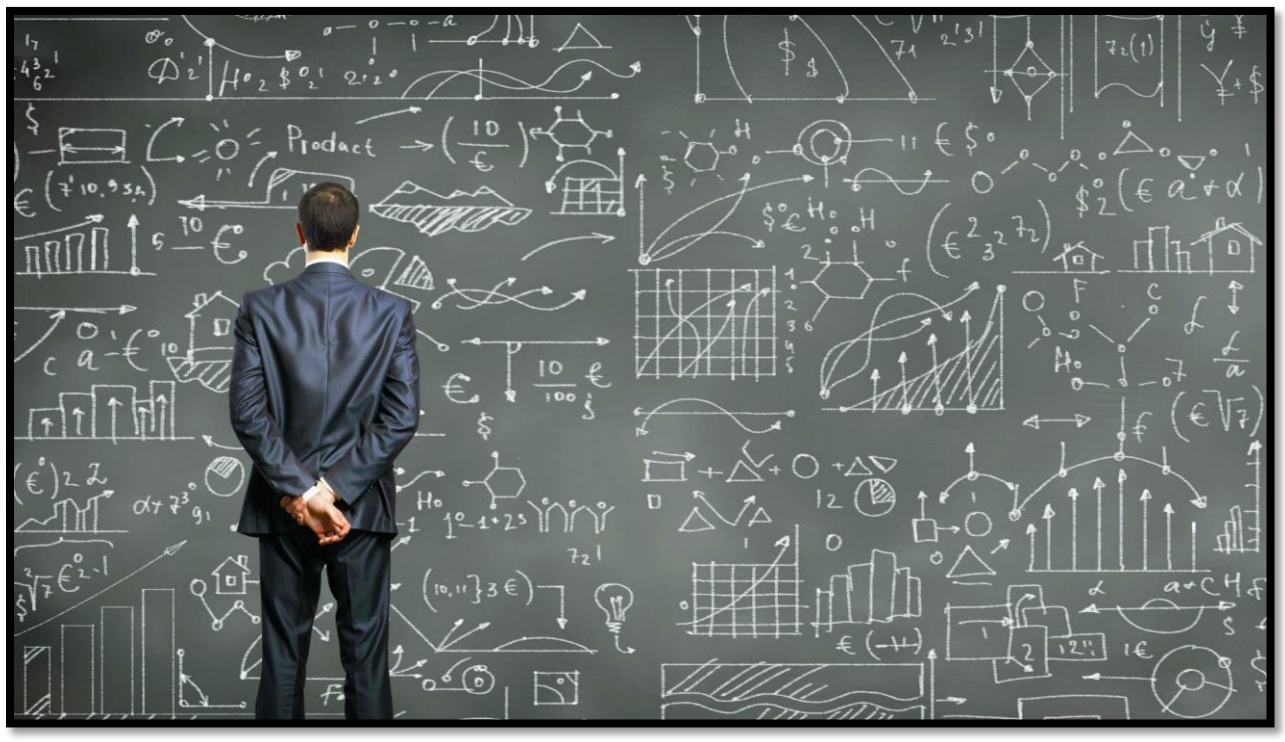
Data Science and Business Analytics

ADVANCED STATISTICS PROJECT REPORT

Vinyas Shreedhar



ANOVA TEST ON SALARY DATA

WHAT IS ANOVA: There are many statistical techniques to identify the systematic sources of variation in a data set. Analysis of Variance (ANOVA) is one of the simplest techniques that identify one or more factors that may contribute to the source of variability. ANOVA is a statistical technique which assumes that two or more population means are same and we conduct hypothesis testing to see if they are same or different for at least one population mean.

There are 2 most common types of ANOVA tests as below:

1. One-Way ANOVA Test – The dependent variable depends on single factor.
2. Two-Way ANOVA Test – The dependent variable depends on two factors and there might be interaction between the two factors themselves or not.

SUMMARY: Salary is hypothesized to depend on educational qualification and occupation. To understand the dependency, the salaries of 40 individuals [SalaryData.csv] are collected and each person's educational qualification and occupation are noted. Educational qualification is at three levels, High School Graduate, Bachelor and Doctorate. Occupation is at four levels, Administrative and Clerical, Sales, Professional or Specialty and Executive or Managerial. A different number of observations are in each level of education – occupation combination.

INTRODUCTION: We will be performing ANOVA (Analysis Of Variance) tests to see the interaction between Education and Occupation levels on Salary.

Assumption is that the data is normally distributed but in reality the normality assumption may not hold good if the sample size is small.

DATA DESCRIPTION:

1. **Salary**
2. **Education (at 3 levels): High School Graduate, Bachelor, Doctorate**
3. **Occupation (at 4 levels): Administrative and Clerical, Sales, Professional or Specialty, Executive or Managerial**

SAMPLE DATASET:

	Education	Occupation	Salary
0	Doctorate	Adm-clerical	153197
1	Doctorate	Adm-clerical	115945
2	Doctorate	Adm-clerical	175935
3	Doctorate	Adm-clerical	220754
4	Doctorate	Sales	170769
5	Doctorate	Sales	219420
6	Doctorate	Sales	237920
7	Doctorate	Sales	160540
8	Doctorate	Sales	180934
9	Doctorate	Prof-specialty	248156
10	Doctorate	Prof-specialty	247724
11	Doctorate	Prof-specialty	249207
12	Doctorate	Prof-specialty	235334
13	Doctorate	Prof-specialty	248871
14	Doctorate	Prof-specialty	257345

Data Types: Salary – int64, Education – object, Occupation - object

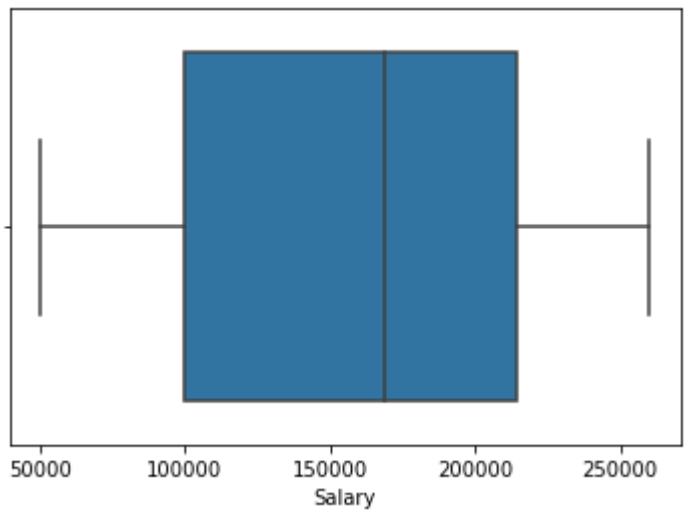
Dataset has 40 rows and 3 columns in total. Salary is the dependent variable whereas Education and Occupation are categorical variables.

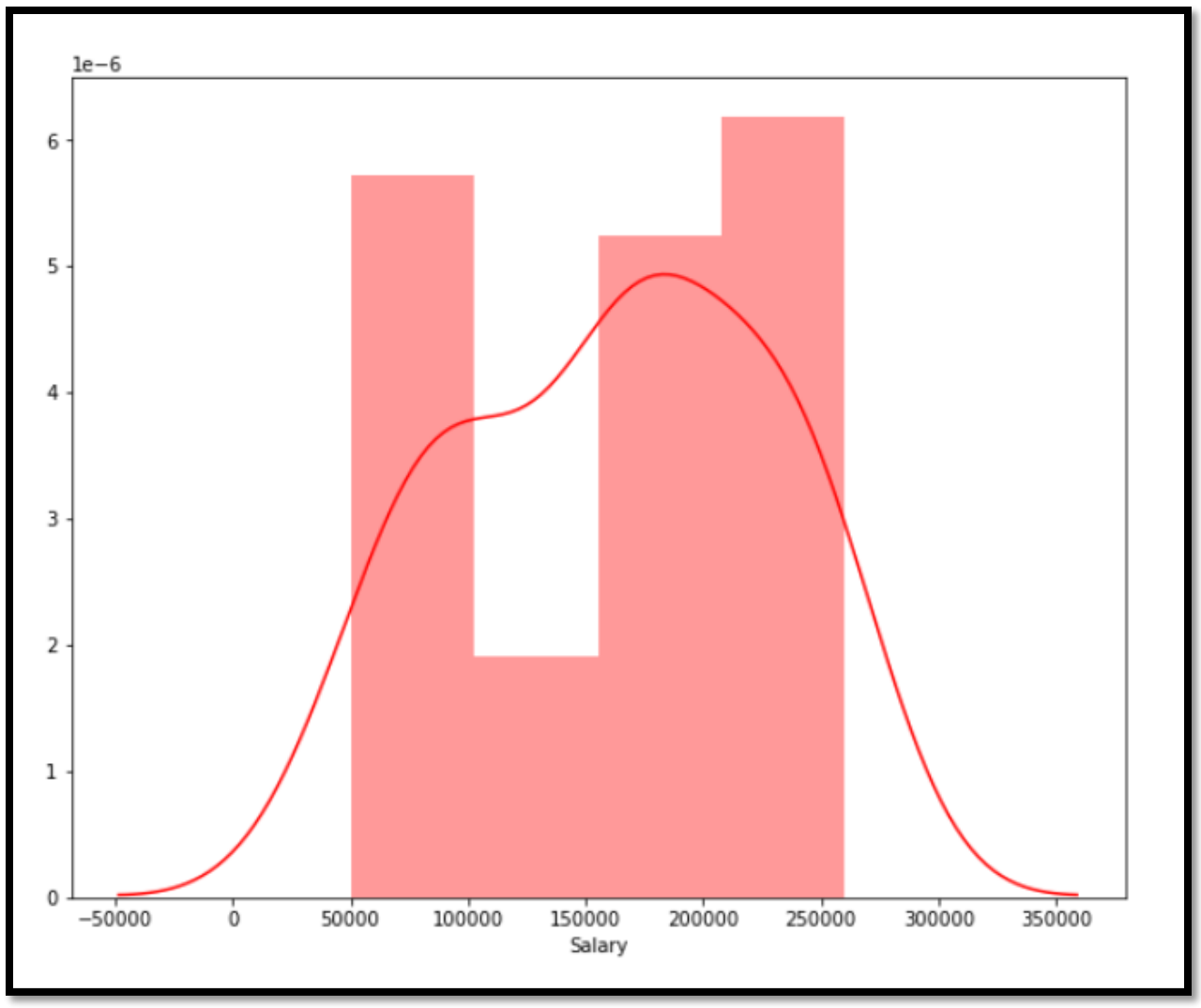
SUMMARY STATISTICS:

	count	mean	std	min	25%	50%	75%	max
Salary	40.0	162186.875	64860.407506	50103.0	99897.5	169100.0	214440.75	260151.0

From the above summary statistics we see that the Mean salary is **162186.875**. Median salary (50%) is **169100**. The salary range is Max – Min which is 260151.0 - 50103.0 = **210048**. The variable salary is close to a **Normal Distribution** as per the below distribution plot.

Let us see how the variable Salary is distributed.





From the above Boxplot and Distribution plot, looks like Salary is normally distributed. However let us test the normality using Shapiro-Wilk's test.

Null Hypothesis - Ho: Salary follows a normal distribution against Alternative Hypothesis - H1: Salary does not follow a normal distribution.

Test Results - **W = 0.9401417970657349** and **P-value = 0.03496258333325386**

Since P-value is smaller than the level of significance 0.05, we reject the null hypothesis and conclude that the feature Salary does not follow a Normal Distribution. However we will still go ahead with our analysis assuming that 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.

Null Hypothesis - Ho: Mean Salary across Education levels are same

Alternative Hypothesis - H1: Mean Salary for at least one of the Education levels are different

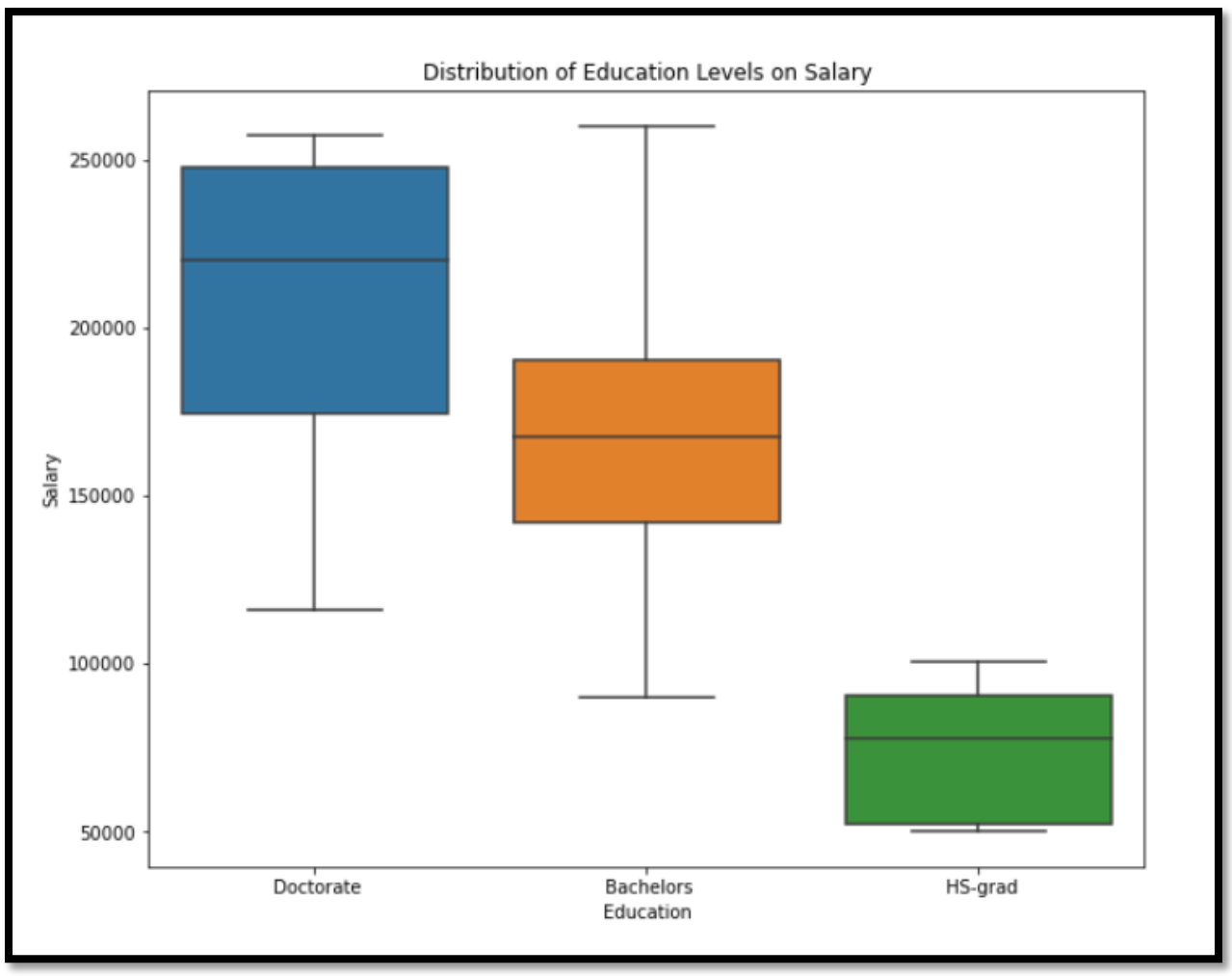
Null Hypothesis - Ho: Mean Salary across Occupation levels are same.

Alternative Hypothesis - H1: Mean Salary for at least one of the Occupation levels are different.

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.

Null Hypothesis - Ho: Mean Salary across Education levels are same

Alternative Hypothesis - H1: Mean Salary for at least one of the Education levels are different



Below are the F-statistic and test results which shows the following:

df = Degrees of Freedom

Sum_sq = Sum of Squares

Mean_sq = Mean Squares

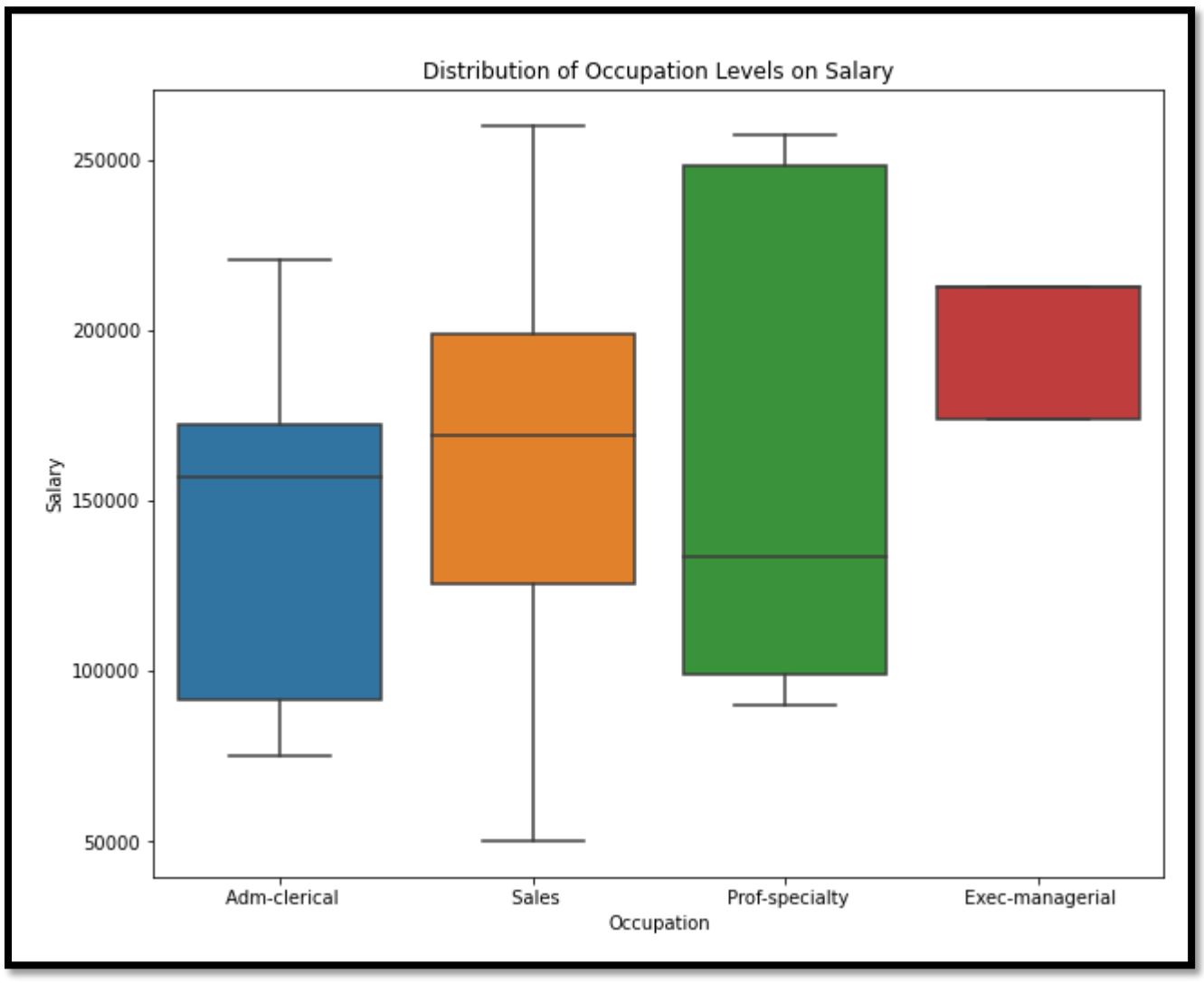
F = F-statistic

PR (>F) = P-value

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

P-value = 1.25770909266291 06e-08 which is less than the level of significance **alpha (0.05)**. Hence we reject the null hypothesis and conclude that the mean Salary across Education levels are not same.

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



Null Hypothesis - Ho: Mean Salary across Occupation levels are same.

Alternative Hypothesis - H1: Mean Salary for at least one of the Occupation levels are different.

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

P-value = 0.45850782664951116 which is greater than the level of significance **alpha (0.05)**. Hence we accept the Null Hypothesis to conclude that Mean Salary across Occupation levels are same.

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.

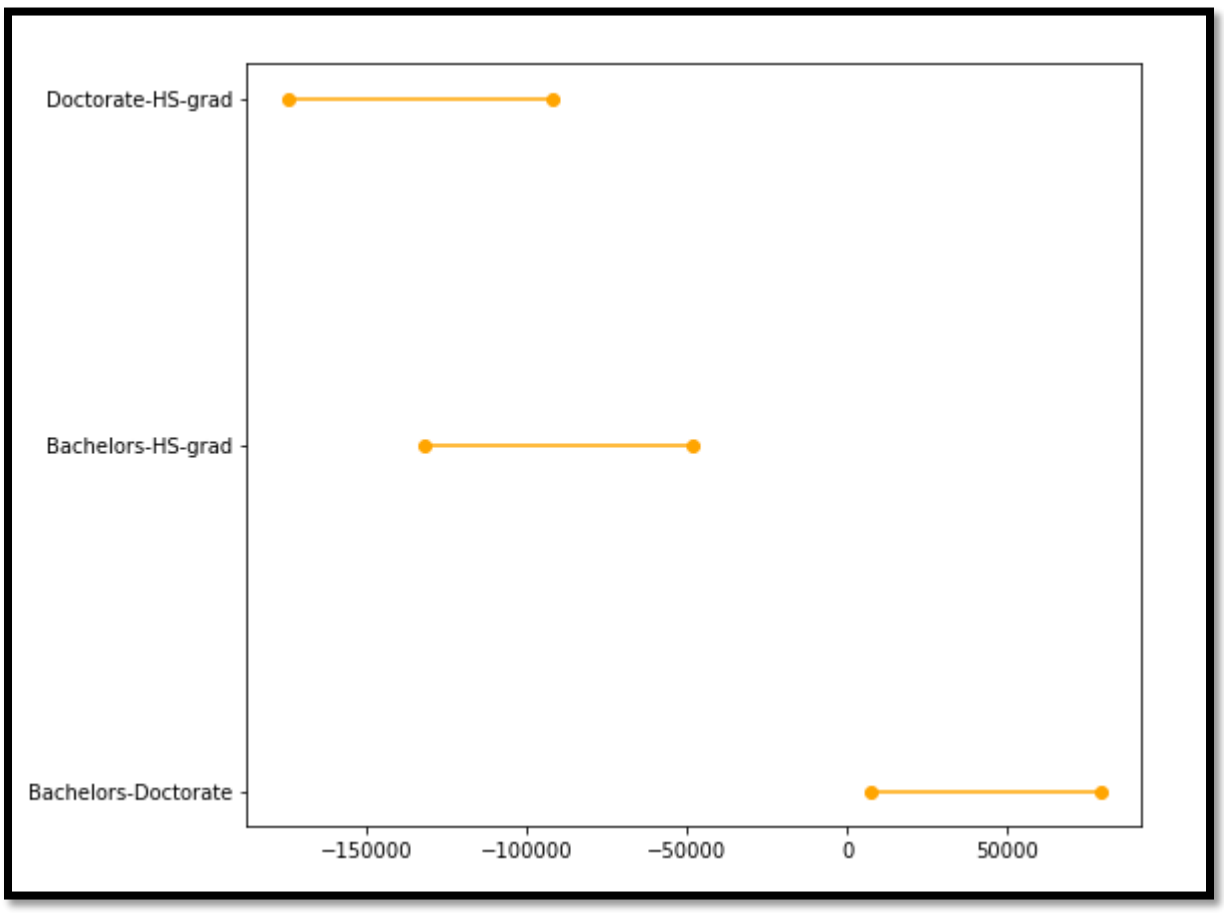
Let's perform the **Multiple Comparison of Means using Tukey's Test** across Education levels.

Null Hypothesis - Ho: Mu Bachelors = Mu Doctorate = Mu Bachelors = Mu HS-grad = Mu Doctorate = Mu HS-grad

Alternative Hypothesis - H1: Mu Bachelors ≠ Mu Doctorate or Mu Bachelors ≠ Mu HS-grad or Mu Doctorate ≠ Mu HS-grad

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



P-value for difference in means between Bachelors and Doctorate is **0.0146**

P-value for difference in means between Bachelors and HS-grad is **0.001**

P-value for difference in means between Doctorate and HS-grad is **0.001**

Since the P-value is smaller than the significance level 0.05 for all groups, we reject the Null Hypothesis and we can thus conclude that there is significant difference in the mean for the above groups.

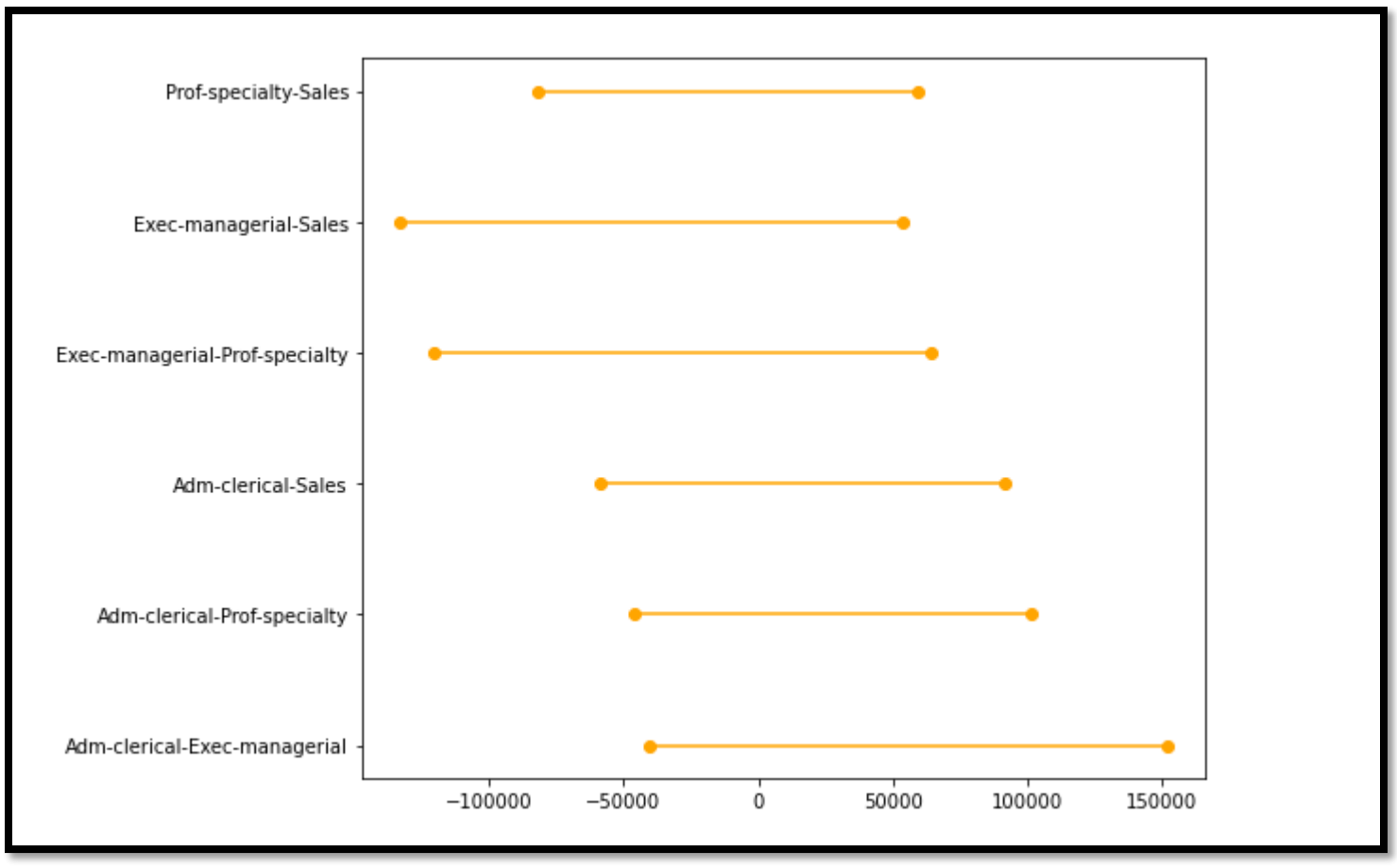
Let's perform the **Multiple Comparison of Means using Tukey's Test** across Occupation levels.

Null Hypothesis - Ho: Mu Adm-clerical = Mu Exec-managerial = Mu Adm-clerical = Mu Prof-specialty = Mu Adm-clerical = Mu Sales = Mu Exec-managerial = Mu Prof-specialty = Mu Exec-managerial = Mu Sales = Mu Prof-specialty = Mu Sales

Alternative Hypothesis - H1: Mu Adm-clerical ≠ Mu Exec-managerial or Mu Adm-clerical ≠ Mu Prof-specialty or Mu Adm-clerical ≠ Mu Sales or Mu Exec-managerial ≠ Mu Prof-specialty or Mu Exec-managerial ≠ Mu Sales or Mu Prof-specialty ≠ Mu Sales

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



P-value for difference in means between Adm-clerical and Exec-managerial is **0.4146**

P-value for difference in means between Adm-clerical and Prof-specialty is **0.7252**

P-value for difference in means between Adm-clerical and Sales is **0.9**

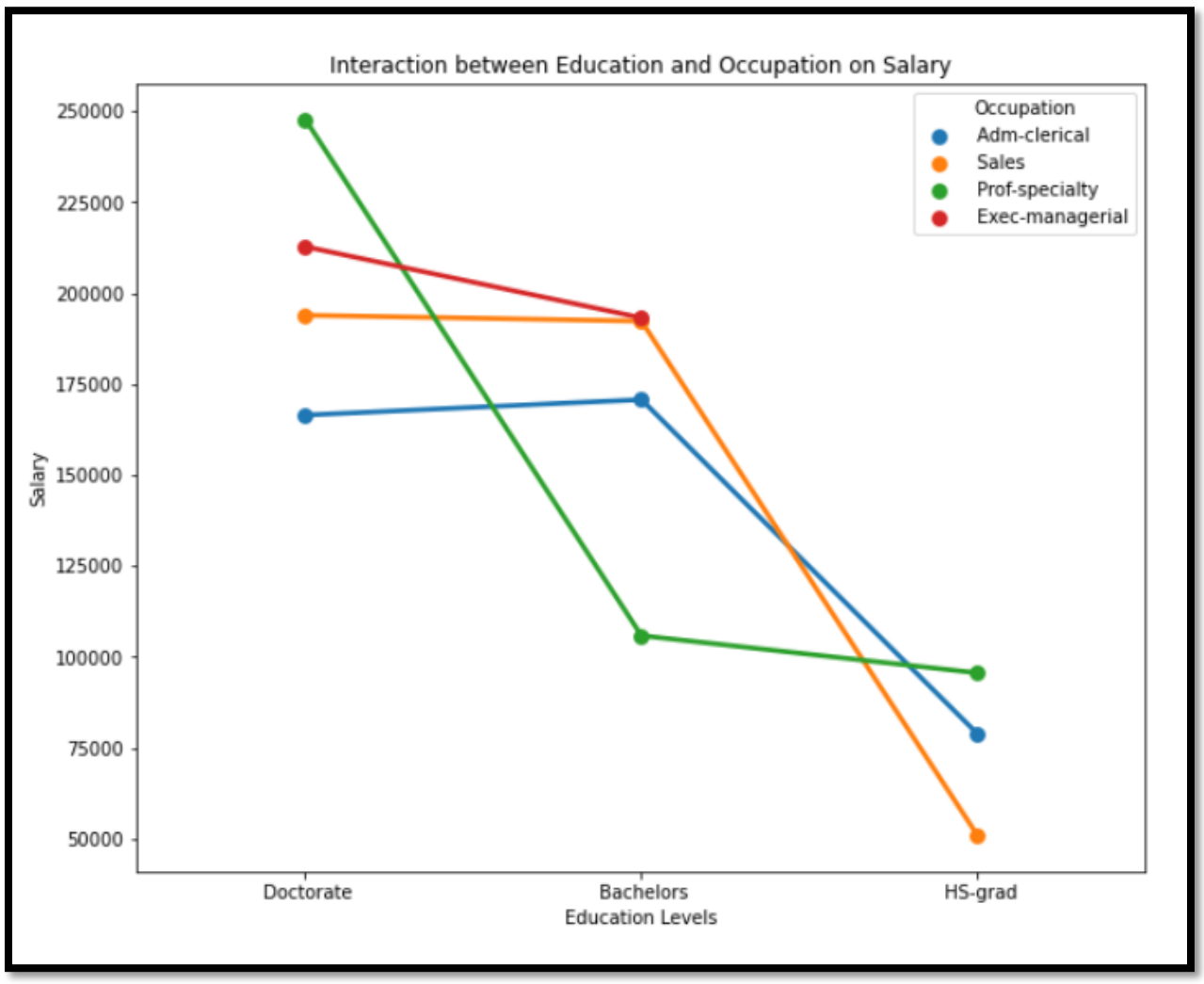
P-value for difference in means between Exec-managerial and Prof-specialty is **0.8263**

P-value for difference in means between Exec-managerial and Sales is **0.6507**

P-value for difference in means between Prof-specialty and Sales is **0.9**

Since the P-value is greater than the level of significance 0.05 we can thus conclude that there is no significant difference in the mean for the above occupation level groups.

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



There is **significant** impact between Education and Occupation with respect to Salary. As we can see there is an increase in the salary from HS-Grad to Bachelors Education and there is a further increase in Salary from Bachelors Education to Doctorate. Among Doctorates, Prof-specialty has the highest salary which is close to **2, 50,000** whereas among HS-Grad, Sales has the lowest salary which is close to **50,000**. Exec-Managerial and Sales intersect at the same point for Bachelors Education at around **2, 00,000**. From the above plot we can say that there is interaction between the two treatments as they are all intersecting at some points.

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?

Null Hypothesis - Ho: Mean Salary across Occupation and Education levels are same.

Alternative Hypothesis - H1: Mean Salary for atleast one Occupation and Education level are not same.

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

P-value for Education is **1.9815392541413873e-08** which is less than alpha **0.05**.

P-value for Occupation is **0.3545824933162919** which is greater than alpha **0.05**.

We can interpret that there is some interaction between Education and Occupation in relation to Salary.

Since there is interaction between Education and Occupation on Salary, let's introduce a new term while performing Two Way ANOVA.

	df	sum_sq	mean_sq	F	PR(>F)
Education	2.0	1.026955e+11	5.134773e+10	72.211958	5.466264e-12
Occupation	3.0	5.519946e+09	1.839982e+09	2.587626	7.211580e-02
Education:Occupation	6.0	3.634909e+10	6.058182e+09	8.519815	2.232500e-05
Residual	29.0	2.062102e+10	7.110697e+08	NaN	NaN

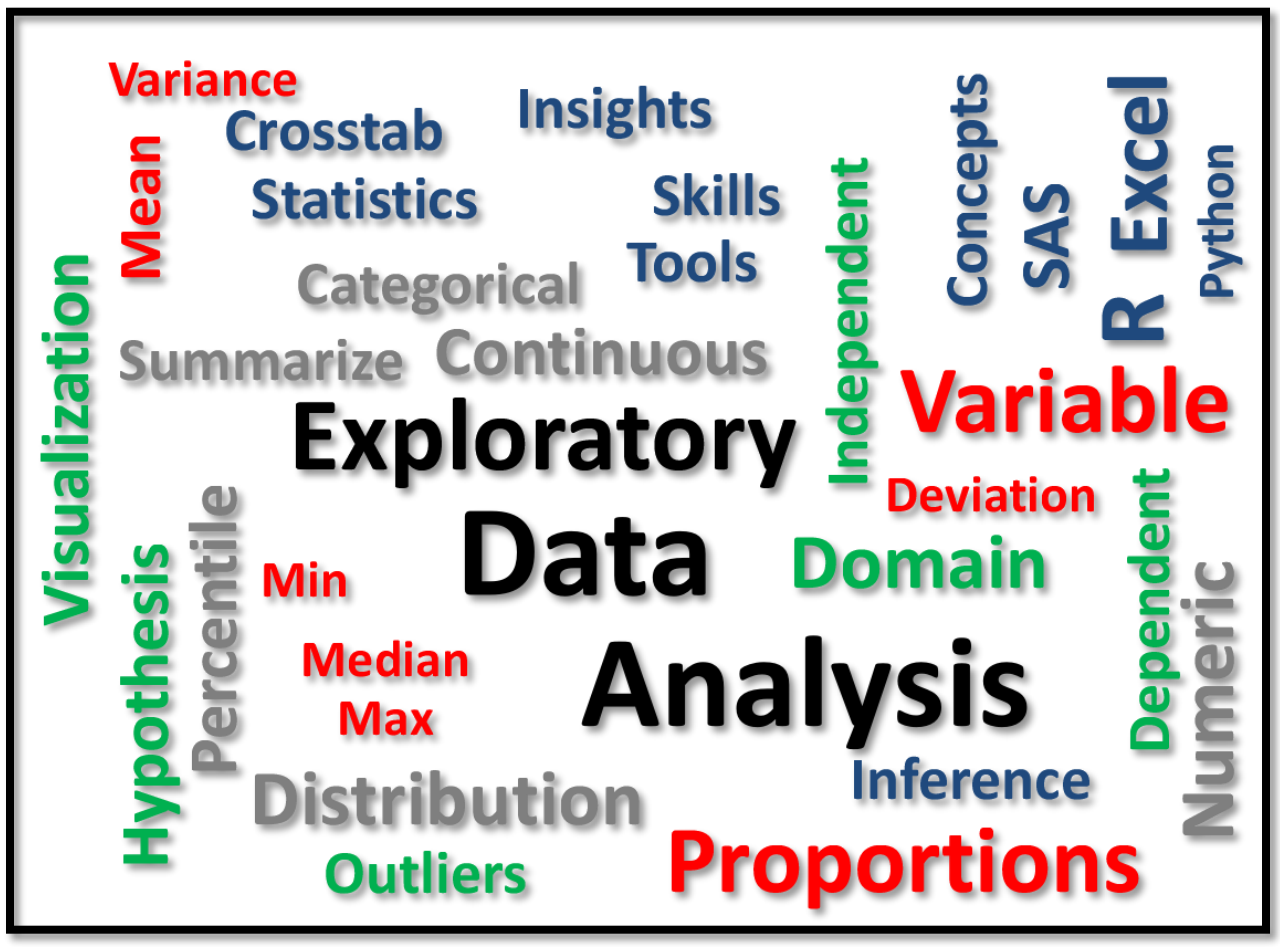
Due to the inclusion of the interaction effect term, we can see a slight change in the P-value of the first two treatments as compared to the Two-Way ANOVA without the interaction effect terms. And we see that the P-value (**2.2325004523478696e-05**) of the interaction effect term of 'Education:Occupation' is less than the level of significance (**0.05**) which suggests that we reject the Null Hypothesis in this case.

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

CONCLUSION: It is observed that the variation in Salary is significantly impacted by different Education and Occupation levels along with their interaction effect. ANOVA in this particular case study helps to understand which independent factor (Education & Occupation) can explain the variation in the Salary.

EDA and PCA on EDUCATION – POST 12TH STANDARD

WHAT IS EDA: EDA refers to Exploratory Data Analysis.



Data Scientists spend **70%** of the time in exploring data, analyzing it and deriving inferences from the provided past data. The remaining **30%** constitutes predictive modelling and prescriptive statistics. This is the most crucial part of Data Science Life Cycle project. EDA forms part of Descriptive Statistics as we are dealing with past data. The famous quote by Jim Bergeson states "***Data will talk to you if you are willing to listen to it.***" Let us perform EDA on the provided dataset to explore what insights we can infer from the same.

EDA follows the below step by step process:

1. Importing Data from an external source (xlsx or csv files)
2. Check the shape of the dataset (no of rows and columns)
3. Information on dataset (no of categorical, object, integer and float variables)
4. Explore for any missing values in the dataset and perform Missing Value Treatment if required
5. Check if there are duplicate values in the dataset
6. Perform Summary Statistics / 5-Point Summary on the numerical variables to extract the data distribution and behavior
7. Identify Outliers and perform Outlier Treatment
8. Univariate Analysis
9. Bivariate Analysis

SUMMARY: The dataset contains information on various colleges with regard to education levels post 12th Standard with the below data dictionary for our reference.

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

INTRODUCTION: We will be performing Exploratory Data Analysis on the dataset to draw insights from the EDA.

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

Importing Dataset and viewing the sample dataset.

	Names	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alum
0	Abilene Christian University	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12
1	Adelphi University	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16
2	Adrian College	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30
3	Agnes Scott College	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37
4	Alaska Pacific University	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
772	Worcester State College	2197	1515	543	4	26	3089	2029	6797	3900	500	1200	60	60	21.0	14
773	Xavier University	1959	1805	695	24	47	2849	1107	11520	4960	600	1250	73	75	13.3	31
774	Xavier University of Louisiana	2097	1915	695	34	61	2793	166	6900	4200	617	781	67	75	14.4	20
775	Yale University	10705	2453	1317	95	99	5217	83	19840	6510	630	2115	96	96	5.8	49
776	York College of Pennsylvania	2989	1855	691	28	63	2988	1726	4990	3560	500	1250	75	75	18.1	28

777 rows × 18 columns

The dataset contains 777 rows and 18 columns.

Information on the dataset.

```
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)
```

We can see that there is **1 object** variable and **17 numerical** variables which includes 1 float and 16 integer variable data types.

Let's check for any missing values.

```
Names          0
Apps           0
Accept         0
Enroll         0
Top10perc      0
Top25perc      0
F.Undergrad    0
P.Undergrad    0
Outstate       0
Room.Board     0
Books          0
Personal       0
PhD            0
Terminal       0
S.F.Ratio      0
perc.alumni    0
Expend         0
Grad.Rate      0
```

The above table shows that there are **Zero** Missing Values in the dataset.

As per the analysis we found that there are **No Duplicate** values in the dataset.

Summary Statistics / 5-Point Summary on the numerical variables to extract the data distribution and behavior.

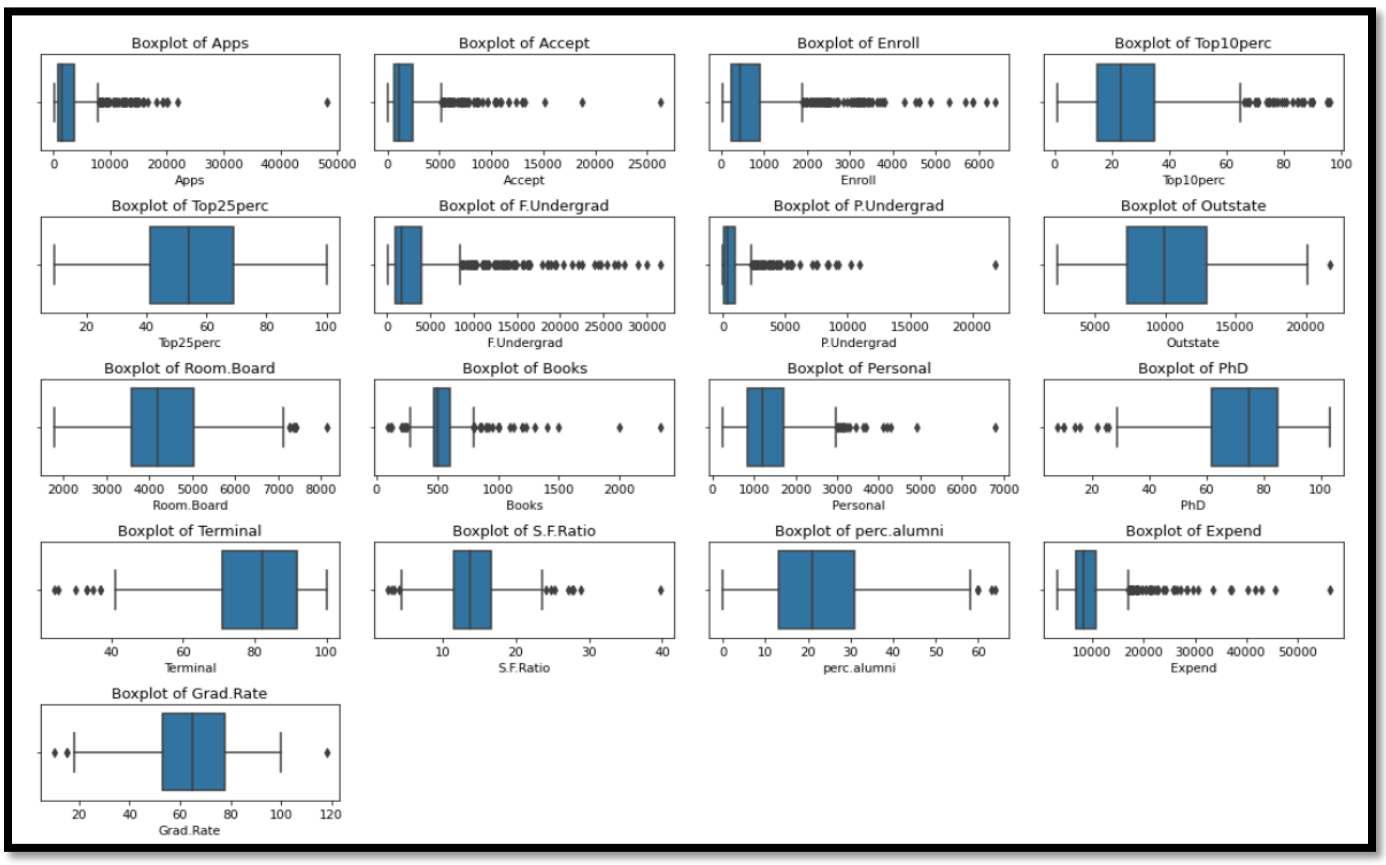
	count	mean	std	min	25%	50%	75%	max
Apps	777.0	3001.64	3870.20	81.0	776.0	1558.0	3624.0	48094.0
Accept	777.0	2018.80	2451.11	72.0	604.0	1110.0	2424.0	26330.0
Enroll	777.0	779.97	929.18	35.0	242.0	434.0	902.0	6392.0
Top10perc	777.0	27.56	17.64	1.0	15.0	23.0	35.0	96.0
Top25perc	777.0	55.80	19.80	9.0	41.0	54.0	69.0	100.0
F.Undergrad	777.0	3699.91	4850.42	139.0	992.0	1707.0	4005.0	31643.0
P.Undergrad	777.0	855.30	1522.43	1.0	95.0	353.0	967.0	21836.0
Outstate	777.0	10440.67	4023.02	2340.0	7320.0	9990.0	12925.0	21700.0
Room.Board	777.0	4357.53	1096.70	1780.0	3597.0	4200.0	5050.0	8124.0
Books	777.0	549.38	165.11	96.0	470.0	500.0	600.0	2340.0
Personal	777.0	1340.64	677.07	250.0	850.0	1200.0	1700.0	6800.0
PhD	777.0	72.66	16.33	8.0	62.0	75.0	85.0	103.0
Terminal	777.0	79.70	14.72	24.0	71.0	82.0	92.0	100.0
S.F.Ratio	777.0	14.09	3.96	2.5	11.5	13.6	16.5	39.8
perc.alumni	777.0	22.74	12.39	0.0	13.0	21.0	31.0	64.0
Expend	777.0	9660.17	5221.77	3186.0	6751.0	8377.0	10830.0	56233.0
Grad.Rate	777.0	65.46	17.18	10.0	53.0	65.0	78.0	118.0

- The average number of applications received are **3001** from a minimum of **81** to a maximum of **48094** applications.
- The average cost of room and board is **4357.53** dollars.
- Personal spending for a student is between **250** to **6800** dollars.
- Average Percentage of faculties with Ph.D.'s is **72.66%.**
- Median Percentage of new students from top 10% of Higher Secondary class is **23%** and the data is almost **normally** distributed since Mean is at **27.56%.**

Identifying Outliers. We only want to look at numerical variables hence I am creating another data frame removing the Names column which is categorical and storing it as **"dfnum"**.

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

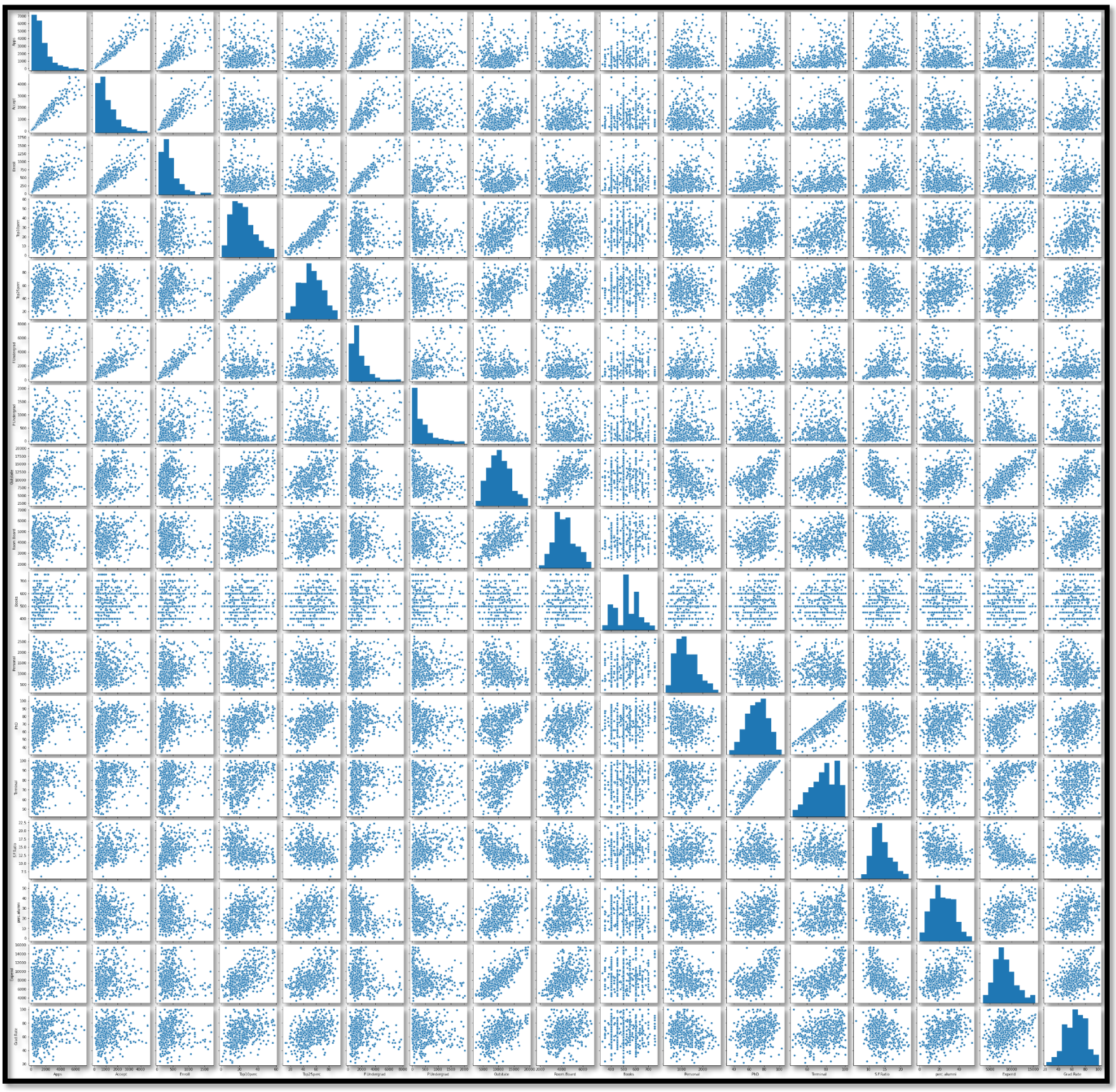
To identify outliers we will use the **Boxplot** technique which shows the distribution of data along with **Inter-Quartile Range (IQR)**. The data points outside the range of the box plot are Outliers at both ends.



From the above plots we can see that almost **every feature has outliers** except for Boxplot of Percentage of new students from top 25% of Higher Secondary class (1st plot from the 2nd row). Some features are **left skewed** and some are **right skewed** distributions.

UNIVARIATE ANALYSIS: This forms the part of descriptive statistics which shows relationship of a **single** variable. Since we have 17 features we will construct a Pair plot to see the relationship of each variable and how they are correlated with other variables.

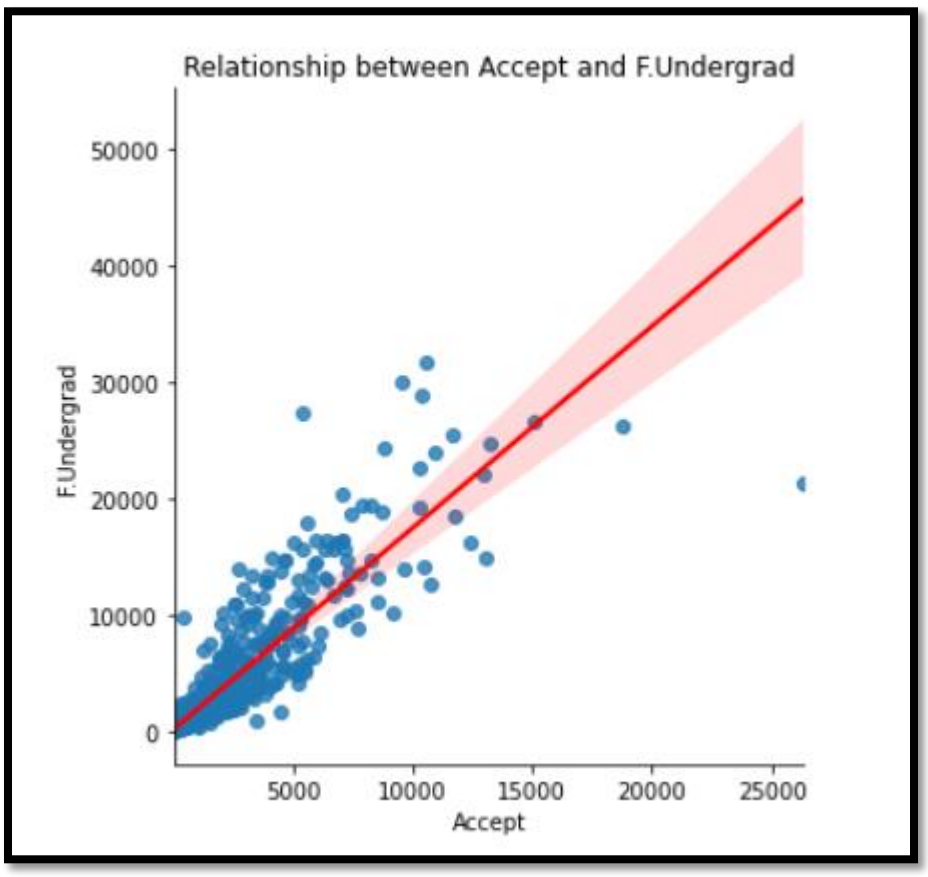
PAIR Plot: Pair plot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

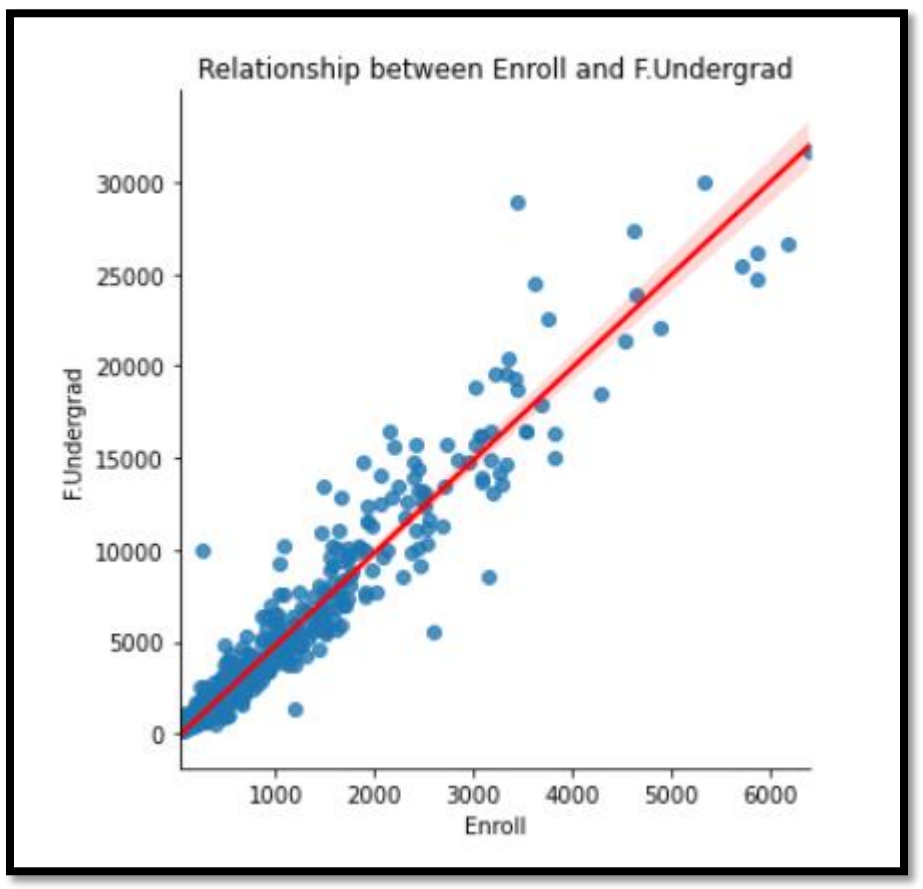


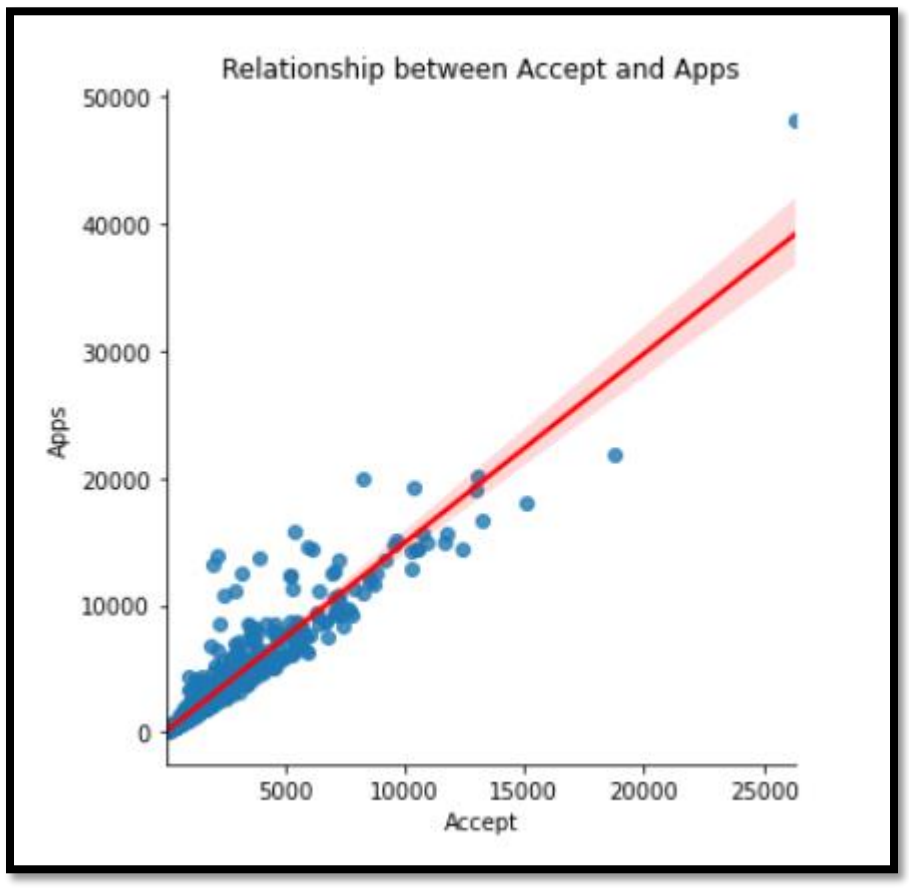
Observations from the above Pair plot:

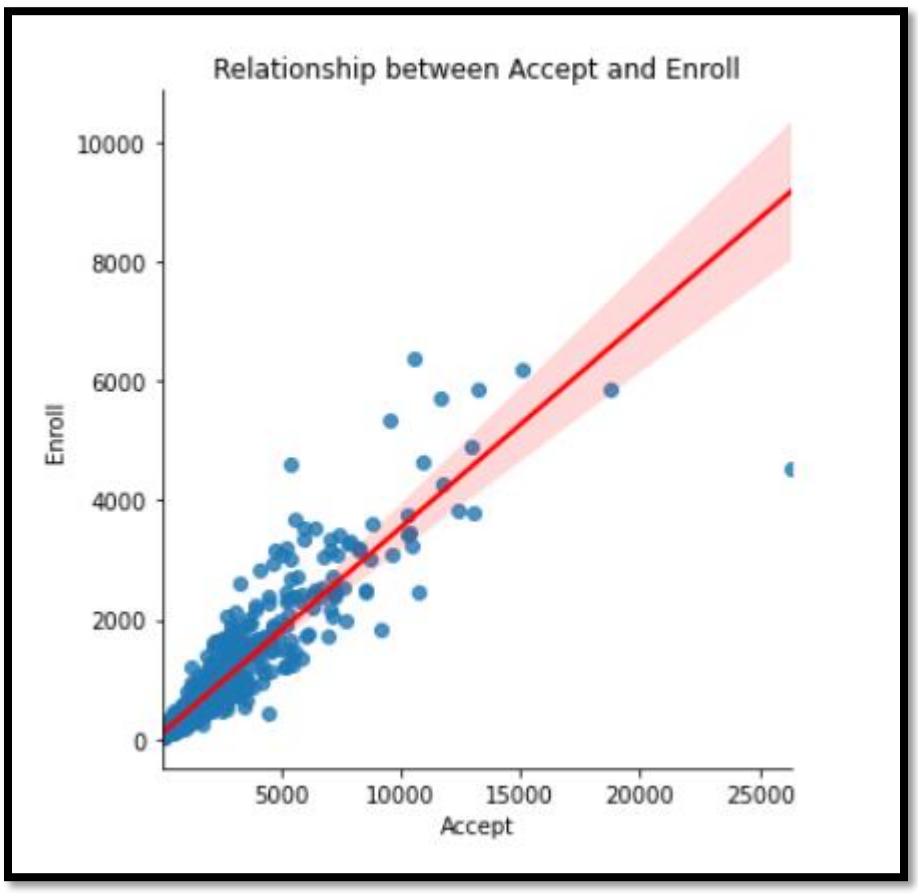
1. Accept, Enroll and F.Undergrad have a positive relationship with Apps.
2. Accept has a positive relationship with F.Undergrad and Enroll.
3. Enroll and F.Undergrad have a strong linear positive relationship and looks as if the data points are almost on the best fit line. This means as Number of new students enrolled increases the Number of full-time undergraduate students also increases.
4. Top 25 Perc also shows a strong positive relationship with Top 10 Perc.
5. Rest of the plots do not show a rich relationship between the variables.

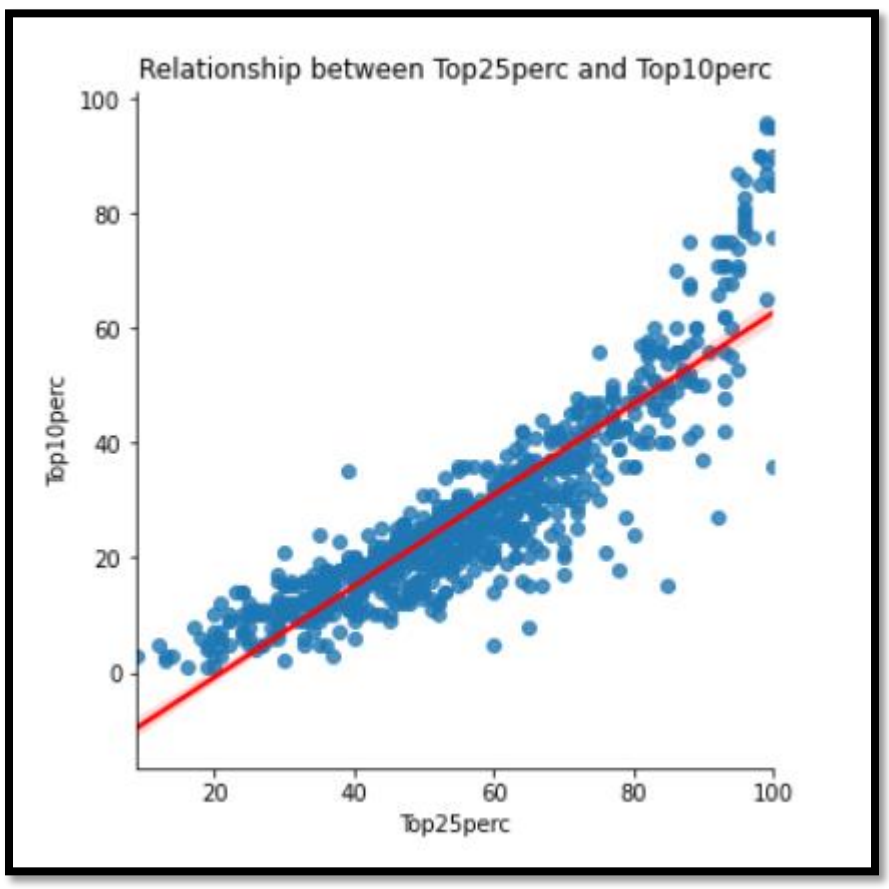
Note: Since we had to accommodate 17 variables in the above pair plot which is not clearly visible here, I have provided below the scatter plot for some of the variables which provide some insights into the relationship.





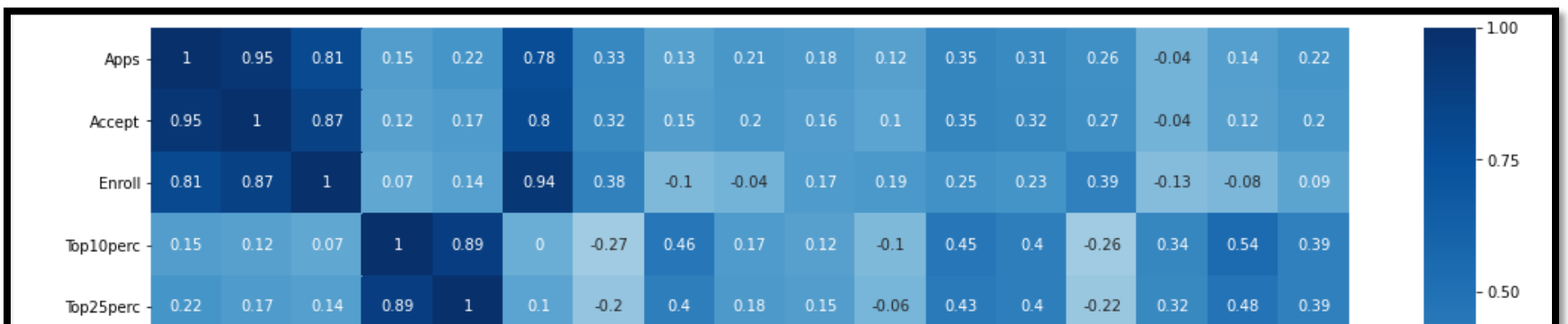



Relationship between Accept and Enroll
Enroll
Accept
10000
8000
6000
4000
2000
0
5000
10000
15000
20000
25000


Relationship between Top25perc and Top10perc
Top10perc
Top25perc
100
80
60
40
20
0
20
40
60
80
100

BIVARIATE ANALYSIS: This involves analyzing the relationship between two variables. One of the best methods to analyze bivariate data is **Correlation Coefficients**. Correlation values are always between **1 and -1**. Those points closest to 1 show a **positive** relationship and those values closest to -1 shows **negative** relationship. Those which are closer to Zero show no relationship between the features. Hence we will first create a Correlation table and then visualize it with a **Heatmap**.

	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	1.00	0.95	0.81	0.15	0.22	0.78	0.33	0.13	0.21	0.18	0.12	0.35	0.31	0.26	-0.04	0.14	0.22
Accept	0.95	1.00	0.87	0.12	0.17	0.80	0.32	0.15	0.20	0.16	0.10	0.35	0.32	0.27	-0.04	0.12	0.20
Enroll	0.81	0.87	1.00	0.07	0.14	0.94	0.38	-0.10	-0.04	0.17	0.19	0.25	0.23	0.39	-0.13	-0.08	0.09
Top10perc	0.15	0.12	0.07	1.00	0.89	0.00	-0.27	0.46	0.17	0.12	-0.10	0.45	0.40	-0.26	0.34	0.54	0.39
Top25perc	0.22	0.17	0.14	0.89	1.00	0.10	-0.20	0.40	0.18	0.15	-0.06	0.43	0.40	-0.22	0.32	0.48	0.39
F.Undergrad	0.78	0.80	0.94	0.00	0.10	1.00	0.48	-0.19	-0.06	0.21	0.21	0.24	0.21	0.44	-0.22	-0.13	0.01
P.Undergrad	0.33	0.32	0.38	-0.27	-0.20	0.48	1.00	-0.26	0.06	0.08	0.13	-0.09	-0.07	0.33	-0.27	-0.31	-0.09
Outstate	0.13	0.15	-0.10	0.46	0.40	-0.19	-0.26	1.00	0.60	-0.04	-0.30	0.42	0.44	-0.45	0.49	0.78	0.51
Room.Board	0.21	0.20	-0.04	0.17	0.18	-0.06	0.06	0.60	1.00	0.06	-0.19	0.32	0.34	-0.24	0.21	0.52	0.35
Books	0.18	0.16	0.17	0.12	0.15	0.21	0.08	-0.04	0.06	1.00	0.17	0.14	0.14	0.05	-0.06	0.08	-0.02
Personal	0.12	0.10	0.19	-0.10	-0.06	0.21	0.13	-0.30	-0.19	0.17	1.00	-0.12	-0.12	0.14	-0.26	-0.21	-0.21
PhD	0.35	0.35	0.25	0.45	0.43	0.24	-0.09	0.42	0.32	0.14	-0.12	1.00	0.83	-0.04	0.24	0.51	0.30
Terminal	0.31	0.32	0.23	0.40	0.40	0.21	-0.07	0.44	0.34	0.14	-0.12	0.83	1.00	-0.09	0.27	0.54	0.26
S.F.Ratio	0.26	0.27	0.39	-0.26	-0.22	0.44	0.33	-0.45	-0.24	0.05	0.14	-0.04	-0.09	1.00	-0.35	-0.54	-0.21
perc.alumni	-0.04	-0.04	-0.13	0.34	0.32	-0.22	-0.27	0.49	0.21	-0.06	-0.26	0.24	0.27	-0.35	1.00	0.43	0.43
Expend	0.14	0.12	-0.08	0.54	0.48	-0.13	-0.31	0.78	0.52	0.08	-0.21	0.51	0.54	-0.54	0.43	1.00	0.37
Grad.Rate	0.22	0.20	0.09	0.39	0.39	0.01	-0.09	0.51	0.35	-0.02	-0.21	0.30	0.26	-0.21	0.43	0.37	1.00



Apps and **Accept** have a strong positive correlation with coefficient **0.95**.

F.Undergrad and **Enroll** have a positive correlation with coefficient **0.94**.

Top 25 perc and **Top 10 perc** have a strong relationship with coefficient **0.89**.

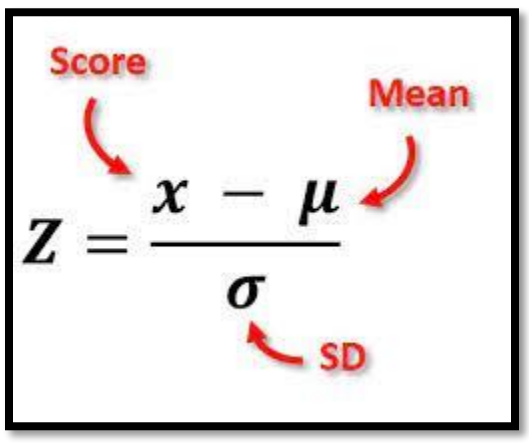
Enroll and **Accept** too have a positive relationship with coefficient **0.87**.

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

Feature Scaling is a technique used in data pre-processing to standardize the scales between different features. Since different features have different scales or weights, the machine learning model would not provide accurate results. Hence scaling is required for PCA in this case.

Two most commonly used Scaling techniques are:

Z-score Scaling: a method in which all the values are converted to z-scores. Z-score brings the mean to Zero and scales the data to unit variance.



Source: Google

Min-Max Scaling: this method linearly rescales every feature to the [0,1] interval. The presence of this bounded range in contrast to z-score scaling is that we will end up with smaller standard deviations, which can suppress the effect of outliers.

$$x_{\text{norm}} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Source: Google

Let's perform Z-score scaling for column standardization.

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Exp
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	1.013776	-0.867574	-0.501
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	-0.477704	-0.544572	0.166
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	-0.300749	0.585935	-0.177
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	-1.615274	1.151188	1.792
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	-0.553542	-1.675079	0.241

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

Since Both covariance and correlation matrices show the relationship between 2 variables. **Covariance** shows the direction of the linear relationship between the variables whereas **Correlation** shows the strength and direction of the linear relationship between variables. I have provided both the tables below.

Correlation table

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \textbf{(Eq.1)}$$

where:

cov is the covariance

σ_X is the standard deviation of X

σ_Y is the standard deviation of Y

Source: Wikipedia

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alun
Apps	1.000000	0.943451	0.846822	0.338834	0.351640	0.814491	0.398264	0.050159	0.164939	0.132559	0.178731	0.390697	0.369491	0.095633	-0.09022
Accept	0.943451	1.000000	0.911637	0.192447	0.247476	0.874223	0.441271	-0.025755	0.090899	0.113525	0.200989	0.355758	0.337583	0.176229	-0.15999
Enroll	0.846822	0.911637	1.000000	0.181294	0.226745	0.964640	0.513069	-0.155477	-0.040232	0.112711	0.280929	0.331469	0.308274	0.237271	-0.18079
Top10perc	0.338834	0.192447	0.181294	1.000000	0.891995	0.141289	-0.105356	0.562331	0.371480	0.118858	-0.093316	0.531828	0.491135	-0.384875	0.455485
Top25perc	0.351640	0.247476	0.226745	0.891995	1.000000	0.199445	-0.053577	0.489394	0.331490	0.115527	-0.080810	0.545862	0.524749	-0.294629	0.417864
F.Undergrad	0.814491	0.874223	0.964640	0.141289	0.199445	1.000000	0.570512	-0.215742	-0.068890	0.115550	0.317200	0.318337	0.300019	0.279703	-0.22946
P.Undergrad	0.398264	0.441271	0.513069	-0.105356	-0.053577	0.570512	1.000000	-0.253512	-0.061326	0.081200	0.319882	0.149114	0.141904	0.232531	-0.28079
Outstate	0.050159	-0.025755	-0.155477	0.562331	0.489394	-0.215742	-0.253512	1.000000	0.654256	0.038855	-0.299087	0.382982	0.407983	-0.554821	0.566262
Room.Board	0.164939	0.090899	-0.040232	0.371480	0.331490	-0.068890	-0.061326	0.654256	1.000000	0.127963	-0.199428	0.329202	0.374540	-0.362628	0.272363
Books	0.132559	0.113525	0.112711	0.118858	0.115527	0.115550	0.081200	0.038855	0.127963	1.000000	0.179295	0.026906	0.099955	-0.031929	-0.04020
Personal	0.178731	0.200989	0.280929	-0.093316	-0.080810	0.317200	0.319882	-0.299087	-0.199428	0.179295	1.000000	-0.010936	-0.030613	0.136345	-0.28596
PhD	0.390697	0.355758	0.331469	0.531828	0.545862	0.318337	0.149114	0.382982	0.329202	0.026906	-0.010936	1.000000	0.849587	-0.130530	0.249009
Terminal	0.369491	0.337583	0.308274	0.491135	0.524749	0.300019	0.141904	0.407983	0.374540	0.099955	-0.030613	0.849587	1.000000	-0.160104	0.267130
S.F.Ratio	0.095633	0.176229	0.237271	-0.384875	-0.294629	0.279703	0.232531	-0.554821	-0.362628	-0.031929	0.136345	-0.130530	-0.160104	1.000000	-0.40292
perc.alumni	-0.090226	-0.159990	-0.180794	0.455485	0.417864	-0.229462	-0.280792	0.566262	0.272363	-0.040208	-0.285968	0.249009	0.267130	-0.402929	1.000000
Expend	0.259592	0.124717	0.064169	0.660913	0.527447	0.018652	-0.083568	0.672779	0.501739	0.112409	-0.097892	0.432762	0.438799	-0.583832	0.417712
Grad.Rate	0.146755	0.067313	-0.022341	0.494989	0.477281	-0.078773	-0.257001	0.571290	0.424942	0.001061	-0.269344	0.305038	0.289527	-0.306710	0.490898

Covariance table

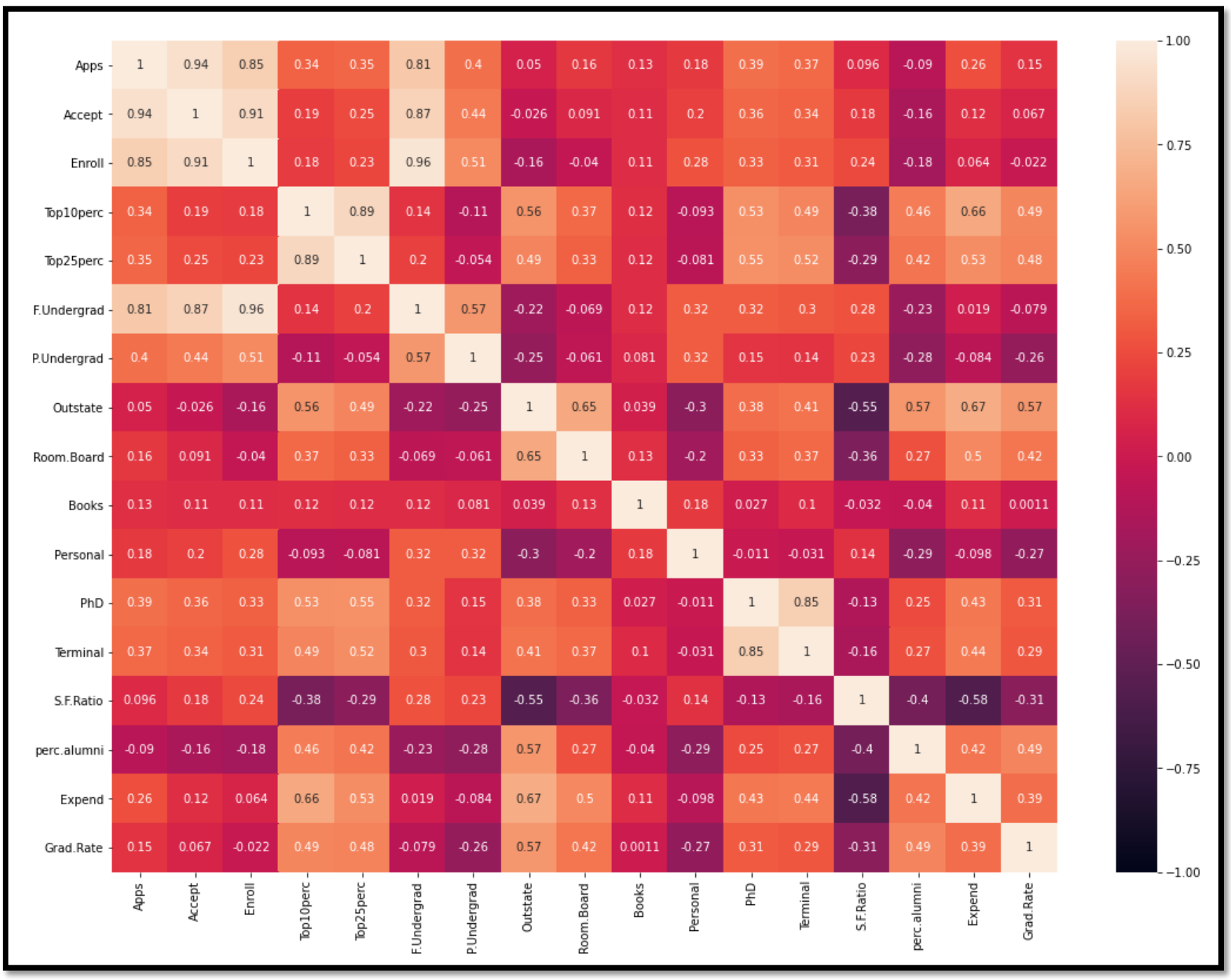
$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \qquad \textbf{(Eq.1)}$$

where $\mathrm{E}[X]$ is the expected value of X, also known as the mean of X.

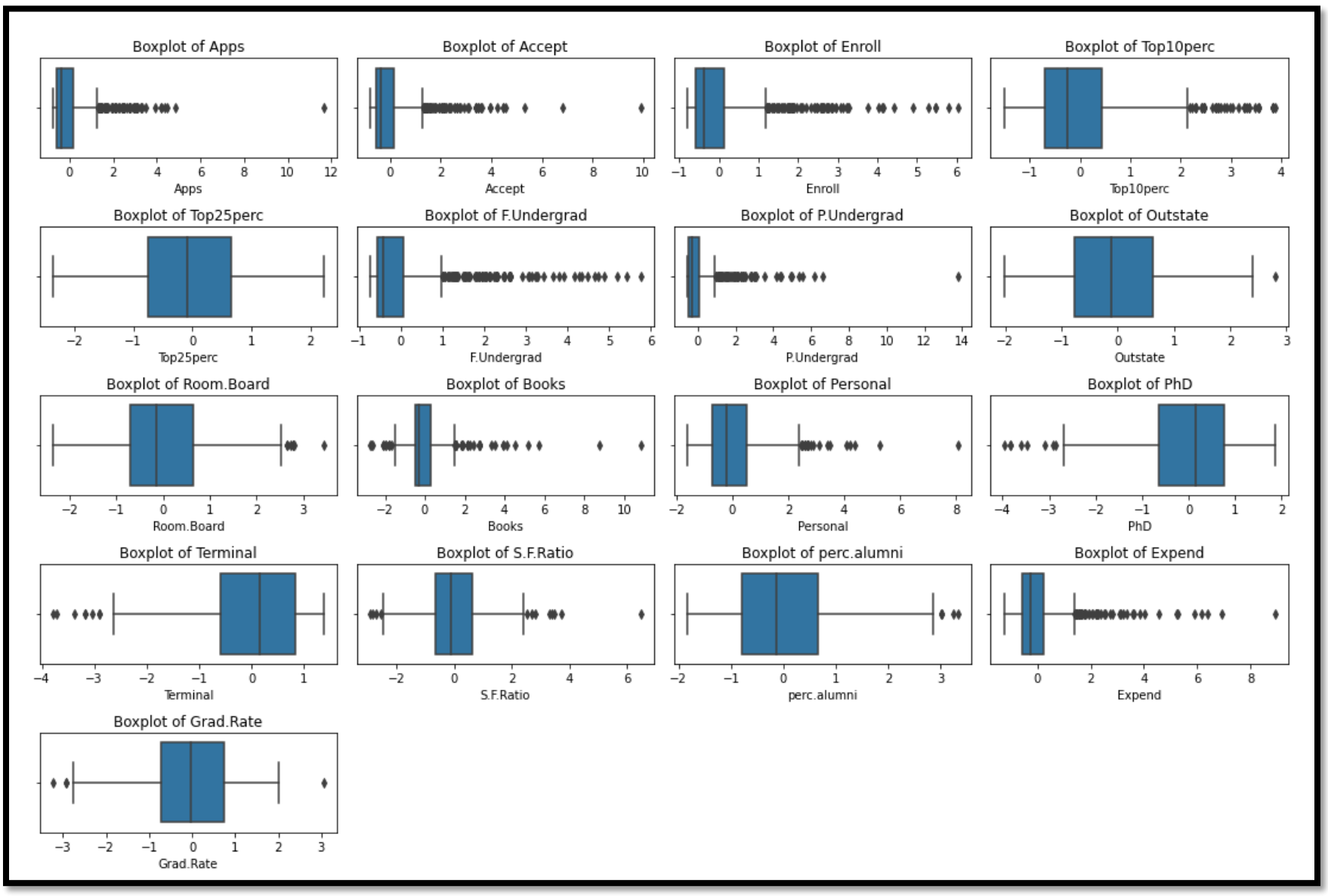
Source: Wikipedia

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alun
Apps	1.001289	0.944666	0.847913	0.339270	0.352093	0.815540	0.398777	0.050224	0.165152	0.132729	0.178961	0.391201	0.369968	0.095756	-0.09034
Accept	0.944666	1.001289	0.912811	0.192695	0.247795	0.875350	0.441839	-0.025788	0.091016	0.113672	0.201248	0.356216	0.338018	0.176456	-0.16019
Enroll	0.847913	0.912811	1.001289	0.181527	0.227037	0.965883	0.513730	-0.155678	-0.040284	0.112856	0.281291	0.331896	0.308671	0.237577	-0.18102
Top10perc	0.339270	0.192695	0.181527	1.001289	0.893144	0.141471	-0.105492	0.563055	0.371959	0.119012	-0.093437	0.532513	0.491768	-0.385370	0.456072
Top25perc	0.352093	0.247795	0.227037	0.893144	1.001289	0.199702	-0.053646	0.490024	0.331917	0.115676	-0.080914	0.546566	0.525425	-0.295009	0.418403
F.Undergrad	0.815540	0.875350	0.965883	0.141471	0.199702	1.001289	0.571247	-0.216020	-0.068979	0.115699	0.317608	0.318747	0.300406	0.280064	-0.22975
P.Undergrad	0.398777	0.441839	0.513730	-0.105492	-0.053646	0.571247	1.001289	-0.253839	-0.061405	0.081304	0.320294	0.149306	0.142086	0.232830	-0.28115
Outstate	0.050224	-0.025788	-0.155678	0.563055	0.490024	-0.216020	-0.253839	1.001289	0.655100	0.038905	-0.299472	0.383476	0.408509	-0.555536	0.566992
Room.Board	0.165152	0.091016	-0.040284	0.371959	0.331917	-0.068979	-0.061405	0.655100	1.001289	0.128128	-0.199685	0.329627	0.375022	-0.363095	0.272714
Books	0.132729	0.113672	0.112856	0.119012	0.115676	0.115699	0.081304	0.038905	0.128128	1.001289	0.179526	0.026940	0.100084	-0.031970	-0.04026
Personal	0.178961	0.201248	0.281291	-0.093437	-0.080914	0.317608	0.320294	-0.299472	-0.199685	0.179526	1.001289	-0.010950	-0.030653	0.136521	-0.28633
PhD	0.391201	0.356216	0.331896	0.532513	0.546566	0.318747	0.149306	0.383476	0.329627	0.026940	-0.010950	1.001289	0.850682	-0.130698	0.249330
Terminal	0.369968	0.338018	0.308671	0.491768	0.525425	0.300406	0.142086	0.408509	0.375022	0.100084	-0.030653	0.850682	1.001289	-0.160310	0.267475
S.F.Ratio	0.095756	0.176456	0.237577	-0.385370	-0.295009	0.280064	0.232830	-0.555536	-0.363095	-0.031970	0.136521	-0.130698	-0.160310	1.001289	-0.40344
perc.alumni	-0.090342	-0.160196	-0.181027	0.456072	0.418403	-0.229758	-0.281154	0.566992	0.272714	-0.040260	-0.286337	0.249330	0.267475	-0.403448	1.001289
Expend	0.259927	0.124878	0.064252	0.661765	0.528127	0.018676	-0.083676	0.673646	0.502386	0.112554	-0.098018	0.433319	0.439365	-0.584584	0.418250
Grad.Rate	0.146944	0.067399	-0.022370	0.495627	0.477896	-0.078875	-0.257332	0.572026	0.425489	0.001062	-0.269691	0.305431	0.289900	-0.307106	0.491530

Below is the Heatmap of Correlation Coefficients on the scaled data.



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



Outliers have not been treated and remains as it is. However the main difference between non-scaled data and scaled data is that the scales have been standardized. Post scaling, the data has now Zero mean and unit variance.

Let's perform **Bartletts Test of Sphericity**

Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

Null Hypothesis - Ho: All variables in the data are uncorrelated

Alternative Hypothesis - Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

If the p-value is small, then we can reject the null hypothesis and agree that there is at least one pair of variables in the data which are correlated hence PCA is recommended.

Test Result: **P-value is 0.00**

Since P-value 0.0 is smaller than the level of significance 0.05 we reject the null hypothesis and conclude that at least one pair of variables in the data are correlated.

Let's perform **Kaiser-Meyer-Olkin (KMO) Test**

The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if **MSA < 0.5**, then PCA is not recommended, since no reduction is expected. On the other hand, if **MSA > 0.7** then it is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Test Result: MSA is **0.8131251200373522**

Since the value is more than 0.7 we can proceed with dimensionality reduction using PCA.

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

Eigen Vectors: PCA Components

```
[[ 2.48765602e-01  2.07601502e-01  1.76303592e-01  3.54273947e-01
   3.44001279e-01  1.54640962e-01  2.64425045e-02  2.94736419e-01
   2.49030449e-01  6.47575181e-02 -4.25285386e-02  3.18312875e-01
   3.17056016e-01 -1.76957895e-01  2.05082369e-01  3.18908750e-01
   2.52315654e-01]
 [ 3.31598227e-01  3.72116750e-01  4.03724252e-01 -8.24118211e-02
  -4.47786551e-02  4.17673774e-01  3.15087830e-01 -2.49643522e-01
  -1.37808883e-01  5.63418434e-02  2.19929218e-01  5.83113174e-02
   4.64294477e-02  2.46665277e-01 -2.46595274e-01 -1.31689865e-01
  -1.69240532e-01]
 [-6.30921033e-02 -1.01249056e-01 -8.29855709e-02  3.50555339e-02
  -2.41479376e-02 -6.13929764e-02  1.39681716e-01  4.65988731e-02
   1.48967389e-01  6.77411649e-01  4.99721120e-01 -1.27028371e-01
  -6.60375454e-02 -2.89848401e-01 -1.46989274e-01  2.26743985e-01
  -2.08064649e-01]
 [ 2.81310530e-01  2.67817346e-01  1.61826771e-01 -5.15472524e-02
  -1.09766541e-01  1.00412335e-01 -1.58558487e-01  1.31291364e-01
   1.84995991e-01  8.70892205e-02 -2.30710568e-01 -5.34724832e-01
  -5.19443019e-01 -1.61189487e-01  1.73142230e-02  7.92734946e-02
   2.69129066e-01]
 [ 5.74140964e-03  5.57860920e-02 -5.56936353e-02 -3.95434345e-01
  -4.26533594e-01 -4.34543659e-02  3.02385408e-01  2.22532003e-01
   5.60919470e-01 -1.27288825e-01 -2.22311021e-01  1.40166326e-01
   2.04719730e-01 -7.93882496e-02 -2.16297411e-01  7.59581203e-02
  -1.09267913e-01]
 [-1.62374420e-02  7.53468452e-03 -4.25579803e-02 -5.26927980e-02
   3.30915896e-02 -4.34542349e-02 -1.91198583e-01 -3.00003910e-02
   1.62755446e-01  6.41054950e-01 -3.31398003e-01  9.12555212e-02
   1.54927646e-01  4.87045875e-01 -4.73400144e-02 -2.98118619e-01
   2.16163313e-01]
 [-4.24863486e-02 -1.29497196e-02 -2.76928937e-02 -1.61332069e-01
  -1.18485556e-01 -2.50763629e-02  6.10423460e-02  1.08528966e-01
   2.09744235e-01 -1.49692034e-01  6.33790064e-01 -1.09641298e-03
  -2.84770105e-02  2.19259358e-01  2.43321156e-01 -2.26584481e-01
   5.59943937e-01]
 [-1.03090398e-01 -5.62709623e-02  5.86623552e-02 -1.22678028e-01
```

Eigen Values: Explained Variance

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

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

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	
Apps	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	-0.176958	0.205082	0.3
Accept	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	0.246665	-0.246595	-0.1
Enroll	-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	-0.289848	-0.146989	0.2
Top10perc	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	-0.161189	0.017314	0.0
Top25perc	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	-0.079388	-0.216297	0.0
F.Undergrad	-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	0.487046	-0.047340	-0.2
P.Undergrad	-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	0.219259	0.243321	-0.2
Outstate	-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	-0.083605	0.678524	-0.0
Room.Board	-0.090227	-0.177865	-0.128561	0.341100	0.403712	-0.059442	0.560673	-0.004573	0.275023	-0.133663	-0.094469	-0.185182	-0.254938	0.274544	-0.255335	-0.0
Books	0.052510	0.041140	0.034488	0.064026	0.014549	0.020847	-0.223106	0.186675	0.298324	-0.082029	0.136028	-0.123452	-0.088578	0.472045	0.423000	0.1
Personal	0.043046	-0.058406	-0.069399	-0.008105	-0.273128	-0.081158	0.100693	0.143221	-0.359322	0.031940	-0.018578	0.040372	-0.058973	0.445001	-0.130728	0.6
PhD	0.024071	-0.145102	0.011143	0.038554	-0.089352	0.056177	-0.063536	-0.823444	0.354560	-0.028159	-0.039264	0.023222	0.016485	-0.011026	0.182661	0.3
Terminal	0.595831	0.292642	-0.444638	0.001023	0.021884	-0.523622	0.125998	-0.141856	-0.069749	0.011438	0.039455	0.127696	-0.058313	-0.017715	0.104088	-0.0
S.F.Ratio	0.080633	0.033467	-0.085697	-0.107828	0.151742	-0.056373	0.019286	-0.034012	-0.058429	-0.066849	0.027529	-0.691126	0.671009	0.041374	-0.027154	0.0
perc.alumni	0.133406	-0.145498	0.029590	0.697723	-0.617275	0.009916	0.020952	0.038354	0.003402	-0.009439	-0.003090	-0.112056	0.158910	-0.020899	-0.008418	-0.2
Expend	0.459139	-0.518569	-0.404318	-0.148739	0.051868	0.560363	-0.052731	0.101595	-0.025929	0.002883	-0.012890	0.029808	-0.027076	-0.021248	0.003334	-0.0
Grad.Rate	0.358970	-0.543427	0.609651	-0.144986	0.080348	-0.414705	0.009018	0.050900	0.001146	0.000773	-0.001114	0.013813	0.006209	-0.002222	-0.019187	-0.0

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]

	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
PC1	0.25	0.33	-0.06	0.28	0.01	-0.02	-0.04	-0.10	-0.09	0.05	0.04	0.02	0.60	0.08	0.13	0.46	0.36
PC2	0.21	0.37	-0.10	0.27	0.06	0.01	-0.01	-0.06	-0.18	0.04	-0.06	-0.15	0.29	0.03	-0.15	-0.52	-0.54
PC3	0.18	0.40	-0.08	0.16	-0.06	-0.04	-0.03	0.06	-0.13	0.03	-0.07	0.01	-0.44	-0.09	0.03	-0.40	0.61
PC4	0.35	-0.08	0.04	-0.05	-0.40	-0.05	-0.16	-0.12	0.34	0.06	-0.01	0.04	0.00	-0.11	0.70	-0.15	-0.14
PC5	0.34	-0.04	-0.02	-0.11	-0.43	0.03	-0.12	-0.10	0.40	0.01	-0.27	-0.09	0.02	0.15	-0.62	0.05	0.08
PC6	0.15	0.42	-0.06	0.10	-0.04	-0.04	-0.03	0.08	-0.06	0.02	-0.08	0.06	-0.52	-0.06	0.01	0.56	-0.41
PC7	0.03	0.32	0.14	-0.16	0.30	-0.19	0.06	0.57	0.56	-0.22	0.10	-0.06	0.13	0.02	0.02	-0.05	0.01
PC8	0.29	-0.25	0.05	0.13	0.22	-0.03	0.11	0.01	-0.00	0.19	0.14	-0.82	-0.14	-0.03	0.04	0.10	0.05
PC9	0.25	-0.14	0.15	0.18	0.56	0.16	0.21	-0.22	0.28	0.30	-0.36	0.35	-0.07	-0.06	0.00	-0.03	0.00
PC10	0.06	0.06	0.68	0.09	-0.13	0.64	-0.15	0.21	-0.13	-0.08	0.03	-0.03	0.01	-0.07	-0.01	0.00	0.00
PC11	-0.04	0.22	0.50	-0.23	-0.22	-0.33	0.63	-0.23	-0.09	0.14	-0.02	-0.04	0.04	0.03	-0.00	-0.01	-0.00
PC12	0.32	0.06	-0.13	-0.53	0.14	0.09	-0.00	-0.08	-0.19	-0.12	0.04	0.02	0.13	-0.69	-0.11	0.03	0.01
PC13	0.32	0.05	-0.07	-0.52	0.20	0.15	-0.03	-0.01	-0.25	-0.09	-0.06	0.02	-0.06	0.67	0.16	-0.03	0.01
PC14	-0.18	0.25	-0.29	-0.16	-0.08	0.49	0.22	-0.08	0.27	0.47	0.45	-0.01	-0.02	0.04	-0.02	-0.02	-0.00
PC15	0.21	-0.25	-0.15	0.02	-0.22	-0.05	0.24	0.68	-0.26	0.42	-0.13	0.18	0.10	-0.03	-0.01	0.00	-0.02
PC16	0.32	-0.13	0.23	0.08	0.08	-0.30	-0.23	-0.05	-0.05	0.13	0.69	0.33	-0.09	0.07	-0.23	-0.04	-0.04
PC17	0.25	-0.17	-0.21	0.27	-0.11	0.22	0.56	-0.01	0.04	-0.59	0.22	0.12	-0.07	0.04	-0.00	-0.01	-0.01

PC1 = **0.25 * SApps + 0.33 * SAccept - 0.06 * SEnroll + 0.28 * STop10perc + 0.01 * STop25perc - 0.02 * SF.Undergrad - 0.04 * SP.Undergrad - 0.10 * SOutstate - 0.09 * SRoom.Board + 0.05 * SBooks + 0.04 * SPersonal + 0.02 * SPhD + 0.60 * STerminal + 0.08 * SS.F.Ratio + 0.13 * Sperc.alumni + 0.46 * SExpend + 0.36 * SGrad.Rate**

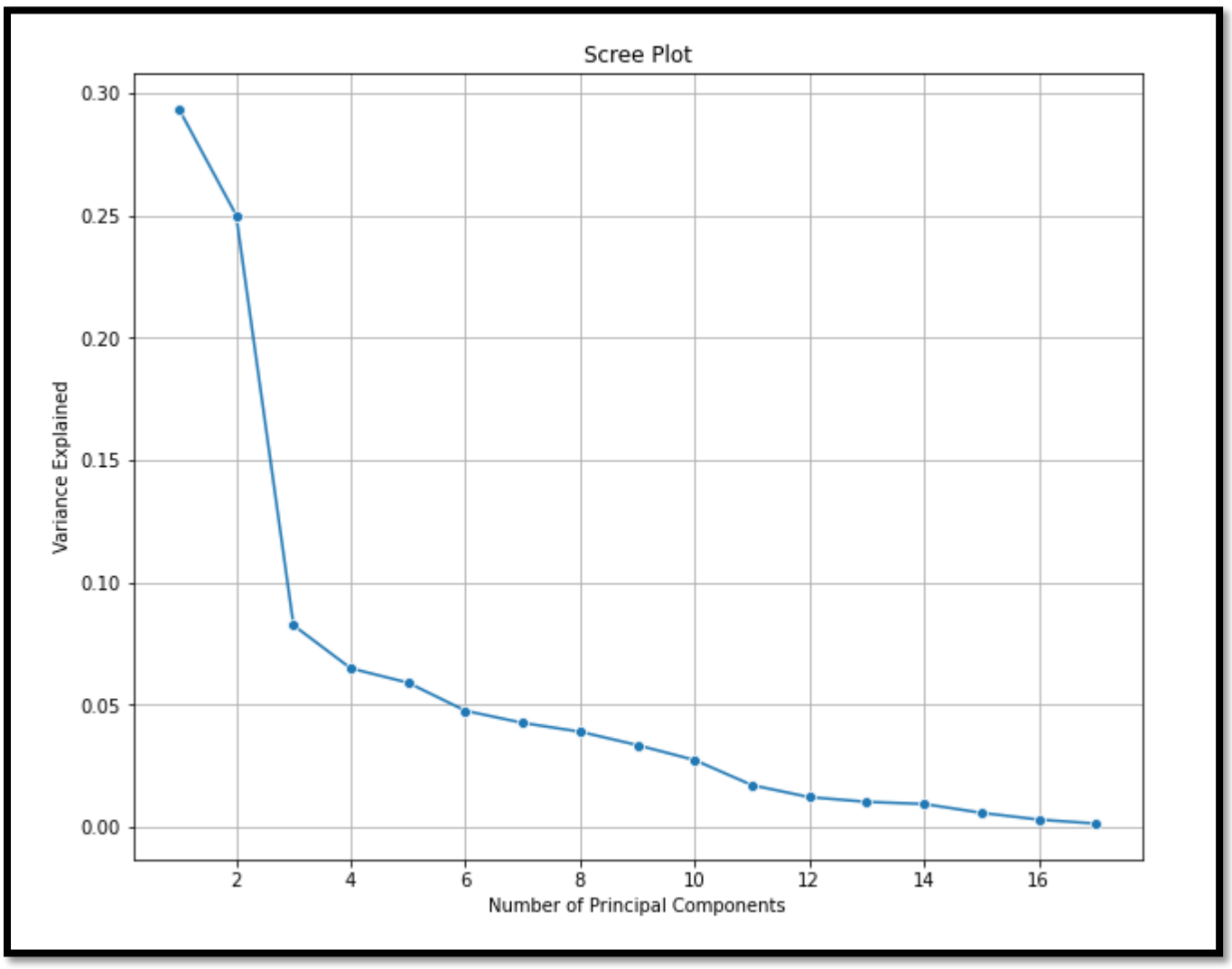
The letter S indicates that the scaled (normalized) variable is used to construct the PCs.

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?

Let's now see the cumulative explained variance ratio to see how many components to be kept for retaining maximum information.

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

I have plotted a **Scree Plot** to see the relationship between Number of Principal Components and their Variance Explained.



From the above plot and the cumulative explained variance ratio table, we can see that we can retain close to **80%** of information with only **6 Principal Components**. The rest of the components can be dropped as it does not add much value to the maximum variance retained.

Eigen vectors or Principal Components are linear combinations of original features. We can now use the Eigen vectors for our further analysis or machine learning algorithms to perform required predictive modelling.

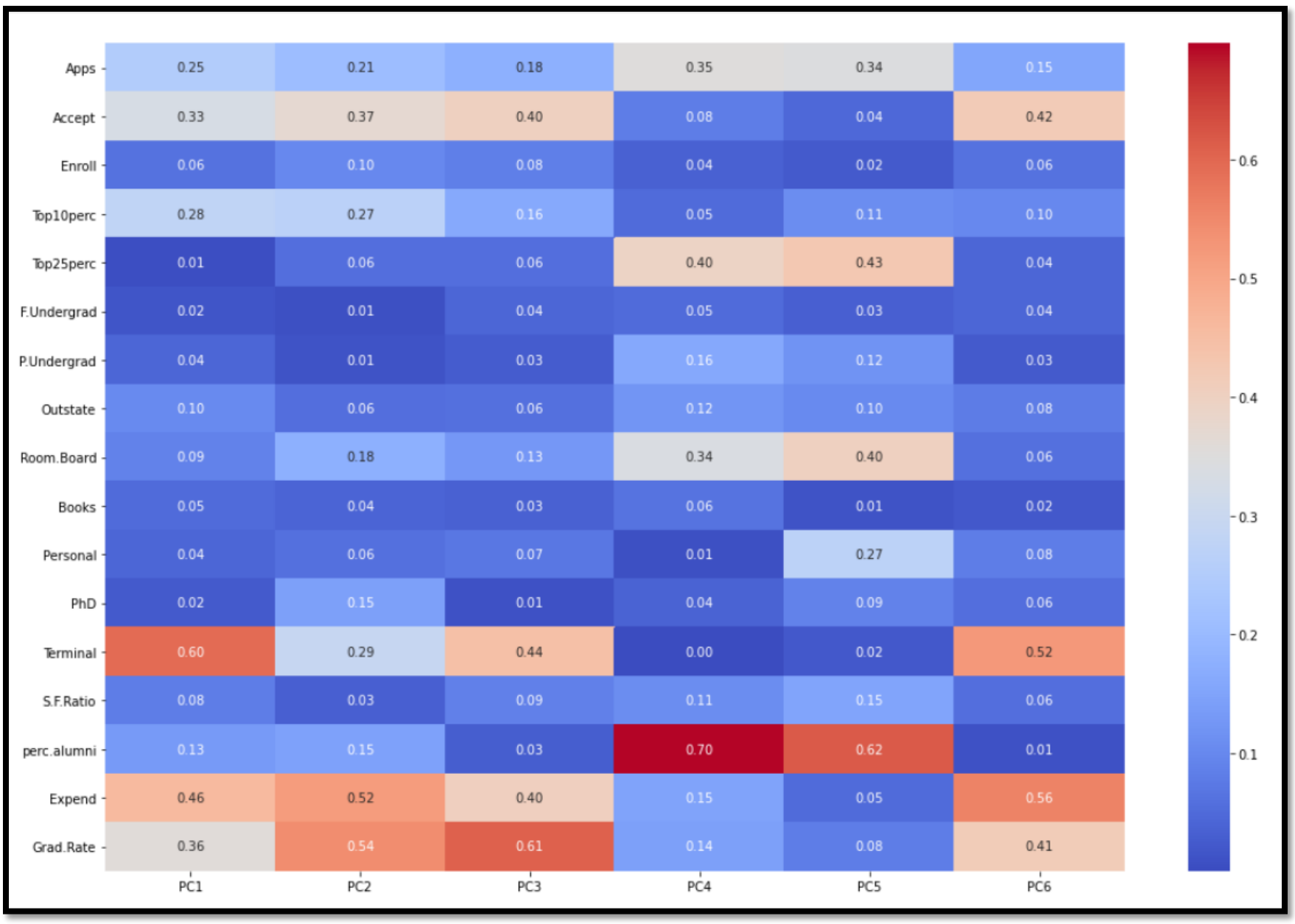
2.9 Explain the business implication of using the Principal Component Analysis for this case study. How may PCs help in the further analysis? [Hint: Write Interpretations of the Principal Components Obtained]

Principal Component Analysis (PCA) is a transformation technique for dimensionality reduction such that maximum variance or information is retained. Below are some of the assumptions and data requirements for PCA.

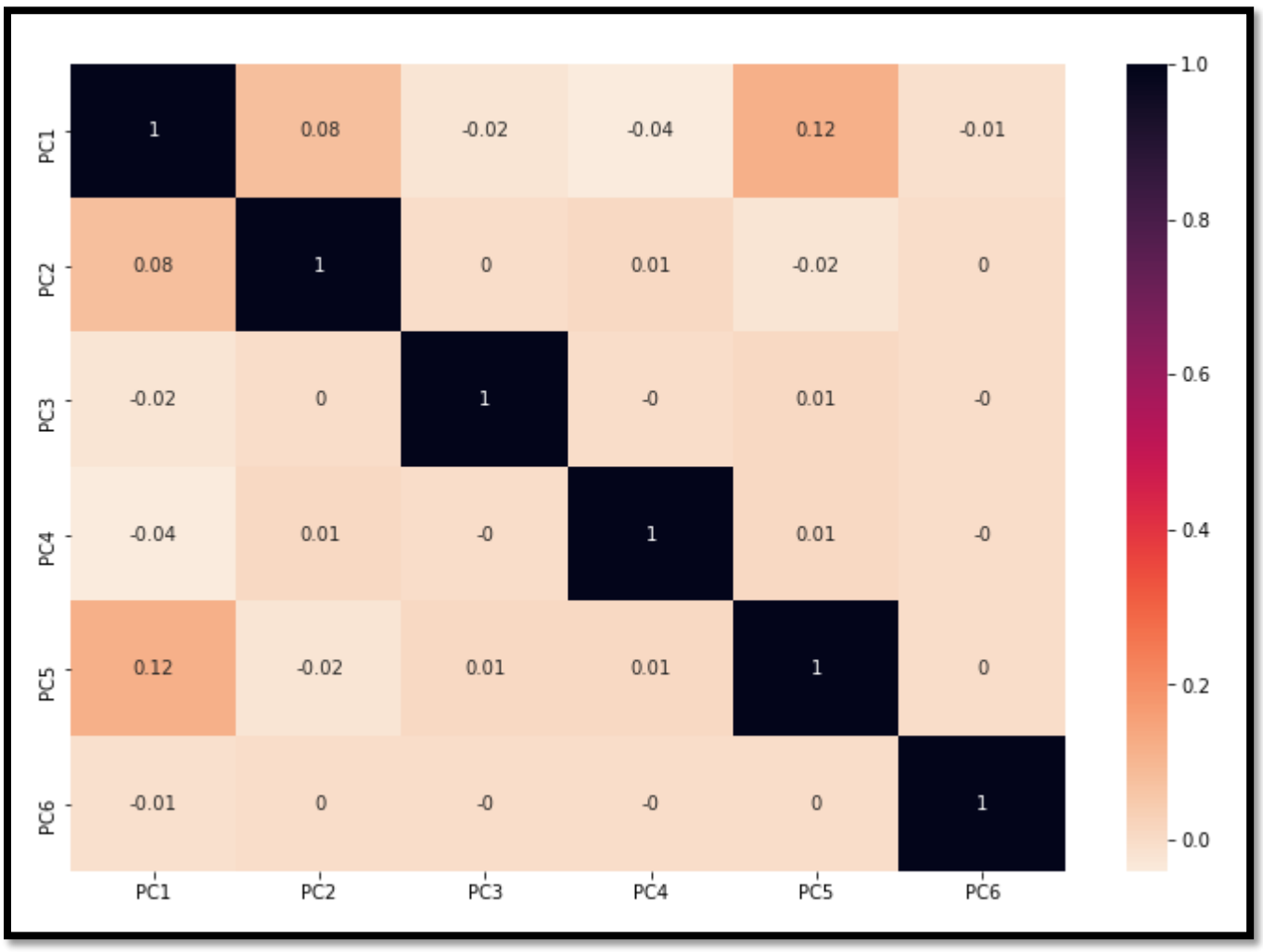
1. At least two independent variables should have **linear** correlation.
2. Data should be **normally** distributed.
3. Sample size should be sufficiently **large**.
4. PCA can be performed only on **numerical** variables.
5. If the features have varying weights or scale, then the data needs to be **scaled** such that the mean is Zero and the data is at unit-variance.

PCA does not mean we drop the features from the dataset. It only means we are trying to see the correlation between features and transform them into new features such that maximum information is retained. If two independent variables are correlated, then performing PCA would provide better results during predictive modelling. PCA is usually used as an intermediary step to further analysis. Outcome of PCA may be used in regression, such as principal component regression or partial least squares regression. It can also be used in extraction of latent factors, which often provides important insights into various business applications, such as customer behavior, ecommerce etc.

For the current dataset provided, out of 17 features, we have reduced the dimensions from 17 to 6 such that **80%** of the variance is retained. These 6 principal components can further be used to perform various regression models.



1. PC1 has strong loading value or score for **Terminal**.
2. PC2 has strong loading value or score for **Grad.Rate and Expend**.
3. PC3 has strong loading value or score for **Grad.Rate**.
4. PC4 has strong loading value or score for **perc.alumni**.
5. PC5 has strong loading value or score for **perc.alumni**.
6. PC6 has strong loading value or score for **Expend**.



In the Education – Post 12th Std data, although there were **17** original attributes, more than **80%** of the total variance has been explained with only the first **6 PC's**, and thus the goal of **dimension reduction** is achieved.