



# ADVANCE STATISTICS PROJECT

M.K. SUGANTHE RAMYA





## **CASE STUDY 1: ANOVA ANALYSIS OF HAY FEVER**



## Table of contents

### Hay Fever

• Project objective .....	5
• Assumptions .....	5
• Exploratory data analysis	
i. Univariate analysis .....	7
ii. Hours of Relief with A compound at three levels .....	7
iii. Hours of Relief with B compound at three levels .....	8
• ANOVA analysis	
i. State Null and alternative hypothesis .....	8
ii. One-way ANOVA analysis	
a) One- way ANOVA analysis of A compound .....	9
b) One- way ANOVA analysis of B compound .....	9
iii. Interaction plot between A and B compounds .....	10
iv. Two-way ANOVA analysis between A and B compounds .....	11
v. The business implications of performing ANOVA for this case study .....	11
• Conclusion .....	12
• Reference .....	12

# HAY FEVER

## Hay Fever: An Infographic

### What is Hay Fever?

"An allergy caused by pollen or dust in which the mucous membranes of the eyes and nose are inflamed, causing running at the nose and watery eyes."

- English Oxford Dictionary

### Hay Fever Symptoms



### Top Tips: to Help with Hay Fever

- Don't cut grass or walk on grass.
- Shower and wash clothes after being outside.
- Don't keep fresh flowers inside the home.
- Don't smoke or be around tobacco smoke, it heightens symptoms.
- Don't dry clothes outside – pollen can stick to wet clothing.
- Try not to spend a lot of time outside.
- Don't let pets inside – they can carry pollen indoors.
- During the summer months continue taking medication on days with no symptoms.
- Buy a high quality fan to keep cool rather than sleeping with the window open.
- Use a wet flannel and place over eyes to relieve irritation.

### Top Medicines to Help with Hay Fever



**Allergy Tablets**  
Chlorphenamine (Piriton) tablets are commonly used for quick relief from hay fever. Non-Drowsy products like Cetirizine (Zirtek, Piriteze) and Loratadine (Claritin) provide all day protection.

**Adult Nasal Drops**  
Ortivine 10ml.



**Eye Drops**  
Sodium Cromoglycate (Opticrom) – regular use to prevent and treat hayfever. A preservative free version is also available for people who wear contact lenses.

**Nasal Spray**  
For adults and children, Beclomethasone (Beconase) helps reduce inflammation in the sinuses.



This infographic was bought to you by Travelpharm – an independent private pharmacy. For more advice, or to learn about our services and products head over to our website. [www.travelpharm.com](http://www.travelpharm.com)



**Sources**  
[nhs.co.uk](http://nhs.co.uk)  
[avogel.co.uk](http://avogel.co.uk)  
[en.oxforddictionaries.com](http://en.oxforddictionaries.com)

 [Travelpharm.com](http://Travelpharm.com)

## Problem 1:

A research laboratory was developing a new compound for the relief of severe cases of hay fever. In an experiment with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each. Randomization was used in assigning four volunteers to each of the nine treatments. The data on hours of relief can be found in the following .csv file: Fever.csv

### Project Objective:

The Objective of the report is to explore the “FEVER.csv” dataset in Python (JUPYTER NOTEBOOK) and generate insights about the dataset. This exploration report will consist of the following:

- ✚ Importing the dataset in jupyter notebook.
- ✚ Understanding the structure of dataset.
- ✚ Exploratory Data analysis
- ✚ Graphical exploration
- ✚ ANOVA analysis
- ✚ Insights from the dataset

### Assumptions:

ANOVA (known as Analysis of Variance) is a technique which is used to check whether the means of two or more sample groups are statistically different or not. In the Healthcare Industry, we can use the ANOVA test to compare different medications and the effect on patients. We are going to use ANOVA analysis on A and B compounds to determine which new compound give the relief for severe cases of hay fever.

### Exploratory Data Analysis:

#### Fever dataset

	A	B	Volunteer	Relief
0	1	1	1	2.4
1	1	1	2	2.7
2	1	1	3	2.3
3	1	1	4	2.5
4	1	2	1	4.6
5	1	2	2	4.2
6	1	2	3	4.9
7	1	2	4	4.7
8	1	3	1	4.8
9	1	3	2	4.5



### Information on Fever dataset:

```
<class 'pandas.core.frame.DataFrame'>
```

**Range Index: 36 entries, 0 to 35**

Data columns (total 4 columns):

#	Column	Non-Null Count	Dtype
0	A	36 non-null	int64
1	B	36 non-null	int64
2	Volunteer	36 non-null	int64
3	Relief	36 non-null	float64

dtypes: float64(1), int64(3)

memory usage: 1.2 KB

### Inference:

- Total number of entries = 36
- Total number of columns = 4
- There are no null values
- There are 3 Int64 and 1 float64 dtypes

### Summary of Relief:

count	36.000000
mean	7.183333
std	3.272090
min	2.300000
25%	4.675000
50%	6.000000
75%	9.325000
max	13.500000

Name: Relief, dtype: float64

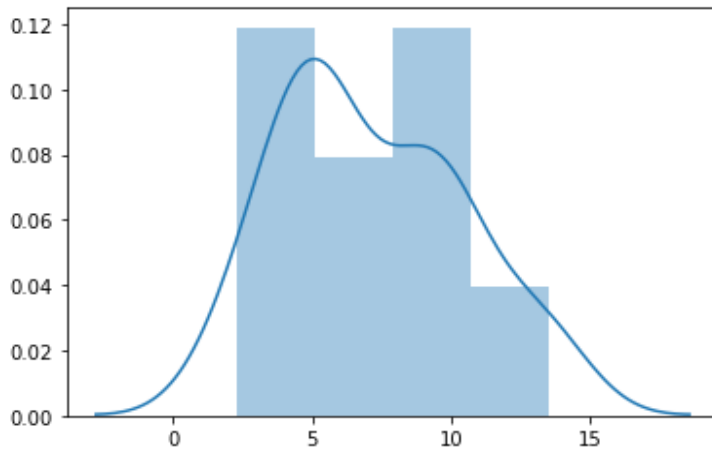
### Inference

- Minimum hours of relief are 2.3 hours
- Maximum hours of relief are 13.5 hours
- Average hours of relief are 7.18 hours

Before Starting analysis, we need to convert A, B and Volunteer columns into categorical variable. It will be helpful for analysis

## Univariate Analysis:

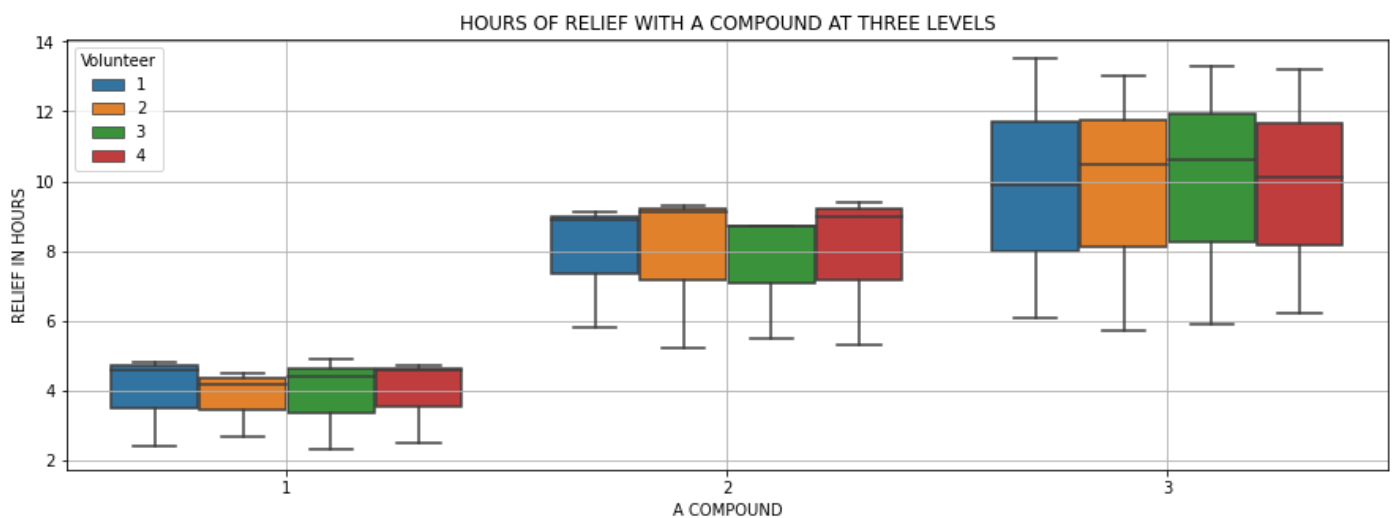
### Distribution plot of Relief:



### Inference:

Slightly left skewed

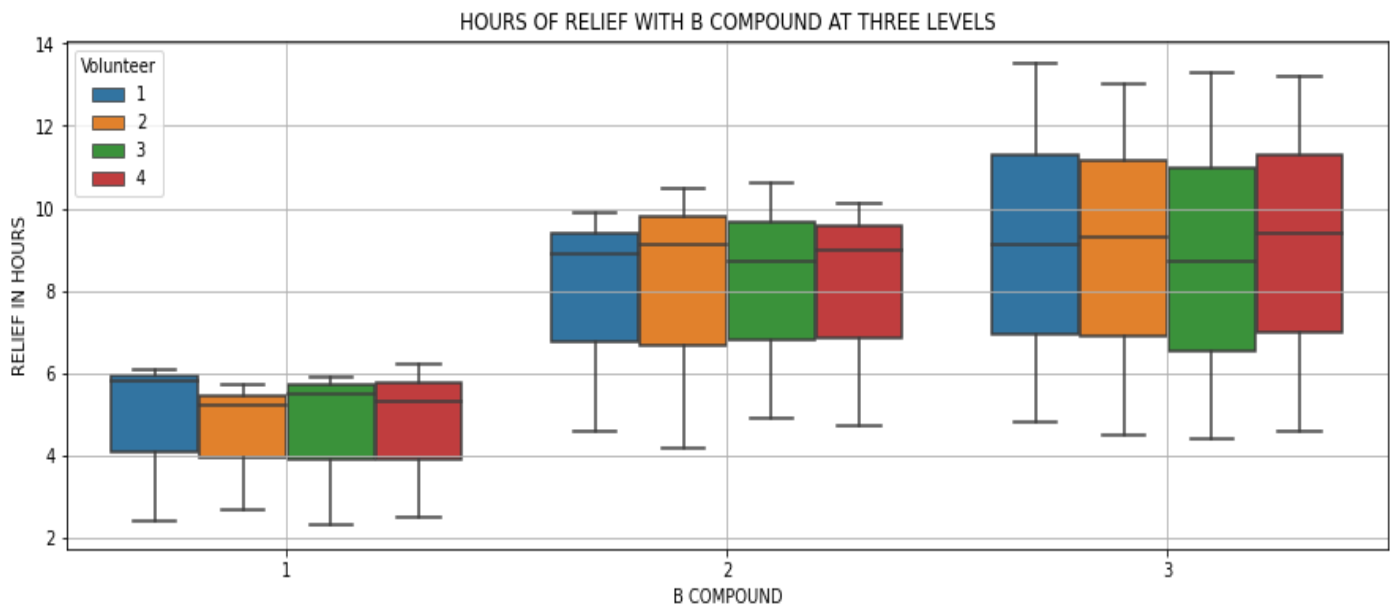
### HOURS OF RELIEF WITH A COMPOUND AT THREE LEVELS:



### INFERENCE:

- Level 1 treatment of A compound seems to give relief within 2.3 - 5 hours for the volunteers
- Level 2 treatment of A compound seems to give relief within 5 - 9 hours for the volunteers
- Level 3 treatment of A compound seems to give relief within 6 - 13.5 hours but taking more time than other two levels

## HOURS OF RELIEF WITH B COMPOUND AT THREE LEVELS:



### INFERENCE:

- Level 1 treatment of B compound seems to give relief within 2.3 - 6 hours for the volunteers
- Level 2 treatment of B compound seems to give relief within 4 - 11 hours for the volunteers
- Level 3 treatment of B compound seems to give relief within 4 - 13.5 hours but taking more time than other two levels

### 1.1) State the Null and Alternate Hypothesis for conducting one-way ANOVA for both the variables 'A' and 'B' individually.

#### A COMPOUND

$H_0$ : The means of 'Relief' variable with respect to each level of 'A' is equal.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k$

$H_1$ : At least one of the means of 'Relief' variable with respect to each level of A is unequal

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots \neq \mu_k$  (Not all the means are equal)

#### B COMPOUND

$H_0$ : The means of 'Relief' variable with respect to each level of 'B' is equal.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k$

$H_1$ : At least one of the means of 'Relief' variable with respect to each level of 'B' is unequal

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots \neq \mu_k$  (Not all the means are equal)



1.2) Perform one-way ANOVA for variable 'A' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

One-way ANOVA with the 'A' variable:

#### A COMPOUND

$H_0$ : The means of 'Relief' variable with respect to each level of 'A' is equal.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k$

$H_1$ : At least one of the means of 'Relief' variable with respect to each level of A is unequal

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots \neq \mu_k$  (Not all the means are equal)

	df	sum_sq	mean_sq	F	PR(>F)
C (A)	2.0	220.02	110.010000	23.465387	4.578242e-07
Residual	33.0	154.71	4.688182	NaN	NaN

#### Inference

- Since the p value in this scenario is less than  $\alpha < (0.05)$ , we can say that we reject the Null Hypothesis( $H_0$ ).
- The hours of relief vary at each level with A compound treatment

1.3) Perform one-way ANOVA for variable 'B' with respect to the variable 'Relief'. State whether the Null Hypothesis is accepted or rejected based on the ANOVA results.

One-way ANOVA with the 'B' variable

#### B COMPOUND

$H_0$ : The means of 'Relief' variable with respect to each level of 'B' is equal.

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \dots = \mu_k$

$H_1$ : At least one of the means of 'Relief' variable with respect to each level of 'B' is unequal

$H_1: \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \dots \neq \mu_k$  (Not all the means are equal)

	df	sum_sq	mean_sq	F	PR(>F)
C (B)	2.0	123.66	61.830000	8.126777	0.00135
Residual	33.0	251.07	7.608182	NaN	NaN

#### Inference

- Since the p value in this scenario is less than  $\alpha < (0.05)$ , we can say that we reject the Null Hypothesis ( $H_0$ ).
- The hours of relief vary at each level with B compound treatment

#### 1.4) Analyse the effects of one variable on another with the help of an interaction plot.

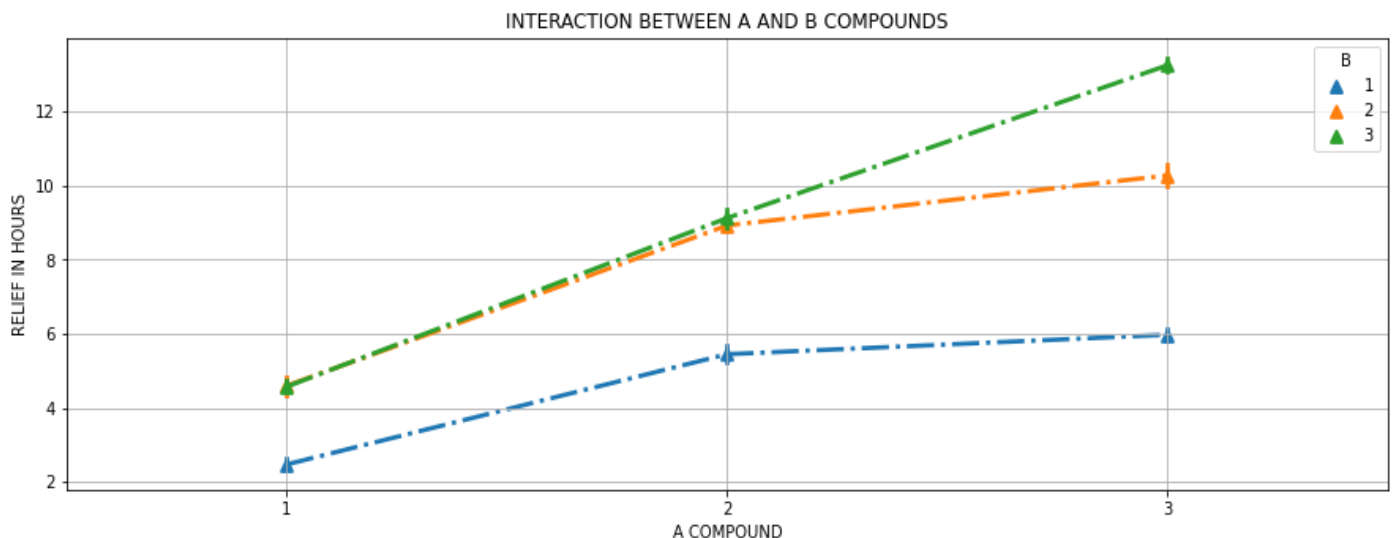
What is an interaction between two treatments?

[hint: use the 'point plot' function from the 'seaborn' function]

Formulate the hypothesis of ANOVA with both 'A' and 'B' variables with respect to the variable 'Relief'.

*H<sub>0</sub>: There is interaction between compounds A and B*

*H<sub>1</sub>: There is no interaction between compounds A and B*



#### Inference of interaction plot:

- An interaction effect can usually be a set of non-parallel lines.
- Based on plot we can interpret that there is no significant interaction between 1 and 2 level treatment in A and B compounds
- There seem to be slight interaction between 2nd and 3rd level treatment in A and B compounds

#### Interaction effect using ANOVA:

	df	sum_sq	mean_sq	F	PR(>F)
C (A)	2.0	220.020	110.010000	1827.858462	1.514043e-29
C (B)	2.0	123.660	61.830000	1027.329231	3.348751e-26
C (A) : C (B)	4.0	29.425	7.356250	122.226923	6.972083e-17
Residual	27.0	1.625	0.060185	NaN	NaN

#### Inference:

As A and B compound interaction is 0.0000000000000000697 which  $< 0.05$ , there seems to be no statistical interaction.

There is no interaction between A and B compounds

1.5) Perform a two-way ANOVA based on the different ingredients (variable 'A' & 'B') with the variable 'Relief' and state your results.

Formulate the hypothesis of ANOVA with both 'A' and 'B' variables with respect to the variable 'Relief'.

*H<sub>0</sub>*: The means of 'Relief' variable with respect to each level of 'A' category and 'B' category is equal.

*H<sub>1</sub>*: At least one of the means of 'Relief' variable with respect to each level of 'A' category and 'B' is unequal.

	df	sum_sq	mean_sq	F	PR(>F)
C (A)	2.0	220.020	110.010000	1827.858462	1.514043e-29
C (B)	2.0	123.660	61.830000	1027.329231	3.348751e-26
C (A) : C (B)	4.0	29.425	7.356250	122.226923	6.972083e-17
Residual	27.0	1.625	0.060185	NaN	NaN

#### Inference:

- Considering both the factors (A and B), Both A and B has a significant factor as the p value is <0.05. so, we can say that we reject the Null Hypothesis (*H<sub>0</sub>*).
- We can conclude that there is a statistically significant difference in the hours of relief between A and B medicines.
- There is no interaction between A and B compounds.

#### 1.6) Mention the business implications of performing ANOVA for this particular case study.

ANOVA (known as Analysis of Variance) is a technique which is used to check whether the means of two or more sample groups are statistically different or not.

Suppose in the Healthcare Industry, we can use the ANOVA test to compare different medications and the effect on patients. We can compare which medication works better for treatment and hence we able to choose the best one.

A research laboratory was developing a new compound for the relief of severe cases of hay fever. hay fever research experiment conducted with 36 volunteers, the amounts of the two active ingredients (A & B) in the compound were varied at three levels each.

After performing one-way ANOVA analysis on A and B separately, we can conclude that

#### A compound:

- Since the p value in this scenario is less than  $\alpha < (0.05)$ , we can say that we reject the Null Hypothesis (*H<sub>0</sub>*).
- The hours of relief vary at each level with A compound treatment

### **B compound:**

- Since the p value in this scenario is less than  $\alpha < (0.05)$ , we can say that we reject the Null Hypothesis ( $H_0$ ).
- The hours of relief vary at each level with B compound treatment

### **Interaction between A and B in the plot:**

- An interaction effect can usually be a set of non-parallel lines.
- Based on plot we can interpret that there is no significant interaction between 1 and 2 level treatment in A and B compounds
- There seem to be slight interaction between 2nd and 3rd level treatment in A and B compounds

### **Interaction effect between A and B components based on ANOVA analysis:**

1. As A and B compound interaction is 0.0000000000000000697 which  $< 0.05$ , there seems to be no statistical interaction.

2. There is no interaction between A and B compounds

### **Two-way ANOVA with A and B components:**

Considering both the factors (A and B), Both A and B has a significant factor as the p value is  $< 0.05$ . so, we can say that we reject the Null Hypothesis ( $H_0$ ).

We can conclude that there is a statistically significant difference in the hours of relief between A and B medicines

### **Conclusion:**

Based on ANOVA analysis of A and B compounds, we can conclude 3rd level of variation in A and B compounds gives more hours of relief for severe cases of hay fever. There is no significant interaction between A and B compounds. We can also conclude that, there is a statistically significant difference in the hours of relief between A and B medicines.

### **Recommendation:**

Patients who suffering from hay fever need more hours of relief. There are many on counter medicines itself gives 24 hours of relief. But A and B compounds gives only 10 -13 hours of relief which is significantly low. Research laboratory should work on efficiency of A and B compounds, so that they result in more hours of relief for the patients



## CASE STUDY 2:

### PRINCIPAL COMPONENT ANALYSIS ON VARIOUS INSTITUTIONS PARAMETERS



## Table of contents

• Project objective .....	15
• Assumptions .....	16
• Exploratory data analysis	
iv. Univariate analysis .....	17
v. Multivariate analysis .....	18
vi. Correlation .....	19
vii. Outlier Treatment .....	20
• Standardisation before processing Principal component analysis	
i. Z-Score standardisation .....	21
ii. MaxMinScaler .....	22
• Principal Component analysis	
i. Covariance Matrix .....	23
ii. Boxplot after Z-score scaling .....	24
iii. Covariance, Eigenvalues & Eigenvectors .....	25
iv. Scree plot .....	28
v. Cumulative variance .....	29
vi. Principal components .....	30
• Principal components analysis Interpretation	
i. First principal component (PCA_1) .....	31
ii. Second principal component (PCA_2) .....	31
iii. Third principal component (PCA_3) .....	31
iv. Fourth principal component (PCA_4) .....	31
• Conclusion .....	32
• Reference .....	32









## Problem 2:

The dataset Education - Post 12th Standard.csv is a dataset which contains the names of various colleges. This particular case study is based on various parameters of various institutions. You are expected to do Principal Component Analysis for this case study according to the instructions given in the following rubric. The data dictionary of the 'Education - Post 12th Standard.csv' can be found in the following file: Data Dictionary.xlsx.

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

## Objective:

The objective of the report is to explore the “ Education - Post 12th Standard.csv ” dataset in python (JUPYTER NOTEBOOK) and generate insights about the dataset. The exploration report will consist of the following:

-  Importing the dataset in jupyter notebook.
-  Understanding the structure of dataset
-  Exploratory Data Analysis
-  Graphical exploration
-  Outlier treatment
-  Scaling the dataset
-  Principal Component analysis step by step
-  Insights from the dataset

## Assumptions:

Principal component analysis or PCA is a dimensionality – reduction method that is often used to reduce the dimensionality of large set of variables into a smaller one that still contains most of the large set. Because smaller datasets are easier to explore and visualize and make analyzing data much easier and faster. In this dataset, Principal components analysis performed on different parameters of the various institutions and the data is explored.



## 2.1) Perform Exploratory Data Analysis [both univariate and multivariate analysis to be performed]. The inferences drawn from this should be properly documented.

### Exploratory Data Analysis:

#### Education - Post 12th Standard dataset:

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

Information of the dataset:

```
<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
```

#### Inference:

- This dataset contains 777 entries and 18 columns
- There is one object, one float and 16 int datatypes.
- There are no null values

## Summary of the dataset:

Column1	count	mean	std	min	25%	50%	75%	max
Apps	777	3001.638	3870.201	81	776	1558	3624	48094
Accept	777	2018.804	2451.114	72	604	1110	2424	26330
Enroll	777	779.973	929.1762	35	242	434	902	6392
Top10perc	777	27.55856	17.64036	1	15	23	35	96
Top25perc	777	55.79665	19.80478	9	41	54	69	100
F.Undergrad	777	3699.907	4850.421	139	992	1707	4005	31643
P.Undergrad	777	855.2986	1522.432	1	95	353	967	21836
Outstate	777	10440.67	4023.016	2340	7320	9990	12925	21700
Room.Board	777	4357.526	1096.696	1780	3597	4200	5050	8124
Books	777	549.381	165.1054	96	470	500	600	2340
Personal	777	1340.642	677.0715	250	850	1200	1700	6800
PhD	777	72.66023	16.32816	8	62	75	85	103
Terminal	777	79.7027	14.72236	24	71	82	92	100
S.F.Ratio	777	14.0897	3.958349	2.5	11.5	13.6	16.5	39.8
perc.alumni	777	22.74389	12.3918	0	13	21	31	64
Expend	777	9660.171	5221.768	3186	6751	8377	10830	56233
Grad.Rate	777	65.46332	17.17771	10	53	65	78	118

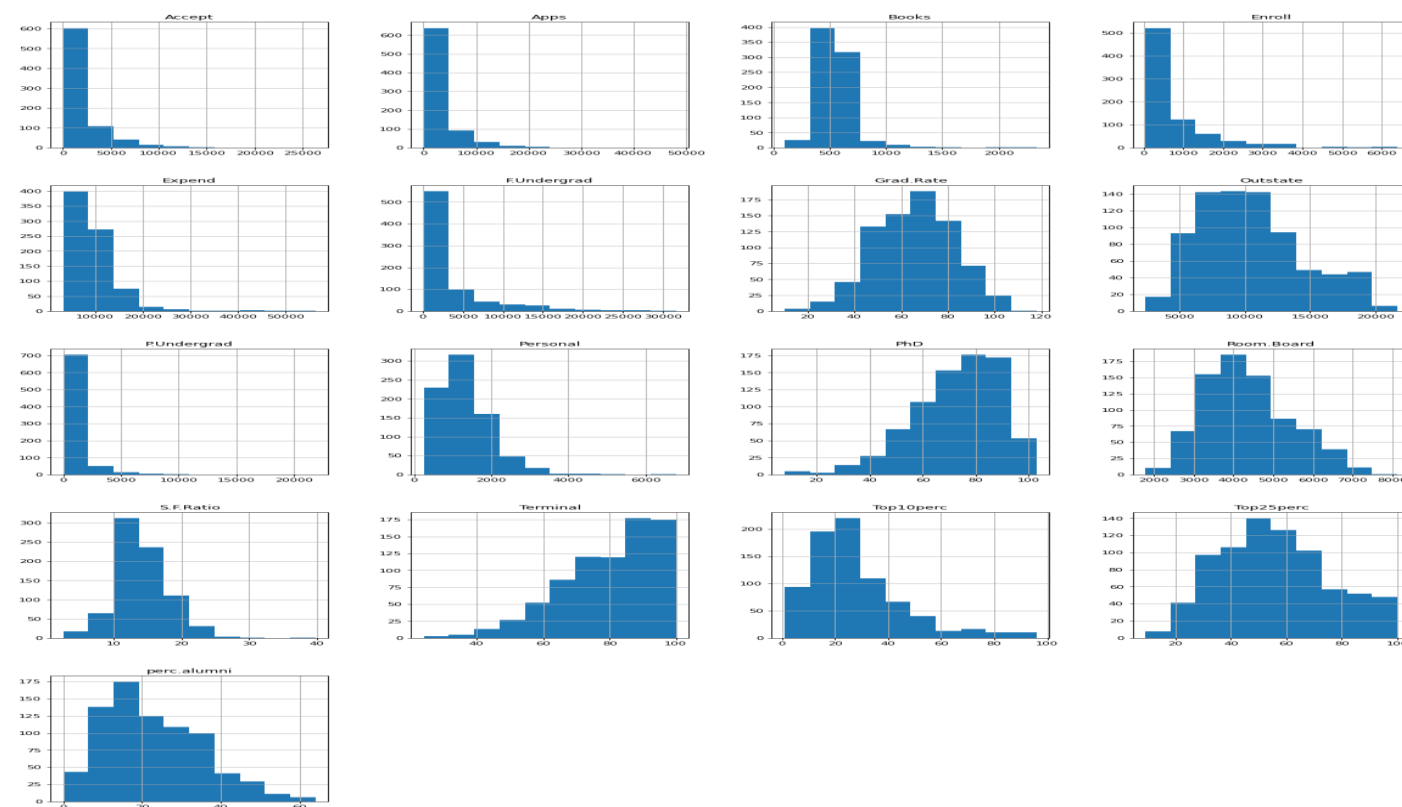
## Inference:

There seems to be outliers. Let us confirm it by further analysis

## Univariate Analysis

Now let us check the distribution of the data.

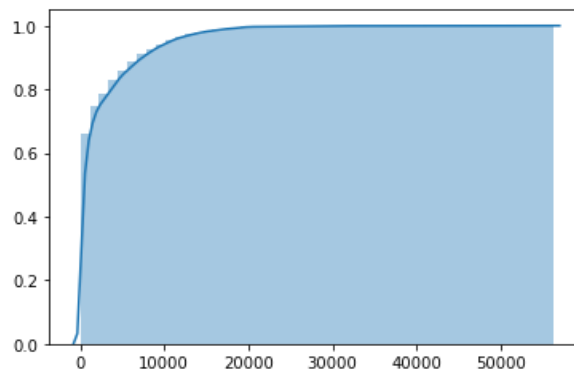
Distribution plot of each variable:



## Inference:

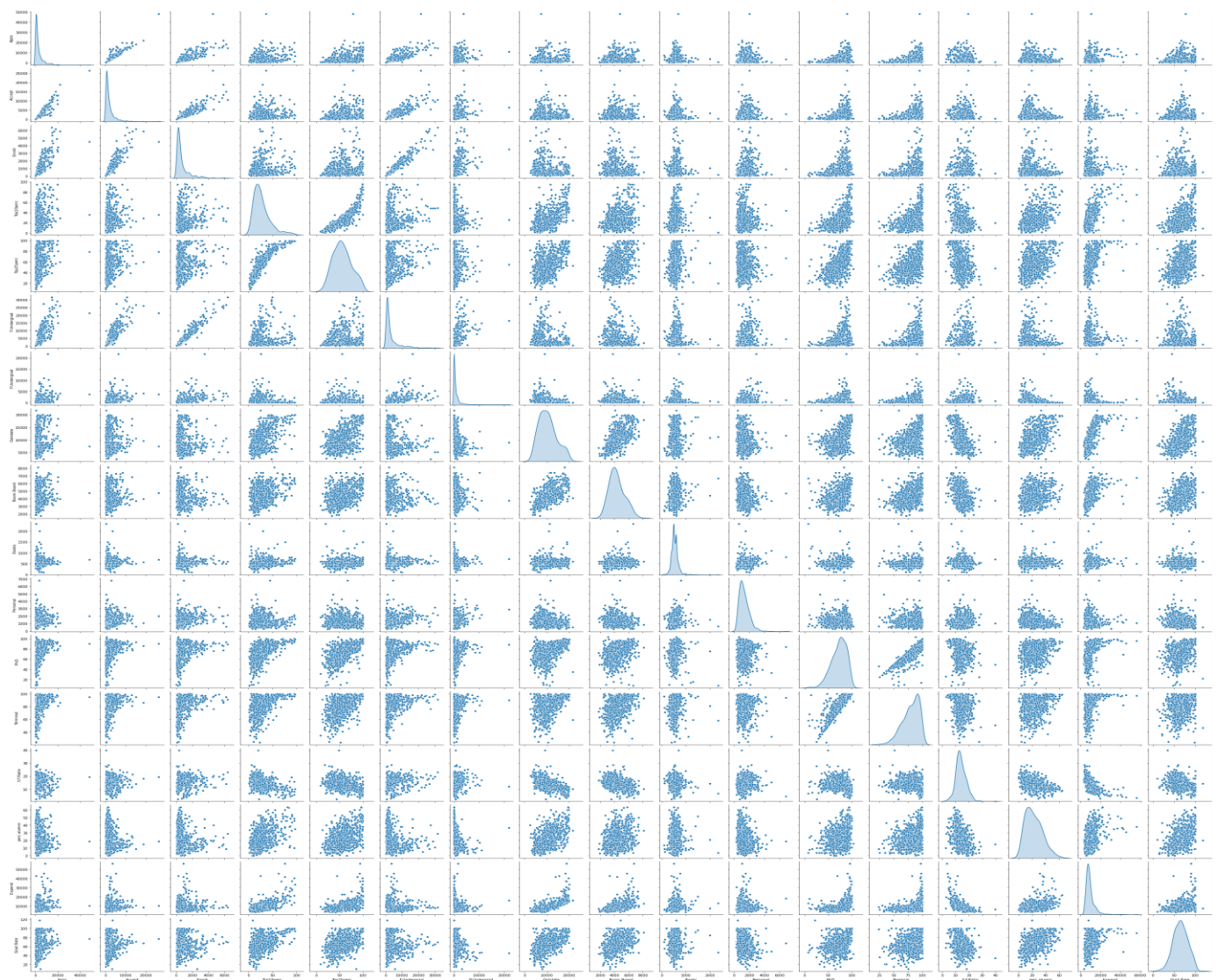
Most of the variable is left skewed and only terminal is right skewed

## Cumulative Distribution:



A cumulative distribution function (CDF) plot shows the empirical cumulative distribution function of the data. CDF plots are useful for comparing the distribution of different sets of data.

## Multivariate analysis



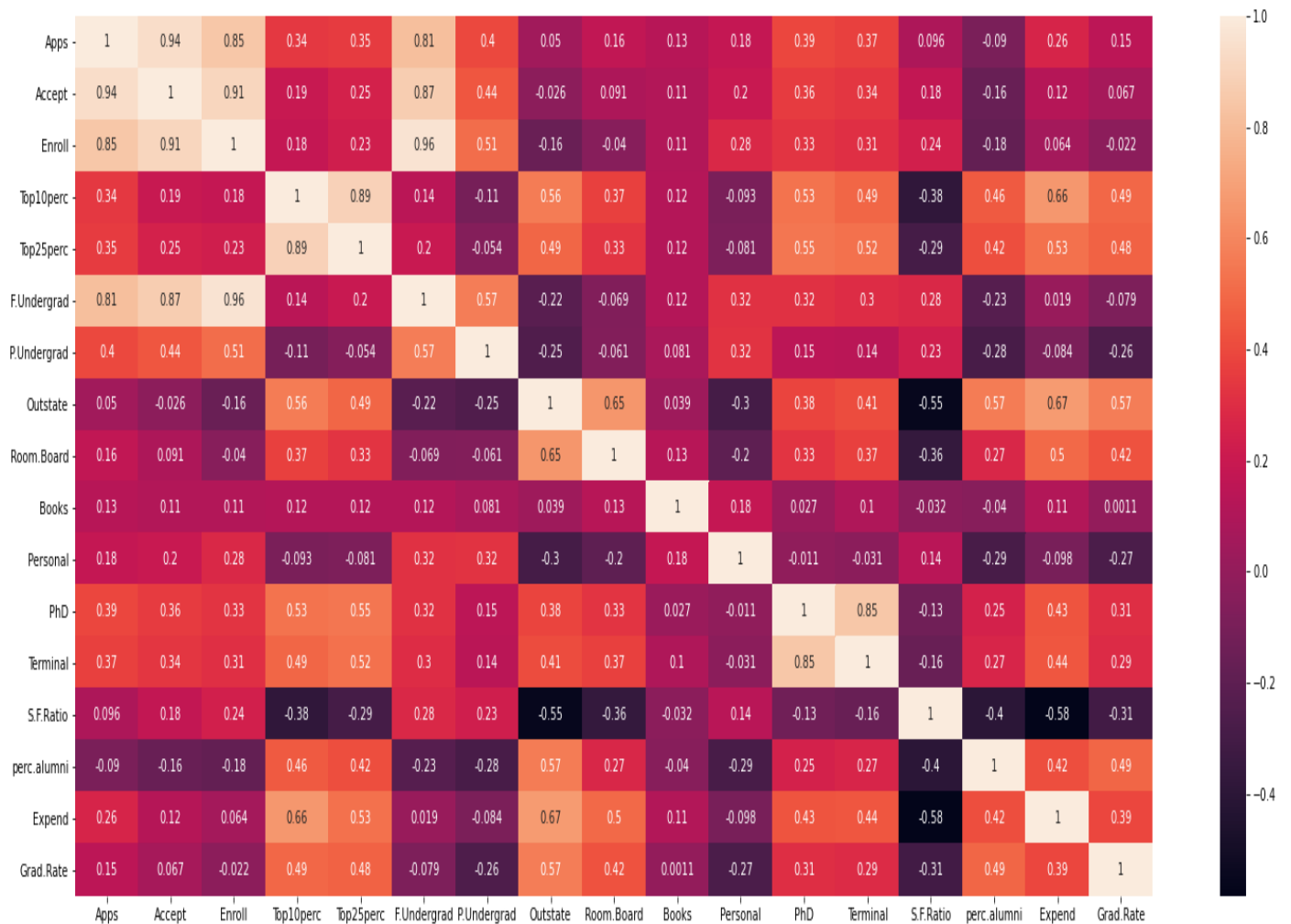
## Inference:

A scatter plot is a visual representation of the degree of correlation between any two columns.

There is more correlation between Apps and Accept variables

Range is different for each variable.

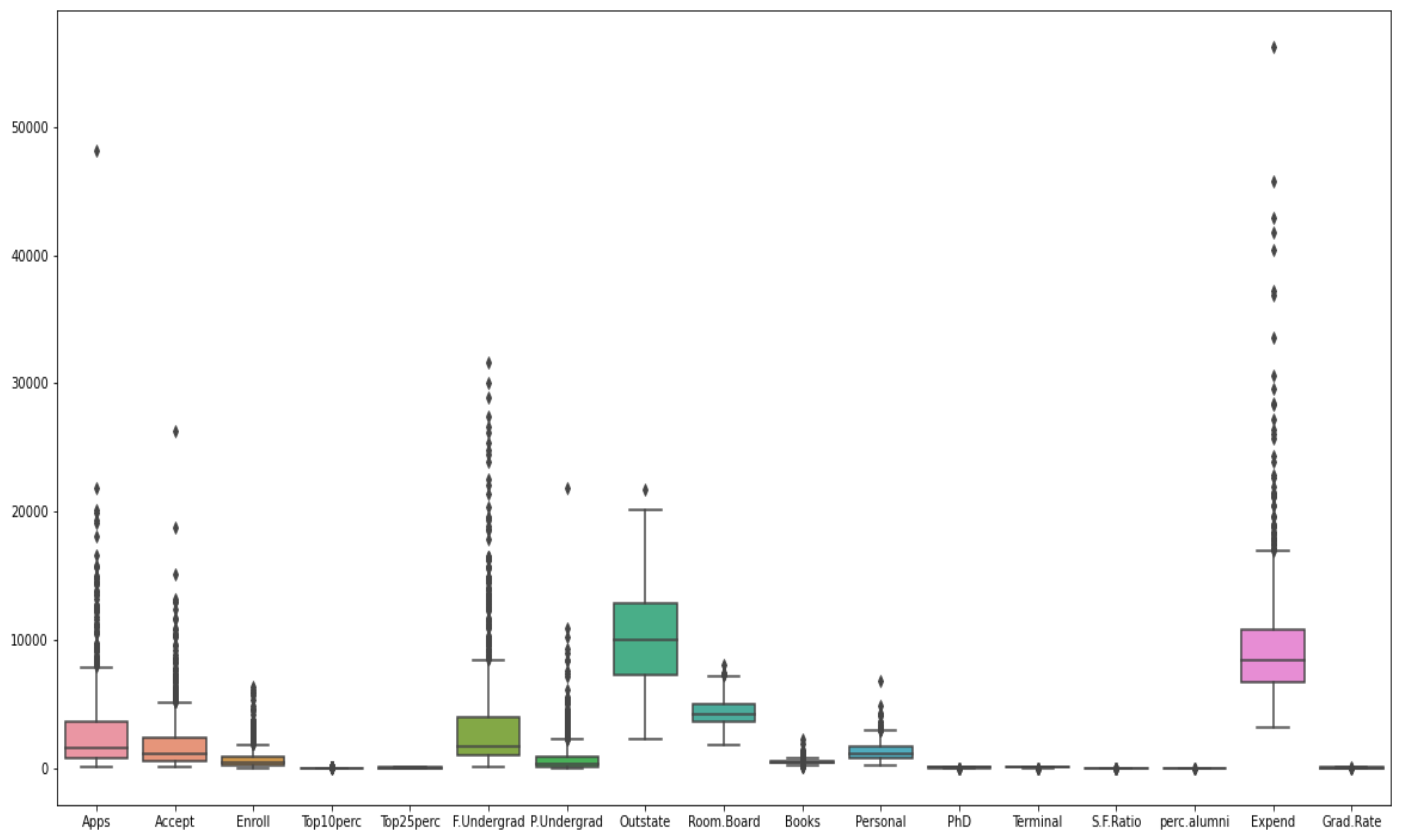
## Correlation between variables:



## Inference:

- There is high negative correlation between the Expend and S.F.Ratio
- There is more correlation between Apps and Accept variables

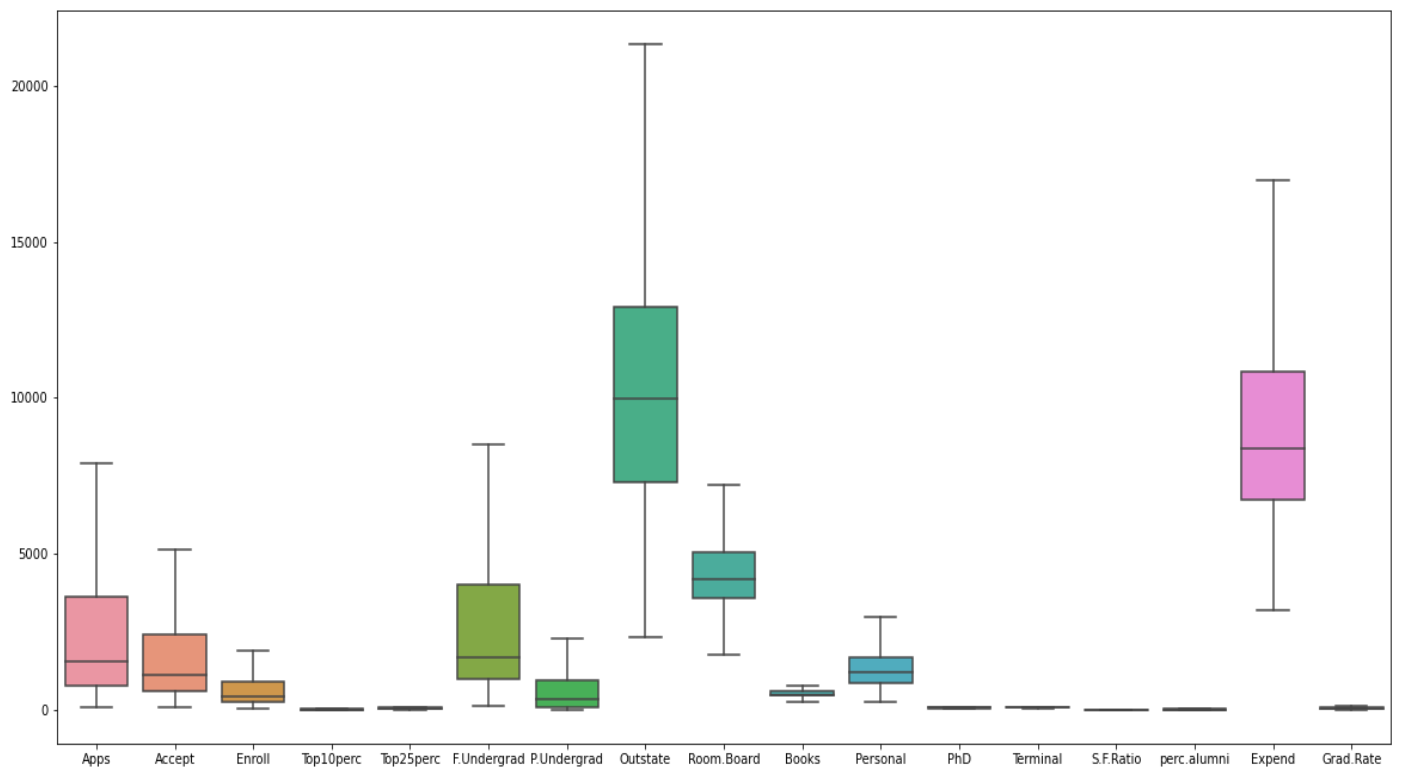
### Before outlier treatment:



### Inference:

There are outliers almost in all variable

### After outlier treatment



### Inference:

There are no outliers in any variables

## 2.2) Scale the variables and write the inference for using the type of scaling function for this case study.

### Standardising before processing PCA

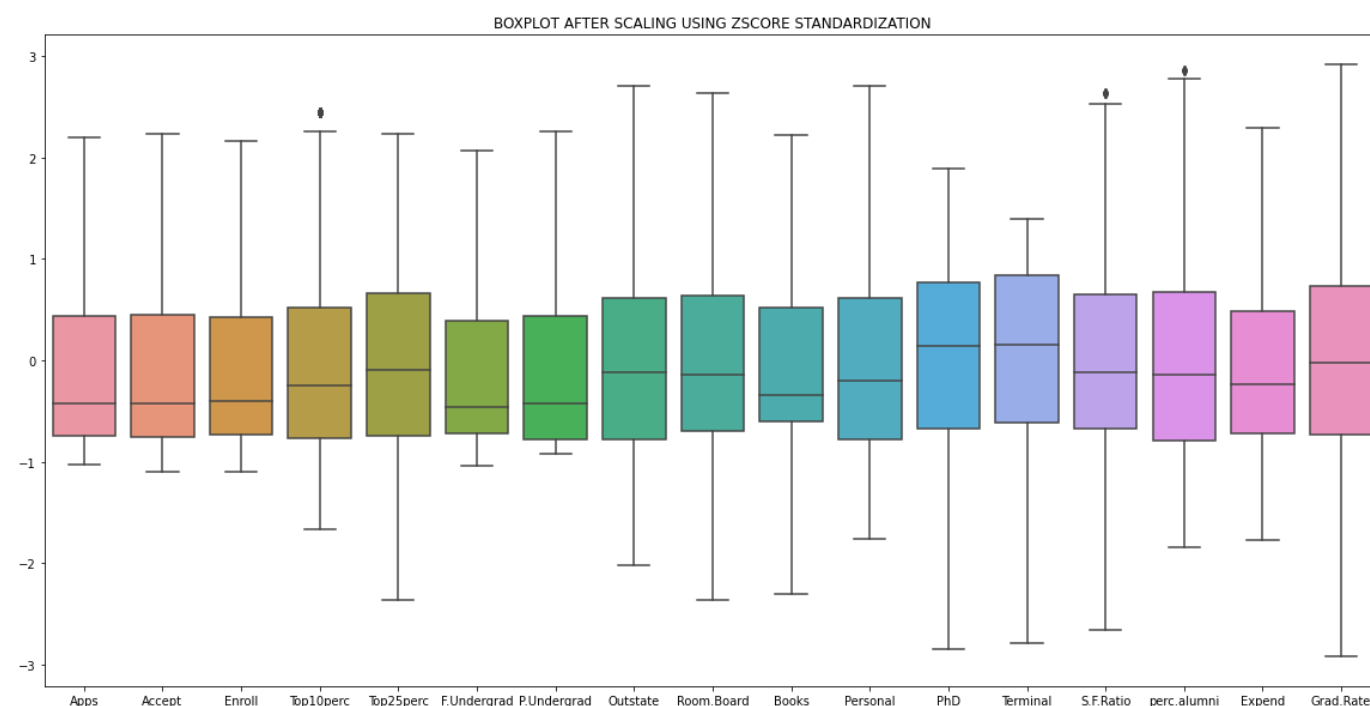
The aim of this step is to standardise the range of the continuous initial variables so that each one of them contributes equality to the analysis. This dataset contains different range of the continuous variable. So, it is important to standardise the variables

### Z score standardisation

Z score standardisation will return a normalized value (z-score) based on the mean and standard deviation. A z-score, or standard score, is used for standardizing scores on the same scale by dividing a score's deviation by the standard deviation in a data set. The result is a standard score. It measures the number of standard deviations that a given data point is from the mean.

A z-score can be negative or positive. A negative score indicates a value less than the mean, and a positive score indicates a value greater than the mean. The average of every z-score for a data set is zero.

Column1	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
0	-0.37649	-0.33783	0.10638	-0.24678	-0.191827	-0.018769	-0.166083	-0.74648	-0.968324	-0.77657	1.4385	-0.17405	-0.123239	1.070602	-0.870466	-0.63092	-0.319205
1	-0.1592	0.116744	-0.26044	-0.69629	-1.353911	-0.093626	0.797856	0.457762	1.92168	1.828605	0.289289	-2.74573	-2.785068	-0.489511	-0.545726	0.396097	-0.552693
2	-0.47234	-0.42651	-0.56934	-0.310996	-0.292878	-0.703966	-0.777974	0.201488	-0.555466	-1.21076	-0.260691	-1.24035	-0.9529	-0.304413	0.590864	-0.13185	-0.669437
3	-0.88999	-0.91787	-0.91861	2.129202	1.677612	-0.898889	-0.828267	0.626954	1.004218	-0.77657	-0.736792	1.205884	1.190391	-1.679429	1.159159	2.28794	-0.377577
4	-0.98253	-1.05122	-1.06253	-0.69629	-0.596031	-0.99561	0.297726	-0.716623	-0.216006	2.219381	0.289289	0.202299	-0.538069	-0.568839	-1.682316	0.512468	-2.916759



### Inference

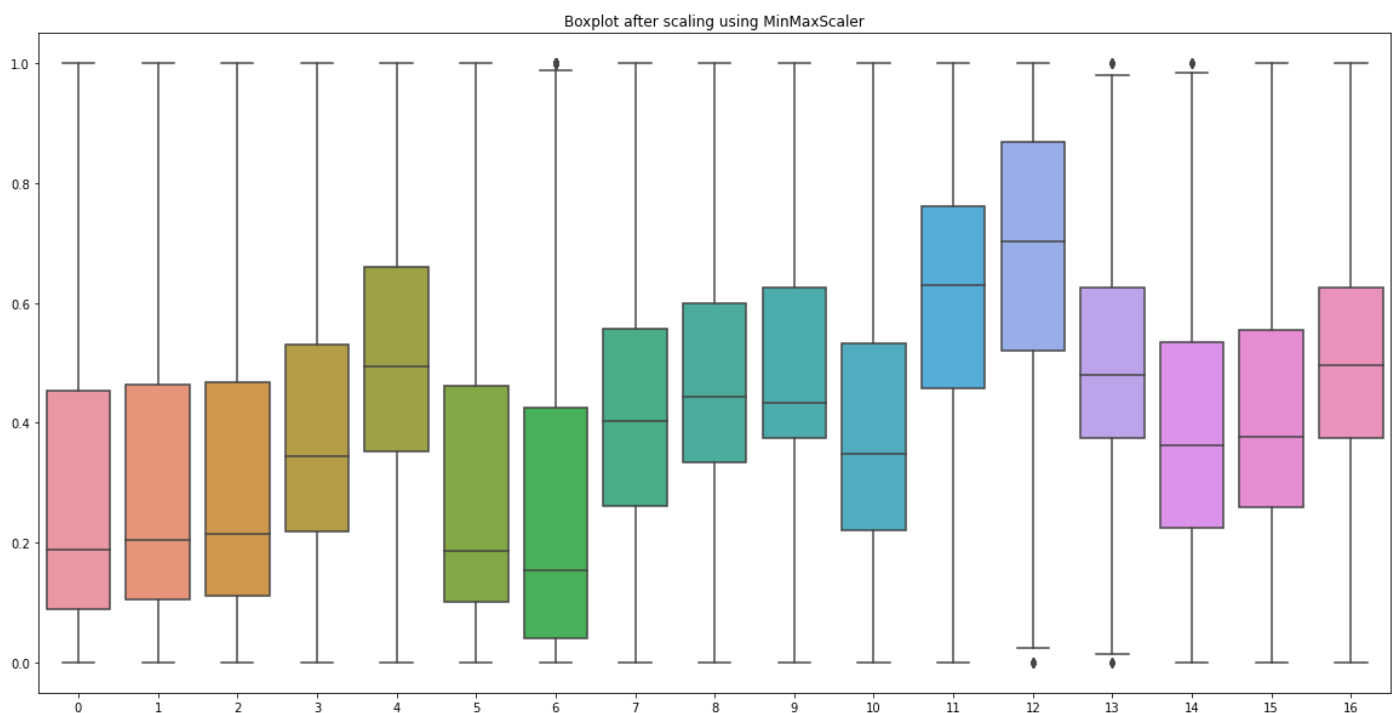
After standardised by z-score standardisation, all variables will be transformed into same scale

## MaxMinscaler:

An alternative approach to Z-score normalization (or standardization) is the so-called Min-Max scaling (often also simply called "normalization" - a common cause for ambiguities). In this approach, the data is scaled to a fixed range - usually 0 to 1. The cost of having this bounded range - in contrast to standardization - is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

After MaxMinscaler Normalisation in the dataset

```
array([[0.20204734, 0.22825659, 0.36941303, ..., 0.20689655, 0.28010899,
        0.445      ],
       [0.26935381, 0.36442346, 0.25686591, ..., 0.27586207, 0.53340599,
        0.405      ],
       [0.17236084, 0.20169225, 0.16208939, ..., 0.51724138, 0.40319709,
        0.385      ],
       ...,
       [0.25796545, 0.3626525 , 0.35541195, ..., 0.34482759, 0.37326067,
        0.335      ],
       [1.      , 0.46851633, 0.6903608 , ..., 0.84482759, 1.      ,
        0.835      ],
       [0.37210493, 0.35084612, 0.35325794, ..., 0.48275862, 0.09613079,
        0.835      ]])
```



## Inference

- MaxMinscaler standardisation, here scales are between 0 and 1.
- since we are interested in the components that maximize the variance, here I used Z score standardised data for rest of the principal component analysis



### 2.3) Comment on the comparison between covariance and the correlation matrix after scaling.

#### Create a covariance matrix for identifying Principal components

The aim of this step is to understand how the variables of the input dataset are varying from the mean with respect to each other or to check any relationship between them.

- If Positive, then the two variables increase or decrease together (correlated)
- If Negative, then one increases when the other decreases (inversely correlated)

#### Covariance Matrix

```
%s [[ 0.39728021 -0.18584152 -0.01909896 ... -0.05025984 -0.41106778
0.17096373]
[-0.18584152 1.61327963 0.08300998 ... 0.04838876 -0.10924942
-0.22225919]
[-0.01909896 0.08300998 0.21261253 ... -0.0385113 0.16013293
-0.07338076]
...
[-0.05025984 0.04838876 -0.0385113 ... 0.21807345 -0.05505902
-0.0563193 ]
[-0.41106778 -0.10924942 0.16013293 ... -0.05505902 1.52015207
-0.48150377]
[ 0.17096373 -0.22225919 -0.07338076 ... -0.0563193 -0.48150377
0.74695292]]
```

#### Comparing Correlation and Covariance Matrix

“Covariance” indicates the direction of the linear relationship between variables. “Correlation” on the other hand measures both the strength and direction of the linear relationship between two variables. Correlation is a function of the covariance.

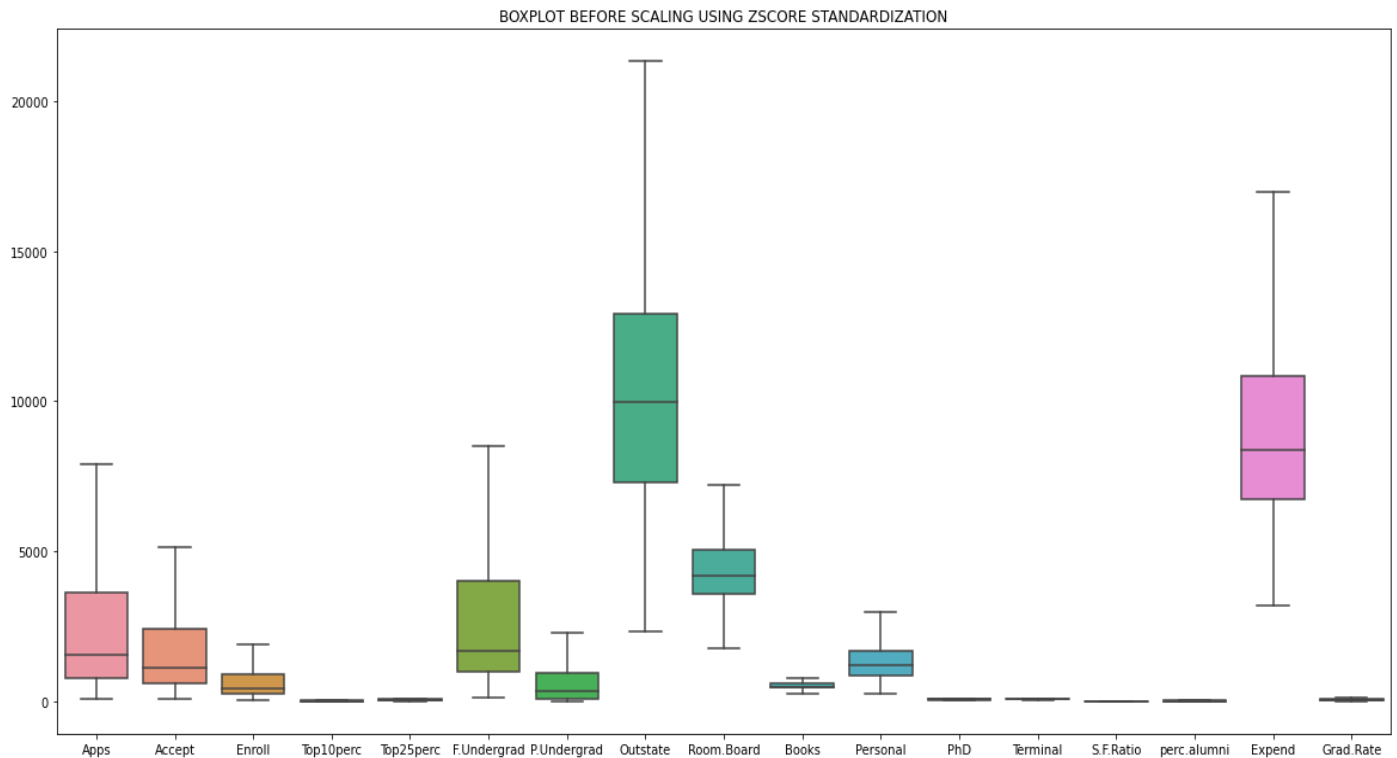
Column1	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Apps	1	0.955307	0.896883	0.321342	0.364491	0.861002	0.519823	0.065337	0.187475	0.236138	0.229948	0.463924	0.434478	0.126411	-0.101158	0.242935	0.150803
Accept	0.955307	1	0.935277	0.223298	0.273681	0.897034	0.572691	-0.005002	0.119586	0.208705	0.256346	0.427341	0.403409	0.188506	-0.165516	0.161808	0.078982
Enroll	0.896883	0.935277	1	0.171756	0.230434	0.967302	0.641595	-0.155655	-0.023846	0.202057	0.339348	0.38154	0.354379	0.274269	-0.222723	0.054221	-0.023251
Top10perc	0.321342	0.223298	0.171756	1	0.913875	0.111215	-0.180009	0.56216	0.357366	0.153452	-0.11673	0.544048	0.506748	-0.387926	0.455797	0.657039	0.49367
Top25perc	0.364491	0.273681	0.230434	0.913875	1	0.181196	-0.099295	0.489569	0.330987	0.169761	-0.08681	0.551461	0.527654	-0.297233	0.416832	0.572905	0.478985
F.Undergrad	0.861002	0.897034	0.967302	0.111215	0.181196	1	0.69613	-0.226166	-0.054476	0.207879	0.359783	0.361564	0.335054	0.324504	-0.285457	0.000371	-0.082239
P.Undergrad	0.519823	0.572691	0.641595	-0.180009	-0.099295	0.69613	1	-0.354216	-0.067638	0.122529	0.344053	0.127663	0.122152	0.370607	-0.419334	-0.201929	-0.265158
Outstate	0.065337	-0.005	-0.15566	0.56216	0.489569	-0.226166	-0.354216	1	0.655489	0.00511	-0.325609	0.391321	0.412579	-0.573683	0.565736	0.775328	0.572458
Room.Board	0.187475	0.119586	-0.02385	0.357366	0.330987	-0.054476	-0.067638	0.655489	1	0.108924	-0.219554	0.341469	0.37927	-0.37643	0.272393	0.580622	0.42579
Books	0.236138	0.208705	0.202057	0.153452	0.169761	0.207879	0.122529	0.00511	0.108924	1	0.239863	0.13639	0.159318	-0.008536	-0.042832	0.149983	-0.008051
Personal	0.229948	0.256346	0.339348	-0.11673	-0.08681	0.359783	0.344053	-0.325609	-0.219554	0.239863	1	-0.011684	-0.031971	0.173913	-0.305753	-0.163271	-0.290894
PhD	0.463924	0.427341	0.38154	0.544048	0.551461	0.361564	0.127663	0.391321	0.341469	0.13639	-0.011684	1	0.862928	-0.12939	0.248877	0.510529	0.310019
Terminal	0.434478	0.403409	0.354379	0.506748	0.527654	0.335054	0.122152	0.412579	0.37927	0.159318	-0.031971	0.862928	1	-0.150993	0.266033	0.524068	0.292803
S.F.Ratio	0.126411	0.188506	0.274269	-0.387926	-0.297233	0.324504	0.370607	-0.573683	-0.37643	-0.008536	0.173913	-0.12939	-0.150993	1	-0.412101	-0.654376	-0.308525
perc.alumni	-0.10116	-0.16552	-0.22272	0.455797	0.416832	-0.285457	-0.419334	0.565736	0.272393	-0.042832	-0.305753	0.248877	0.266033	-0.412101	1	0.462922	0.491408
Expend	0.242935	0.161808	0.054221	0.657039	0.572905	0.000371	-0.201929	0.775328	0.580622	0.149983	-0.163271	0.510529	0.524068	-0.654376	0.462922	1	0.415291
Grad.Rate	0.150803	0.078982	-0.02325	0.49367	0.478985	-0.082239	-0.265158	0.572458	0.42579	-0.008051	-0.290894	0.310019	0.292803	-0.308525	0.491408	0.415291	1

#### Inference:

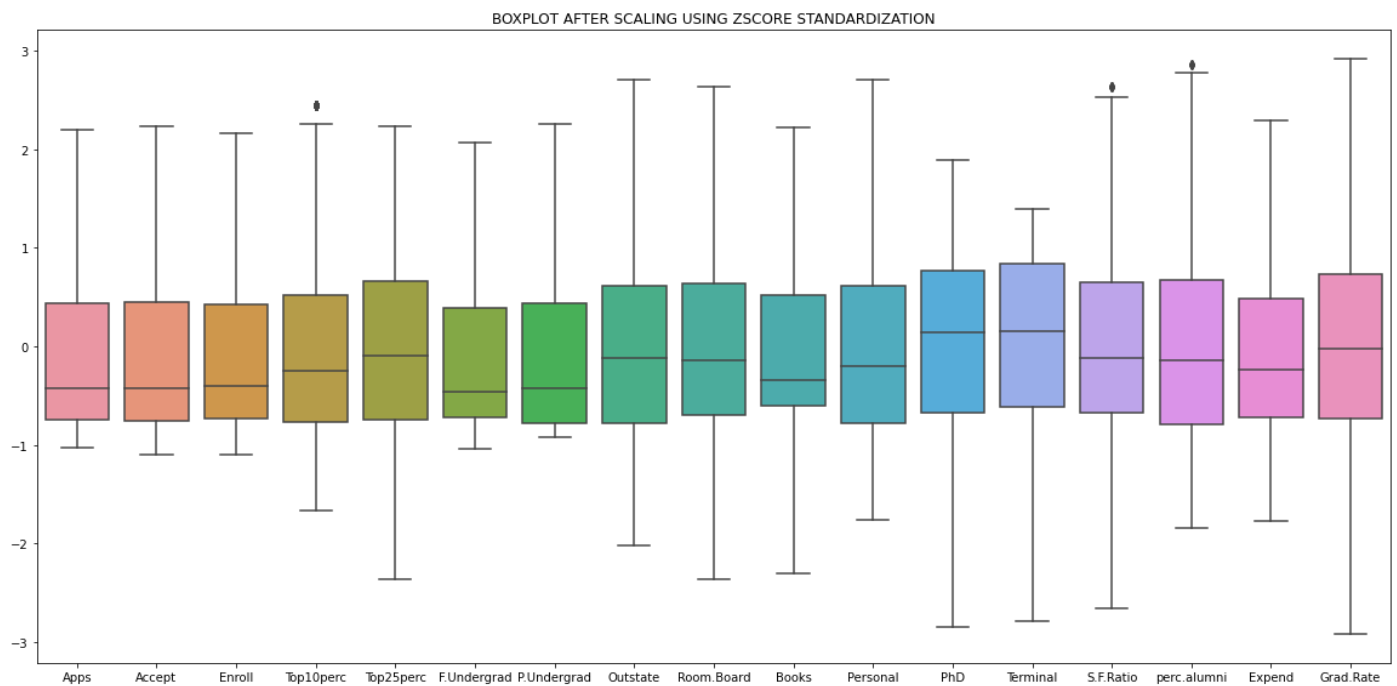
- Covariance matrix is not more than a table that summaries the correlations between all the possible pairs of variables. With standardisation also, correlation matrix yields same result
- Correlation is a normalized form of covariance and not affected by scale. Both covariance and correlation measure the linear relationship between variables but cannot be used interchangeably.

2.4) Check the dataset for outliers before and after scaling. Draw your inferences from this exercise.

### BOXPLOT BEFORE SCALING USING ZSCORE STANDARDIZATION



### BOXPLOT AFTER SCALING USING ZSCORE STANDARDIZATION



#### Inference:

- Before scaling, each variable has different Range.
- After scaling, all the variables have almost same range between -3 to 3

## 2.5) Build the covariance matrix, eigenvalues, and eigenvector.

The covariance matrix is a measure of how the variables correlate with each other

### Covariance Matrix

```
[ [ 1.00128866e+00  9.56537704e-01  8.98039052e-01  3.21756324e-01
 3.64960691e-01  8.62111140e-01  5.20492952e-01  6.54209711e-02
 1.87717056e-01  2.36441941e-01  2.30243993e-01  4.64521757e-01
 4.35037784e-01  1.26573895e-01 -1.01288006e-01  2.43248206e-01
 1.50997775e-01]
[ 9.56537704e-01  1.00128866e+00  9.36482483e-01  2.23586208e-01
 2.74033187e-01  8.98189799e-01  5.73428908e-01 -5.00874847e-03
 1.19740419e-01  2.08974091e-01  2.56676290e-01  4.27891234e-01
 4.03929238e-01  1.88748711e-01 -1.65728801e-01  1.62016688e-01
 7.90839722e-02]
[ 8.98039052e-01  9.36482483e-01  1.00128866e+00  1.71977357e-01
 2.30730728e-01  9.68548601e-01  6.42421828e-01 -1.55856056e-01
 -2.38762560e-02  2.02317274e-01  3.39785395e-01  3.82031198e-01
 3.54835877e-01  2.74622251e-01 -2.23009677e-01  5.42906862e-02
 -2.32810071e-02]
[ 3.21756324e-01  2.23586208e-01  1.71977357e-01  1.00128866e+00
 9.15052977e-01  1.11358019e-01 -1.80240778e-01  5.62884044e-01
 3.57826139e-01  1.53650150e-01 -1.16880152e-01  5.44748764e-01
 5.07401238e-01 -3.88425719e-01  4.56384036e-01  6.57885921e-01
 4.94306540e-01]
[ 3.64960691e-01  2.74033187e-01  2.30730728e-01  9.15052977e-01
 1.00128866e+00  1.81429267e-01 -9.94231153e-02  4.90200034e-01
 3.31413314e-01  1.69979808e-01 -8.69219644e-02  5.52172085e-01
 5.28333659e-01 -2.97616423e-01  4.17369123e-01  5.73643193e-01
 4.79601950e-01]
[ 8.62111140e-01  8.98189799e-01  9.68548601e-01  1.11358019e-01
 1.81429267e-01  1.00128866e+00  6.97027420e-01 -2.26457040e-01
 -5.45459528e-02  2.08147257e-01  3.60246460e-01  3.62030390e-01
 3.35485771e-01  3.24921933e-01 -2.85825062e-01  3.71119607e-04
 -8.23447851e-02]
[ 5.20492952e-01  5.73428908e-01  6.42421828e-01 -1.80240778e-01
 -9.94231153e-02  6.97027420e-01  1.00128866e+00 -3.54672874e-01
 -6.77252009e-02  1.22686416e-01  3.44495974e-01  1.27827147e-01
 1.22309141e-01  3.71084841e-01 -4.19874031e-01 -2.02189396e-01
 -2.65499420e-01]
[ 6.54209711e-02 -5.00874847e-03 -1.55856056e-01  5.62884044e-01
 4.90200034e-01 -2.26457040e-01 -3.54672874e-01  1.00128866e+00
 6.56333564e-01  5.11656377e-03 -3.26028927e-01  3.91824814e-01
 4.13110264e-01 -5.74421963e-01  5.66465309e-01  7.76326650e-01
 5.73195743e-01]
[ 1.87717056e-01  1.19740419e-01 -2.38762560e-02  3.57826139e-01
 3.31413314e-01 -5.45459528e-02 -6.77252009e-02  6.56333564e-01
 1.00128866e+00  1.09064551e-01 -2.19837042e-01  3.41908577e-01
 3.79759015e-01 -3.76915472e-01  2.72743761e-01  5.81370284e-01
 4.26338910e-01]
[ 2.36441941e-01  2.08974091e-01  2.02317274e-01  1.53650150e-01
 1.69979808e-01  2.08147257e-01  1.22686416e-01  5.11656377e-03
 1.09064551e-01  1.00128866e+00  2.40172145e-01  1.36566243e-01
 1.59523091e-01 -8.54689129e-03 -4.28870629e-02  1.50176551e-01
 -8.06107505e-03]
[ 2.30243993e-01  2.56676290e-01  3.39785395e-01 -1.16880152e-01
 -8.69219644e-02  3.60246460e-01  3.44495974e-01 -3.26028927e-01
 -2.19837042e-01  2.40172145e-01  1.00128866e+00 -1.16986124e-02
 -3.20117803e-02  1.74136664e-01 -3.06146886e-01 -1.63481407e-01
 -2.91268705e-01]
[ 4.64521757e-01  4.27891234e-01  3.82031198e-01  5.44748764e-01
 5.52172085e-01  3.62030390e-01  1.27827147e-01  3.91824814e-01
 3.41908577e-01  1.36566243e-01 -1.16986124e-02  1.00128866e+00
 8.64040263e-01 -1.29556494e-01  2.49197779e-01  5.11186852e-01
 3.10418895e-01]
[ 4.35037784e-01  4.03929238e-01  3.54835877e-01  5.07401238e-01
 5.28333659e-01  3.35485771e-01  1.22309141e-01  4.13110264e-01
 3.79759015e-01  1.59523091e-01 -3.20117803e-02  8.64040263e-01
 1.00128866e+00 -1.51187934e-01  2.66375402e-01  5.24743500e-01
 2.93180212e-01]
[ 1.26573895e-01  1.88748711e-01  2.74622251e-01 -3.88425719e-01
 -2.97616423e-01  3.24921933e-01  3.71084841e-01 -5.74421963e-01
 -3.76915472e-01 -8.54689129e-03  1.74136664e-01 -1.29556494e-01
 -1.51187934e-01  1.00128866e+00 -4.12632056e-01 -6.55219504e-01
 -3.08922187e-01]
[ -1.01288006e-01 -1.65728801e-01 -2.23009677e-01  4.56384036e-01
 4.17369123e-01 -2.85825062e-01 -4.19874031e-01  5.66465309e-01
 2.72743761e-01 -4.28870629e-02 -3.06146886e-01  2.49197779e-01
 2.66375402e-01 -4.12632056e-01  1.00128866e+00  4.63518674e-01
 4.92040760e-01]
[ 2.43248206e-01  1.62016688e-01  5.42906862e-02  6.57885921e-01
 5.73643193e-01  3.71119607e-04 -2.02189396e-01  7.76326650e-01
 5.81370284e-01  1.50176551e-01 -1.63481407e-01  5.11186852e-01
 5.24743500e-01 -6.55219504e-01  4.63518674e-01  1.00128866e+00
 4.15826026e-01]
[ 1.50997775e-01  7.90839722e-02 -2.32810071e-02  4.94306540e-01
 4.79601950e-01 -8.23447851e-02 -2.65499420e-01  5.73195743e-01
 4.26338910e-01 -8.06107505e-03 -2.91268705e-01  3.10418895e-01
 2.93180212e-01 -3.08922187e-01  4.92040760e-01  4.15826026e-01
 1.00128866e+00]]
```

## Identify eigen values and eigen vector

The eigenvectors and eigenvalues of a covariance (or correlation) matrix represent the “core” of a PCA: The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes.

To decide which eigenvector(s) can be dropped without losing too much information for the construction of lower-dimensional subspace, we need to inspect the corresponding eigenvalues: The eigenvectors with the lowest eigenvalues bear the least information about the distribution of the data; those are the ones that can be dropped

### Eigen Values

```
%s [5.6625219 4.89470815 1.12636744 1.00397659 0.87218426 0.7657541  
0.58491404 0.5445048 0.42352336 0.38101777 0.24701456 0.02239369  
0.03789395 0.14726392 0.13434483 0.09883384 0.07469003]
```

### Eigen Vectors

```
%s [[-2.62171542e-01 3.14136258e-01 8.10177245e-02 -9.87761685e-02  
-2.19898081e-01 2.18800617e-03 -2.83715076e-02 -8.99498102e-02  
1.30566998e-01 -1.56464458e-01 -8.62132843e-02 1.82169814e-01  
-5.99137640e-01 8.99775288e-02 8.88697944e-02 5.49428396e-01  
5.41453698e-03]  
[-2.30562461e-01 3.44623583e-01 1.07658626e-01 -1.18140437e-01  
-1.89634940e-01 -1.65212882e-02 -1.29584896e-02 -1.37606312e-01  
1.42275847e-01 -1.49209799e-01 -4.25899061e-02 -3.91041719e-01  
6.61496927e-01 1.58861886e-01 4.37945938e-02 2.91572312e-01  
1.44582845e-02]  
[-1.89276397e-01 3.82813322e-01 8.55296892e-02 -9.30717094e-03  
-1.62314818e-01 -6.80794143e-02 -1.52403625e-02 -1.44216938e-01  
5.08712481e-02 -6.48997860e-02 -4.38408622e-02 7.16684935e-01  
2.33235272e-01 -3.53988202e-02 -6.19241658e-02 -4.17001280e-01  
-4.97908902e-02]  
[-3.38874521e-01 -9.93191661e-02 -7.88293849e-02 3.69115031e-01  
-1.57211016e-01 -8.88656824e-02 -2.57455284e-01 2.89538833e-01  
-1.22467790e-01 -3.58776186e-02 1.77837341e-03 -5.62053913e-02  
2.21448729e-02 -3.92277722e-02 6.99599977e-02 8.79767299e-03  
-7.23645373e-01]  
[-3.34690532e-01 -5.95055011e-02 -5.07938247e-02 4.16824361e-01  
-1.44449474e-01 -2.76268979e-02 -2.39038849e-01 3.45643551e-01  
-1.93936316e-01 6.41786425e-03 -1.02127328e-01 1.96735274e-02  
3.22646978e-02 1.45621999e-01 -9.70282598e-02 -1.07779150e-02  
6.55464648e-01]  
[-1.63293010e-01 3.98636372e-01 7.37077827e-02 -1.39504424e-02  
-1.02728468e-01 -5.16468727e-02 -3.11751439e-02 -1.08748900e-01  
1.45452749e-03 -1.63981359e-04 -3.49993487e-02 -5.42774834e-01  
-3.67681187e-01 -1.33555923e-01 -8.71753137e-02 -5.70683843e-01  
2.53059904e-02]  
[-2.24797091e-02 3.57550046e-01 4.03568700e-02 -2.25351078e-01  
9.56790178e-02 -2.45375721e-02 -1.00138971e-02 1.23841696e-01  
-6.34774326e-01 5.46346279e-01 2.52107094e-01 2.95029745e-02  
2.62494456e-02 5.02487566e-02 4.45537493e-02 1.46321060e-01  
-3.97146972e-02]  
[-2.83547285e-01 -2.51863617e-01 1.49394795e-02 -2.62975384e-01  
-3.72750885e-02 -2.03860462e-02 9.45370782e-02 1.12721477e-02  
-8.36648339e-03 -2.31799759e-01 5.93433149e-01 1.03393587e-03  
-8.14247697e-02 5.60392799e-01 6.72405494e-02 -2.11561014e-01  
-1.59275617e-03]
```

```

[-2.44186588e-01 -1.31909124e-01 -2.11379165e-02 -5.80894132e-01
 6.91080879e-02 2.37267409e-01 9.45210745e-02 3.89639465e-01
-2.20526518e-01 -2.55107620e-01 -4.75297296e-01 9.85725168e-03
2.67779296e-02 -1.07365653e-01 1.77715010e-02 -1.00935084e-01
-2.82578388e-02]
[-9.67082754e-02 9.39739472e-02 -6.97121128e-01 3.61562884e-02
-3.54056654e-02 6.38604997e-01 -1.11193334e-01 -2.39817267e-01
2.10246624e-02 9.11624912e-02 4.35697999e-02 4.36086500e-03
1.04624246e-02 5.16224550e-02 3.54343707e-02 -2.86384228e-02
-8.06259380e-03]
[ 3.52299594e-02 2.32439594e-01 -5.30972806e-01 1.14982973e-01
4.75358244e-04 -3.81495854e-01 6.39418106e-01 2.77206569e-01
1.73715184e-02 -1.27647512e-01 1.51627393e-02 -1.08725257e-02
4.54572099e-03 9.39409228e-03 -1.18604404e-02 3.38197909e-02
1.42590097e-03]
[-3.26410696e-01 5.51390195e-02 8.11134044e-02 1.47260891e-01
5.50786546e-01 3.34444832e-03 8.92320786e-02 -3.42628480e-02
1.66510079e-01 1.00975002e-01 -3.91865961e-02 1.33146759e-02
1.25137966e-02 -7.16590441e-02 7.02656469e-01 -6.38096394e-02
8.31471932e-02]
[-3.23115980e-01 4.30332048e-02 5.89785929e-02 8.90079921e-02
5.90407136e-01 3.54121294e-02 9.16985445e-02 -9.03076644e-02
1.12609034e-01 8.60363025e-02 -8.48575651e-02 7.38135022e-03
-1.79275275e-02 1.63820871e-01 -6.62488717e-01 9.85019644e-02
-1.13374007e-01]
[ 1.63151642e-01 2.59804556e-01 2.74150657e-01 2.59486122e-01
1.42842546e-01 4.68752604e-01 1.52864837e-01 2.42807562e-01
-1.53685343e-01 -4.70527925e-01 3.63042716e-01 8.85797314e-03
1.83059753e-02 -2.39902591e-01 -4.79006197e-02 6.19970446e-02
3.83160891e-03]
[-1.86610828e-01 -2.57092552e-01 1.03715887e-01 2.23982467e-01
-1.28215768e-01 1.25669415e-02 3.91400512e-01 -5.66073056e-01
-5.39235753e-01 -1.47628917e-01 -1.73918533e-01 -2.40534190e-02
-8.03169296e-05 -4.89753356e-02 3.58875507e-02 2.80805469e-02
-7.32598621e-03]
[-3.28955847e-01 -1.60008951e-01 -1.84205687e-01 -2.13756140e-01
2.24240837e-02 -2.31562325e-01 -1.50501305e-01 -1.18823549e-01
2.42371616e-02 -8.04154875e-02 3.93722676e-01 1.05658769e-02
5.60069250e-02 -6.90417042e-01 -1.26667522e-01 1.28739213e-01
1.45099786e-01]
[-2.38822447e-01 -1.67523664e-01 2.45335837e-01 3.61915064e-02
-3.56843227e-01 3.13556243e-01 4.68641965e-01 1.80458508e-01
3.15812873e-01 4.88415259e-01 8.72638706e-02 -2.51028410e-03
1.48410810e-02 -1.59332164e-01 -6.30737002e-02 -7.09643331e-03
-3.29024228e-03]]

```

### Inference:

- The first eigen value is 5.6625219 consist of maximum information, remaining information in the second and to other eigen values.
- The last eigen value is 0.07469003 which has very less information, we can drop these values

## 2.6) Write the explicit form of the first PC (in terms of Eigen Vectors).

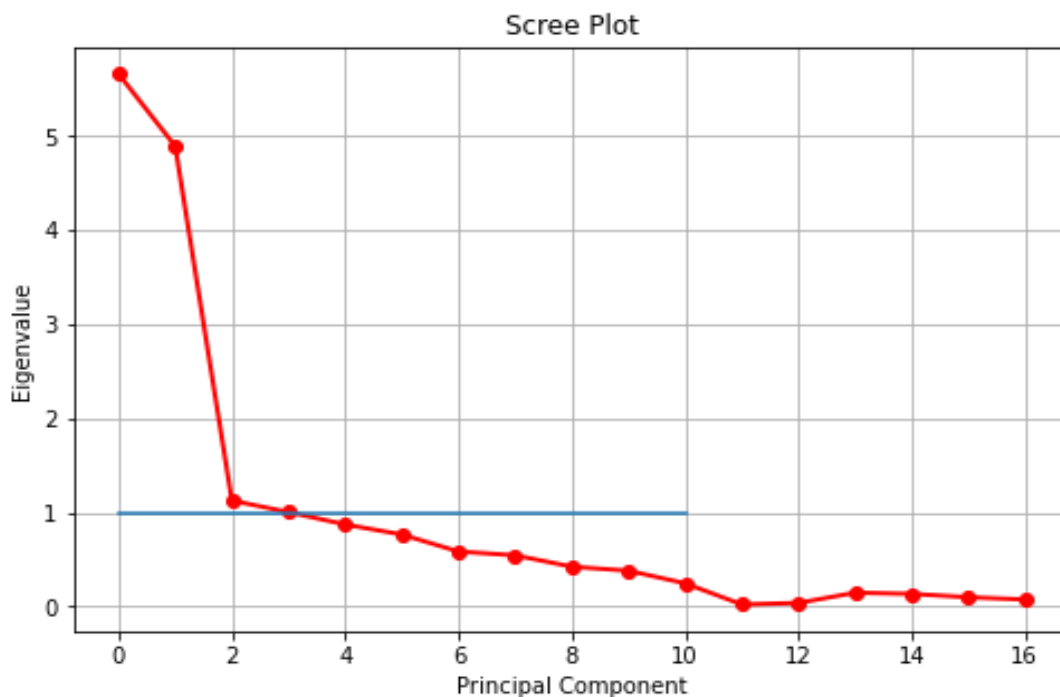
Eigenvectors and eigenvalues exist in pairs: every eigenvector has a corresponding eigenvalue. An eigenvector is a direction, and an eigenvalue is a number, telling you how much variance there is in the data in that direction. The eigenvector with the highest eigenvalue is therefore become principal component. the number of eigenvectors/values that exist equals the number of dimensions the data set has.

First eigen vector

```
[-0.26217154  0.31413626  0.08101772 -0.09877617 -0.21989808  0.00218801  
-0.02837151 -0.08994981  0.130567   -0.15646446 -0.08621328  0.18216981  
-0.59913764  0.08997753  0.08886979  0.5494284   0.00541454]
```

In this data, first eigenvector has high eigenvalue and therefore it forms first principal component. The number of eigenvectors / values is 17 that has 17 number of dimensions. Later in analysis, we can drop least values

**As per given Scree Plot, how many principal components are preferred?**

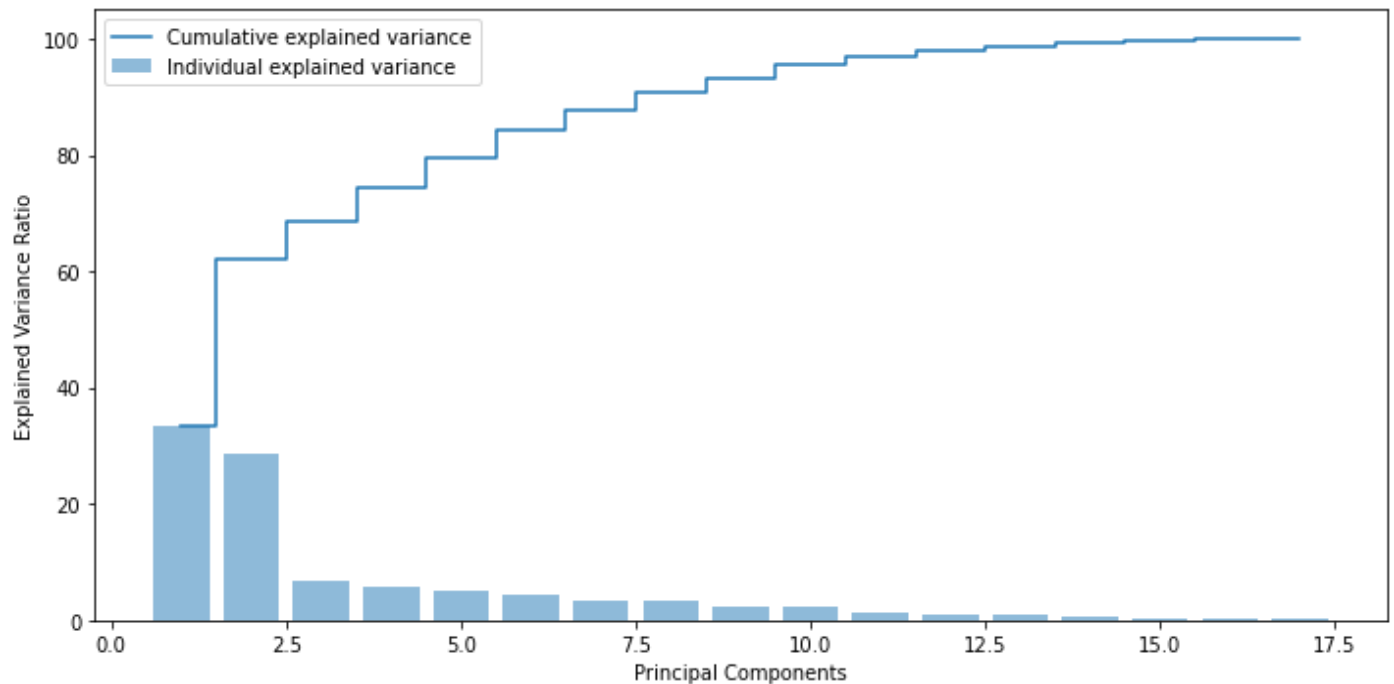


### Inference:

- A scree plot visualizes the dimensionality of the data.
- The scree plot shows the cumulative variance explained by each principal component.
- Visually we can observe that there is steep drop in variance explained with increase in number of PC's.
- We will proceed with 8 components here. But depending on requirement 80% variation or 6 components will also do good

## 2.7) Discuss 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 Variance Explained [ 33.26608367 62.02142867 68.63859223 74.53673619  
79.66062886 84.15926753 87.59551019 90.79435736 93.28246491 95.52086136  
96.97201814 97.83716159 98.62640821 99.20703552 99.64582321  
99.86844192 100. ]



### Inference

- cumulative explained variance ratio has an ability to estimate how many components are needed to describe the data.
- first Individual explained variance covers 33.26% of data.
- second Individual explained variance covers 62.02% of data.
- sixth Individual explained variance covers nearly 84.159% of data.



## Perform PCA and export the data of the Principal Component scores into a data frame

	PCA_1	PCA_2	PCA_3	PCA_4	PCA_5	PCA_6
Apps	0.262172	0.314136	-0.08102	0.098776	0.219898	0.002188
Accept	0.230562	0.344624	-0.10766	0.11814	0.189635	-0.016521
Enroll	0.189276	0.382813	-0.08553	0.009307	0.162315	-0.068079
Top10perc	0.338875	-0.09932	0.078829	-0.369115	0.157211	-0.088866
Top25perc	0.334691	-0.05951	0.050794	-0.416824	0.144449	-0.027627
F.Undergrad	0.163293	0.398636	-0.07371	0.01395	0.102728	-0.051647
P.Undergrad	0.02248	0.35755	-0.04036	0.225351	-0.095679	-0.024538
Outstate	0.283547	-0.25186	-0.01494	0.262975	0.037275	-0.020386
Room.Board	0.244187	-0.13191	0.021138	0.580894	-0.069108	0.237267
Books	0.096708	0.093974	0.697121	-0.036156	0.035406	0.638605
Personal	-0.03523	0.23244	0.530973	-0.114983	-0.000475	-0.381496
PhD	0.326411	0.055139	-0.08111	-0.147261	-0.550787	0.003344
Terminal	0.323116	0.043033	-0.05898	-0.089008	-0.590407	0.035412
S.F.Ratio	-0.16315	0.259805	-0.27415	-0.259486	-0.142843	0.468753
perc.alumni	0.186611	-0.25709	-0.10372	-0.223982	0.128216	0.012567
Expend	0.328956	-0.16001	0.184206	0.213756	-0.022424	-0.231562
Grad.Rate	0.238822	-0.16752	-0.24534	-0.036192	0.356843	0.313556

Here, we have six Principal components which has 84.16% of data

## Correlation of Principal component data frame



## Inference:

- In PCA\_1, Top10perc, Top25perc are positively correlated, P.Undergrad is negatively correlated
- In PCA\_3, Books is positively correlated

## **2.8) Mention the business implication of using the Principal Component Analysis for this case study.**

### **First Principal analysis - PCA\_1:**

The first principal component accounts for the largest possible variable in the dataset

The first principal component is moderately correlated with five of its original values and it has largest possible variance. The first principal component increases with increasing Top10Perc, Top25perc, Expend, PhD and Terminal scores. This suggests that these five criteria vary together. If one increases, then the remaining ones tend to increase as well

Top10perc (Percentage of new from top 10% of Higher secondary class) is more correlated. Universities concentrate more on top 10% of high scored students in Higher secondary Class and then they concentrate on the top 25% high scored Students. so, For the university acceptance, they are looking for high scored students of higher secondary class

instructional spending is the amount each institution spends on the units that run its educational programs. here it shows that instructional spending per student is positively correlated with the portion of the budget devoted to instruction.

The percentage of faculties of PhD and terminal degree increases as top10 percentage and top25 percentage students increases

### **Second Principal analysis - PCA\_2:**

The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., Perpendicular to) the first principal component and its accounts for next highest variance

The second principal component is moderately correlated with Five scores. They are F.Undergrad, Enroll, P.Undergrad, Accept and apps features. no of Application received and no of acceptance and enrolled are correlated. If the no of application increases, there is chance of increase in number of Acceptance and enrollment in full time and part time undergraduate program.

There are some negative correlated scores, one feature increases when the other features decrease.

### **Third Principal analysis - PCA\_3:**

The third principal components are highly correlated with books and Personal features. Estimated Book cost and estimated personnel spending should be increased for a student

### **Fourth Principal analysis - PCA\_4:**

The fourth principal component is moderately correlated with Room. Board Lightly correlated with Outstate

Fifth and sixth Principal components has some less positively correlated variable

## Conclusion:

Principal Component Analysis (PCA) is a dimensionality reduction technique. The principal application of PCA is dimension reduction. If you have high dimensional data, PCA allows you to reduce the dimensionality of your data so the bulk of the variation that exists in your data across many high dimensions is captured in fewer dimensions.

Based on Principal component analysis done on various parameters of various institutions:

In this Education \_Post 12th standard dataset, there are 17 parameters. After Principal components analysis, dimension reduction leads to main five parameters. They are Top10perc, Top25perc, Expend, PhD and Terminal.

During Universities acceptance for Undergrad program, they should concentrate mainly on these parameters like the top 10 %, total 25% of higher secondary class, the instructional expenditure per student, percentage of facilities with PhD and terminal degrees. Because, if one the parameter increases, other parameters also increases.

## Reference:

- <https://datascience.stackexchange.com/questions/45900/when-to-use-standard-scaler-and-when-normalizer>
- <https://www.pharmaceutical-journal.com/news-and-analysis/infographics/hay-fever-otc-management/20206982.article?firstPass=false>
- <https://www.ncbi.nlm.nih.gov/books/NBK279486/>
- <https://www.pharmaceutical-journal.com/news-and-analysis/infographics/hay-fever-otc-management/20206982.article>
- <https://cowboylifestylenetwork.com/best-colleges-team-roping/>