum of squares of deviation of values around the mean can be divided into two additive and independent parts.

  A. SST = SSC + SSE
  B. nj Σ i=1 C Σ j=1 (xij - xbar)** 2 = C Σ j=1 nj(xj bar – x bar)**2 + nj Σ i=1 C Σ j=1 (xij - xj bar)**2

  where:

  i. SST = total sum of squares;
  ii. SSC = sum of squares column (treatment) ;
  iii. SSE = sum of squares error;
  iv. i = particular member of a treatment level ;
  v. j = a treatment level ;
  vi. C = number of treatment levels;
  vii. nj = number of observations in a given treatment level;
  viii. x bar= grand mean;
  ix. xj bar= mean of a treatment group or level;
  x. xij = individual value

- Degrees of freedom (DF) indicate the number of independent values that can vary in an analysis without breaking any constraints.
- Which is used throughout statistics including hypothesis tests, probability distributions, and regression analysis.
- The mean squared error (MSE) tells us how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them.
- The squaring is necessary to remove any negative signs. It also gives more weight to larger differences.
- An F statistic is a value we get when we run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different.

- It's similar to a T statistic from a T-Test; A T-test will tell us if a single variable is statistically significant and an F test will tell us if a group of variables are jointly significant.

❖ <u>Framing the Null Hypothesis and the Alternate Hypothesis.</u>

✓ H0 = Mean salary is the same for all levels of education.
✓ H1= Mean salary differs for at least one level of education.
✓ Level of significance(alpha) = 0.05
✓ ANOVA using stats model's package. Python output for the same:

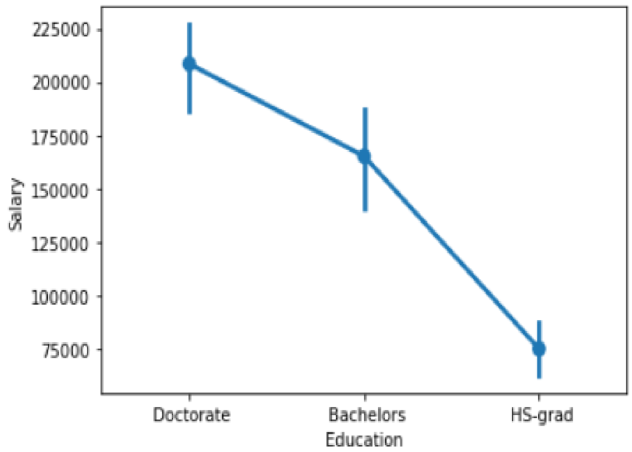


Fig.2

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

Table.1

✓ P value = 1.257709e-08. Since P value < alpha, we reject the null hypothesis based on the ANOVA results. Hence, the mean salary differs for at least one level of education.

- We use a point plot to visualize the graph of Salary vs Education.
- From above plot we can see that the Salary varies with respect to Education.
- It seems to be highest for individuals with a Doctorate, followed by those with only Bachelors and it is least for individuals with only HS-grad.

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.

- Degrees of freedom (DF) indicate the number of independent values that can vary in an analysis without breaking any constraints.
- Which is used throughout statistics including hypothesis tests, probability distributions, and regression analysis.
- The mean squared error (MSE) tells us how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line (these distances are the "errors") and squaring them.
- The squaring is necessary to remove any negative signs. It also gives more weight to larger differences.
- An F statistic is a value we get when we run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different.
- It's similar to a T statistic from a T-Test; A T-test will tell us if a single variable is statistically significant and an F test will tell us if a group of variables are jointly significant.

❖ Framing the Null Hypothesis and the Alternate Hypothesis.

✓ H0 = Mean salary is the same for all levels of occupation
✓ H1= Mean salary differs for at least one level of occupation.
✓ Level of significance(alpha) = 0.05
✓ ANOVA using stats model's package. Python output for the same:

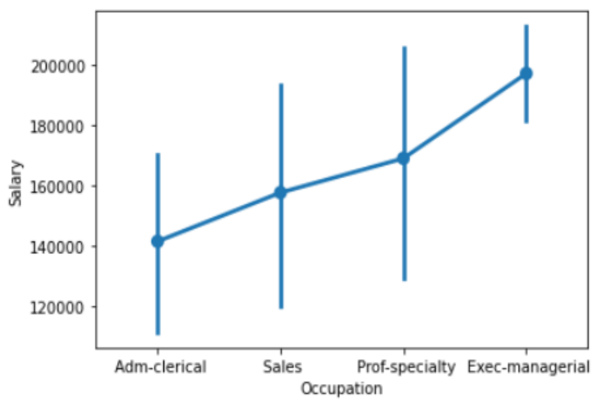


Fig.3

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

Table.2

✓ P value = 0.458508. Since p value >alpha, we fail to reject the null hypothesis based on the ANOVA results. Hence, the mean salary does not vary with occupation.

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Adm-clerical	Exec-managerial	55693.3	0.4146	-40415.1459	151801.7459	False
Adm-clerical	Prof-specialty	27528.8538	0.7252	-46277.4011	101335.1088	False
Adm-clerical	Sales	16180.1167	0.9	-58951.3115	91311.5449	False
Exec-managerial	Prof-specialty	-28164.4462	0.8263	-120502.4542	64173.5618	False
Exec-managerial	Sales	-39513.1833	0.6507	-132913.8041	53887.4374	False
Prof-specialty	Sales	-11348.7372	0.9	-81592.6398	58895.1655	False

Table.2.1

- We use a point plot to visualize the graph of Occupation vs Education.
- From the graph, we can conclude that Salary varying for different occupations (reject: False).

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.

- The null hypothesis is rejected in (1.2).
- An ANOVA test can tell us if our results are significant overall, but it won't tell exactly where those differences lie.
- After running an ANOVA and found significant results, then we can run Tukey's HSD to find out which specific group's means (compared with each other) are different. The test compares all possible pairs of means.
- To test all pairwise comparisons among means using the Tukey HSD, calculate HSD for each pair of means using the following formula:

  ❖ Mi-Mj/sqrt(MSw/nh)

i. Mi – Mj is the difference between the pair of means. to calculate this, M,i should be larger than Mj

ii. MSw is the Mean Square Within, and n is the number in the group or treatment.

❖ Following is the python output for the test conducted on python:

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

Table.3

❖ Inference:

- By comparing two group levels we can say mean difference varies.
- Reject column: Since all three rows say "True", we conclude that Salary varies for individuals with different levels of Education.

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

- To understand the interaction between the two treatments, we use point plot.

❖ Below are the graphs illustrating the interaction:

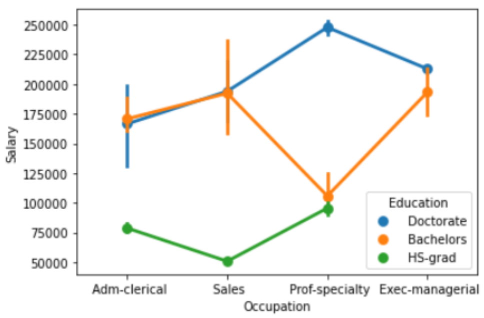


Fig.4

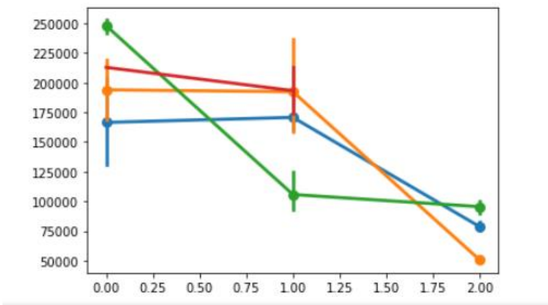


Fig.5

- From the above graphs, it is clear that there is some level of interaction between the two treatments.

Occupation	Education	Salary	
Sales	HS-grad	52242	1
		50122	1
Prof-specialty	Bachelors	100135	1
		99185	1
		90135	1
Exec-managerial	Doctorate	212781	1
	Bachelors	212760	1
		212448	1
		173935	1
		173664	1
Adm-clerical	HS-grad	83203	1
		77743	1
		75333	1
	Doctorate	220754	1
		175935	1
		153197	1
		115945	1
	Bachelors	188729	1
		162494	1
Prof-specialty	Bachelors	133696	1
	Doctorate	235334	1
		247724	1
Sales	Bachelors	191712	1
	HS-grad	50103	1
	Doctorate	237920	1
		219420	1
		180934	1
		170769	1
		160540	1
	Bachelors	260151	1
		167431	1
Prof-specialty	Doctorate	248156	1
Sales	Bachelors	149909	1
Prof-specialty	HS-grad	100678	1
		95469	1
		90456	1
	Doctorate	257345	1
		249207	1
		248871	1
Adm-clerical	Bachelors	160910	1

	Salary
Education	
Doctorate	208427.000000
Bachelors	165152.933333
HS-grad	75038.777778

	Salary
Occupation	
Exec-managerial	197117.600000
Prof-specialty	168953.153846
Sales	157604.416667
Adm-clerical	141424.300000

Fig.6

- The ones working in Adm-Clerical occupation earn the least.
- The ones working in Exec-managerial occupation earn the most.
- Individuals with a Bachelors, the ones working in Sales and Exec-managerial occupations earn almost the same, followed by those in Adm-Clerical.
- Doctorate level of education have the highest salary and those working in Sales with a high school graduation (HS-grad) have the lowest salary.
- With only high school level of education have the least salary and those with a Doctorate earn the most.

- The ones working in Sales earn the least. For individuals with a Bachelors, the ones working in Sales and Exec-managerial occupations earn almost the same, followed by those in Adm-Clerical.
- Among those with Doctorate level of education, the ones in Adm-Clerical earn the least and the ones in Prof-specialty earn the most.
- We can see that there are no individuals with only high school graduation in Exec-managerial. Those with a doctorate in this occupation earn more than those with a Bachelors.
- Those in Prof-specialty with a Bachelors seem to earn the least as compared to people with Bachelors having other occupations.
- Individuals with a Doctorate working in Prof-Specialty earn significantly more than the rest in the same occupation.
- Among individuals with the occupation, Adm-clerical, those with only a high school graduation have the least salary and individuals in Adm-clerical with other two levels of education earn similar salaries.
- The same trend is observed for individuals in Sales as well. Among individuals in Profspecialty, the ones with a high school graduation and a Bachelors seem to earn almost the same.

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?

- A two-way ANOVA test is a statistical test used to determine the effect of two nominal predictor variables on a continuous outcome variable.
- A two-way ANOVA tests the effect of two independent variables on a dependent variable. A two-way ANOVA test analyses the effect of the independent variables on the expected outcome along with their relationship to the outcome itself.
- Random factors would be considered to have no statistical influence on a data set, while systematic factors would be considered to have statistical significance.
- There are two main types of analysis of variance: one-way (or unidirectional) and two-way (bidirectional).
- One-way or two-way refers to the number of independent variables in our analysis of variance test. A one-way ANOVA evaluates the impact of a sole factor on a sole response variable.
- It determines whether the observed differences between the means of independent (unrelated) groups are explainable by chance alone, or whether there are any statistically significant differences between groups.
- A two-way ANOVA is an extension of the one-way ANOVA.
- With a one-way, we have one independent variable affecting a dependent variable. With a two-way ANOVA, there are two independents.
- It is utilized to observe the interaction between the two factors. It tests the effect of two factors at the same time.

❖ Framing the Null Hypothesis and the Alternate Hypothesis.

✓ H0 = Mean salary is the same for all levels of Education and Occupation.
✓ H1 = Mean salary differs for at least one level of Education and Occupation.
✓ Level of significance(alpha) = 0.05

❖ Two Way ANOVA. Python output for the same:

```
                             df        sum_sq       mean_sq          F  \
C(Education)                2.0  1.026955e+11  5.134773e+10  72.211958
C(Occupation)               3.0  5.519946e+09  1.839982e+09   2.587626
C(Education):C(Occupation)  6.0  3.634909e+10  6.058182e+09   8.519815
Residual                   29.0  2.062102e+10  7.110697e+08        NaN

                                 PR(>F)
C(Education)               5.466264e-12
C(Occupation)              7.211580e-02
C(Education):C(Occupation) 2.232500e-05
Residual                            NaN
```

Fig.7

- p value < alpha.
- Hence, we reject the null hypothesis. We can conclude that the mean salary differs for at least one level of education and occupation.

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

❖ Business Implications of performing ANOVA:

- From the Dataset we can see that there are no null values, assumption is normally distributed data.
- Observations represent random samples from the populations.
- Variances of the populations are equal.
- Further, the value counts for Education: Doctorate 16; bachelors 15; HS-grad 9. the value counts for Occupation prof-speciality 13, sales 12, adm-clerical 10, exec-managerial 5.
- Using point plot, we concluded that the Salary varies with respect to Education and varying for different occupations.
- This was further proven by ANOVA.
- Analysis of variance (ANOVA) is a statistical test for detecting differences in group means when there is one parametric dependent variable and one or more independent variables.

- ANOVA regarding Salary with respect to Education, we can say that Salary is dependent on Education. It is different for different levels of education. This is further confirmed by Tukey's HSD test.
- ANOVA regarding Salary with respect to Occupation, we can say that apart from Slight variation mean Salary is more or less same for all categories of Occupation.
- ANOVA regarding Salary with respect to Education and Occupation including the interaction effect of the two treatments, we can say that the main variable affecting Salary is Education. the salary is different for different levels of Education and Occupation.

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

Variables:

- Names – Names of various universities 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 books cost for a student
- Personal – Estimated personal spending for a student.
- PhD – Percentage of faculties with PhDs
- Terminal – Percentage of faculties with terminal degree.
- S.F. Ratio – Student to faculty ratio.
- Perc. Alumni – Percentage of alumni who donate
- Expend – The instructional expenditure per student
- Grad. Rate – Graduation Rate

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

Univariate Analysis:

1.Dataset Info:

```
<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
```

Fig.8

2.Statistical summary of 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

Table.4

3.Duplicates:

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

Fig.9

4.Skewness:

Apps	3.723750
Accept	3.417727
Enroll	2.690465
Top10perc	1.413217
Top25perc	0.259340
F.Undergrad	2.610458
P.Undergrad	5.692353
Outstate	0.509278
Room.Board	0.477356
Books	3.485025
Personal	1.742497
PhD	-0.768170
Terminal	-0.816542
S.F.Ratio	0.667435
perc.alumni	0.606891
Expend	3.459322
Grad.Rate	-0.113777
dtype: float64	

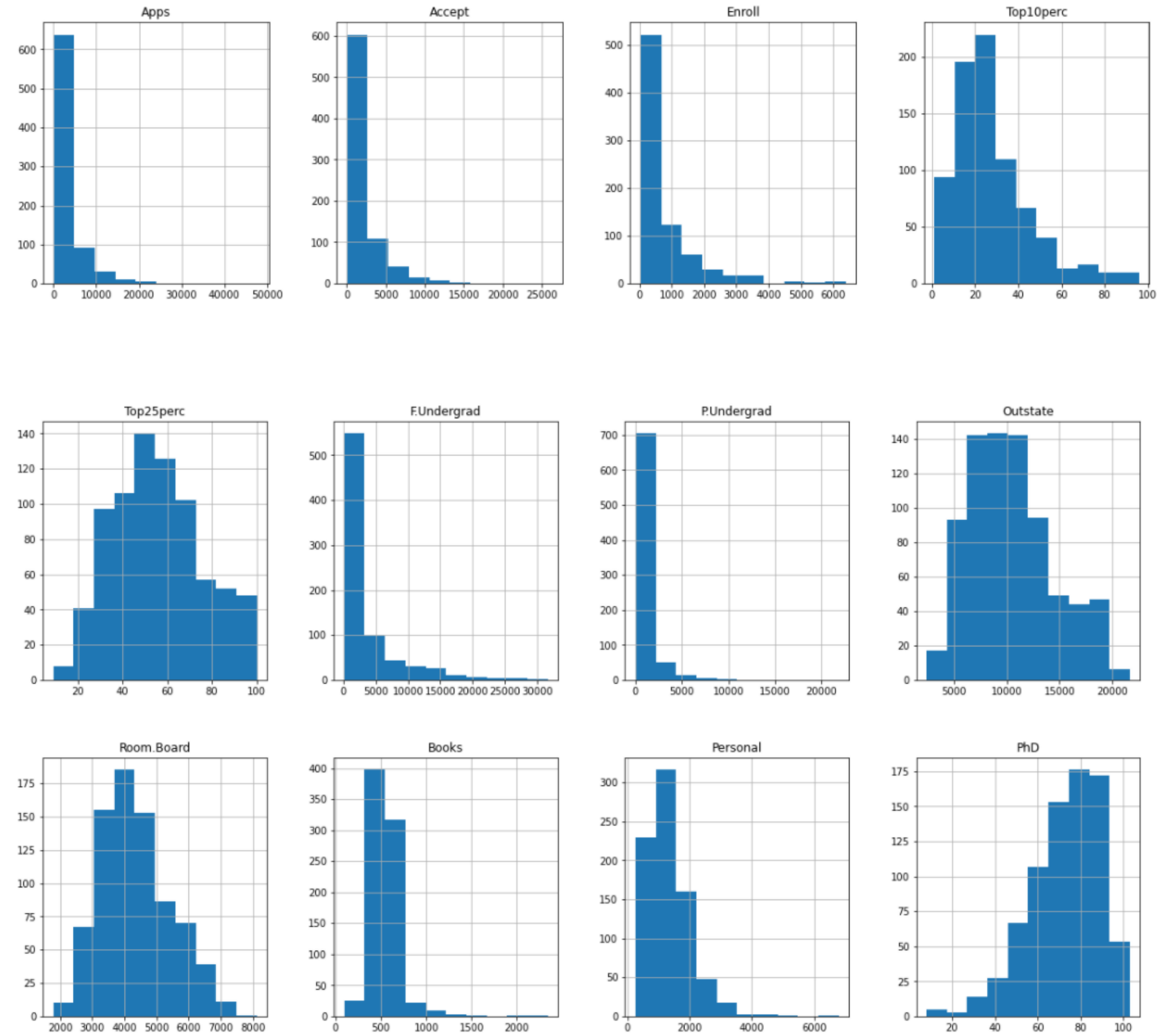
Table.5

5.Kurtosis:

Apps	26.774253
Accept	18.938099
Enroll	8.831544
Top10perc	2.208065
Top25perc	-0.564121
F.Undergrad	7.696586
P.Undergrad	55.034518
Outstate	-0.413832
Room.Board	-0.187553
Books	28.333097
Personal	7.124017
PhD	0.564773
Terminal	0.242019
S.F.Ratio	2.561209
perc.alumni	-0.096807
Expend	18.771500
Grad.Rate	-0.205226
dtype: float64	

Table.6

6.Histograms:



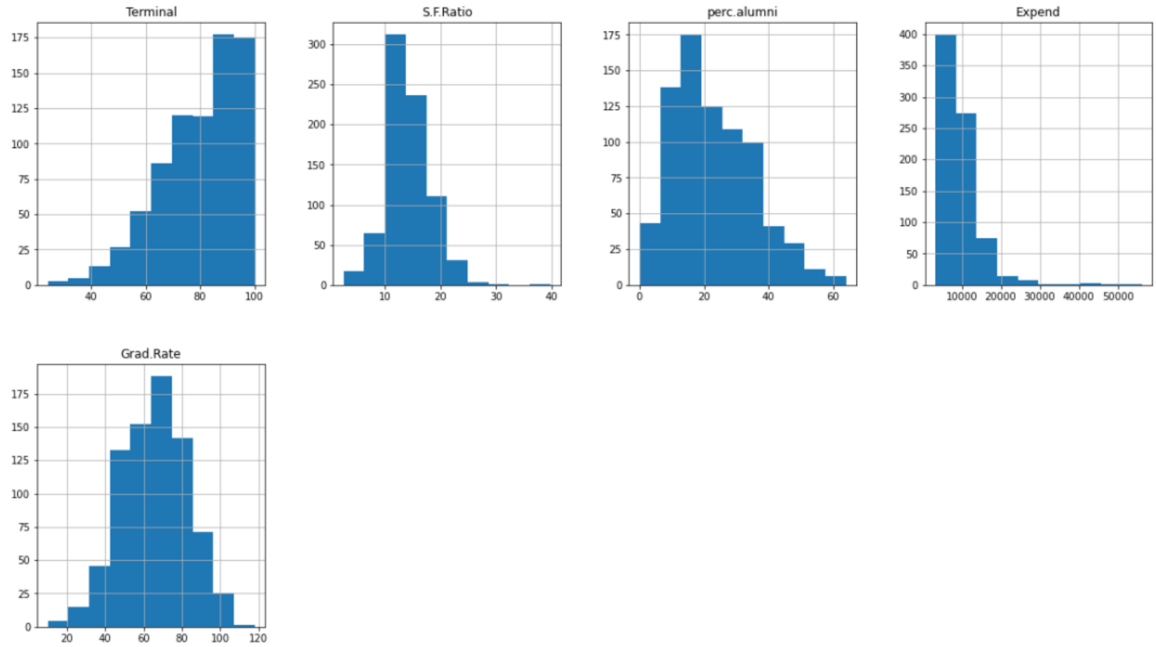


Fig.10

7.Boxplot:

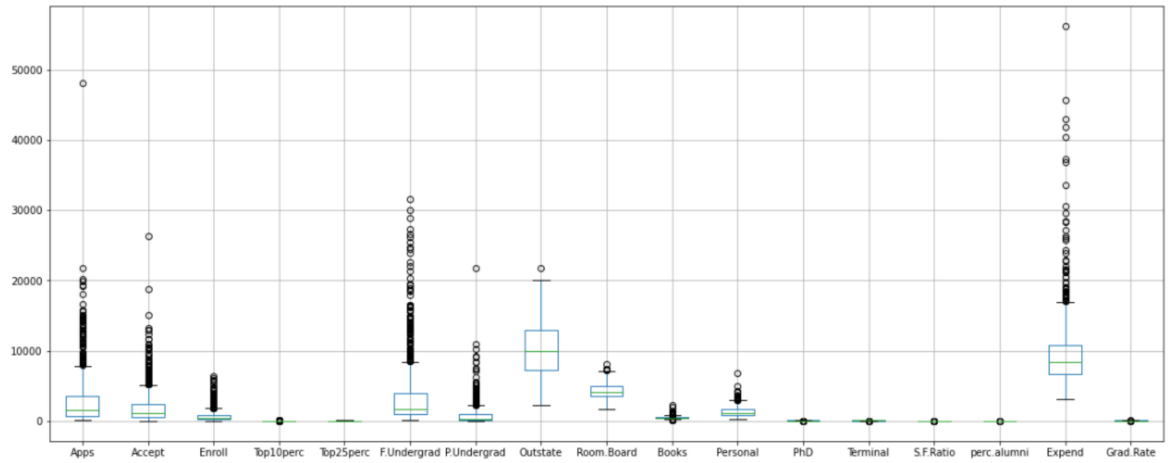


Fig.11

❖ Inferences from this summary.

- Total number of entries = 777
- Total number of columns = 18
- Number of null values = 0
- Data types encountered = float64(1) int64(16), object(1)
- From statistical summary we can see that the data is not normally distributed. We can see that the mean and median are significantly different from each other
- No duplicate values found
- Dropped Names column and created new data frame
- Skewness calculated: lack of symmetry
- Kurtosis calculated: Heavily tailed data concentrated at certain point
- large number of outliers found in the data
- Apps (Right skewed with outliers present in the data)
- Accept (Right skewed with outliers present in the data)
- Enroll (Right skewed with outliers present in the data)
- Top10perc (Slightly Right skewed with outliers present in the data)
- Top25perc (No outliers present in the data)
- F.Undergrad (Right skewed with outliers present in the data)
- P.Undergrad (Slightly Right skewed with outliers present in the data)
- Outsate (Slightly Right skewed with outliers present in the data)
- Room.Board (Slightly Right skewed with outliers present in the data)
- Books (Right skewed with outliers present in the data)
- Personal (Right skewed with outliers present in the data)
- PhD (Left skewed with outliers present in the data)
- Terminal (Left skewed with outliers present in the data)
- S.F.Ratio (Right skewed with outliers present in the data)
- Perc.alumni (Right skewed with outliers present in the data)
- Expend (Right skewed with outliers present in the data)
- Grad.Rate (outliers present in the data)

❖ <u>Multivariate Analysis:</u>

- Using a pairplot/heatmap for multivariate analysis to look at the covariance of the columns in the dataset.

❖ <u>Pair plot:</u>

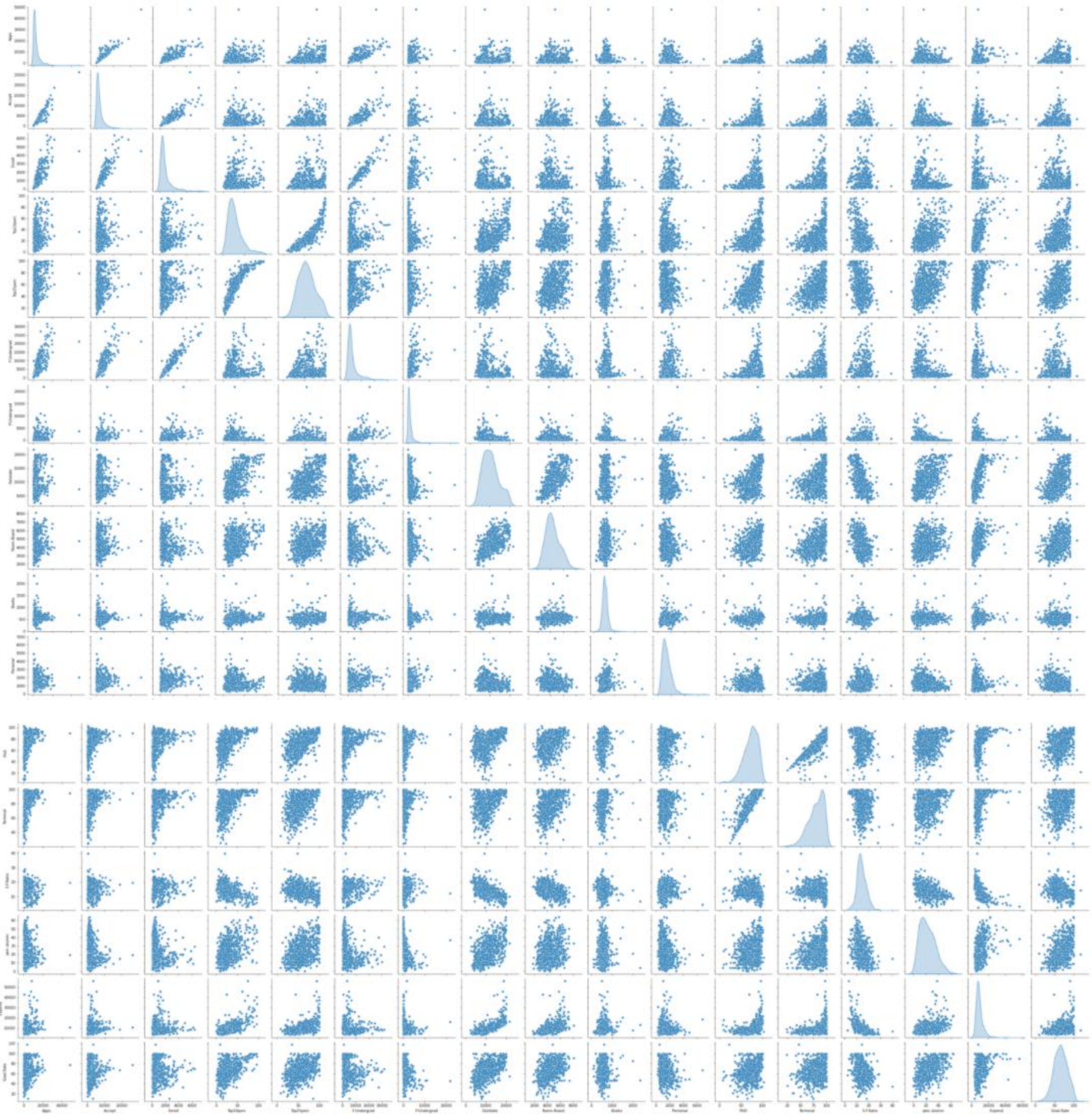


Fig.12

❖ <u>Heatmap:</u>

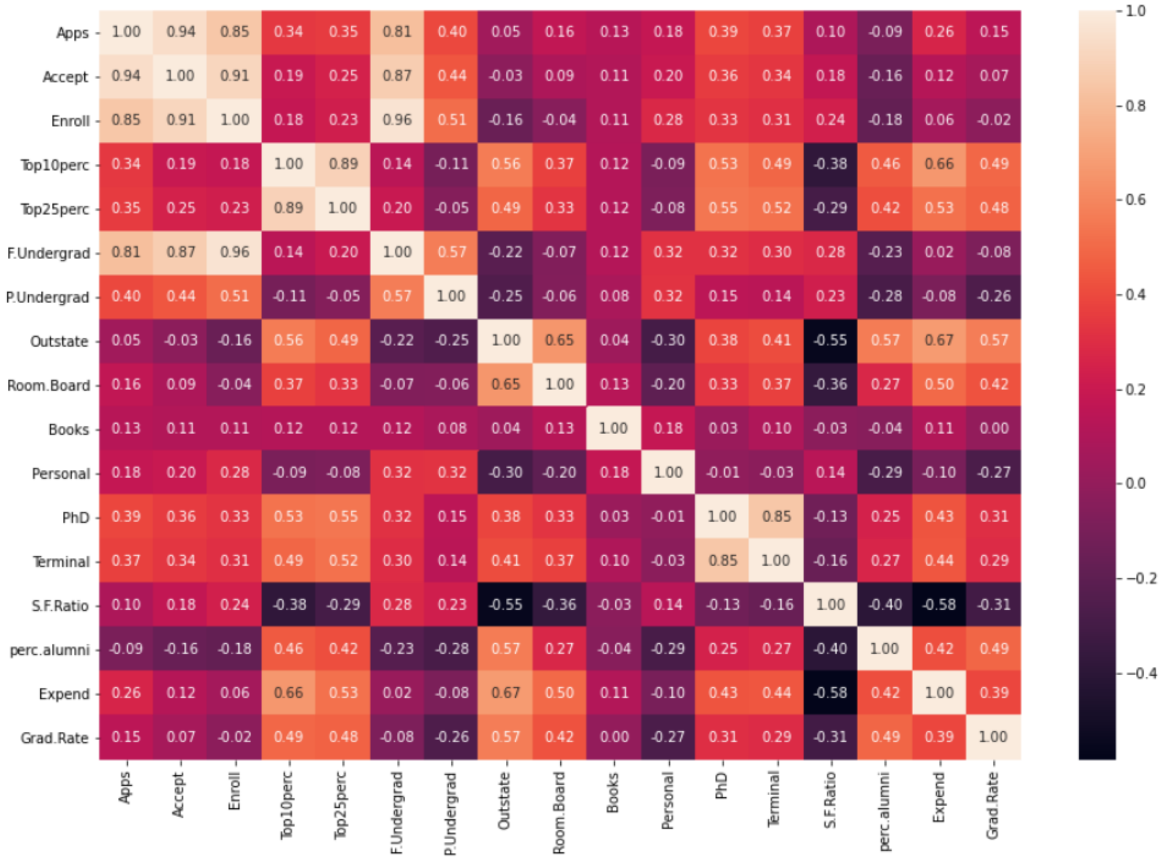


Fig.13

- Used pair plot for multiple pairwise bivariate distributions in a dataset which shows the relationship for (n, 2) combination of variable in a Data Frame as a matrix of plots and the diagonal plots are the univariate plots.
- Used heat map, where square shows the correlation between the variables on each axis.
- Correlation ranges from -1 to +1. Values closer to zero indicate that there is no linear trend between the two variables.
- Closer to 1 the correlation is, more positively correlated are the variables that is as one increases so does the other.
- A correlation closer to -1 is similar, but instead of both increasing one variable will decrease as the other increases

For example:

- Apps – Accept (Strong correlation)
- Enroll – F.undergrad(Strong correlation)
- Top10perc – Top25perc(Strong correlation)
- Phd – terminal (Strong correlation) etc.

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

- Scaling also called as centring is very important for PCA because of the way that the principal components are calculated
- Yes, scaling is necessary for PCA in this case.
- PCA is the solution to a data compression problem,
- where "error" is quantified by total error variance.
- Variables in different units must be scaled.
- Variables in the same units but with very different variances
- are usually scaled.
- Simplest scaling: divide each variable by its standard deviation ⇒ covariances are correlations.
- In other words: use eigen structure of correlation matrix R,not covariance matrix Σ.
- It calculates a new projection of the dataset and the axis for this is dependent on the standard deviation of the data.
- Scaling helps to standardize the data. The standard deviation of the data becomes 1 after scaling.
- At the end of this process, the data is in the form of an array.
- Hence, it is converted into a data frame and used for further analysis.

❖ Scaled data:

```
array([[-3.46881819e-01, -3.21205453e-01, -6.35089011e-02, ...,
        -8.67574189e-01, -5.01910084e-01, -3.18251941e-01],
       [-2.10884040e-01, -3.87029908e-02, -2.88584214e-01, ...,
        -5.44572203e-01,  1.66109850e-01, -5.51261842e-01],
       [-4.06865631e-01, -3.76317928e-01, -4.78121319e-01, ...,
         5.85934748e-01, -1.77289956e-01, -6.67766793e-01],
       ...,
       [-2.33895071e-01, -4.23771558e-02, -9.15087008e-02, ...,
        -2.21570217e-01, -2.56241250e-01, -9.59029170e-01],
       [ 1.99171118e+00,  1.77256262e-01,  5.78332661e-01, ...,
         2.12019418e+00,  5.88797079e+00,  1.95359460e+00],
       [-3.26765760e-03, -6.68715889e-02, -9.58163623e-02, ...,
         4.24433755e-01, -9.87115613e-01,  1.95359460e+00]])
```

Table.7

❖ <u>Scaled data converted into a data frame:</u>

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	-0.346882	-0.321205	-0.063509	-0.258583	-0.191827	-0.168116	-0.209207	-0.746356	-0.964905	-0.602312	1.270045	-0.163028	-0.115729
1	-0.210884	-0.038703	-0.288584	-0.655656	-1.353911	-0.209788	0.244307	0.457496	1.909208	1.215880	0.235515	-2.675646	-3.378176
2	-0.406866	-0.376318	-0.478121	-0.315307	-0.292878	-0.549565	-0.497090	0.201305	-0.554317	-0.905344	-0.259582	-1.204845	-0.931341
3	-0.668261	-0.681682	-0.692427	1.840231	1.677612	-0.658079	-0.520752	0.626633	0.996791	-0.602312	-0.688173	1.185206	1.175657
4	-0.726176	-0.764555	-0.780735	-0.655656	-0.596031	-0.711924	0.009005	-0.716508	-0.216723	1.518912	0.235515	0.204672	-0.523535

S.F.Ratio	perc.alumni	Expend	Grad.Rate
1.013776	-0.867574	-0.501910	-0.318252
-0.477704	-0.544572	0.166110	-0.551262
-0.300749	0.585935	-0.177290	-0.667767
-1.615274	1.151188	1.792851	-0.376504
-0.553542	-1.675079	0.241803	-2.939613

Table.8

❖ <u>Statistical summary of the scaled dataset:</u>

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books
count	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
mean	6.355797e-17	6.774575e-17	-5.249269e-17	-2.753232e-17	-1.546739e-16	-1.661405e-16	-3.029180e-17	6.515595e-17	3.570717e-16	-2.192583e-16
std	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
min	-7.551337e-01	-7.947645e-01	-8.022728e-01	-1.506526e+00	-2.364419e+00	-7.346169e-01	-5.615022e-01	-2.014878e+00	-2.351778e+00	-2.747779e+00
25%	-5.754408e-01	-5.775805e-01	-5.793514e-01	-7.123803e-01	-7.476067e-01	-5.586426e-01	-4.997191e-01	-7.762035e-01	-6.939170e-01	-4.810994e-01
50%	-3.732540e-01	-3.710108e-01	-3.725836e-01	-2.585828e-01	-9.077663e-02	-4.111378e-01	-3.301442e-01	-1.120949e-01	-1.437297e-01	-2.992802e-01
75%	1.609122e-01	1.654173e-01	1.314128e-01	4.221134e-01	6.671042e-01	6.294077e-02	7.341765e-02	6.179271e-01	6.318245e-01	3.067838e-01
max	1.165867e+01	9.924816e+00	6.043678e+00	3.882319e+00	2.233391e+00	5.764674e+00	1.378992e+01	2.800531e+00	3.436593e+00	1.085230e+01

Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02	7.770000e+02
4.765243e-17	5.954768e-17	-4.481615e-16	-2.057556e-17	-6.022638e-17	1.213101e-16	3.886495e-16
1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00	1.000644e+00
-1.611860e+00	-3.962596e+00	-3.785982e+00	-2.929799e+00	-1.836580e+00	-1.240641e+00	-3.230876e+00
-7.251203e-01	-6.532948e-01	-5.915023e-01	-6.546598e-01	-7.868237e-01	-5.574826e-01	-7.260193e-01
-2.078552e-01	1.433889e-01	1.561419e-01	-1.237939e-01	-1.408197e-01	-2.458933e-01	-2.698956e-02
5.310950e-01	7.562224e-01	8.358184e-01	6.093067e-01	6.666852e-01	2.241735e-01	7.302926e-01
8.068387e+00	1.859323e+00	1.379560e+00	6.499390e+00	3.331452e+00	8.924721e+00	3.060392e+00

Table.9

From “std” column we can conclude that the data has been scaled (1.000644e+00).

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

```
Covariance matrix
[[ 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]]
```

Table.10

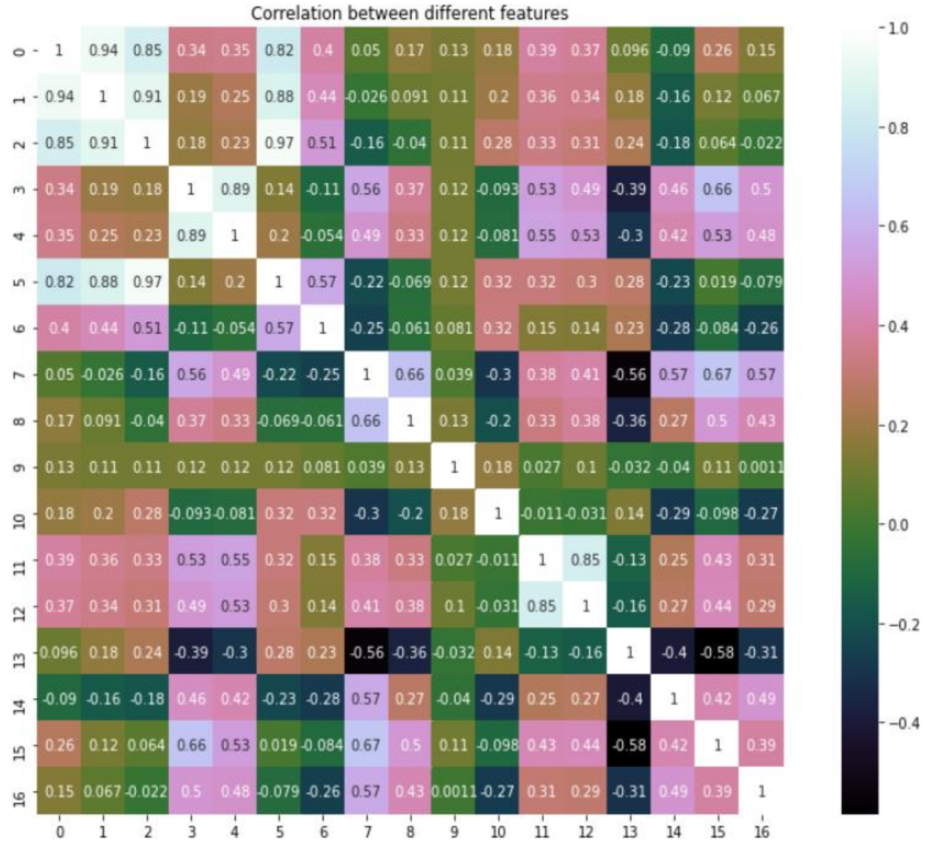


Fig.14

- A principal component analysis is used to reduce the dimensionality of large data sets. An eigen decomposition is performed on the covariance matrix to perform principal component analysis.
- Correlation and Covariance both measure only the linear relationships between two variables. This means that when the correlation coefficient is zero, the covariance is also zero. Both correlation and covariance measures are also unaffected by the change in location.
- However, when it comes to making a choice between covariance vs correlation to measure relationship between variables, correlation is preferred over covariance because it does not get affected by the change in scale.
- The covariance matrix for this data indicates the extent to which two variables of the data change alongside each other.
- Whereas the correlation matrix shows how strongly the variables of the data are related to each other.
- Covariance matrix shows the linear relationship between any two chosen variables from the dataset.
- Correlation matrix measures the linear relationship as well as the strength of this linear relationship. In the covariance matrix, the diagonal elements represent the variances and the other elements indicate the covariance.

- The variance is measured within the dimensions and the covariance is measured among the dimensions.
- Correlation matrix is a scaled form of the covariance matrix. The variances are standardized.

From the data we can say that:

- The avg number of applications received by the listed universities is around 3,001.
- The number of applications accepted ranges from 72 to 26,330.
- Average student enrolment is around 779.
- Median of new students from top 10%,25% of higher secondary class is 23% ,54% respectively.
- Average F.Undergrad is around 3700.
- Average p.Undergrad is around 856.
- Outstate students ranges from 2340 to 2177.
- Average book cost is around 550
- The minimum S.F. ratio is around 2.5
- Average percentage of faculties with Ph.D.'s is 72.6
- Average Gradrate is around 66.

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

Before Scaling:

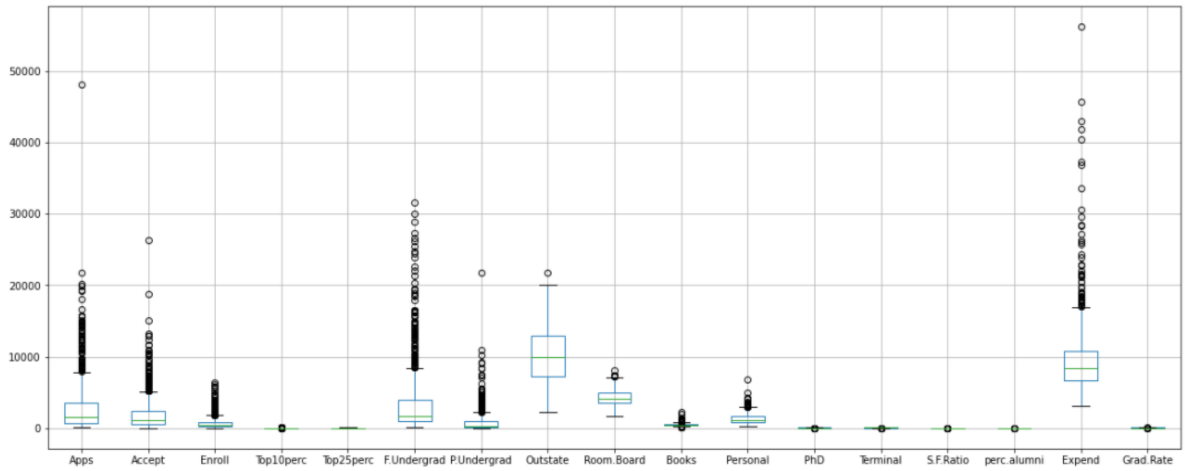


Fig.15

After Scaling:

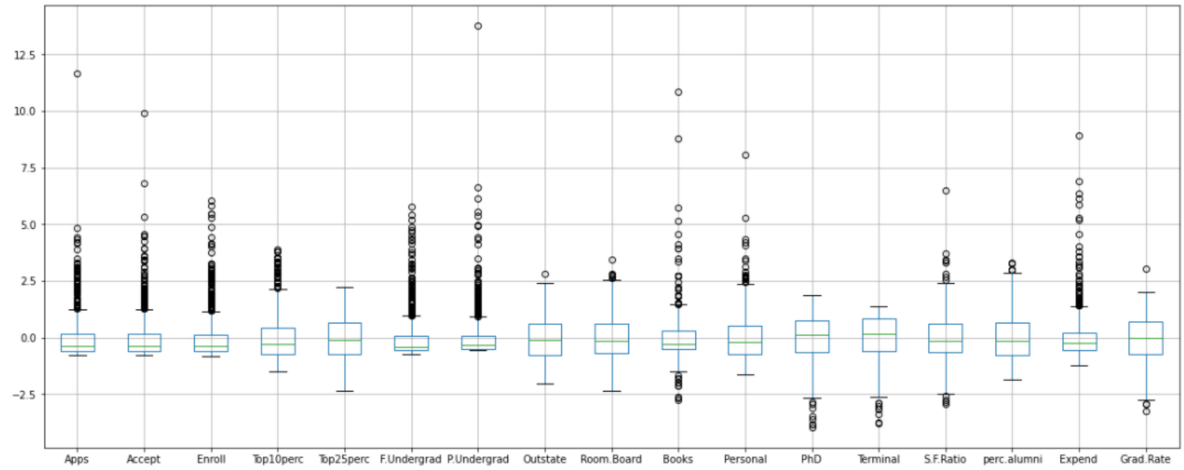


Fig.16

❖ Insights:

- Standard Scaler removes the mean and scales the data to unit variance.
- The scaling shrinks the range of the feature values as shown in the above figure.
- However, the outliers have an influence when computing the empirical mean and standard deviation.
- The outliers on each feature have different magnitudes, the spread of the transformed data on each feature is very different: most of the data lie in the high range for the transformed median income feature while the same data is squeezed in the smaller [-2, 4] range for the transformed number of data.
- Standard Scaler therefore cannot guarantee balanced feature scales in the presence of outliers.
- Outliers are present in the data both before and after scaling. However, after scaling, the median of all the columns is quite close to each other.

2.5 Extract the eigenvalues and eigenvectors. [print both]

- The covariance matrix is used to compute the eigenvectors and eigenvalues.
- The eigenvectors (principal components) determine the directions of the new feature space and the eigenvalues determine their magnitude.

- Python output for eigenvalues:

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

Table.11

- Python output for eigenvectors:

```
Number of Eigenvectors :  17
Eigenvectors
[[-2.48765602e-01  3.31598227e-01  6.30921033e-02 -2.81310530e-01
   5.74140964e-03  1.62374420e-02  4.24863486e-02  1.03090398e-01
   9.02270802e-02 -5.25098025e-02  3.58970400e-01 -4.59139498e-01
   4.30462074e-02 -1.33405806e-01  8.06328039e-02 -5.95830975e-01
   2.40709086e-02]
 [-2.07601502e-01  3.72116750e-01  1.01249056e-01 -2.67817346e-01
   5.57860920e-02 -7.53468452e-03  1.29497196e-02  5.62709623e-02
   1.77864814e-01 -4.11400844e-02 -5.43427250e-01  5.18568789e-01
  -5.84055850e-02  1.45497511e-01  3.34674281e-02 -2.92642398e-01
  -1.45102446e-01]
 [-1.76303592e-01  4.03724252e-01  8.29855709e-02 -1.61826771e-01
  -5.56936353e-02  4.25579803e-02  2.76928937e-02 -5.86623552e-02
   1.28560713e-01 -3.44879147e-02  6.09651110e-01  4.04318439e-01
  -6.93988831e-02 -2.95896092e-02 -8.56967180e-02  4.44638207e-01
   1.11431545e-02]
 [-3.54273947e-01 -8.24118211e-02 -3.50555339e-02  5.15472524e-02
  -3.95434345e-01  5.26927980e-02  1.61332069e-01  1.22678028e-01
  -3.41099863e-01 -6.40257785e-02 -1.44986329e-01  1.48738723e-01
  -8.10481404e-03 -6.97722522e-01 -1.07828189e-01 -1.02303616e-03
   3.85543001e-02]
 [-3.44001279e-01 -4.47786551e-02  2.41479376e-02  1.09766541e-01
  -4.26533594e-01 -3.30915896e-02  1.18485556e-01  1.02491967e-01
  -4.03711989e-01 -1.45492289e-02  8.03478445e-02 -5.18683400e-02
  -2.73128469e-01  6.17274818e-01  1.51742110e-01 -2.18838802e-02
  -8.93515563e-02]
 [-1.54640962e-01  4.17673774e-01  6.13929764e-02 -1.00412335e-01
  -4.34543659e-02  4.34542349e-02  2.50763629e-02 -7.88896442e-02
   5.94419181e-02 -2.08471834e-02 -4.14705279e-01 -5.60363054e-01
  -8.11578181e-02 -9.91640992e-03 -5.63728817e-02  5.23622267e-01
   5.61767721e-02]
 [-2.64425045e-02  3.15087830e-01 -1.39681716e-01  1.58558487e-01
   3.02385408e-01  1.91198583e-01 -6.10423460e-02 -5.70783816e-01
  -5.60672902e-01  2.23105808e-01  9.01788964e-03  5.27313042e-02
   1.00693324e-01 -2.09515982e-02  1.92857500e-02 -1.25997650e-01
  -6.35360730e-02]
 -
```

```
[-2.94736419e-01 -2.49643522e-01 -4.65988731e-02 -1.31291364e-01
  2.22532003e-01  3.00003910e-02 -1.08528966e-01 -9.84599754e-03
  4.57332880e-03 -1.86675363e-01  5.08995918e-02 -1.01594830e-01
  1.43220673e-01 -3.83544794e-02 -3.40115407e-02  1.41856014e-01
 -8.23443779e-01]
[-2.49030449e-01 -1.37808883e-01 -1.48967389e-01 -1.84995991e-01
  5.60919470e-01 -1.62755446e-01 -2.09744235e-01  2.21453442e-01
 -2.75022548e-01 -2.98324237e-01  1.14639620e-03  2.59293381e-02
 -3.59321731e-01 -3.40197083e-03 -5.84289756e-02  6.97485854e-02
  3.54559731e-01]
[-6.47575181e-02  5.63418434e-02 -6.77411649e-01 -8.70892205e-02
 -1.27288825e-01 -6.41054950e-01  1.49692034e-01 -2.13293009e-01
  1.33663353e-01  8.20292186e-02  7.72631963e-04 -2.88282896e-03
  3.19400370e-02  9.43887925e-03 -6.68494643e-02 -1.14379958e-02
 -2.81593679e-02]
[ 4.25285386e-02  2.19929218e-01 -4.99721120e-01  2.30710568e-01
 -2.22311021e-01  3.31398003e-01 -6.33790064e-01  2.32660840e-01
  9.44688900e-02 -1.36027616e-01 -1.11433396e-03  1.28904022e-02
 -1.85784733e-02  3.09001353e-03  2.75286207e-02 -3.94547417e-02
 -3.92640266e-02]
[-3.18312875e-01  5.83113174e-02  1.27028371e-01  5.34724832e-01
  1.40166326e-01 -9.12555212e-02  1.09641298e-03  7.70400002e-02
  1.85181525e-01  1.23452200e-01  1.38133366e-02 -2.98075465e-02
  4.03723253e-02  1.12055599e-01 -6.91126145e-01 -1.27696382e-01
  2.32224316e-02]
[-3.17056016e-01  4.64294477e-02  6.60375454e-02  5.19443019e-01
  2.04719730e-01 -1.54927646e-01  2.84770105e-02  1.21613297e-02
  2.54938198e-01  8.85784627e-02  6.20932749e-03  2.70759809e-02
 -5.89734026e-02 -1.58909651e-01  6.71008607e-01  5.83134662e-02
  1.64850420e-02]
[ 1.76957895e-01  2.46665277e-01  2.89848401e-01  1.61189487e-01
 -7.93882496e-02 -4.87045875e-01 -2.19259358e-01  8.36048735e-02
 -2.74544380e-01 -4.72045249e-01 -2.22215182e-03  2.12476294e-02
  4.45000727e-01  2.08991284e-02  4.13740967e-02  1.77152700e-02
 -1.10262122e-02]
```

```
[-2.05082369e-01 -2.46595274e-01  1.46989274e-01 -1.73142230e-02
 -2.16297411e-01  4.73400144e-02 -2.43321156e-01 -6.78523654e-01
  2.55334907e-01 -4.22999706e-01 -1.91869743e-02 -3.33406243e-03
 -1.30727978e-01  8.41789410e-03 -2.71542091e-02 -1.04088088e-01
  1.82660654e-01]
[-3.18908750e-01 -1.31689865e-01 -2.26743985e-01 -7.92734946e-02
  7.59581203e-02  2.98118619e-01  2.26584481e-01  5.41593771e-02
  4.91388809e-02 -1.32286331e-01 -3.53098218e-02  4.38803230e-02
  6.92088870e-01  2.27742017e-01  7.31225166e-02  9.37464497e-02
  3.25982295e-01]
[-2.52315654e-01 -1.69240532e-01  2.08064649e-01 -2.69129066e-01
 -1.09267913e-01 -2.16163313e-01 -5.59943937e-01  5.33553891e-03
 -4.19043052e-02  5.90271067e-01 -1.30710024e-02  5.00844705e-03
  2.19839000e-01  3.39433604e-03  3.64767385e-02  6.91969778e-02
  1.22106697e-01]]
```

Table.12

- Eigenvector is a vector that does not change when a transformation is applied to it, except that it becomes a scaled version of the original vector.
- It can help us calculating an approximation of a large matrix as a smaller vector.
- Eigenvectors are used to make linear transformation understandable.
- Eigenvalue: The scalar that is used to transform (stretch) an Eigenvector.
- Eigenvalues and Eigenvectors have their importance in linear differential equations where we want to find a rate of change or when we want to maintain relationships between two variables.
- They are used to reduce dimension space and also used in regularisation and they can be used to prevent overfitting.
- Since the value of the first eigenvalue is the greatest, we can see that the first principal component captures the maximum variance

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

STEP 1: STANDARDIZATION

- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

STEP 2: COVARIANCE MATRIX COMPUTATION

- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because sometimes, variables are highly correlated in such a way that they contain redundant information.
- So, in order to identify these correlations, we compute the covariance matrix.

STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the principal components of the data.

STEP 4: FEATURE VECTOR

- As we saw in the previous step, computing the eigenvectors and ordering them by their eigenvalues in descending order, allow us to find the principal components in order of significance.
- In this step, we choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), and form with the remaining ones a matrix of vectors that we call Feature vector.

STEP 5: FEATURE SPACE

- Export the data of the Principal Component (eigenvectors) into a data frame with the original features

We can perform PCA on the scaled data set by importing PCA from sklearn.decomposition. We get following component output:

```
array([[-1.59285540e+00, -2.19240180e+00, -1.43096371e+00, ...,
        -7.32560596e-01,  7.91932735e+00, -4.69508066e-01],
       [ 7.67333510e-01, -5.78829984e-01, -1.09281889e+00, ...,
        -7.72352397e-02, -2.06832886e+00,  3.66660943e-01],
       [-1.01073537e-01,  2.27879812e+00, -4.38092811e-01, ...,
        -4.05641899e-04,  2.07356368e+00, -1.32891515e+00],
       ...,
       [ 4.85927473e-02,  9.65153628e-01,  6.40659846e-01, ...,
         3.00443253e-01,  9.24891818e-01, -1.20894825e+00],
       [ 3.99747188e-01, -2.12508888e-01, -1.54993325e-01, ...,
        -4.71930880e-01,  2.24220692e+00,  2.07168553e-01],
       [-8.96897315e-02,  9.72392862e-02, -3.44730716e-01, ...,
         4.48231630e-01,  1.36325139e+00,  7.85627715e-01]])
```

Table.13

An array of cumsum of var_exp:

```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```

Table.14

Scree plot: Plot of eigen values:

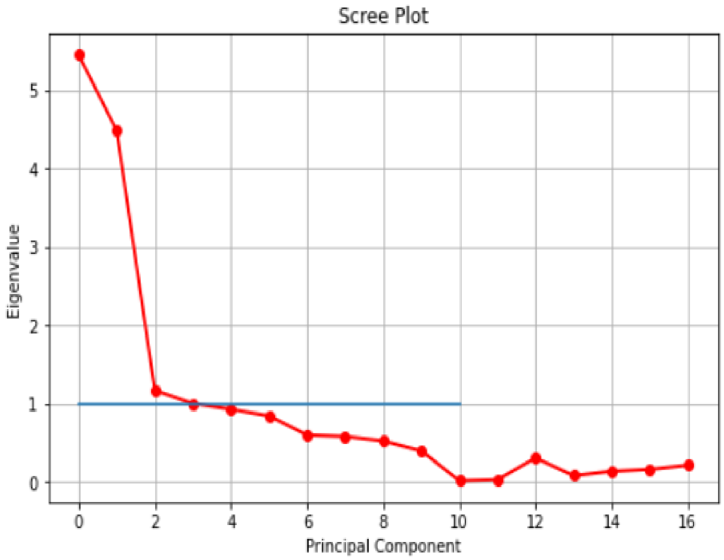


Fig.17

PCA Explained Variance: (Taking up to PCA 8 decending order):

```
array([5.45052162, 4.48360686, 1.17466761, 1.00820573, 0.93423123,
       0.84849117, 0.6057878 , 0.58787222])
```

Table.15

Exporting the data of the Principal Component (eigenvectors) into a data frame with the original features:

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal
0	0.248766	0.207602	0.176304	0.354274	0.344001	0.154641	0.026443	0.294736	0.249030	0.064758	-0.042529	0.318313	0.317056
1	0.331598	0.372117	0.403724	-0.082412	-0.044779	0.417674	0.315088	-0.249644	-0.137809	0.056342	0.219929	0.058311	0.046429
2	-0.063092	-0.101249	-0.082986	0.035056	-0.024148	-0.061393	0.139682	0.046599	0.148967	0.677412	0.499721	-0.127028	-0.066038
3	0.281311	0.267817	0.161827	-0.051547	-0.109767	0.100412	-0.158558	0.131291	0.184996	0.087089	-0.230711	-0.534725	-0.519443
4	0.005741	0.055786	-0.055694	-0.395434	-0.426534	-0.043454	0.302385	0.222532	0.560919	-0.127289	-0.222311	0.140166	0.204720
5	-0.016237	0.007535	-0.042558	-0.052693	0.033092	-0.043454	-0.191199	-0.030000	0.162755	0.641055	-0.331398	0.091256	0.154928
6	-0.042486	-0.012950	-0.027693	-0.161332	-0.118486	-0.025076	0.061042	0.108529	0.209744	-0.149692	0.633790	-0.001096	-0.028477
7	-0.103090	-0.056271	0.058662	-0.122678	-0.102492	0.078890	0.570784	0.009846	-0.221453	0.213293	-0.232661	-0.077040	-0.012161

S.F.Ratio	perc.alumni	Expend	Grad.Rate
-0.176958	0.205082	0.318909	0.252316
0.246665	-0.246595	-0.131690	-0.169241
-0.289848	-0.146989	0.226744	-0.208065
-0.161189	0.017314	0.079273	0.269129
-0.079388	-0.216297	0.075958	-0.109268
0.487046	-0.047340	-0.298119	0.216163
0.219259	0.243321	-0.226584	0.559944
-0.083605	0.678524	-0.054159	-0.005336

Table.16

- Calculated the variance explained and the cumulative variance explained. This helps to decide the number of principal components required.
- From the scree plot it is observed that 8 components are sufficient. From the cumulative variance explained computed before, we see that 8 components cover 88% of the variance of the data.
- The pca components are then calculated using sklearn.
- The array seen in the above image is converted into a data frame using pandas. The helps us to better understand the data.

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]

- PCA aims to fit straight lines to the data points. We call these straight lines "principal components".
- There are as many principal components as there are variables. The first principal component is the best straight line we can fit to the data.
- The second principal component is the best straight line we can fit to the errors from the first principal component.
- The third principal component is the best straight line we can fit to the errors from the first and second principal components, etc.

linear equation of PC in terms of eigenvectors and corresponding features:

```
The Linear eq of 1st component:
0.25 * Apps + 0.21 * Accept + 0.18 * Enroll + 0.35 * Top10perc + 0.34 * Top25perc + 0.15 * F.Undergrad + 0.03 * P.Undergrad
+ 0.29 * Outstate + 0.25 * Room.Board + 0.06 * Books + -0.04 * Personal + 0.32 * PhD + 0.32 * Terminal + -0.18 * S.F.Ratio +
0.21 * perc.alumni + 0.32 * Expend + 0.25 * Grad.Rate +
```

Fig.18

The linear eq of 1st component:

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

2.8 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?

- To compute the cumulative variance experienced, the variance experienced is first calculated using the eigen values.
- The cumulative variance explained is then computed. Below is the python output of cumulative variance experienced:

```
array([ 32.0206282 ,  58.36084263,  65.26175919,  71.18474841,
        76.67315352,  81.65785448,  85.21672597,  88.67034731,
        91.78758099,  94.16277251,  96.00419883,  97.30024023,
        98.28599436,  99.13183669,  99.64896227,  99.86471628,
       100.        ])
```

Fig.19

- Eigenvalue associated with each principal component tells us how much variation in the data set
- They are usually expressed as a percentage of the total variation in the data set.
- Eigenvectors are just the linear combinations of the original variables (in the simple or rotated factor space); they described how variables "contribute" to each factor axis.
- PCA as way to construct new axes that point to the directions of maximal variance (in the original variable space), as expressed by the eigenvalue, and how variables contributions are weighted or linearly transformed in this new space.
- By looking at the cumulative variance experienced, we can understand how much variance is captured by particular number of principal components.
- For example, if we wanted to work with 91% variance captured, the number of components will be 9.
- Similarly, we can decide upon what percentage of variance we want to work with and choose the number of components accordingly.
- The eigenvectors determine the directions of the new feature space. They indicate the principal components.

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]

- Initially we performed dimension reduction (No. of features/Attributes/variables/columns have been reduced)
- Dropped the least important variables by ensuring no loss of crucial information (Names)
- To have less but more informative attributes
- To get rid od redundant or highly correlated attributes
- Less attributes means: Less storage space, Less computation time, Easier training of a model
- Curse of dimensionality is the most important reason to go for dimensionality reduction.
- Number of features increases: Amount of data we need increases exponentially too.
- Because we need to have all combinations of attribute values to be considered in our dataset

Business implication of using Principal Component Analysis for this case study:

- In this particular case study, we deal with 8 numerical columns.

Fig 20

Following are the interpretations from the obtained PC's:

- PC1: Shows the no. of students for whom the particular university is Out-of-state and instructional expenditure per student
- PC2: Represents the highly correlated variables such as Apps, Enroll,F.undergrad and Accept
- PC3: Highlights the estimated cost of books for a student
- PC4: Represents the average no. of faculties with Ph.D.'s and terminal degree
- PC5: Explains percentage of new students from top 10% and 25% of higher secondary class including cost of room and board
- PC6: info about student-faculty ratio
- PC7: Highlights estimated personal spending for a student and graduation rate
- PC8: Explains number of part-time undergraduate students and alumni who donate
- This is a high dimensional data. When it comes to high dimensional data, it is usually difficult to recognise and interpret patterns.
- This renders it difficult to work with data and gain any kind of insights. The way PCA works is that based on the original data, it calculates a set of variables that describe as much variance as possible in the data.
- For example, in this case study, the first principal component captures 32.02% of the variance in the data, followed by the second principal component, which captures 26.34% of the variance in the data.
- Depending on how much variance in the data we want to work with, we can choose the number of principal components. In the case study at hand, we reduce the dimensions from 17 to 8.

- PCA’s key advantages are its low noise sensitivity, the decreased requirements for capacity and memory, and increased efficiency given the processes taking place in a smaller dimensions.
- Lack of redundancy of data given the orthogonal components.
- Reduced complexity in original data grouping with the use of PCA
- Smaller database representation stored in the form of their projections on a reduced basis.